The effects of biological knowledge graph topology on embedding-based link prediction

Michael S. Bradshaw, [1] Alisa Gaskell,[2] Ryan M. Layer [1*]

[1]Department of Computer Science, University of Colorado Boulder, Boulder, CO, United States of America

[2] Precision Medicine Institute, Children's Hospital Colorado, Aurora, CO, United States of America

[*]Correspondence: ryan.layer@colorado.edu (RL)

Abstract

Due to the limited information available about rare diseases and their causal variants, knowledge graphs are often used to augment our understanding and make inferences about new gene-disease connections. Knowledge graph embedding methods have been successfully applied to various biomedical link prediction tasks but have yet to be adopted for rare disease variant prioritization. Here, we explore the effect of knowledge graph topology on Knowledge graph embedding link prediction performance and challenge the assumption that massively aggregating knowledge graphs is beneficial in deciphering rare disease cases and improving outcomes. We find that using a filtered version of the Monarch knowledge graph with only 11% of the size of the full knowledge graph results in improved model predictive performance. Additionally, we found that as information is aggregated and re-added to the knowledge graph, performance improvements are driven by the quality of information, not the quantity.

Introduction

One of the difficulties in diagnosing rare and undiagnosed diseases is the limited information about the disease's progression and its underlying genetic causes. Many diseases are still characterized solely by the phenotype-driven method, which is hindered by symptoms and clinical findings often shared between syndromes (Liu et al. 2022). Similarly, understanding the disease mechanism for a specific variant is difficult when few to no other cases have ever been reported. In such cases, it is useful to apply existing knowledge about biology more broadly to make inferences about genes and diseases we know little about by using knowledge graphs (KGs) (Robinson et al. 2014).

KGs represent the interactions and relationships between a wide range of biological entities, such as genes, pathways, diseases, anatomy, and therapeutics, modeled as nodes in a graph. Edges between nodes represent a relationship between them, creating a web of interconnected nodes.

Like our collective understanding of the natural world, KGs are inherently incomplete and grow as discoveries are made and published. One heavily researched topic for filling in missing information in KGs is link prediction, with a recent emphasis on KG embedding (KGE) methods (Mohamed, Nounu, and Nováček 2021). KGE offers several advantages over alternative methods for link prediction, including edge type awareness, the use of higher-order patterns, scalability, and improved accuracy over feature-based approaches (Mohamed, Nounu, and Nováček 2021). KGEs have been successfully used in drug discovery tasks (Mohamed, Nounu, and Nováček 2019; Bonner, Kirik, et al. 2022) (Zitnik, Agrawal, and Leskovec 2018) (Prabhakar et al. 2023). The same methods used for these pharmaceutical-specific tasks can also be directly applied to rare disease variant prioritization.

While KGs have been used in the rare disease space and KGE link predictors have been deployed in highly-related fields, there have yet to be any stand-alone variant prioritization tools based on KGE link prediction. Sanjak et al. 2023 (Sanjak et al. 2023), showed that embeddings could be used to cluster KGs and find rare disease modules, though they did not publish a stand-alone tool. Vilela et al. 2023 (Vilela et al. 2023) showed that KGE link prediction could be used for variant prioritization in common conditions like Autism Spectrum Disorder but also did not publish a tool.

As the field of variant prioritization circles KGE link prediction, certain limitations of KGE link prediction models must be addressed. Among these, node-degree bias stands out. This is the tendency for a model to preferentially score high-degree nodes over low-degree nodes in link prediction tasks (Bonner et al. 2022). This is a major technical challenge standing in the way of adopting KGE models for rare disease variant prioritization. Given that the end goal of link prediction is to make inferences about less-studied genes and diseases, a bias towards high-degree, well-characterized topics is counterproductive (Zietz et al. 2024).

Beyond their impact on individual analysis, degree-biased models are part of a troubling cyclical pattern in which the overemphasis of certain genes in experimental studies gives rise to topologically imbalanced KGs (Lucchetta et al. 2023). Training on these KGs results in degree-biased models that give inflated importance to certain biological entities and falsely lowly rank others (Bonner, Kirik, et al. 2022) (Figure 1). We call this the bias cycle.

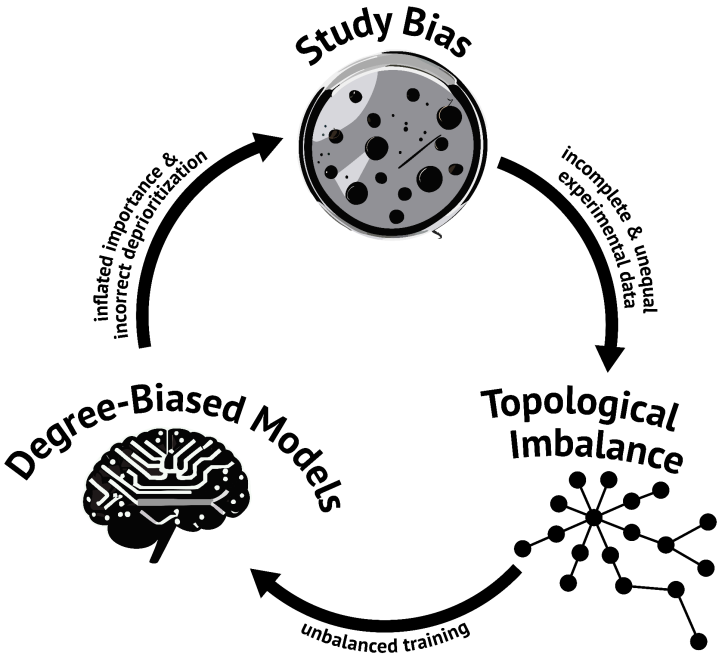

Figure 1: A visual description of the bias cycle. Study bias creates incomplete and unequal data; this data is incorporated into KG which then exhibits topological imbalance. Training predictive models on imbalanced data creates biased models that give inflated importance to certain entities while incorrectly lowly ranking others. Model predictions are used to inform study design, beginning the cycle again.

Elements of this cycle are well-known and active research areas (Bonner, Kirik, et al. 2022; Erten et al. 2011). It has been shown that this imbalance originates not from biology itself but from the interest of researchers in certain proteins from which the unbalanced study design ensues (Lucchetta et al. 2023). This coincides with findings that certain genes are over-studied while others are ignored (Kustatscher et al. 2022; Stoeger et al. 2018; Yılmaz, Yorgancioglu, and Koyutürk 2023), 19% of genes receive 49% of attention (Stoeger et al. 2018). This is unfortunate since a broad impact on human diseases has been made by studying rare genetic diseases (Lee et al. 2020).

KGs are often employed to fill in the gaps in our understanding of rare and undiagnosed disease cases. An unspoken assumption in KG construction seems to be that the more information one integrates into these KGs, the better – that all additional knowledge, be it anatomical, chemical, human, or model organism, will help decipher rare disease cases and improve outcomes. Based on this assumption, great success and many high-quality KG-based methods have been made (Cipriani et al. 2020; Boudellioua et al. 2019; Putman et al. 2024). However, in the context of KGE models, this assumption remains untested and could be counterproductive as topological imbalance increases as information is aggregated.

