Expanding a Database-derived Biomedical Knowledge Graph via Multi-relation Extraction from Biomedical Abstracts

A DOI-citable version of this manuscript is available at https://doi.org/10.1101/730085.

This manuscript (permalink) was automatically generated from greenelab/text_mined_hetnet_manuscript@855f0b5 on April 25, 2022.

Authors

- **David N. Nicholson**
  0000-0003-0002-5761 · danich1
  Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania · Funded by GBMF4552

- **Daniel S. Himmelstein**
  0000-0002-3012-7446 · dhimmel · dhimmel
  Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania · Funded by GBMF4552

- **Casey S. Greene**
  0000-0001-8713-9213 · cgreene · GreeneScientist
  Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania · Funded by GBMF4552 and R01 HG010067

Abstract

Knowledge graphs support multiple research efforts by providing contextual information for biomedical entities, constructing networks, and supporting the interpretation of high-throughput analyses. These databases are populated via some form of manual curation, which is difficult to scale in the context of an increasing publication rate. Data programming is a paradigm that circumvents this arduous manual process by combining databases with simple rules and heuristics written as label functions, which are programs designed to automatically annotate textual data. Unfortunately, writing a useful label function requires substantial error analysis and is a nontrivial task that takes multiple days per function. This makes populating a knowledge graph with multiple nodes and edge types practically infeasible. We sought to accelerate the label function creation process by evaluating the extent to which label functions could be re-used across multiple edge types. We used a subset of an existing knowledge graph centered on disease, compound, and gene entities to evaluate label function re-use. We determined the best label function combination by comparing a baseline database-only model with the same model but added edge-specific or edge-mismatch label functions. We confirmed that adding additional edge-specific rather than edge-mismatch label functions often improves text annotation and shows that this approach can incorporate novel edges into our source knowledge graph. We expect that continued development of this strategy has the potential to swiftly populate knowledge graphs with new discoveries, ensuring that these resources include cutting-edge results.

Introduction

Knowledge bases are essential resources that hold complex structured and unstructured information. These resources have been used to construct networks for drug repurposing discovery [1,2,3] or as a source of training labels for text mining systems [4,5,6]. Populating knowledge bases often requires highly trained scientists to read biomedical literature and summarize the results through manual curation [7]. In 2007, researchers estimated that filling a knowledge base via manual curation would require approximately 8.4 years to complete [8]. As the rate of publications increases exponentially [9], using only manual curation to populate a knowledge base has become nearly impractical.

Relationship extraction is one of several solutions to the challenge posed by an exponentially growing body of literature [7]. This process creates an expert system to automatically scan, detect, and extract relationships from textual sources. These expert systems fall into three types: unsupervised, rule-based, and supervised systems.

Unsupervised systems extract relationships without the need for annotated text. These approaches utilize linguistic patterns such as the frequency of two entities appearing in a sentence together more often than chance, commonly referred to as co-occurrence [10,11,12,13,14,15,16,17,18]. For example, a possible system would say gene X is associated with disease Y because gene X and disease Y appear together more often than chance [10]. Besides frequency, other systems can utilize grammatical structure to identify relationships [19]. This information is modeled in the form of a tree data structure, termed a dependency tree. Dependency trees depict words as nodes, and edges represent a word's grammatical relationship with one another. Through clustering on these generated trees, one can identify patterns that indicate a biomedical relationship [19]. Unsupervised systems are desirable since they do not require well-annotated training data; however, precision may be limited compared to supervised machine learning systems.

Rule-based systems rely heavily on expert knowledge to perform relationship extraction. These systems use linguistic rules and heuristics to identify critical sentences or phrases that suggest the presence of a biomedical relationship [20,21,22,23,24,25]. For example, a hypothetical extractor

focused on protein phosphorylation events would identify sentences containing the phrase “gene X phosphorylates gene Y” [20]. These approaches provide exact results, but the quantity of positive results remains modest as sentences consistently change in form and structure. For this project, we constructed our label functions without the aid of these works; however, the approaches mentioned in this section provide substantial inspiration for novel label functions in future endeavors.

Supervised systems depend on machine learning classifiers to predict the existence of a relationship using biomedical text as input. These classifiers can range from linear methods such as support vector machines [26,27] to deep learning [28,29,30,31,32,33], which all require access to well-annotated datasets. Typically, these datasets are usually constructed via manual curation by individual scientists [34,35,36,37,38] or through community-based efforts [39,40,41]. Often, these datasets are well annotated but are modest in size, making model training hard as these algorithms become increasingly complex.

Distant supervision is a paradigm that quickly sidesteps manual curation to generate large training datasets. This technique assumes that positive examples have been previously established in selected databases, implying that the corresponding sentences or data points are also positive [4]. The central problem with this technique is that generated labels are often of low quality, resulting in many false positives [42]. Despite this caveat there have been notable effort using this technique [43,44,45].

Data programming is one proposed solution to amend the false positive problem in distant supervision. This strategy combines labels obtained from distant supervision with simple rules and heuristics written as small programs called label functions [46]. These outputs are consolidated via a noise-aware model to produce training labels for large datasets. Using this paradigm can dramatically reduce the time required to obtain sufficient training data; however, writing a helpful label function requires substantial time and error analysis. This dependency makes constructing a knowledge base with a myriad of heterogenous relationships nearly impossible as tens or hundreds of label functions are necessary per relationship type.

This paper seeks to accelerate the label function creation process by measuring how label functions can be reused across different relationship types. We hypothesized that sentences describing one relationship type might share linguistic features such as keywords or sentence structure with sentences describing other relationship types. If this hypothesis were to, one could drastically reduce the time needed to build a relation extractor system and swiftly populate large databases like Hetionet v1. We conducted a series of experiments to estimate how label function reuse enhances performance over distant supervision alone. We focused on relationships that indicated similar types of physical interactions (i.e., gene-binds-gene and compound-binds-gene) and two more distinct types (i.e., disease-associates-gene and compound-treats-disease).

Methods and Materials

Hetionet

Hetionet v1 [3] is a heterogeneous network that contains pharmacological and biological information. This network depicts information in the form of nodes and edges of different types. Nodes in this network represent biological and pharmacological entities, while edges represent relationships between entities. Hetionet v1 contains 47,031 nodes with 11 different data types and 2,250,197 edges that represent 24 different relationship types (Figure 1). Edges in Hetionet v1 were obtained from open databases, such as the GWAS Catalog [47], Human Interaction database [48] and DrugBank [49]. For this project, we analyzed performance over a subset of the Hetionet v1 edge types: disease associates with a gene (DaG), compound binds to a gene (CbG), compound treating a disease (CtD), and gene interacts with gene (GiG) (bolded in Figure 1).

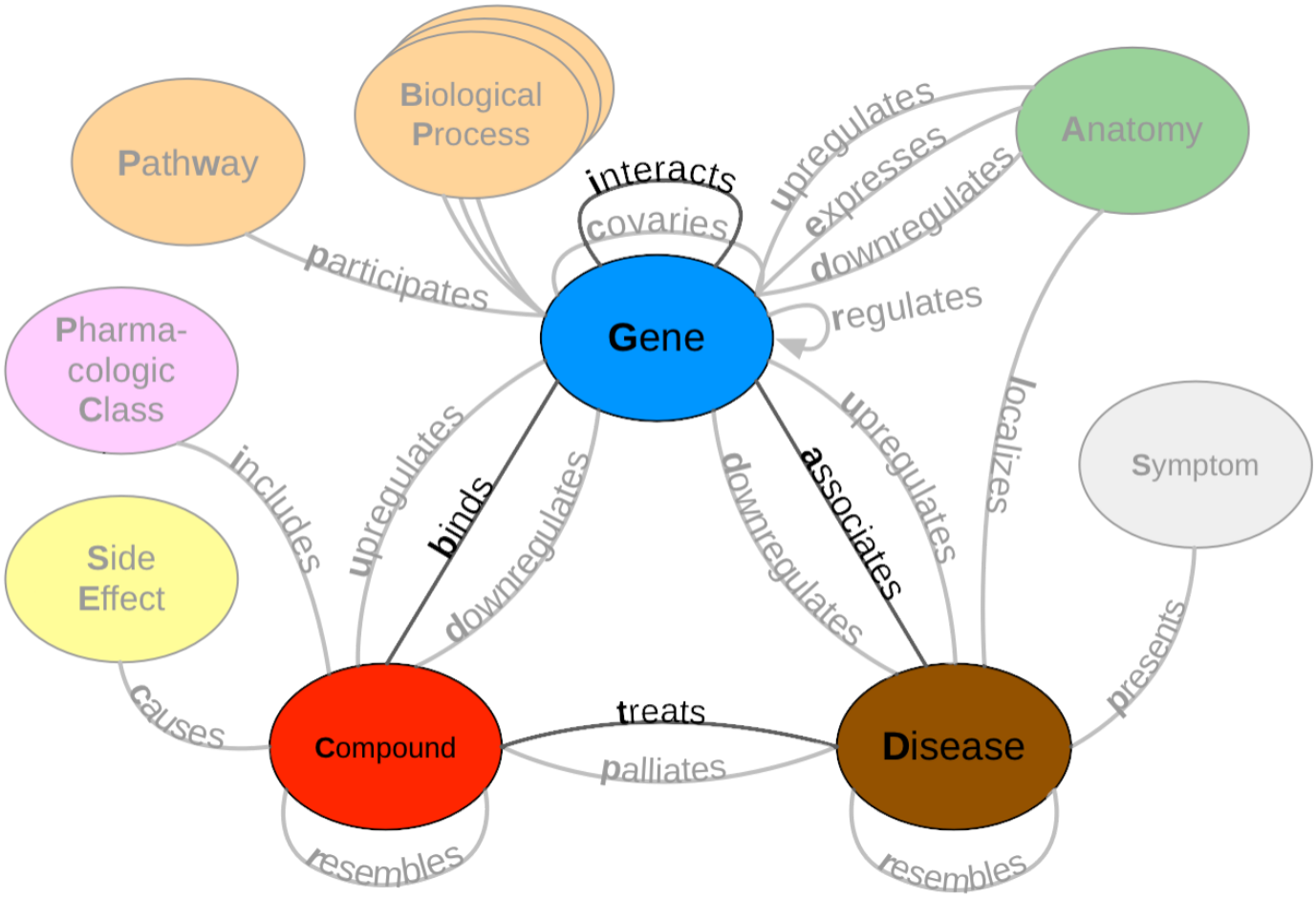

Figure 1: A metagraph (schema) of Hetionet v1 where biomedical entities are represented as nodes and the relationships between them are represented as edges. We examined performance on the highlighted subgraph; however, the long-term vision is to capture edges for the entire graph.