Here, we investigated the implications of degree-bias and KG aggregation when predicting connections among specific groups of genes and diseases and explored if more is always better when training KGE link prediction models. Using KGE models trained on the Monarch KG, we found that heritable cancer genes rank better than non-cancer genes at a link prediction task and that rare diseases, ultra-rare diseases, and diseases generally lag behind heritable cancer genes in predictive performance. We show that using all types of information in a highly heterogeneous KG can degrade the performance of KGE link predictors, suggesting that the set of metanodes used from a KG ideal should undergo a feature selection-like process to identify what types of information are actually helpful for the problem. Additionally, we found that there is a balance to be struck between the number and quality of edges trained on; using all is non-optimal, but overly strict filtering criteria can be worse. These findings will help improve KGE-based link predictors as the field adopts them for rare disease variant prioritization.

Methods

To investigate KGE link prediction performance on various groups of genes and diseases, we optimized and trained a TransE KGE model on the Monarch KG (Putman et al. 2024) and evaluated it at link prediction tasks stratified by five groups of genes (heritable cancer genes and non-cancer genes) and diseases (rare diseases, ultra-rare disease, control diseases). Following the initial results, we tuned RotatE and ComplEx models for the same task.

To test the hypothesis that more data is always better, we created various filtered KGs, starting with one containing only gene and disease nodes and edges. We then tested performance when one additional metanode (of 22 possible types) was reintroduced to the KG. We created an additional version of a KG using protein-protein interaction (PPI) information from HuRI and STRING to explore the effect of edge quality versus quantity on link prediction performance as well.

Knowledge graphs

Monarch knowledge graph

We used the 2023-12-16 version of Monarch KG, a large and highly heterogeneous graph containing medically relevant data about genes, proteins, disease, phenotypes, and anatomy in humans and numerous model organisms. Edges in Monarch KG are structured as semantic triples in the format subject → predicate → object, where subject and object are nodes and the predicate describes the type of edge, or relationship, connecting the two nodes. For example, NBEA → interacts_with → NOTCH1, indicating a physical interaction between the two, or NBEA → causes → epilepsy to indicate there is a causal link between a variant in NBEA and epilepsy. Monarch KG has 519,303

nodes across 63 metanodes (types of nodes often originating from different sources) (Figure 2.A), and 11,412,471 edges across 23 edge types (Figure 2.B).

Filtered KGs

Filtered variations of the Monarch KG were made by removing all nodes and edges that were not MONDO (nodes originating from the Monarch Disease Ontology) or HGNC (HUGO Gene Nomenclature Committee identifiers) (the HGNC and MONDO metanodes). This filtered KG contains the minimum information to train and evaluate KGE models at link prediction on our gene and disease groups. The filtering process resulted in a 92% reduction in edges (Table 1).

Additional filtered KGs included HGNC and MONDO, plus one additional meta-node. There were 22 metanodes from the full KG incident to the two in our filtered graph; we created a new edgelist for each of these.

HuRI

HuRI (Luck et al. 2020) is a PPI interaction network derived from a massive all-by-all binding experiment that should be free from the experimental study design bias that affects aggregate PPI networks (Lucchetta et al. 2023). We hypothesized that eliminating a step in the bias cycle from the data used to train KGE models would improve predictive performance. Thus, we created a version of the filtered KG where all HGNC nodes and edges were removed and replaced with those from HuRI. This reduced the graph by 39,120 nodes and 6,210,908 edges.

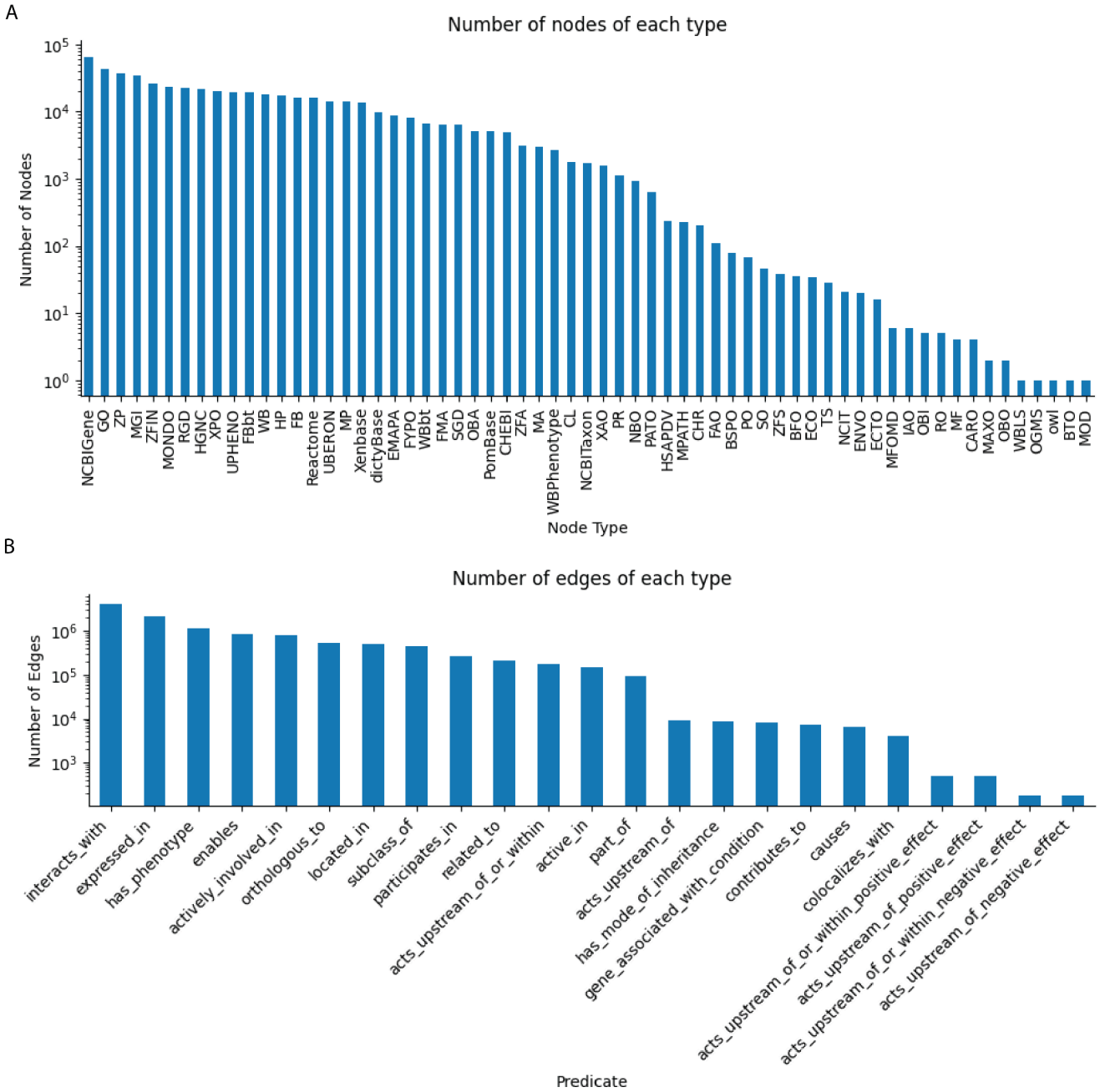

Figure 2: **A**. The number of nodes of each type in Monarch KG: y-axis is on a log scale. **B**. The number of edges of each type: y-axis is on a log scale.