Dataset

We used PubTator Central [50] as input to our analysis. PubTator Central provides MEDLINE abstracts that have been annotated with well-established entity recognition tools including Tagger One [51] for disease, chemical and cell line entities, tmVar [52] for genetic variation tagging, GNormPlus [53] for gene entities and SR4GN [54] for species entities. We downloaded PubTator Central on March 1, 2020, at which point it contained approximately 30,000,000 documents. After downloading, we filtered out annotated entities that were not contained in Hetionet v1. We extracted sentences with two or more annotations and termed these sentences as candidate sentences. We used the Spacy's English natural language processing (NLP) pipeline (en_core_web_sm) [55] to generate dependency trees and parts of speech tags for every extracted candidate sentence. Each candidate sentence was stratified by their corresponding abstract ID to produce a training set, tuning set, and a testing set. We used random assortment to assign dataset labels to each abstract. Every abstract had a 70% chance of being labeled training, 20% chance of being labeled tuning, and 10% chance of being labeled testing. Despite the power of data programming, all text mining systems need to have ground truth labels to be well-calibrated. We hand-labeled five hundred to a thousand candidate sentences of each edge type to obtain a ground truth set (Table 1).

Table 1: Statistics of Candidate Sentences. We sorted each abstract into a training, tuning and testing set. Numbers in parentheses show the number of positives and negatives that resulted from the hand-labeling process.

Relationship	Train	Tune	Test
Disease Associates Gene	2.49 M	696K (397+, 603-)	348K (351+, 649-)

Relationship	Train	Tune	Test
Compound Binds Gene	2.4M	684K (37+, 463-)	341k (31+, 469-)
Compound Treats Disease	1.5M	441K (96+, 404-)	223K (112+, 388-)
Gene Interacts Gene	11.2M	2.19M (60+, 440-)	1.62M (76+, 424-)

Label Functions for Annotating Sentences

The challenge of having too few ground truth annotations is familiar to many natural language processing applications, even when unannotated text is abundant. Data programming circumvents this issue by quickly annotating large datasets using multiple noisy signals emitted by label functions [46]. Label functions are simple pythonic functions that emit: a positive label (1), a negative label (0), or abstain from emitting a label (-1). These functions can use different approaches or techniques to emit a label; however, these functions can be grouped into simple categories discussed below. Once constructed, these functions are combined using a generative model to output a single annotation. This single annotation is a consensus probability score bounded between 0 (low chance of mentioning a relationship) and 1 (high chance of mentioning a relationship). We used these annotations to train a discriminative model for the final classification step.

Label Function Categories

Label functions can be constructed in various ways; however, they also share similar characteristics. We grouped functions into databases and text patterns. The majority of our label functions fall into the text pattern category (Supplemental Table 2). Further, we described each label function category and provided an example that refers to the following candidate sentence: "PTK6 may be a novel therapeutic target for pancreatic cancer".

Databases: These label functions incorporate existing databases to generate a signal, as seen in distant supervision [4]. These functions detect if a candidate sentence's co-mention pair is present in a given database. Our label function emits a positive label if the pair is present and abstains otherwise. If the pair is not present in any existing database, a separate label function emits a negative label. We used a separate label function to prevent a label imbalance problem, which can occur when a single function labels every possible sentence despite being correct or not. If this problem isn't handled correctly, the generative model could become biased and only emit one prediction (solely positive or solely negative) for every sentence.

$$\Lambda_{DB}(D, G) = \begin{cases} 1 & (D, G) \in DB \\ 0 & otherwise \end{cases}$$

$$\Lambda_{\neg DB}(D, G) = \begin{cases} -1 & (D, G) \notin DB \\ 0 & otherwise \end{cases}$$

Text Patterns: These label functions are designed to use keywords or sentence context to generate a signal. For example, a label function could focus on the number of words between two mentions and emit a label if two mentions are too close. Alternatively, a label function could focus on the parts of speech contained within a sentence and ensures a verb is present. Besides parts of speech, a label function could exploit dependency parse trees to emit a label. These trees are akin to the tree data structure where words are nodes and edges are how each word modifies each other. Label functions that use these parse trees will test if the generated tree matches a pattern and emits a positive label if true. For our analysis, we used previously identified patterns designed for biomedical text to generate our label functions [tag:global_network?].

$$\Lambda_{TP}(D, G) = \begin{cases} 1 & \text{" } target \text{ "} \in Candidate\ Sentence \\ -1 & otherwise \end{cases}$$

$$\Lambda_{TP}(D, G) = \begin{cases} 0 & \text{" } VB \text{ "} \notin pos\_tags(Candidate\ Sentence) \\ -1 & otherwise \end{cases}$$

$$\Lambda_{TP}(D, G) = \begin{cases} 1 & dep(Candidate\ Sentence) \in Cluster\ Theme \\ -1 & otherwise \end{cases}$$

Each text pattern label function was constructed via manual examination of sentences within the training set. For example, using the candidate sentence above, one would identify the phrase “novel therapeutic target” and incorporate this phrase into a global list that a label function would use to check if present in a sentence. After initial construction, we tested and augmented the label function using sentences in the tune set. We repeated this process for every label function in our repertoire.

Table 2: The distribution of each label function per relationship.

Relationship	Databases (DB)	Text Patterns (TP)	
DaG	7	30	
CtD	3	22	
CbG	9	20	
GiG	9	28	

Training Models

Generative Model

The generative model is a core part of this automatic annotation framework. It integrates multiple signals emitted by label functions to assign each candidate sentence the most appropriate training class. This model takes as input a label function output in the form of a matrix where rows represent candidate sentences, and columns represent each label function (Λ^{nxm}). Once constructed, this model treats the true training class (Y) as a latent variable and assumes that each label function is independent of one another. Under these two assumptions, the model finds the optimal parameters by minimizing a loglikelihood function marginalized over the latent training class.

$$\hat{\theta} = argmin_{\theta} \sum_{Y} -log(P_{\theta}(\Lambda, Y))$$

Following optimization, the model emits a probability estimate that each sentence belongs to the positive training class. At this step, each probability estimate can be discretized via a chosen threshold into a positive or negative class. We used a threshold of 0.5 for discretizing our training classes within our analysis. For more information on how the likelihood function is constructed and minimized, refer to [56].

Discriminative Model

The discriminative model is the final step in this framework. This model uses training labels generated from the generative model combined with sentence features to classify the presence of a biomedical relationship. Typically, the discriminative model is a neural network. We used BioBERT [31], a BERT

[57] model trained on all papers and abstracts within Pubmed Central [58], as our discriminative model. BioBERT provides its own set of word embeddings, dense vectors representing words that models such as neural networks can use to construct sentence features. We downloaded a pre-trained version of this model using huggingface's transformer python package [59] and fine-tuned it using our generated training labels. Our fine-tuning approach involved freezing all downstream layers except for the classification head of this model. Next, we trained this model for 10 epochs using the Adam optimizer [60] with huggingface's default parameter settings and a learning rate of 0.001.

Experimental Design

Reusing label functions across edge types would substantially reduce the number of label functions required to extract multiple relationships from biomedical literature. We first established a baseline by training a generative model using only distant supervision label functions designed for the target edge type (see Supplemental Methods). Then we compared the baseline model with models that incorporated a set number of text pattern label functions. Using a sampling with replacement approach, we sampled these text pattern label functions from three different groups: within edge types, across edge types, and from a pool of all label functions. We compared within-edge-type performance to across-edge-type and all-edge-type performance. We sampled a fixed number of label functions for each edge type consisting of five evenly spaced numbers between one and the total number of possible label functions. We repeated this sampling process 50 times for each point. Furthermore, we also trained the discriminative model using annotations from the generative model trained on edge-specific label functions at each point. We report the performance of both models in terms of the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPR). Ensuing model evaluations, we quantified the number of edges we could incorporate into Hetionet v1. We used our best performing discriminative model to score every candidate sentence within our dataset and grouped candidates based on their mention pair. We took the max score within each candidate group, and this score represents the probability of the existence of an edge. We established edges using a cutoff score that produced an equal error rate between the false positives and false negatives. Lastly, we report the number of preexisting edges we could recall and the number of novel edges we can incorporate.

Results

Generative Model Using Randomly Sampled Label Functions

Creating label functions is a labor-intensive process that can take days to accomplish. We sought to accelerate this process by measuring how well label functions can be reused. We evaluated this by performing an experiment where label functions are sampled on an individual (edge vs. edge) level and a global (collective pool of sources) level. We observed that performance increased when edge-specific label functions were added to an edge-specific baseline model, while label function reuse usually provided less benefit (AUROC Figure 2, AUPR Supplemental Figure 6). The quintessential example of this overarching trend is the Compound-treats-Disease (CtD) edge type, where edge-specific label functions consistently outperformed transferred label functions. However, there is evidence that label function transferability may be feasible for selected edge types and label function sources. Performance increases as more Gene-interacts-Gene (GiG) label functions are incorporated into the Compound-binds-Gene (CbG) baseline model and vice versa. This trend suggests that sentences for GiG and CbG may share similar linguistic features or terminology that allows for label functions to be reused, which could relate to both describing physical interaction relationships. Perplexingly, edge-specific Disease-associates-Gene (DaG) label functions did not improve performance over label functions drawn from other edge types. Overall, only CbG and GiG showed significant signs of reusability. This pattern suggests that label function transferability may be possible for these two edge types.

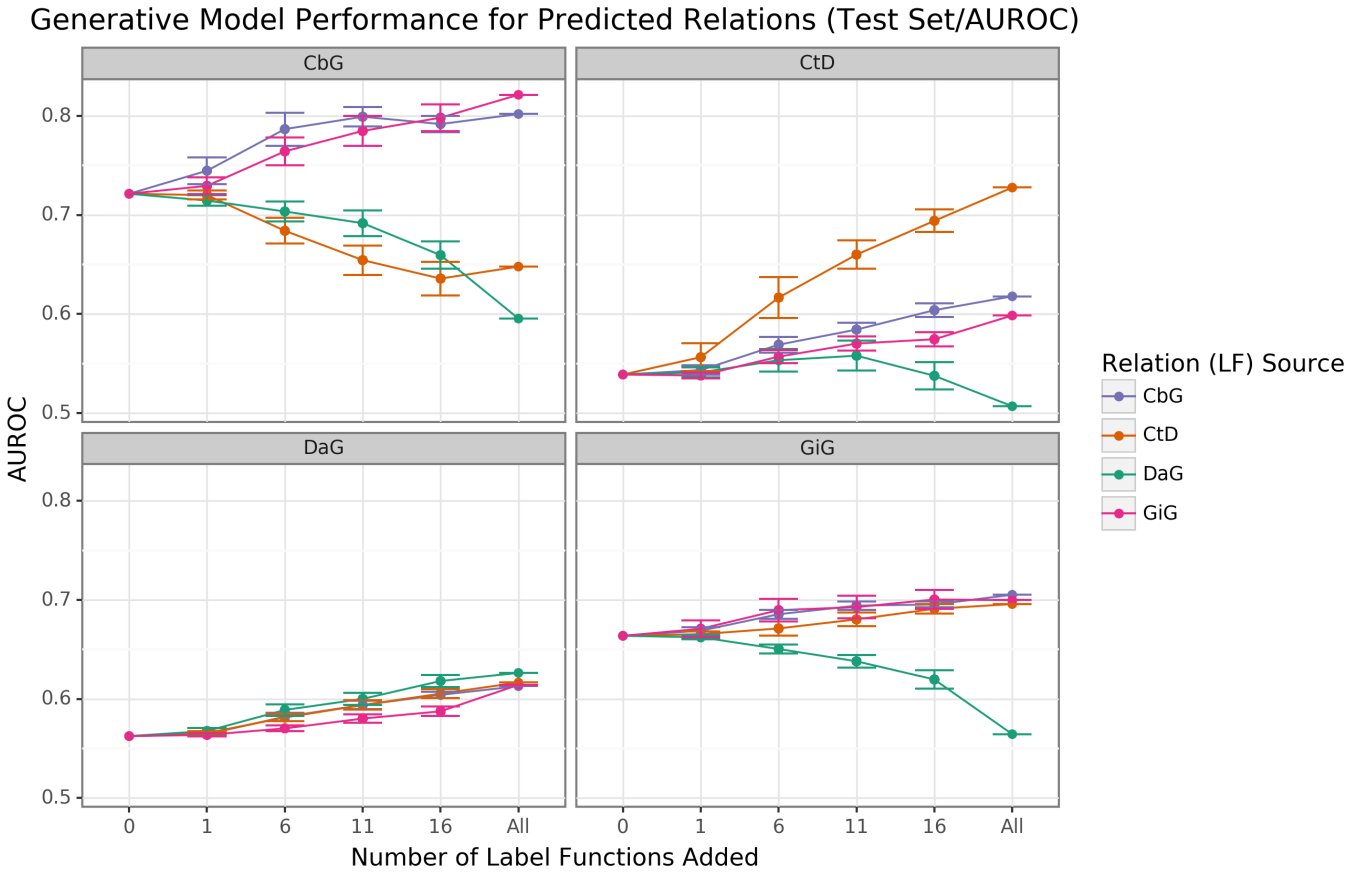

Figure 2: Edge-specific label functions perform better than edge-mismatch label functions, but certain mismatch situations show signs of successful transfer. Each line plot header depicts the edge type the generative model is trying to predict, while the colors represent the source of label functions. For example, orange represents sampling label functions designed to predict the Compound-treats-Disease (CtD) edge type. The x-axis shows the number of randomly sampled label functions incorporated as an addition to the database-only baseline model (the point at 0). The y-axis shows the area under the receiver operating curve (AUROC). Each point on the plot shows the average of 50 sample runs, while the error bars show the 95% confidence intervals of all runs. The baseline and "All" data points consist of sampling from the entire fixed set of label functions.

We found that sampling from all label function sources at once usually underperformed relative to edge-specific label functions (Figure 3 and Supplemental Figure 7). The gap between edge-specific sources and all sources widened as we sampled more label functions. CbG is a prime example of this trend (Figure 3 and Supplemental Figure 7), while CtD and GiG show a similar but milder trend. DaG was the exception to the general rule. The pooled set of label functions improved performance over the edge-specific ones, which aligns with the previously observed results for individual edge types (Figure 2). When pooling all label functions, the decreasing trend supports the notion that label functions cannot simply transfer between edge types (exception being CbG on GiG and vice versa).

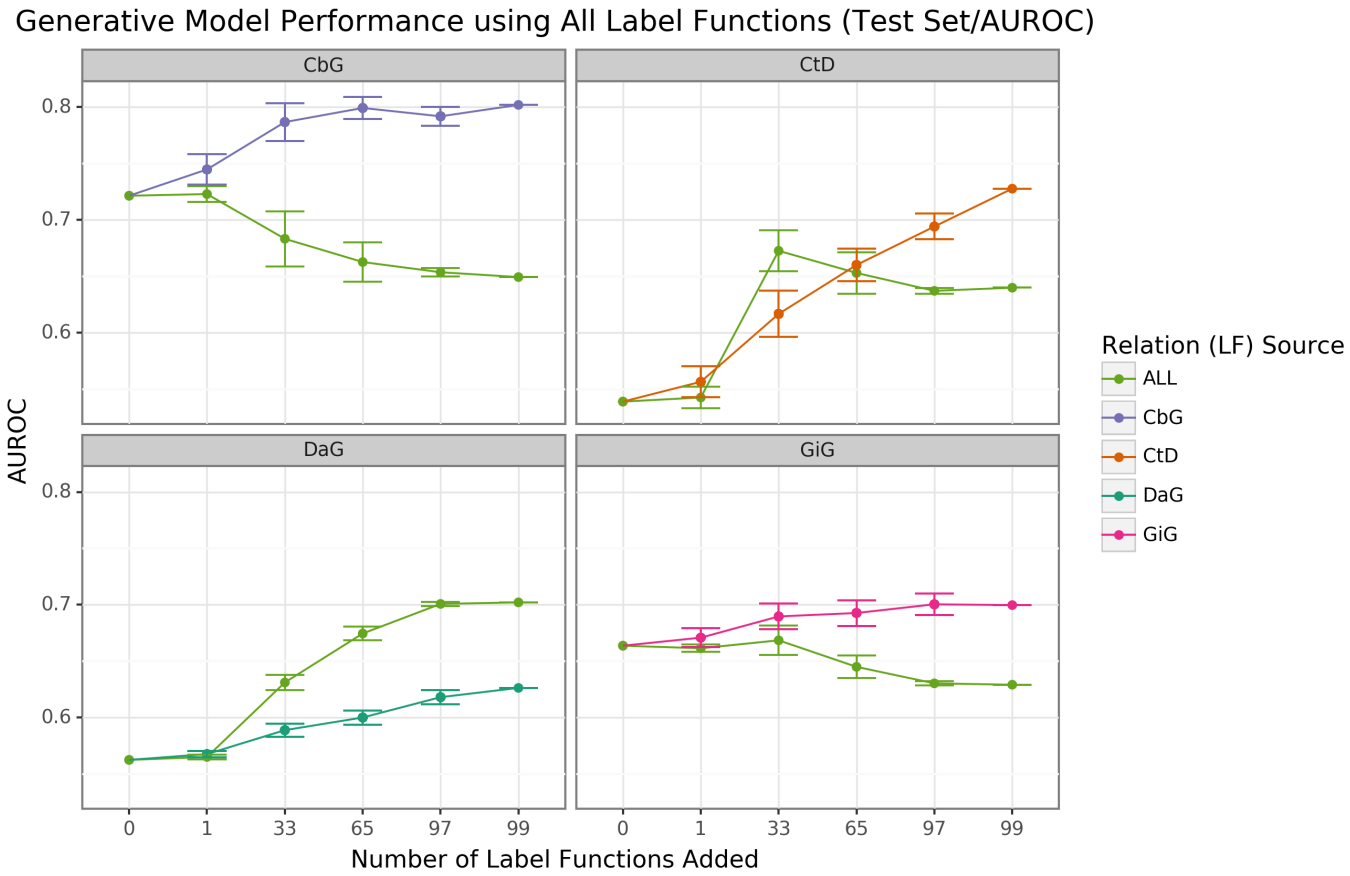

Figure 3: Using all label functions generally hinders generative model performance. Each line plot header depicts the edge type the generative model is trying to predict, while the colors represent the source of label functions. For example, orange represents sampling label functions designed to predict the Compound-treats-Disease (CtD) edge type. The x-axis shows the number of randomly sampled label functions incorporated as an addition to the database-only baseline model (the point at 0). The y-axis shows the area under the receiver operating curve (AUROC). Each point on the plot shows the average of 50 sample runs, while the error bars show the 95% confidence intervals of all runs. The baseline and "All" data points consist of sampling from the entire fixed set of label functions.

Discriminative Model Performance

The discriminative model is intended to augment performance over the generative model by incorporating textual features together with estimated training labels. We found that the discriminative model generally outperformed the generative model with respect to AUROC as more edge-specific label functions were incorporated (Figure 4). Regarding AUPR, this model outperformed the generative model for the Disease-associates-Gene (DaG) edge type. At the same time, it had close to par performance for the rest of the edge types (Supplemental Figure 8). The discriminative model's performance was often poorest when very few edge-specific label functions were incorporated into the baseline model (seen in DaG, CbG, and Gene-interacts-Gene (GiG)). This example suggests that training generative models with more label functions produces better outputs for training for discriminative models. Compound-treats-Disease (CtD) was an exception to this trend, where the discriminative model outperformed the generative model at all sampling levels in regards to AUROC. We observed the opposite trend with the Compound-binds-Gene (CbG) edges as the discriminative model was always worse or indistinguishable from the generative model. Interestingly, the AUPR for CbG plateaus below the generative model and decreases when all edge-specific label functions are used (Supplemental Figure 8). This trend suggests that the discriminative model might have predicted more false positives in this setting. Overall, incorporating more edge-specific label functions usually improved performance for the discriminative model over the generative model.