STRING

We constructed an additional version of the filtered KG using edges from a much larger PPI network (1,152,555 edges vs 97,679), STRING (Szklarczyk et al. 2019). STRING is an aggregate PPI network that derives link information from other databases, experimental results, co-expression networks, and text-mining. STRING also provided edge confidence scores describing the amount of evidence behind an edge, stratified by the type(s) of sources it originates from. Working with the human version of STRING 12.0, we removed all edges that are not experimentally derived, edges with a non-zero value in the "experiments" column of the file 9606.protein.links.full.v12.0.txt. Based on the confidence scores in the experiments, we created three edge lists containing the top 25%, 50%, and 100% most

confident edges. We created three versions of our filtered KG with these edge lists, replacing all HGNC nodes and edges with those filtered STRING edge lists.

KG	Edges	Nodes
Monarch KG	11412471	519303
Monarch KG filtered	1289563	43737
HuRI	97679	31579
STRING 100%	1152555	40843
STRING 50%	608734	40094
STRING 25%	284992	37923
CHEBI + Filtered KG	1301537	48660
CHR + Filtered KG	1289929	43940
CL + Filtered KG	1292776	45353
dictyBase + Filtered KG	1301690	48890
FB + Filtered KG	1593952	56464
GO + Filtered KG	1689454	87452
HP + Filtered KG	1788336	60417
HSAPDV + Filtered KG	1290264	43975
MGI + Filtered KG	1773282	64649
NCBIGene + Filtered KG	2053721	86872
NCBITaxon + Filtered KG	1291965	45434
NCIT + Filtered KG	1289610	43758
PATO + Filtered KG	1290794	44368
PomBase + Filtered KG	1295589	46932
Reactome + Filtered KG	1334419	46040
RGD + Filtered KG	1619008	63424
SGD + Filtered KG	1296195	47057
SO + Filtered KG	1289640	43776
UBERON + Filtered KG	1334609	58196
WB + Filtered KG	1545733	57466
Xenbase + Filtered KG	1440666	56984
ZFIN + Filtered KG	1601905	60856

Table 1: The number of nodes and edges in each KG. KG names from CHEBI - ZFIN are the names of the source added back to the filtered KG in the re-addition experiments.

KGE models and optimization

We used three models, TransE (Bordes et al. 2013), RotatE (Sun et al. 2019), and ComplEx(Trouillon et al. 20--22 Jun 2016), as implemented by pykeen (Ali et al. 2020). Models were optimized and trained for each KG version. For each KG, edges were randomly divided into an 80/10/10 train/validation/test split. For all model and KG pairings, hyperparameters were chosen using a hyperparameter optimization pipeline using a quasi-Monte Carlo (QMC) (Bergstra and Bengio 2012) sampler to sample parameters from the ranges listed in Supplemental Table 1. Parameter ranges were informed by previous work on these models in a similar learning task (Bonner, Barrett, et al. 2022). Hyperparameter optimization for each model was run on a single dedicated 80Gb nvidia A100 GPU. Trial run times were collected from the hyperparameter optimization trial logs for all 30 iterations for each model-KG combination. The accompanying file, supplemental_table_1.tsv, contains a tabulated list of all model—KG pairs and their respective hyperparameters.

Datasets

Heritable cancer genes

A list of 25 genes with mutations linked to hereditary cancer risk associated with breast, colorectal, endometrial, fallopian tube, ovarian, primary peritoneal, gastric, pancreatic, skin, and prostate cancer. This list of genes was compiled by the FORCE - Facing Our Risk of Cancer Empowered (FORCE, n.d.)

Non-cancer genes

A set of 500 genes was chosen randomly from all genes in HGNC, excluding those in the heritable cancer genes set.

Rare and ultra-rare diseases

A list of 300 ultra-rare diseases (worldwide prevalence of < 1 / 1,000,000) was selected from the Orphanet Epidemiology report from July 2023 (https://www.orphadata.com/data/xml/en_product9_prev.xml). We selected disorders from this list that had a "PrevalenceGeographic" of "Worldwide", "PrevalenceValidationStatus" of "Validated", "PrevalenceClass" of "< 1 / 1 000 000" and those that had cross-references to MONDO identifiers that were also present in our chosen version of Monarch KG. This resulted in 3,007 disorders, of which we chose 300 uniformly at random. Using all diseases with "PrevalenceGeographic" of "Worldwide" but > 1 / 1,000,000, we made a list of 290 rare diseases - all listed that did not fall into the ultra-rare category.

Control diseases

A set of 300 MONDO disease identifiers was randomly sampled from all disease identifiers listed in MONDO to serve as a set of control diseases for general comparison.

Results

TransE trained on the full Monarch KG showed evidence of node-degree bias. Of the five groups of genes and diseases we tested, link prediction for four groups performed poorly relative to the heritable cancer genes – highlighting how this stock KGE model would be ill-adapted in rare disease variant prioritization. RotatE and ComplEx, which consider more robust types of information, improved performance but still exhibited the pattern of the four groups lagging well behind heritable cancer genes. Models trained on filtered versions of the Monarch KG improved predictive performance for all gene groups and some disease groups, demonstrating that using all information in a KG is non-optimal. Reintroducing certain metanodes to the filtered KG further improved performance - revealing the need to explore the metanode combination space. Finally, experiments into the quality of PPI edges showed that there is a balance to be struck between using all (including lower quality) edges and using fewer high-quality edges.

Degree bias

All KGE models we trained on Monarch KG exhibited degree bias. The degree of a gene and its disease-relatedness score had a Pearson correlation coefficient of as high as 0.48, 0.55, 0.42, with medians of 0.20, 0.37, and 0.23 with TransE, RotatE, and ComplEx, respectively (Figure 3.A). The degree bias we observed was not as pronounced as that of the same models trained on Hetionet (Bonner, Kirik, et al. 2022).

Primary results were produced with a TransE model - one of the most widely used KGE models. With TransE, there was poor performance in non-cancer genes (median normalized rank (MNR) = 61), rare diseases (MNR = 71), ultra-rare diseases (MNR = 65), and control diseases (MNR = 62), but much better performance on the heritable cancer genes (MNR = 82) (Figure 3.B).

Since the publication of TransE in 2013, numerous more complex models have been published(Cao et al. 2024). We tested two of these, ComplEx and RotatE. We found that RotatE made the best improvements over our original TransE model in all groups: +12% cancer, +19% rare diseases, +28% ultra-rare diseases, +19% non-cancer genes, +26% control diseases (Figure 3.C&D). This highlights the importance of careful model selection. Regardless of the model, we observed that the heritable cancer genes had higher scores than the other groups (Figure 3.E).

From these results, we concluded that these node-degree biased models successfully rank well-studied, higher-degree nodes but often fail to correctly prioritize results for less studied entities such as non-cancer genes, rare and ultra-rare diseases, and genetic diseases at large. Using these skewed results to inform experimental design completes the bias cycle and propagates these issues.

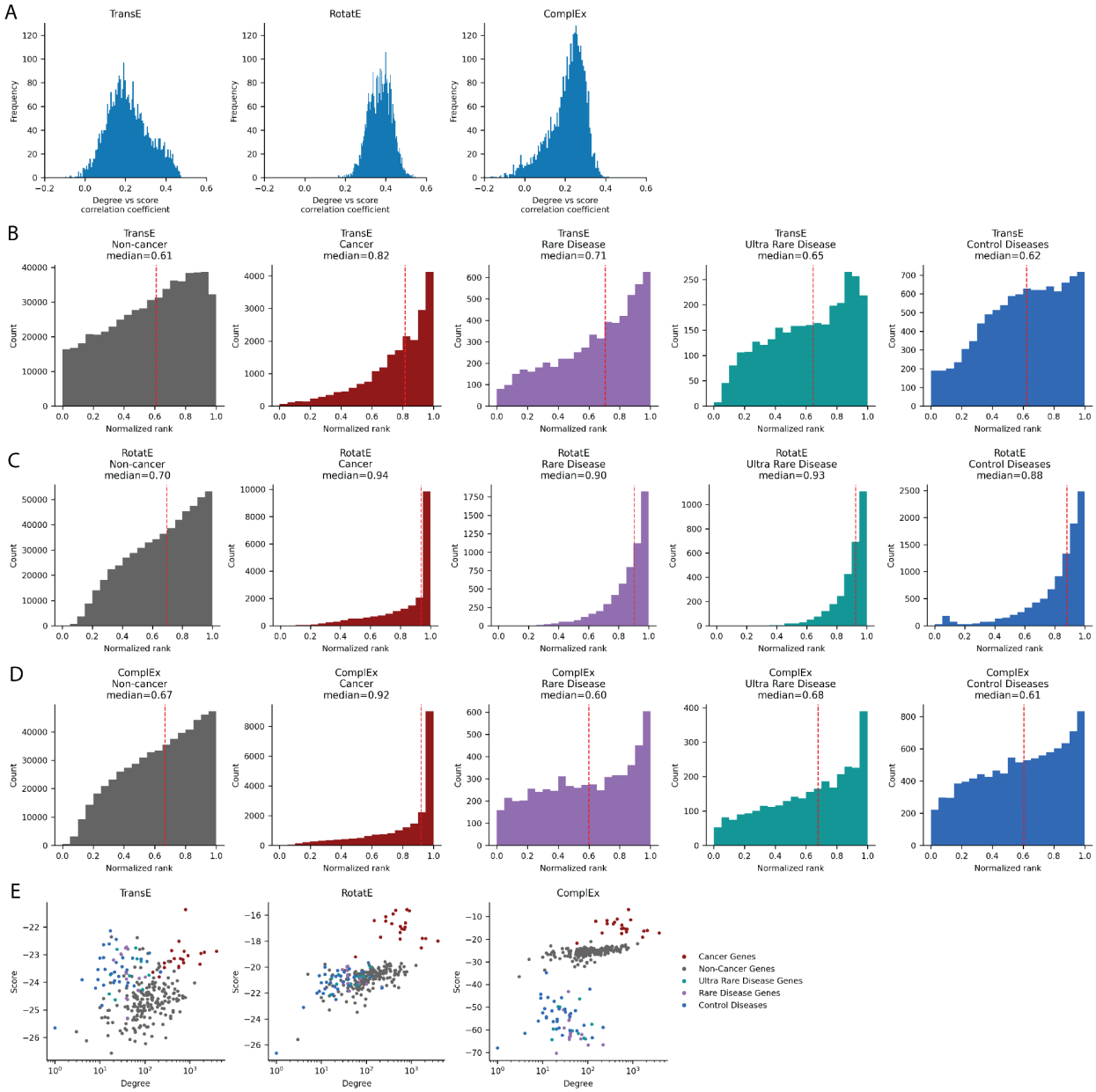

Figure 3: **A** Distribution of correlation coefficients of the relationship between degree and scores; scores were generated for whether a gene was connected to each MONDO disease in Monarch KG - every gene was measured for one disease at a time, and correlation coefficients were calculated. Distribution of normalized ranks of test set edges within group and across the models **B** TransE **C** RotatE **D** ComplEx. **E** Relationship between degree and score in test set edges divided by groups and models.

KG filtration

Our results show that using the full KG for training can be detrimental to both predictive performance and run time. Monarch KG contained 63 metanodes representing the relationships between entities, including diseases, phenotypes, genes, proteins, anatomy, chemicals/drugs, and more in both humans and model organisms. There is a correlation between the degree of genes in gene/protein networks and their degree in the rest of the KG, Pearson's R = 0.34, suggesting nodes that were already highly connected become evermore connected in the aggregate KG, potentially exasperating the effects of degree-bias. To investigate if this aggregation affected predictive performance, we

removed all but the bare minimum metanodes from the graph required to carry out link prediction tasks with the five groups - creating a filter KG made of only HGNC and MONDO nodes.

After removing the other 61 metanodes and retraining the three KGE models, we found all models improved gene-gene link prediction results over the original TransE model, cancer predictions improved by 9-10%, and non-cancer genes improved by 6-14% (Figure 4.A-C). Predictions of ultra-rare disease improved with the TransE and RotatE models by 6-12%. Rare diseases and control diseases performed better with the TransE and the full KG, except in the case of Rotate + filtered KG, where control disease MNR was slightly better (1% increase in RotatE + filtered KG).

Another benefit of filtering the KG was that it dramatically reduced the time required to optimize and train the KGE models (Figure 4.D). Filtered KG models needed only 0.8%-13.8% as much time per trial as their full KG counterparts. This took hyperparameter optimization down from >30 days to <1 day.