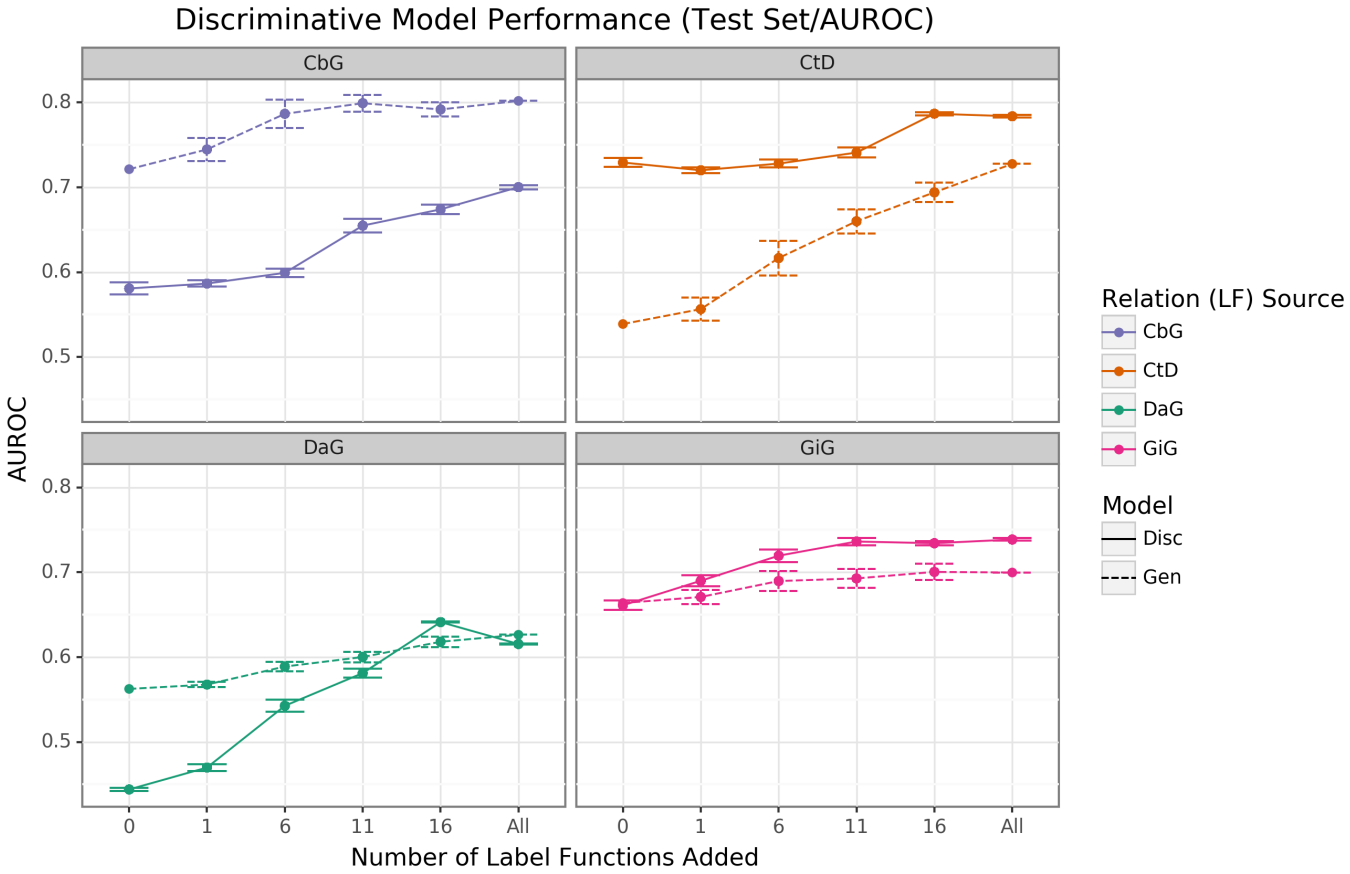

Figure 4: The discriminative model usually improves faster than the generative model as more edge-specific label functions are included. The line plot headers represent the specific edge type the discriminative model is trying to predict. The x-axis shows the number of randomly sampled label functions incorporated as an addition to the baseline model (the point at 0). The y axis shows the area under the receiver operating curve (AUROC). Each data point represents the average of 3 sample runs for the discriminator model and 50 sample runs for the generative model. The error bars represent each run's 95% confidence interval. The baseline and "All" data points consist of sampling from the entire fixed set of label functions.

Text Mined Edges Can Expand a Database-derived Knowledge Graph

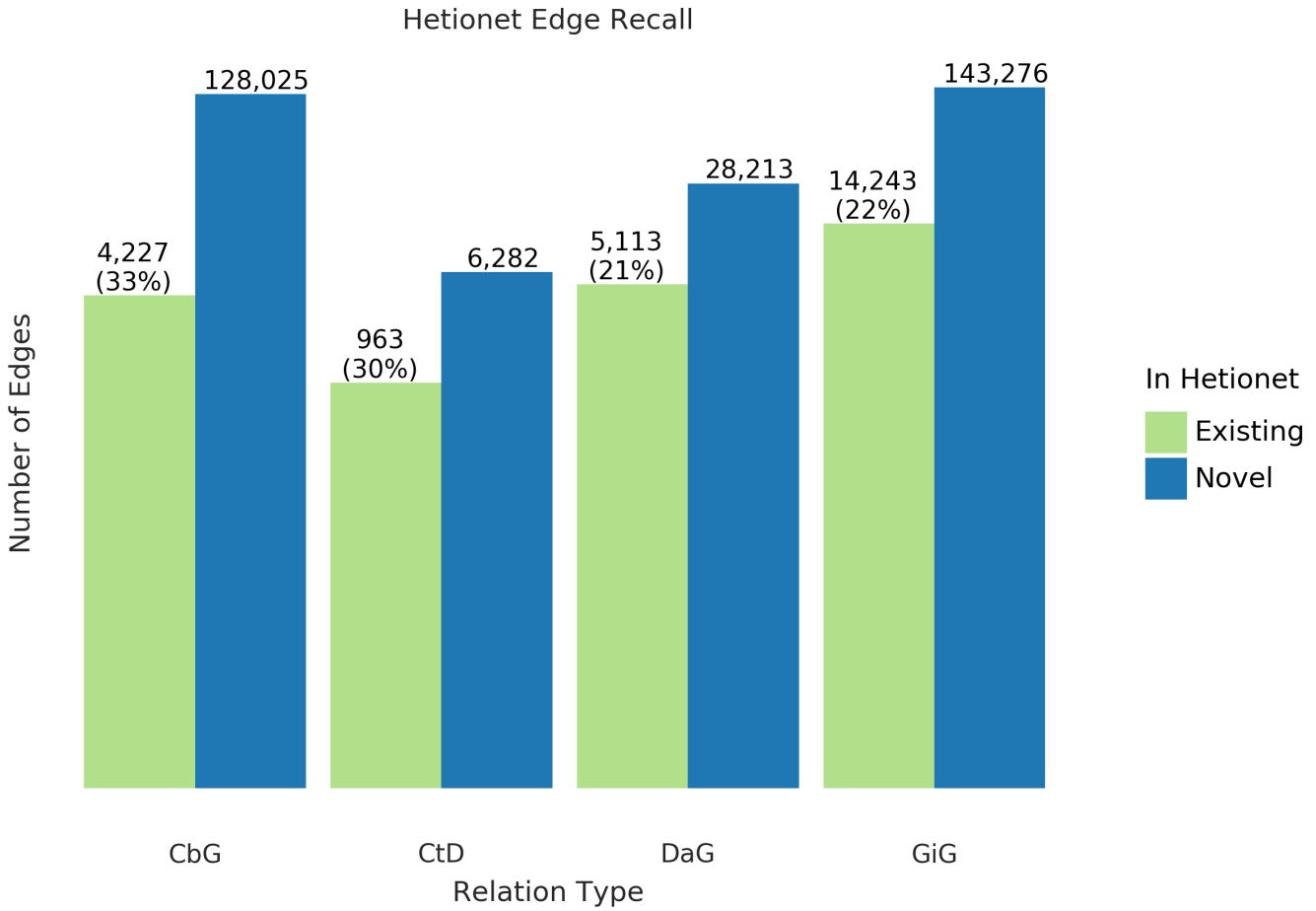

Figure 5: Text-mined edges recreate a substantial fraction of an existing knowledge graph and include new predictions. This bar chart shows the number of edges we can successfully recall in green and indicates the number of new edges in blue.
The recall for the Hetionet v1 knowledge graph is shown as a percentage in parentheses. For example, for the Compound-treats-Disease (CtD) edge, our method recalls 30% of existing edges and can add 6,282 new ones.

One of the goals of our work is to measure the extent to which learning multiple edge types could construct a biomedical knowledge graph. Using Hetionet v1 as an evaluation set, we measured this framework's recall and quantified the number of edges that may be incorporated with high confidence. Overall, we were able to recall about thirty percent of the preexisting edges for all edge types (Figure 5) and report our top ten scoring sentences for each edge type in Supplemental Table 3. Our best recall was with the Compound-binds-Gene (CbG) edge type, where we retained 33% of preexisting edges. In contrast, we only recalled close to 30% for Compound-treats-Disease (CtD), while the other two categories achieved a recall score close to 22%. Despite the modest recall level, the amount of novel edge types remains elevated. This notion highlights that Hetionet v1 is missing a compelling amount of biomedical information, and relationship extraction is a viable way to close the information gap.

Discussion

We measured the extent to which label functions can be re-used across multiple edge types to extract relationships from literature. Through our sampling experiment, we found that adding edge-specific label functions increases performance for the generative model (Figure 2). We found that label functions designed from relatively related edge types can increase performance (Gene interacts Gene (GiG) label functions predicting the Compound binds Gene (CbG) edge and vice versa), while the Disease associates Gene (DaG) edge type remained agnostic to label function sources (Figure 2 and Supplemental Figure 6). Furthermore, we found that using all label functions at once generally hurts performance with the exception being the DaG edge type (Supplemental Figures 3 and 7). One

possibility for this observation is that DaG is a broadly defined edge type. For example, DaG may contain many concepts related to other edge types such as Disease (up/down) regulating a Gene, which makes it more agnostic to label function sources (examples highlighted in our annotated sentences).

Regarding the discriminative model, adding edge-specific label function substantially improved performance for two out of the four edge types (Compound treats Disease (CtD) and Disease associates Gene (DaG)) (Figure 4 and Supplemental Figure 8). Gene interacts Gene (GiG) and Compound binds Gene (CbG) discriminative models showed minor improvements compared to the generative model, but only when nearly all edge-specific label functions are included (Figure 4 and Supplemental Figure 8). We came across a large amount of spurious gene mentions when working with the discriminative model and believe that these mentions contributed to CbG and GiG's hindered performance. We encountered difficulty in calibrating each discriminative model (Supplemental Figure ??). The temperature scaling algorithm appears to improve calibration for the highest scores for each model but did not successfully calibrate throughout the entire range of predictions. Improving performance for all predictions may require more labeled examples or may be a limitation of the approach in this setting. Even with these limitations, this early-stage approach could recall many existing edges from an existing knowledge base, Hetionet v1, and suggest many new high-confidence edges for inclusion (Supplemental Figure 5). Our findings suggest that further work, including an expansion of edge types and a move to full text from abstracts, may make this approach suitable for building continuously updated knowledge bases to address drug repositioning and other biomedical challenges.

Conclusion and Future Direction

Filling out knowledge bases via manual curation can be an arduous and erroneous task [8]. As the rate of publications increases, relying on manual curation alone becomes impractical. Data programming, a paradigm that uses label functions as a means to speed up the annotation process, can be used as a solution for this problem. An obstacle for this paradigm, however, is creating useful label functions, which takes a considerable amount of time. We tested the feasibility of reusing label functions as a way to reduce the total number of label functions required for strong prediction performance. We conclude that label functions may be re-used with closely related edge types, but that re-use does not improve performance for most pairings. The discriminative model's performance improves as more edge-specific label functions are incorporated into the generative model; however, we did notice that performance greatly depends on the annotations provided by the generative model.