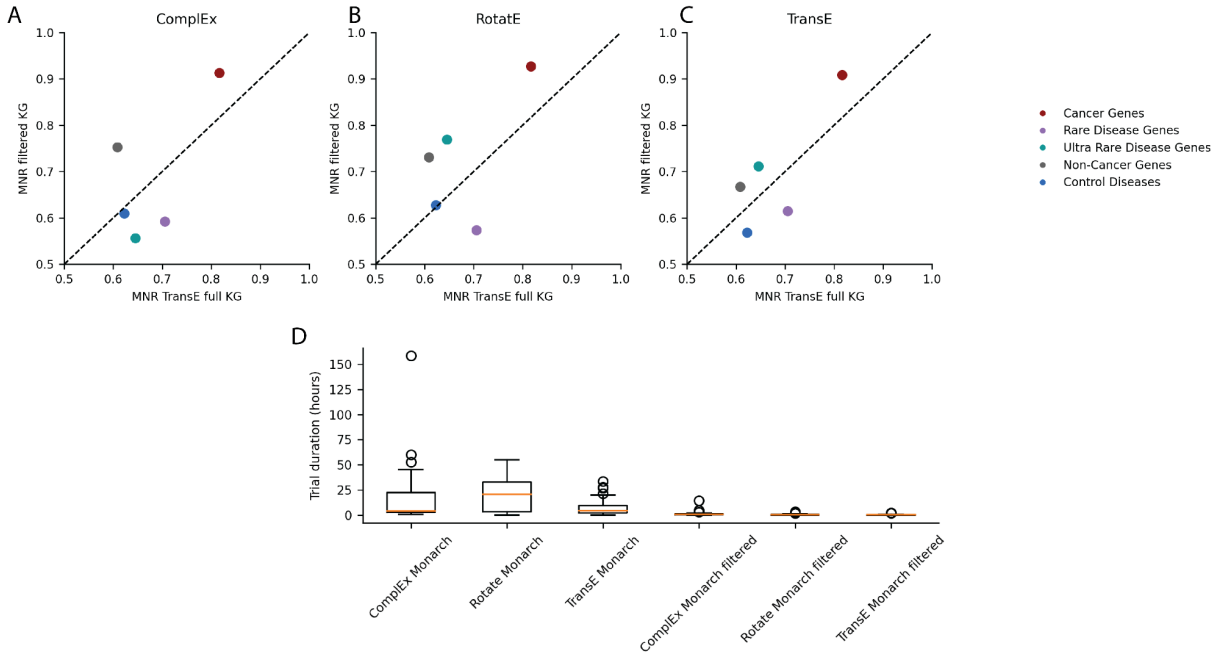

Figure 4**: A,B,C** Change in MNR in models trained on the TransE full KG vs. filtered KG, x-axis is the MNR of our original TransE + full Monarch KG model, diagonal line marks slope=1 passing through the intercept. Groups with points above the diagonal performed better with the filtered KG and below the diagonal indicates performance was better in the original TransE + full KG model. **D** Run time in hours of a single trial during hyperparameter optimization divided by model and KG filtration status.

KG filtration + single re-addition

Reintroducing the Human Phenotype Ontology (HPO) to the filtered KG increased performance on rare, ultra-rare, and control diseases by 10-22%. Reducing the full Monarch KG from 63 metanodes to just two showed that the full KG was non-optimal, suggesting an optimal combination of metanodes

exists between using all 63 and two. We took the first step into exploring this metanode combination space by reintroducing a single metanode to our filtered KG.

An additional 22 metanodes are connected to the two in our filtered KG; we found that, when reintroduced, several metanodes improved predictive performance (Figure 5.A). The reintroduction of Saccharomyces Genome Database (SGD) (Cherry et al. 2012) improved the performance of the heritable cancer genes group the most (3%), dictyBase (Fey et al. 2013) (1%) improved non-cancer gene performance the most. All disease groups (rare, ultra-rare, and control) improved the most when HPO was reintroduced (10-22%). There did not seem to be any relationship between the number of edges added to the network and predictive performance improvement. For example, HPO increased MNR by 23% in rare diseases and also added 110% more edges to the filtered KG, whereas Chemical Entities of Biological Interest (CHEBI) (Hastings et al. 2016), a chemical interaction database, improved MNR by almost the same amount, but only added 80% more edges (Figure 5.B) or how NCBIGene added 149% more edges but decreased performance for non-cancer genes, ultra-rare, and control diseases. The finding that the reintroduction of HPO increased performance the most for all disease groups is not surprising, given it was created specifically to increase understanding and communication around human disease phenotypes, thus highlighting its continued importance.

Parts of the KG should be removed/added depending on the type of disease or genes one is interested in making predictions around. Only 5/22 of the metanodes, when reintroduced, increased MNR for the ultra-rare disease group, illuminating the need for careful consideration of all parts of a KG.

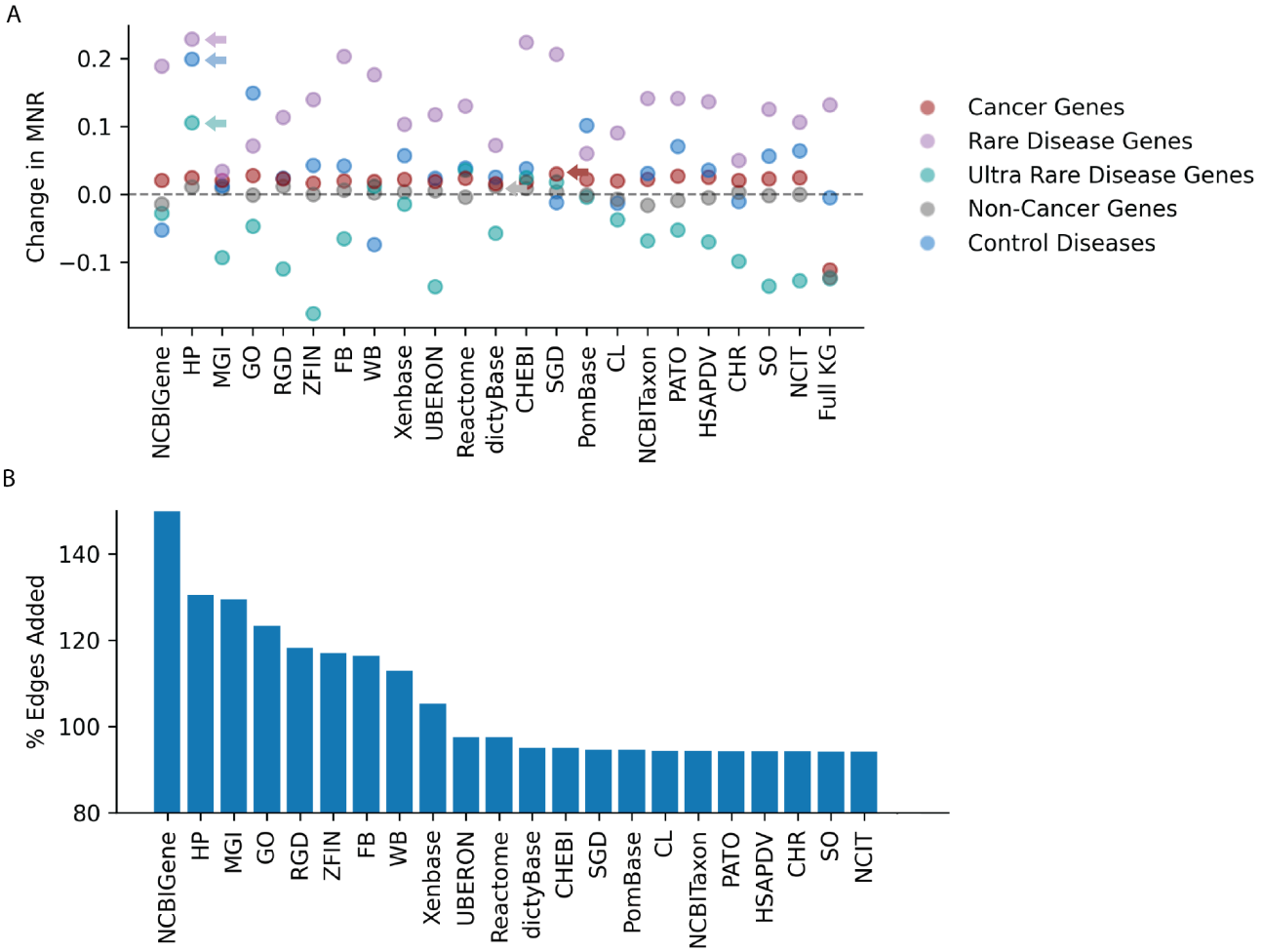

Figure 5: **A** Change in MNR when each of the 22 metanode was added back into the KG and trained on. Only RotatE models were trained and reported; changes are relative to the RotatE + filtered KG model. **B** The percent of edges added into the network by each metanode. The number of edges added by including the metanode divided by the filtered Monarch KG edge count (1,603,267).

PPI edge filtration

The results of Lucchetta (Lucchetta et al. 2023) call into question the reliability of PPI networks due to the effect of aggregating many small study-bias experiments on network topology. To test if controlling for study bias using a PPI network originating from an all-by-all experiment would improve results for our otherwise underperforming groups, we replaced all the HGNC nodes and edges in Monarch KG with those from the HuRI network. This replacement had an overall detrimental effect on performance; MNR decreased for non-cancer and by -16% and -17% in TransE and ComplEx, RotatE increased by +5; and for cancer genes -3% and -8% for TransE and ComplEx, and an increase of +2% for RotatE. Rare diseases followed a similar pattern, -4% in TransE, -6% in RotatE, but +7 for ComplEx. Control disease universally improved by 8-13%. Ultra-rare diseases were not even measurable due to the removal of too many nodes. This variability and loss in performance is likely due to the drastic reduction in the number of edges. HuRI contains only 3% of the edges in the full Monarch KG and only 10,000 proteins as opposed to the 21,000 in Monarch KG.

To see if the drop in performance was a result of excessive edge removal, we created KGs with varying degrees of edge filtration based on experimentally derived edge confidence scores from

STRING; we used the top 25%, 50%, and 100% highest quality edges, which equated to 15%, 36% and 71% of all PPI edges in Monarch KG. We found that peak MNR performance for non-cancer gene prediction was achieved with only 36% (the top 50% of edges in STRING) of the edges in the ComplEx and RotatE models (in TransE MNR continually decreased as the percent of edges shrunk), suggesting that reducing the KG to only 3% of edges with HuRI was too strict. We also found, that performance could be improved by removing less confident edges from the system.

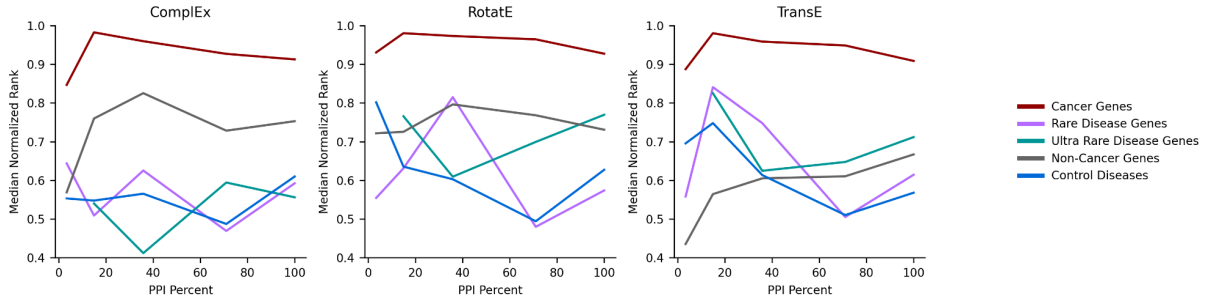

Figure 5: Median normalized rank of the various groups divided by model, **A**. Complex, **B**. RotatE, **C**. TransE. The x-axis is the portion of HGNC to HGNC edges in the training split of the KG relative to the total number in the Monarch KG train split.

Best combinations - percent edges at K

Of all the combinations of metanodes and models we evaluated, the two best-performing pairings for rare and ultra-rare diseases were RotatE + Full Monarch KG (MNR = 0.90 and 0.93), and Rotate + Filterer KG with HPO re-added (MNR = 0.80 and 0.88). While MNR gives a high-level view of how these models perform on all edges, realistically, only edges ranked in the top 100 would ever be considered for further research and evaluation, which equates to > 0.99 normalized rank. We found that when looking at only the top-ranked k edges (up to k = 100), RotatE trained on the filtered KG with HPO re-added substantially outperformed the full KG model at predicting rare disease edges (7.7% of edges were rediscovered a k=100 as opposed to 5.6%) despite having a lowe MNR (Figure 6).

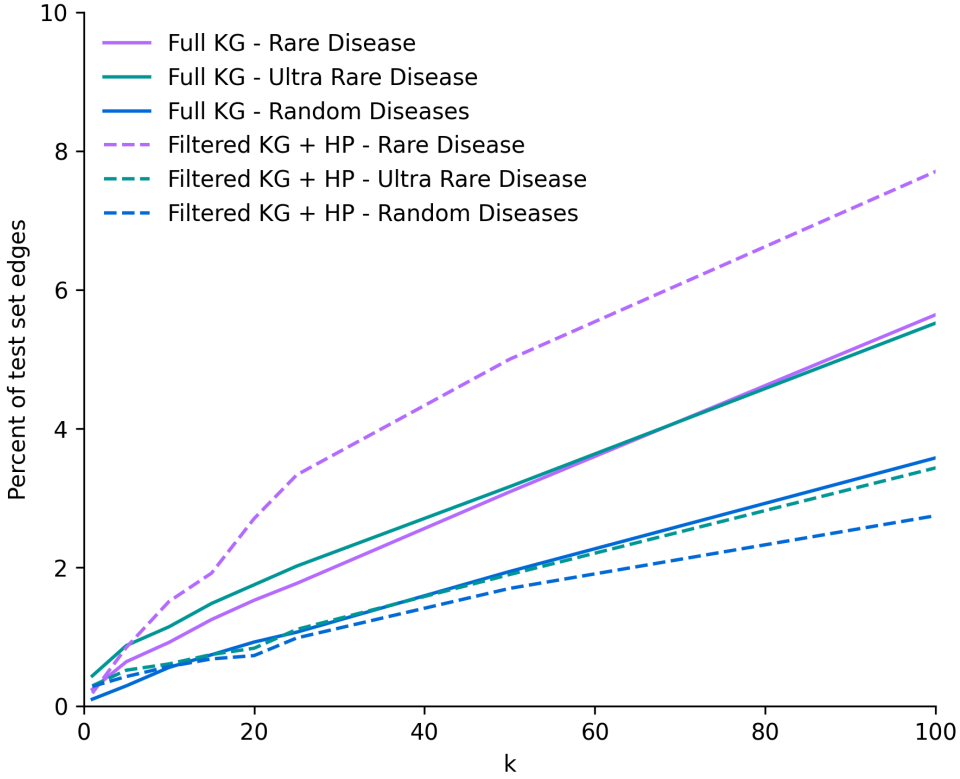

Figure 6: Percent of test edges ranked in the top k. Dashed lines are the results for Rotate trained on the filtered KG with HPO re-added, solid lines are from RotatE trained on the full KG.

Discussion

A small set of genes receives disproportionate attention in the scientific literature (Stoeger et al. 2018). This study bias results in topological imbalance (Lucchetta et al. 2023), creating degree-biased predictive models (Bonner, Kirik, et al. 2022). We have shown here that these degree-biased models struggle to prioritize understudied entities like rare diseases correctly; however, their performance and run time can be improved by filtering the KG. Relying on the predictions of degree-biased KGE models can give priority to studying already well-characterized genes and restarting the bias cycle.