This work sets up the foundation for creating a common framework that mines text to create edges. Within this framework we would continuously incorporate new knowledge as novel findings are published, while providing a single confidence score for an edge via sentence score consolidation. As opposed to many existing knowledge graphs (for example, Hetionet v1 where text-derived edges generally cannot be exactly attributed to excerpts from literature [3,61]), our approach has the potential to annotate each edge based on its source sentences. In addition, edges generated with this approach would be unencumbered from upstream licensing or copyright restrictions, enabling openly licensed hetnets at a scale not previously possible [62,63,64]. New multitask learning [65] strategies may make it even more practical to reuse label functions to construct continuously updating literature-derived knowledge graphs.

Supplemental Information

An online version of this manuscript is available at https://greenelab.github.io/text_mined_hetnet_manuscript/. Labeled sentences are available at https://github.com/greenelab/text_mined_hetnet_manuscript/tree/master/supplementary_materials/a

nnotated_sentences. Source code for this work is available under open licenses at: https://github.com/greenelab/snorkeling/.

Acknowledgements

The authors would like to thank Christopher Ré's group at Stanford University, especially Alex Ratner and Steven Bach, for their assistance with this project. We also want to thank Graciela Gonzalez-Hernandez for her advice and input with this project. This work was support by Grant GBMF4552 from the Gordon Betty Moore Foundation.

References

1. **Graph Theory Enables Drug Repurposing – How a Mathematical Model Can Drive the Discovery of Hidden Mechanisms of Action**
Ruggero Gramatica, T Di Matteo, Stefano Giorgetti, Massimo Barbiani, Dorian Bevec, Tomaso Aste
PLoS ONE (2014-01-09) https://doi.org/gf45zp
DOI: 10.1371/journal.pone.0084912 · PMID: 24416311 · PMCID: PMC3886994

2. **Drug repurposing through joint learning on knowledge graphs and literature**
Mona Alshahrani, Robert Hoehndorf
Bioinformatics (2018-08-06) https://doi.org/gf45zk
DOI: 10.1101/385617

3. **Systematic integration of biomedical knowledge prioritizes drugs for repurposing**
Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, Sergio E Baranzini
eLife (2017-09-22) https://doi.org/cdfk
DOI: 10.7554/elife.26726 · PMID: 28936969 · PMCID: PMC5640425

4. **Distant supervision for relation extraction without labeled data**
Mike Mintz, Steven Bills, Rion Snow, Dan Jurafsky
Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - ACL-IJCNLP '09 (2009) https://doi.org/fg9q43
DOI: 10.3115/1690219.1690287 · ISBN: 9781932432466

5. **CoCoScore: Context-aware co-occurrence scoring for text mining applications using distant supervision**
Alexander Junge, Lars Juhl Jensen
Bioinformatics (2018-10-16) https://doi.org/gf45zm
DOI: 10.1101/444398

6. **Knowledge-guided convolutional networks for chemical-disease relation extraction**
Huiwei Zhou, Chengkun Lang, Zhuang Liu, Shixian Ning, Yingyu Lin, Lei Du
BMC Bioinformatics (2019-12) https://doi.org/gf45zn
DOI: 10.1186/s12859-019-2873-7 · PMID: 31113357 · PMCID: PMC6528333

7. **Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies?**
R Winnenburg, T Wachter, C Plake, A Doms, M Schroeder
Briefings in Bioinformatics (2008-07-11) https://doi.org/bfsnwg
DOI: 10.1093/bib/bbn043 · PMID: 19060303

8. **Manual curation is not sufficient for annotation of genomic databases**
William A Baumgartner, KBretonnel Cohen, Lynne M Fox, George Acquaah-Mensah, Lawrence Hunter
Bioinformatics (2007-07-01) https://doi.org/dtck86
DOI: 10.1093/bioinformatics/btm229 · PMID: 17646325 · PMCID: PMC2516305

9. **Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references: Growth Rates of Modern Science: A Bibliometric Analysis Based on the Number of Publications and Cited References**
Lutz Bornmann, Rüdiger Mutz

Journal of the Association for Information Science and Technology (2015-11)
https://doi.org/gfj5zc
DOI: 10.1002/asi.23329

10. **DISEASES: Text mining and data integration of disease–gene associations**
Sune Pletscher-Frankild, Albert Pallejà, Kalliopi Tsafou, Janos X Binder, Lars Juhl Jensen
Methods (2015-03) https://doi.org/f3mn6s
DOI: 10.1016/j.ymeth.2014.11.020 · PMID: 25484339

11. **PolySearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more**
Yifeng Liu, Yongjie Liang, David Wishart
Nucleic Acids Research (2015-07-01) https://doi.org/f7nzn5
DOI: 10.1093/nar/gkv383 · PMID: 25925572 · PMCID: PMC4489268

12. **The research on gene-disease association based on text-mining of PubMed**
Jie Zhou, Bo-quan Fu
BMC Bioinformatics (2018-12) https://doi.org/gf479k
DOI: 10.1186/s12859-018-2048-y · PMID: 29415654 · PMCID: PMC5804013

13. **A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts**
David Westergaard, Hans-Henrik Stærfeldt, Christian Tønsberg, Lars Juhl Jensen, Søren Brunak
PLOS Computational Biology (2018-02-15) https://doi.org/gcx747
DOI: 10.1371/journal.pcbi.1005962 · PMID: 29447159 · PMCID: PMC5831415

14. **Literature Mining for the Discovery of Hidden Connections between Drugs, Genes and Diseases**
Raoul Frijters, Marianne van Vugt, Ruben Smeets, René van Schaik, Jacob de Vlieg, Wynand Alkema
PLoS Computational Biology (2010-09-23) https://doi.org/bhrw7x
DOI: 10.1371/journal.pcbi.1000943 · PMID: 20885778 · PMCID: PMC2944780

15. **Analyzing a co-occurrence gene-interaction network to identify disease-gene association**
Amira Al-Aamri, Kamal Taha, Yousof Al-Hammadi, Maher Maalouf, Dirar Homouz
BMC Bioinformatics (2019-12) https://doi.org/gf49nm
DOI: doi:10.1186/s12859-019-2634-7

16. **COMPARTMENTS: unification and visualization of protein subcellular localization evidence**
JX Binder, S Pletscher-Frankild, K Tsafou, C Stolte, SI O'Donoghue, R Schneider, LJ Jensen
Database (2014-02-25) https://doi.org/btbm
DOI: 10.1093/database/bau012 · PMID: 24573882 · PMCID: PMC3935310

17. **A new method for prioritizing drug repositioning candidates extracted by literature-based discovery**
Majid Rastegar-Mojarad, Ravikumar Komandur Elayavilli, Dingcheng Li, Rashmi Prasad, Hongfang Liu
2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (2015-11)
https://doi.org/gf479j
DOI: 10.1109/bibm.2015.7359766 · ISBN: 9781467367998

18. **Comprehensive comparison of large-scale tissue expression datasets**
Alberto Santos, Kalliopi Tsafou, Christian Stolte, Sune Pletscher-Frankild, Seán I O'Donoghue, Lars Juhl Jensen

PeerJ (2015-06-30) https://doi.org/f3mn6p
DOI: 10.7717/peerj.1054 · PMID: 26157623 · PMCID: PMC4493645

19. **A global network of biomedical relationships derived from text**
Bethany Percha, Russ B Altman
Bioinformatics (2018-08-01) https://doi.org/gc3ndk
DOI: 10.1093/bioinformatics/bty114 · PMID: 29490008 · PMCID: PMC6061699

20. **RLIMS-P 2.0: A Generalizable Rule-Based Information Extraction System for Literature Mining of Protein Phosphorylation Information**
Manabu Torii, Cecilia N Arighi, Gang Li, Qinghua Wang, Cathy H Wu, K Vijay-Shanker
IEEE/ACM Transactions on Computational Biology and Bioinformatics (2015-01-01) https://doi.org/gf8fpv
DOI: 10.1109/tcbb.2014.2372765 · PMID: 26357075 · PMCID: PMC4568560

21. **Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing**
Rong Xu, QuanQiu Wang
BMC Bioinformatics (2013-12) https://doi.org/gb8v3k
DOI: 10.1186/1471-2105-14-181 · PMID: 23742147 · PMCID: PMC3702428

22. **Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text**
Yael Garten, Russ B Altman
BMC Bioinformatics (2009-02) https://doi.org/df75hq
DOI: 10.1186/1471-2105-10-s2-s6 · PMID: 19208194 · PMCID: PMC2646239

23. **LimTox: a web tool for applied text mining of adverse event and toxicity associations of compounds, drugs and genes**
Andres Cañada, Salvador Capella-Gutierrez, Obdulia Rabal, Julen Oyarzabal, Alfonso Valencia, Martin Krallinger
Nucleic Acids Research (2017-07-03) https://doi.org/gf479h
DOI: 10.1093/nar/gkx462 · PMID: 28531339 · PMCID: PMC5570141

24. **PPInterFinder—a mining tool for extracting causal relations on human proteins from literature**
Kalpana Raja, Suresh Subramani, Jeyakumar Natarajan
Database (2013-01-01) https://doi.org/gf479b
DOI: 10.1093/database/bas052 · PMID: 23325628 · PMCID: PMC3548331

25. **PKDE4J: Entity and relation extraction for public knowledge discovery.**
Min Song, Won Chul Kim, Dahee Lee, Go Eun Heo, Keun Young Kang
Journal of biomedical informatics (2015-08-12)
https://www.ncbi.nlm.nih.gov/pubmed/26277115
DOI: 10.1016/j.jbi.2015.08.008 · PMID: 26277115

26. **Automatic extraction of gene-disease associations from literature using joint ensemble learning**
Balu Bhasuran, Jeyakumar Natarajan
PLOS ONE (2018-07-26) https://doi.org/gdx63f
DOI: 10.1371/journal.pone.0200699 · PMID: 30048465 · PMCID: PMC6061985

27. **DTMiner: identification of potential disease targets through biomedical literature mining**
Dong Xu, Meizhuo Zhang, Yanping Xie, Fan Wang, Ming Chen, Kenny Q Zhu, Jia Wei
Bioinformatics (2016-08-09) https://doi.org/f9nw36

DOI: 10.1093/bioinformatics/btw503 · PMID: 27506226 · PMCID: PMC5181534

28. **Extracting chemical–protein relations using attention-based neural networks**
Sijia Liu, Feichen Shen, Ravikumar Komandur Elayavilli, Yanshan Wang, Majid Rastegar-Mojarad, Vipin Chaudhary, Hongfang Liu
Database (2018-01-01) https://doi.org/gfdz8d
DOI: doi:10.1093/database/bay102