Expecting understudied genes and diseases to coincide with common cases of health inequity, we investigated if there was a disparity in the predictive performance between male and female differentially expressed genes and diseases caused by ancestry-specific variants in European, African, East Asian, and Latino ancestral populations. We did not find evidence of disparities in KGE link prediction performance between female or male differentially expressed genes nor between diseases caused by European, African, East Asian, or Latino ancestry-specific variants (Supplemental Figure 1). However, we found that all of these groups underperform in KGE link prediction tasks compared to well-studied topics like genes related to heritable cancer genes. But, we do note a substantial difference in the number of annotations available to these ancestry-specific variants, with European ancestry having 2.5, 2.2, and 1.6 times as many annotations compared to more than the African, East Asian, or Latino ancestry groups, respectively.

The untested assumption of KG construction seems to have been that bigger is better, that aggregating more types of information and more edges is beneficial. We have shown that to be false.

Predictive performance improved when a drastically filtered version of the KG was used, and the time required to optimize and train the model was reduced to only a fraction of the original time. Training a KGE model on a large KG like Monarch KG can be prohibitively resource intensive; our RotatE + full Monarch KG took >30 days to complete hyperparameter optimization running on a dedicated NVIDIA a100 80G GPU. We were fortunate to have access to this resource through a shared compute cluster since to use a Google Compute Engine with similar specifications would cost an estimated $3,701.22 per month, and since we trained and compared 40 models. Careful filtration of KGs will improve predictive performance and the amount of time spent waiting to have a model to use and greatly reduce the costs associated with such research.

Our results show that the quality and type of information included in a KG are more important for KGE link prediction performance. In our metanode re-addition experiments, we found that the information that improved performance for disease-gene link prediction was not the source that added the most edges to the network; it was HPO, a source constructed via expert curation to improve our understanding of human diseases.

High-degree nodes from proportionally over-studied genes and diseases give a false sense of high performance in link prediction tasks (Zietz et al. 2024). To achieve a more realistic and representative measure of model performance, we need to independently measure classes and types of nodes that are masked when considering the KG as a whole. Models that cannot make accurate predictions about the sparse regions of the KG are of limited utility, especially regarding rare diseases. It is less important to know more about genes and diseases that are already well characterized with multiple therapeutics; rather we need reliable predictions about the genes and diseases we know little or nothing about and that have no approved therapeutics.

Compared to the $2.3 * 10^{18}$ possible combinations of the 61 remaining metanodes in Monarch KG, we explored very few. Yet we found that not all of the KG is beneficial in training and that better performance and speed can be achieved with a subset of metanodes. It may seem logical to expect graph-based learning models to learn to ignore superfluous graph structure, but that is not the case (Bechler-Speicher et al. 2023). Thus, methods for exploring the meta-node combination space to filter KGs appear to be an open and useful area of research.

Ratajczak et al. also explored KG filtration (Ratajczak et al. 2022) by removing nodes "that are counter productive" which reduced KG size by 60% and increased performance by 40.8%. However, this filtering process also caused a 65% increase in the average degree of the KG (increasing the average degree from 24 to 49), suggesting that many of these counterproductive nodes were of low degree. Suppose the primary aim of KGE link prediction is simply to develop a high-performance model or improve predictions regarding well-studied genes or diseases. In that case, this method reportedly excels at achieving those objectives. Still, if the goal is to predict less studied entities and advance research on new topics, our findings and those of other researchers suggest that models biased against these less connected nodes are actually hindering progress, not helping it (Zietz et al. 2024).

Topological imbalance and subsequent degree bias can be and have been addressed from many angles, including new computational tools (Erten et al. 2011) and training regimes (Eid et al. 2021). We have added to this body of knowledge, highlighting the importance of inspecting and curating the input data, the KGs themselves. It is unlikely that the best solution to degree bias will come from any one of these methodologies on their own (Sardina, Costabello, and Guéret 2024), but from integrating

many solutions addressing different aspects of the problem that we will achieve KGE-based link predictors capable of reliable variant prioritization on sparse regions of the KG such as rare diseases.

Key Points

- Knowledge graph embedding models exhibit node-degree bias that negatively affects rare diseases.
- Using a massively aggregate knowledge graph is non-optimal; better performance and runtime can be achieved with fewer types of information.
- The quality of information and edges added to a knowledge graph is more important than the quantity for improving predictive performance knowledge graph embedding models.
- The combination of metanodes or types of information used in training a knowledge graph embedding model is task-specific and should be chosen empirically.

Study funding

This work was supported by a grant from the Children's Hospital of Colorado awarded to RL.

Data Availability

No new data were generated or analyzed in support of this research. We used the 2023-12-16 version of Monarch KG: https://data.monarchinitiative.org/monarch-kg-dev/2023-12-16/index.html. The code used to generate our results can be found on Github: https://github.com/MSBradshaw/LinkPrediction.

References

Ali, Mehdi, Max Berrendorf, Charles Tapley Hoyt, Laurent Vermue, Sahand Sharifzadeh, Volker Tresp, and Jens Lehmann. 2020. "PyKEEN 1.0: A Python Library for Training and Evaluating Knowledge Graph Embeddings." *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2007.14175.

Bechler-Speicher, Maya, Ido Amos, Ran Gilad-Bachrach, and Amir Globerson. 2023. "Graph Neural Networks Use Graphs When They Shouldn't." *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2309.04332.

Bergstra, James, and Yoshua Bengio. 2012. "Random Search for Hyper-Parameter Optimization." *Journal of Machine Learning Research: JMLR* 13 (10): 281–305.

Bonner, Stephen, Ian P. Barrett, Cheng Ye, Rowan Swiers, Ola Engkvist, Charles Tapley Hoyt, and William L. Hamilton. 2022. "Understanding the Performance of Knowledge Graph Embeddings in Drug Discovery." *Artificial Intelligence in the Life Sciences* 2 (December): 100036.

Bonner, Stephen, Ufuk Kirik, Ola Engkvist, Jian Tang, and Ian P. Barrett. 2022. "Implications of Topological Imbalance for Representation Learning on Biomedical Knowledge Graphs." *Briefings in Bioinformatics* 23 (5). https://doi.org/10.1093/bib/bbac279.

Bordes, Antoine, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. 2013. "Translating Embeddings for Modeling Multi-Relational Data." In *Proceedings of the 26th*

International Conference on Neural Information Processing Systems - Volume 2, 2787–95. NIPS'13. Red Hook, NY, USA: Curran Associates Inc.

Boudellioua, Imane, Maxat Kulmanov, Paul N. Schofield, Georgios V. Gkoutos, and Robert Hoehndorf. 2019. "DeepPVP: Phenotype-Based Prioritization of Causative Variants Using Deep Learning." *BMC Bioinformatics* 20 (1): 65.

Cherry, J. Michael, Eurie L. Hong, Craig Amundsen, Rama Balakrishnan, Gail Binkley, Esther T. Chan, Karen R. Christie, et al. 2012. "Saccharomyces Genome Database: The Genomics Resource of Budding Yeast." *Nucleic Acids Research* 40 (Database issue): D700–705.