29. **Deep learning in neural networks: An overview**
Jürgen Schmidhuber
Neural Networks (2015-01) https://doi.org/f6v78n
DOI: 10.1016/j.neunet.2014.09.003 · PMID: 25462637

30. **Probing Biomedical Embeddings from Language Models**
Qiao Jin, Bhuwan Dhingra, William W Cohen, Xinghua Lu
arXiv (2019-04-05) https://arxiv.org/abs/1904.02181

31. **BioBERT: a pre-trained biomedical language representation model for biomedical text mining**
Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang
arXiv (2019-10-21) https://arxiv.org/abs/1901.08746
DOI: 10.1093/bioinformatics/btz682

32. **Attention Is All You Need**
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, Illia Polosukhin
arXiv (2017-12-07) https://arxiv.org/abs/1706.03762

33. **Chemical–gene relation extraction using recursive neural network**
Sangrak Lim, Jaewoo Kang
Database (2018-01-01) https://doi.org/gdss6f
DOI: 10.1093/database/bay060 · PMID: 29961818 · PMCID: PMC6014134

34. **Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research**
Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, Laura I Furlong
BMC Bioinformatics (2015-12) https://doi.org/f7kn8s
DOI: 10.1186/s12859-015-0472-9 · PMID: 25886734 · PMCID: PMC4466840

35. **The EU-ADR corpus: Annotated drugs, diseases, targets, and their relationships**
Erik M van Mulligen, Annie Fourrier-Reglat, David Gurwitz, Mariam Molokhia, Ainhoa Nieto, Gianluca Trifiro, Jan A Kors, Laura I Furlong
Journal of Biomedical Informatics (2012-10) https://doi.org/f36vn6
DOI: 10.1016/j.jbi.2012.04.004 · PMID: 22554700

36. **Comparative experiments on learning information extractors for proteins and their interactions**
Razvan Bunescu, Ruifang Ge, Rohit J Kate, Edward M Marcotte, Raymond J Mooney, Arun K Ramani, Yuk Wah Wong
Artificial Intelligence in Medicine (2005-02) https://doi.org/dhztpn
DOI: 10.1016/j.artmed.2004.07.016 · PMID: 15811782

37. **BioInfer: a corpus for information extraction in the biomedical domain**
Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, Tapio Salakoski

BMC Bioinformatics (2007-12) https://doi.org/b7bhhc
DOI: 10.1186/1471-2105-8-50 · PMID: 17291334 · PMCID: PMC1808065

38. **RelEx--Relation extraction using dependency parse trees**
K Fundel, R Kuffner, R Zimmer
Bioinformatics (2007-02-01) https://doi.org/cz7q4d
DOI: 10.1093/bioinformatics/btl616 · PMID: 17142812

39. **BioCreative V CDR task corpus: a resource for chemical disease relation extraction**
Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, Zhiyong Lu
Database (2016) https://doi.org/gf5hfw
DOI: 10.1093/database/baw068 · PMID: 27161011 · PMCID: PMC4860626

40. **Overview of the biocreative vi chemical-protein interaction track**
Martin Krallinger, Obdulia Rabal, Saber A Akhondiothers
Proceedings of the sixth biocreative challenge evaluation workshop (2017)
https://www.semanticscholar.org/paper/Overview-of-the-BioCreative-VI-chemical-protein-Krallinger-Rabal/eed781f498b563df5a9e8a241c67d63dd1d92ad5

41. **Comparative analysis of five protein-protein interaction corpora**
Sampo Pyysalo, Antti Airola, Juho Heimonen, Jari Björne, Filip Ginter, Tapio Salakoski
BMC Bioinformatics (2008-04) https://doi.org/fh3df7
DOI: 10.1186/1471-2105-9-s3-s6 · PMID: 18426551 · PMCID: PMC2349296

42. **Revisiting distant supervision for relation extraction**
Tingsong Jiang, Jing Liu, Chin-Yew Lin, Zhifang Sui
Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018) (2018-05) https://aclanthology.org/L18-1566

43. **Large-scale extraction of gene interactions from full-text literature using DeepDive**
Emily K Mallory, Ce Zhang, Christopher Ré, Russ B Altman
Bioinformatics (2015-09-03) https://doi.org/gb5g7b
DOI: 10.1093/bioinformatics/btv476 · PMID: 26338771 · PMCID: PMC4681986

44. **Distant Supervision for Large-Scale Extraction of Gene–Disease Associations from Literature Using DeepDive**
Balu Bhasuran, Jeyakumar Natarajan
International Conference on Innovative Computing and Communications (2019)
https://doi.org/gf5hfv
DOI: 10.1007/978-981-13-2354-6_39 · ISBN: 9789811323539

45. **CoCoScore: context-aware co-occurrence scoring for text mining applications using distant supervision**
Alexander Junge, Lars Juhl Jensen
Bioinformatics (2020-01-01) https://doi.org/gf4789
DOI: 10.1093/bioinformatics/btz490 · PMID: 31199464 · PMCID: PMC6956794

46. **Data Programming: Creating Large Training Sets, Quickly**
Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, Christopher Ré
arXiv (2018-12-10) https://arxiv.org/abs/1605.07723

47. **The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)**
Jacqueline MacArthur, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, … Helen Parkinson

Nucleic Acids Research (2017-01-04) https://doi.org/f9v7cp
DOI: 10.1093/nar/gkw1133 · PMID: 27899670 · PMCID: PMC5210590

48. **A Proteome-Scale Map of the Human Interactome Network**
Thomas Rolland, Murat Taşan, Benoit Charloteaux, Samuel J Pevzner, Quan Zhong, Nidhi Sahni, Song Yi, Irma Lemmens, Celia Fontanillo, Roberto Mosca, … Marc Vidal
Cell (2014-11) https://doi.org/f3mn6x
DOI: 10.1016/j.cell.2014.10.050 · PMID: 25416956 · PMCID: PMC4266588

49. **DrugBank 5.0: a major update to the DrugBank database for 2018**
David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, … Michael Wilson
Nucleic Acids Research (2018-01-04) https://doi.org/gcwtzk
DOI: 10.1093/nar/gkx1037 · PMID: 29126136 · PMCID: PMC5753335

50. **PubTator central: automated concept annotation for biomedical full text articles**
Chih-Hsuan Wei, Alexis Allot, Robert Leaman, Zhiyong Lu
Nucleic Acids Research (2019-07-02) https://doi.org/ggzfsc
DOI: 10.1093/nar/gkz389 · PMID: 31114887 · PMCID: PMC6602571

51. **TaggerOne: joint named entity recognition and normalization with semi-Markov Models**
Robert Leaman, Zhiyong Lu
Bioinformatics (2016-09-15) https://doi.org/f855dg
DOI: 10.1093/bioinformatics/btw343 · PMID: 27283952 · PMCID: PMC5018376

52. **tmVar 2.0: integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine**
Chih-Hsuan Wei, Lon Phan, Juliana Feltz, Rama Maiti, Tim Hefferon, Zhiyong Lu
Bioinformatics (2018-01-01) https://doi.org/gbzsmc
DOI: 10.1093/bioinformatics/btx541 · PMID: 28968638 · PMCID: PMC5860583

53. **GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains**
Chih-Hsuan Wei, Hung-Yu Kao, Zhiyong Lu
BioMed Research International (2015) https://doi.org/gb85jb
DOI: 10.1155/2015/918710 · PMID: 26380306 · PMCID: PMC4561873

54. **SR4GN: A Species Recognition Software Tool for Gene Normalization**
Chih-Hsuan Wei, Hung-Yu Kao, Zhiyong Lu
PLoS ONE (2012-06-05) https://doi.org/gpq498
DOI: 10.1371/journal.pone.0038460 · PMID: 22679507 · PMCID: PMC3367953

55. **spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing**
Matthew Honnibal, Ines Montani
(2017)

56. **Snorkel: rapid training data creation with weak supervision**
Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, Christopher Ré
The VLDB Journal (2020-05) https://doi.org/ghbw5f
DOI: 10.1007/s00778-019-00552-1 · PMID: 32214778 · PMCID: PMC7075849

57. **[No title found]**
Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova
Proceedings of the 2019 Conference of the North (2019) https://doi.org/ggbwf6
DOI: 10.18653/v1/n19-1423

58. **PubMed Central: The GenBank of the published literature**
Richard J Roberts
Proceedings of the National Academy of Sciences (2001-01-16) https://doi.org/bbn9k8
DOI: 10.1073/pnas.98.2.381 · PMID: 11209037 · PMCID: PMC33354

59. **Transformers: State-of-the-Art Natural Language Processing**
Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Perric Cistac, Clara Ma, Yacine Jernite, Julien Plu, … Alexander M Rush
Association for Computational Linguistics (2020-10)
https://www.aclweb.org/anthology/2020.emnlp-demos.6

60. **Adam: A Method for Stochastic Optimization**
Diederik P Kingma, Jimmy Ba
arXiv (2017-01-31) https://arxiv.org/abs/1412.6980

61. **[No title found]** https://doi.org/f3mqwp
DOI: 10.15363/thinklab.d67

62. **[No title found]** https://doi.org/bfmk
DOI: 10.15363/thinklab.d107

63. **Legal confusion threatens to slow data science**
Simon Oxenham
Nature (2016-08-04) https://doi.org/bndt
DOI: 10.1038/536016a · PMID: 27488781

64. **An analysis and metric of reusable data licensing practices for biomedical resources**
Seth Carbon, Robin Champieux, Julie A McMurry, Lilly Winfree, Letisha R Wyatt, Melissa A Haendel
PLOS ONE (2019-03-27) https://doi.org/gf5m8v
DOI: 10.1371/journal.pone.0213090 · PMID: 30917137 · PMCID: PMC6436688

65. **Snorkel MeTaL: Weak Supervision for Multi-Task Learning**
Alex Ratner, Braden Hancock, Jared Dunnmon, Roger Goldman, Christopher Ré
Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning (2018-06-15) https://doi.org/gf3xk7
DOI: 10.1145/3209889.3209898 · PMID: 30931438 · PMCID: PMC6436830 · ISBN: 9781450358286