Cipriani, Valentina, Nikolas Pontikos, Gavin Arno, Panagiotis I. Sergouniotis, Eva Lenassi, Penpitcha Thawong, Daniel Danis, et al. 2020. "An Improved Phenotype-Driven Tool for Rare Mendelian Variant Prioritization: Benchmarking Exomiser on Real Patient Whole-Exome Data." *Genes* 11 (4). https://doi.org/10.3390/genes11040460.

Eid, Fatma-Elzahraa, Haitham A. Elmarakeby, Yujia Alina Chan, Nadine Fornelos, Mahmoud ElHefnawi, Eliezer M. Van Allen, Lenwood S. Heath, and Kasper Lage. 2021. "Systematic Auditing Is Essential to Debiasing Machine Learning in Biology." *Communications Biology* 4 (1): 183.

Erten, Sinan, Gurkan Bebek, Rob M. Ewing, and Mehmet Koyutürk. 2011. "DADA: Degree-Aware Algorithms for Network-Based Disease Gene Prioritization." *BioData Mining* 4 (June): 19.

Fey, Petra, Robert J. Dodson, Siddhartha Basu, and Rex L. Chisholm. 2013. "One Stop Shop for Everything Dictyostelium: dictyBase and the Dicty Stock Center in 2012." In *Dictyostelium Discoideum Protocols*, edited by Ludwig Eichinger and Francisco Rivero, 59–92. Totowa, NJ: Humana Press.

FORCE. n.d. "Table of Different Genes and the Hereditary Cancers That They Cause." Table of Different Genes and the Hereditary Cancers That They Cause. Accessed May 21, 2024. https://www.facingourrisk.org/info/hereditary-cancer-and-genetic-testing/genes-by-cancer-types.

Hastings, Janna, Gareth Owen, Adriano Dekker, Marcus Ennis, Namrata Kale, Venkatesh Muthukrishnan, Steve Turner, Neil Swainston, Pedro Mendes, and Christoph Steinbeck. 2016. "ChEBI in 2016: Improved Services and an Expanding Collection of Metabolites." *Nucleic Acids Research* 44 (D1): D1214–19.

Kustatscher, Georg, Tom Collins, Anne-Claude Gingras, Tiannan Guo, Henning Hermjakob, Trey Ideker, Kathryn S. Lilley, et al. 2022. "Understudied Proteins: Opportunities and Challenges for Functional Proteomics." *Nature Methods* 19 (7): 774–79.

Lucchetta, Marta, Markus List, David B. Blumenthal, and Martin H. Schaefer. 2023. "Emergence of Power-Law Distributions in Protein-Protein Interaction Networks through Study Bias." *bioRxiv*. https://doi.org/10.1101/2023.03.17.533165.

Luck, Katja, Dae-Kyum Kim, Luke Lambourne, Kerstin Spirohn, Bridget E. Begg, Wenting Bian, Ruth Brignall, et al. 2020. "A Reference Map of the Human Binary Protein Interactome." *Nature* 580 (7803): 402–8.

Mohamed, Sameh K., Aayah Nounu, and Vít Nováček. 2019. "Drug Target Discovery Using Knowledge Graph Embeddings." In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, 11–18. SAC '19. New York, NY, USA: Association for Computing Machinery.

———. 2021. "Biological Applications of Knowledge Graph Embedding Models." *Briefings in Bioinformatics* 22 (2): 1679–93.

Oliva, Meritxell, Manuel Muñoz-Aguirre, Sarah Kim-Hellmuth, Valentin Wucher, Ariel D. H. Gewirtz, Daniel J. Cotter, Princy Parsana, et al. 2020. "The Impact of Sex on Gene Expression across Human Tissues." *Science* 369 (6509). https://doi.org/10.1126/science.aba3066.

Prabhakar, Vignesh, Chau Vu, Jennifer Crawford, Joseph Waite, and Kai Liu. 2023. "An Ensemble Learning Approach to Perform Link Prediction on Large Scale Biomedical Knowledge Graphs for Drug Repurposing and Discovery." *bioRxiv*. https://doi.org/10.1101/2023.03.19.533306.

Putman, Tim E., Kevin Schaper, Nicolas Matentzoglu, Vincent P. Rubinetti, Faisal S. Alquaddoomi, Corey Cox, J. Harry Caufield, et al. 2024. "The Monarch Initiative in 2024: An Analytic Platform Integrating Phenotypes, Genes and Diseases across Species." *Nucleic Acids Research* 52 (D1): D938–49.

Ratajczak, Florin, Mitchell Joblin, Martin Ringsquandl, and Marcel Hildebrandt. 2022. “Task-Driven Knowledge Graph Filtering Improves Prioritizing Drugs for Repurposing.” *BMC Bioinformatics* 23 (1): 84.

Sanjak, Jaleal, Qian Zhu, and Ewy A. Mathé. 2023. “Clustering Rare Diseases within an Ontology-Enriched Knowledge Graph.” *bioRxiv*. https://doi.org/10.1101/2023.02.15.528673.

Sardina, Jeffrey, Luca Costabello, and Christophe Guéret. 2024. “Veni, Vidi, Vici: Solving the Myriad of Challenges before Knowledge Graph Learning.” *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2402.06098.

Stoeger, Thomas, Martin Gerlach, Richard I. Morimoto, and Luís A. Nunes Amaral. 2018. “Large-Scale Investigation of the Reasons Why Potentially Important Genes Are Ignored.” *PLoS Biology* 16 (9): e2006643.

Sun, Zhiqing, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. “RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1902.10197.

Szklarczyk, Damian, Annika L. Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, et al. 2019. “STRING v11: Protein-Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-Wide Experimental Datasets.” *Nucleic Acids Research* 47 (January): D607–13.

Trouillon, Théo, Johannes Welbl, Sebastian Riedel, Eric Gaussier, and Guillaume Bouchard. 20--22 Jun 2016. “Complex Embeddings for Simple Link Prediction.” In *Proceedings of The 33rd International Conference on Machine Learning*, edited by Maria Florina Balcan and Kilian Q. Weinberger, 48:2071–80. Proceedings of Machine Learning Research. New York, New York, USA: PMLR.

Vilela, Joana, Muhammad Asif, Ana Rita Marques, João Xavier Santos, Célia Rasga, Astrid Vicente, and Hugo Martiniano. 2023. “Biomedical Knowledge Graph Embeddings for Personalized Medicine: Predicting Disease‐gene Associations.” *Expert Systems* 40 (5). https://doi.org/10.1111/exsy.13181.

Yılmaz, Serhan, Kaan Yorgancioglu, and Mehmet Koyutürk. 2023. “Are under-Studied Proteins under-Represented? How to Fairly Evaluate Link Prediction Algorithms in Network Biology.” *bioRxiv*. https://doi.org/10.1101/2022.10.13.511953.

Zietz, Michael, Daniel S. Himmelstein, Kyle Kloster, Christopher Williams, Michael W. Nagle, and Casey S. Greene. 2024. “The Probability of Edge Existence due to Node Degree: A Baseline for Network-Based Predictions.” *GigaScience* 13 (January). https://doi.org/10.1093/gigascience/giae001.

Zitnik, Marinka, Monica Agrawal, and Jure Leskovec. 2018. “Modeling Polypharmacy Side Effects with Graph Convolutional Networks.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1802.00543.