Supplemental Figures

Generative Model Using Randomly Sampled Label Functions

Individual Sources

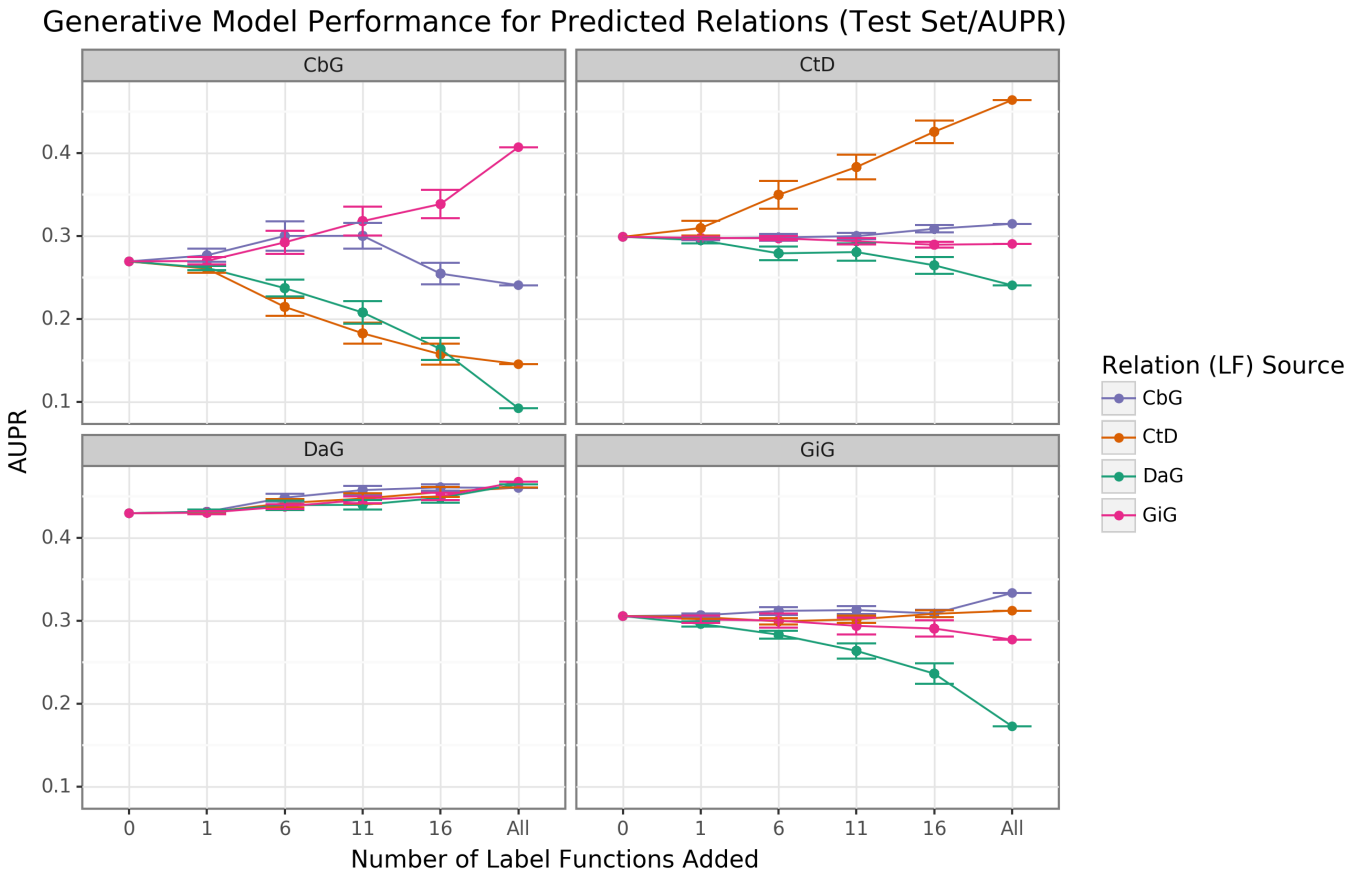

Figure 6: Edge-specific label functions improve performance over edge-mismatch label functions. Each line plot header depicts the edge type the generative model is trying to predict, while the colors represent the source of label functions. For example, orange represents sampling label functions designed to predict the Compound treats Disease (CtD) edge type. The x-axis shows the number of randomly sampled label functions incorporated as an addition to the database-only baseline model (the point at 0). The y-axis shows the area under the precision-recall curve (AUPR). Each point on the plot shows the average of 50 sample runs, while the error bars show the 95% confidence intervals of all runs. The baseline and "All" data points consist of sampling from the entire fixed set of label functions.

Collective Pool of Sources

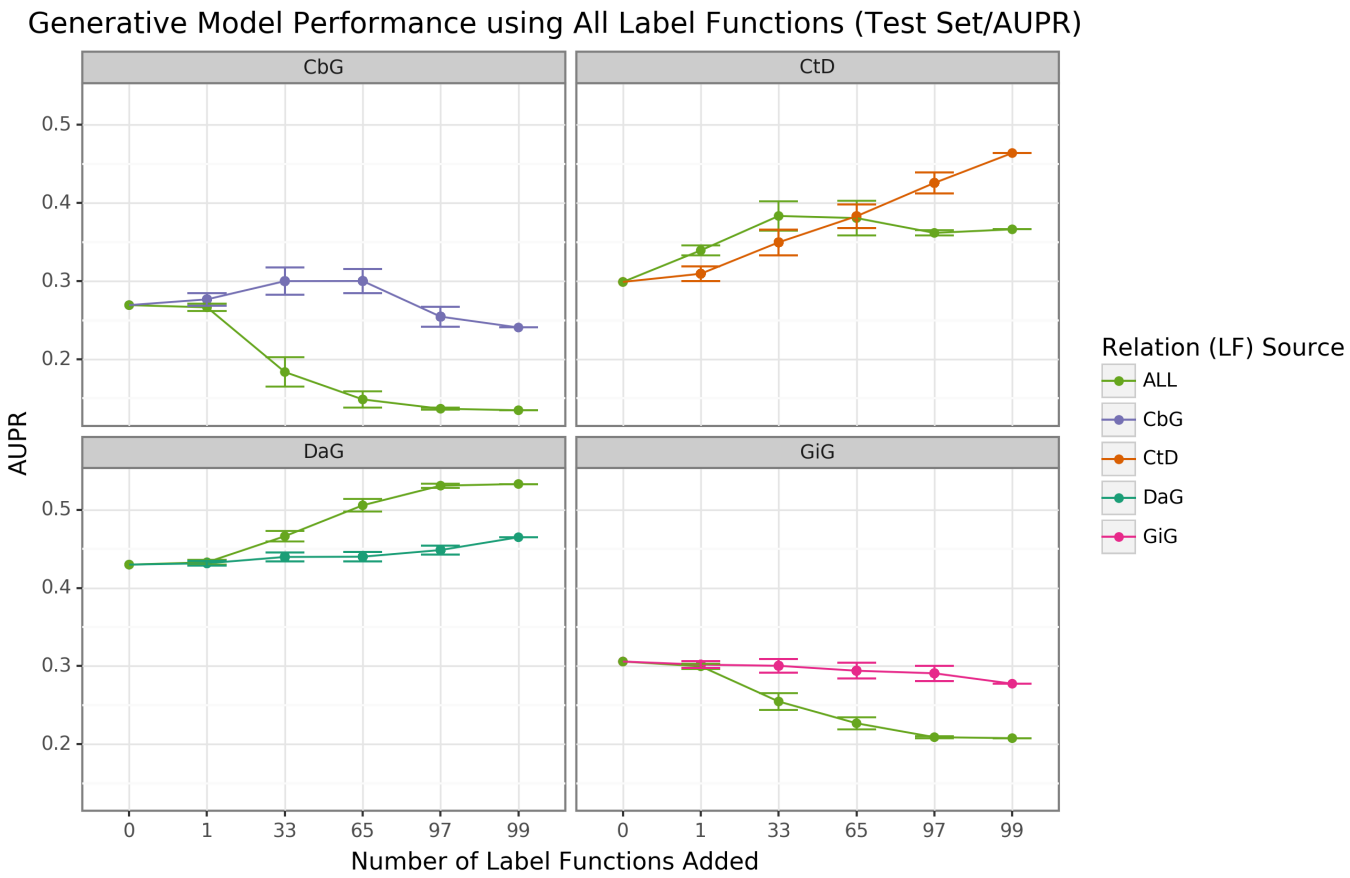

Figure 7: Using all label functions generally hinders generative model performance. Each line plot header depicts the edge type the generative model is trying to predict, while the colors represent the source of label functions. For example, orange represents sampling label functions designed to predict the Compound treats Disease (CtD) edge type. The x-axis shows the number of randomly sampled label functions incorporated as an addition to the database-only baseline model (the point at 0). The y-axis shows the area under the precision-recall curve (AUPR). Each point on the plot shows the average of 50 sample runs, while the error bars show the 95% confidence intervals of all runs. The baseline and "All" data points consist of sampling from the entire fixed set of label functions.

Discriminative Model Performance

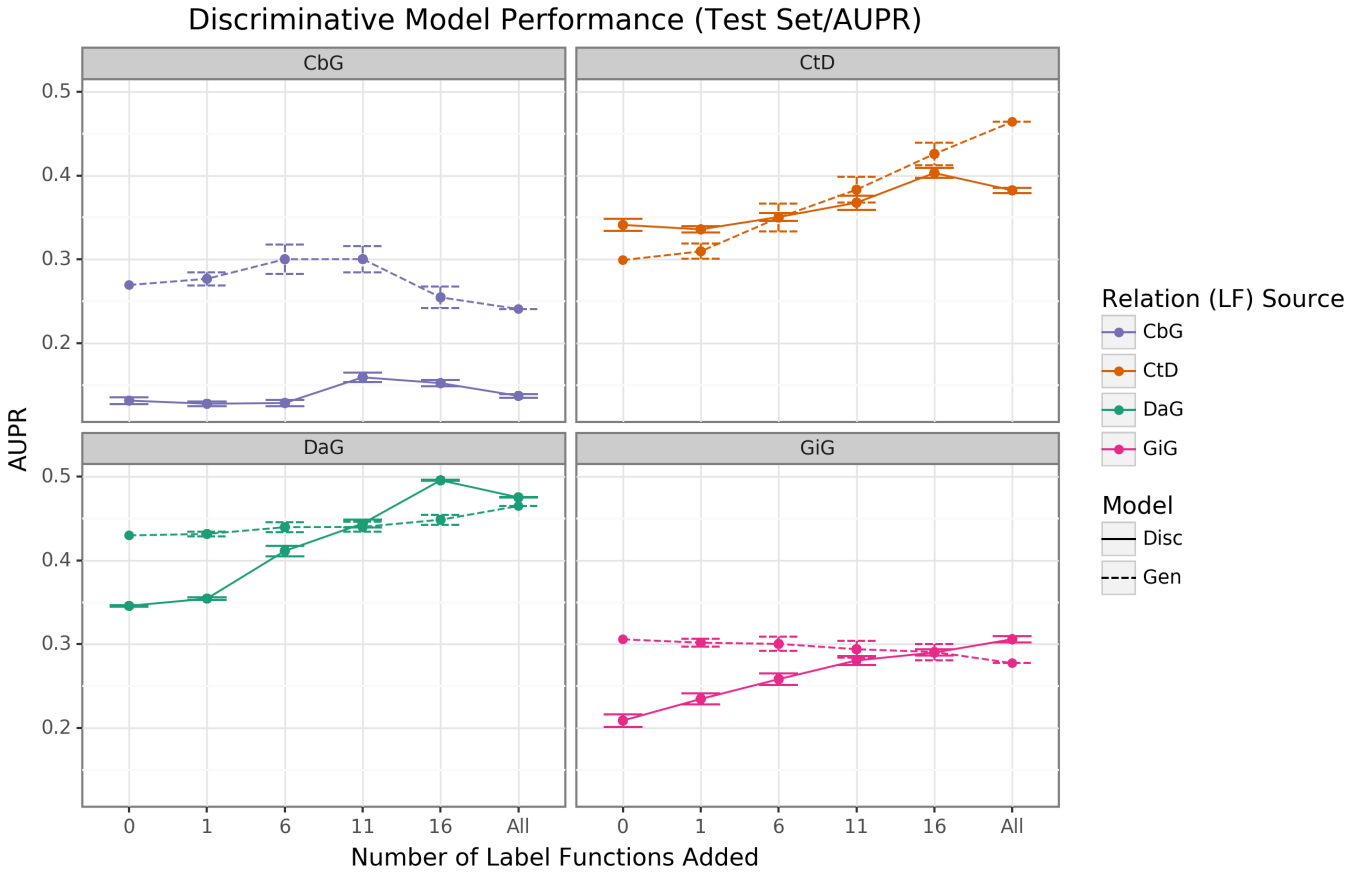

Figure 8: The discriminator model improves performance as the number of edge-specific label functions is added to the baseline model. The line plot headers represent the specific edge type the discriminator model is trying to predict. The x-axis shows the number of randomly sampled label functions incorporated as an addition to the baseline model (the point at 0). The y axis shows the area under the precision-recall curve (AUPR). Each data point represents the average of 3 sample runs for the discriminator model and 50 sample runs for the generative model. The error bars represent each run's 95% confidence interval. The baseline and "All" data points consist of sampling from the entire fixed set of label functions.

Supplemental Tables

Top Ten Sentences for Each Edge Type

Table 3: Contains the top ten predictions for each edge type. Highlighted words represent entities mentioned within the given sentence.

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text

DaG	hematologic cancer	STMN1	1.000	0.979	83	Novel	the stathmin1 mrna expression level in de novo al patient be high than that in healthy person (p < 0.05) , the [stathmin1].{gene_color} mrna expression level in relapse patient with al be high than that in de novo patient (p < 0.05) , and there be no significant difference of stathmin1 mrna expression between patient with [aml].{disease_color} and patient with all .
DaG	breast cancer	INSIG2	1.000	0.979	4	Novel	in analysis of [idc].{disease_color} cell , the level of [insig2].{gene_color} mrna expression be significantly high in late - stage patient than in early - stage patient .
DaG	lung cancer	GNAO1	1.000	0.979	104	Novel	high [numb].{disease_color} expression be associate with favorable prognosis in patient with [lung adenocarcinoma].{gene_color} , but not in those with squamous cell carcinoma .
DaG	breast cancer	TTF1	1.000	0.977	88	Novel	significant [ttf-1].{gene_color} overexpression be observe in adenocarcinomas harbor egfr mutation (p = 0.008) , and no or significantly low level expression of ttf-1 be observe in [adenocarcinomas].{disease_color} harbor kras mutation (p = 0.000) .

DaG	breast cancer	BUB1B	1.000	0.977	13	Novel	elevated [bubr1].{gene_color} expression be associate with poor survival in early stage [breast cancer].{disease_color} patient .
DaG	Alzheimer's disease	SERPINA3	1.000	0.977	182	Existing	a common polymorphism within act and il-1beta gene affect plasma level of [act].{gene_color} or il-1beta , and [ad].{disease_color} patient with the act t , t or il-1beta t , t genotype show the high level of plasma act or il-1beta , respectively .
DaG	esophageal cancer	TRAF6	1.000	0.976	15	Novel	expression of traf6 be highly elevated in [esophageal cancer].{disease_color} tissue , and patient with high [traf6].{gene_color} expression have a significantly short survival time than those with low traf6 expression .
DaG	hypertension	TBX4	1.000	0.975	146	Novel	the proportion of circulate [th1].{gene_color} cell and the level of t - bet , ifng mrna be increase in [ht].{disease_color} patient , the expression of ifng - as1 be upregulated and positively correlate with the proportion of circulate th1 cell or t - bet , and ifng expression , or serum level of anti - thyroglobulin antibody / thyroperoxidase antibody in ht patient .

DaG	breast cancer	TP53	1.000	0.975	3481	Existing	hormone receptor status rather than her2 status be significantly associate with increase ki-67 and [p53].{gene_color} expression in triple [- negative]. {disease_color} breast carcinoma , and high expression of ki-67 but not p53 be significantly associate with axillary nodal metastasis in triple - negative and high - grade non - triple - negative breast carcinoma .
DaG	esophageal cancer	COL17A1	1.000	0.975	32	Novel	high [cd147].{gene_color} expression in patient with [esophageal cancer]. {disease_color} be associate with bad survival outcome and common clinicopathological indicator of poor prognosis .
CtD	Docetaxel	prostate cancer	0.996	0.964	5614	Existing	docetaxel and atrasentan versus [docetaxel].{compound_color} and placebo for man with advanced castration - resistant [prostate cancer].{disease_color} (swog s0421) : a randomised phase 3 trial
CtD	E7389	breast cancer	0.999	0.957	862	Novel	clinical effect of prior trastuzumab on combination [eribulin mesylate]. {compound_color} plus trastuzumab as first - line treatment for human epidermal growth factor receptor 2 positive locally recurrent or metastatic [breast cancer]. {disease_color} : result from a phase ii , single - arm , multicenter study

CtD	Zoledronate	bonecancer	0.996	0.955	226	Novel	[zoledronate].{compound_color} in combination with chemotherapy and surgery to treat [osteosarcoma].{disease_color} (os2006) : a randomised , multicentre , open - label , phase 3 trial .
CtD			0.878	0.954	484	Existing	the role of [ixazomib].{compound_color} as an augment conditioning therapy in salvage autologous stem cell transplant (asct) and as a post - asct consolidation and maintenance strategy in patient with relapse multiple myeloma (accord [uk - mra [myeloma].{disease_color} xii] trial) : study protocol for a phase iii randomise controlled trial
CtD	Topotecan	lungcancer	1.000	0.954	315	Existing	combine chemotherapy with cisplatin , etoposide , and irinotecan versus [topotecan].{compound_color} alone as second - line treatment for patient with [sensitive relapse small].{disease_color} - cell lung cancer (jcog0605) : a multicentre , open - label , randomised phase 3 trial .
CtD	Epirubicin	breastcancer	0.999	0.953	21447	Existing	accelerate versus standard [epirubicin].{compound_color} follow by cyclophosphamide , methotrexate , and fluorouracil or capecitabine as adjuvant therapy for [breast cancer].{disease_color} in the randomised uk tact2 trial (cruk/05/19) : a multicentre , phase 3 , open - label , randomise , control trial
CtD	Paclitaxel	breastcancer	1.000	0.952	10255	Existing	sunitinib plus [paclitaxel].{compound_color} versus bevacizumab plus paclitaxel for first - line treatment of patients with [advanced breast cancer].{disease_color} : a phase iii , randomized , open - label trial

CtD	Anastrozole	breast cancer	0.996	0.952	2364	Existing	a european organisation for research and treatment of cancer randomize , double - blind , placebo - control , multicentre [phase].{disease_color} ii trial of anastrozole in combination with [gefitinib or placebo in hormone].{compound_color} receptor - positive advanced breast cancer (nct00066378) .
CtD	Gefitinib	lung cancer	1.000	0.950	11860	Existing	[gefitinib].{compound_color} versus placebo as maintenance therapy in patient with locally advanced or metastatic [non - small].{disease_color} - cell lung cancer (inform ; c - tong 0804) : a multicentre , double - blind randomise phase 3 trial .
CtD	Docetaxel	prostate cancer	1.000	0.949	5614	Existing	ipilimumab versus placebo after radiotherapy in patient with metastatic castration - resistant [prostate cancer].{disease_color} that have progress after [docetaxel].{compound_color} chemotherapy (ca184 - 043) : a multicentre , randomised , double - blind , phase 3 trial
CtD	Sulfamethazine	lung cancer	0.611	0.949	4	Novel	[tmp].{compound_color} / smz (320/1600 mg / day) treatment be compare to placebo in a double - blind , randomized trial in [patient with newly diagnose].{disease_color} small cell carcinoma of the lung during the initial course of chemotherapy with cyclophosphamide , doxorubicin , and etoposide .
CbG	D-Tyrosine	EGFR	0.601	0.876	34423	Novel	amphiregulin (ar) and heparin - binding egf - like growth factor (hb - [egf].{gene_color}) bind and activate the egfr while heregulin (hrg [) act].{compound_color} through the p185erbb-2 and p180erbb-4 tyrosine kinase .

CbG	Phosphonotyrosine	ANK3	0.004	0.865	1	Novel	at least two domain of p85 can bind to [ank3].{gene_color} , and the interaction involve the p85 c - sh2 domain be find to be [phosphotyrosine].{compound_color} - independent .
CbG	Adenosine	ABCC8	0.891	0.860	353	Novel	sulfonylurea act by inhibition of [beta - cell].{compound_color} adenosine triphosphate - dependent potassium (k(atp)) channel after bind to the sulfonylurea subunit 1 [receptor ( ].{gene_color} sur1) .
CbG	D-Tyrosine	AREG	0.891	0.857	22	Novel	amphiregulin ([ar)].{gene_color} and heparin - binding egf - like growth factor (hb - egf) bind and activate the egfr while heregulin (hrg [) act].{compound_color} through the p185erbb-2 and p180erbb-4 tyrosine kinase .
CbG	D-Tyrosine	EGF	0.602	0.856	389	Novel	upon activation of the receptor for the epidermal growth factor ([egfr)].{gene_color} , sprouty2 undergoe phosphorylation at a conserve [tyrosine].{compound_color} that recruit the src homology 2 domain of c - cbl .
CbG	D-Tyrosine	CSF1	0.101	0.854	106	Novel	as a member of the subclass iii family of receptor [tyrosine].{compound_color} kinase , kit be closely relate to the receptor for platelet derive growth factor alpha and beta (pdgf - a and b [) , macrophage colony].{gene_color} stimulate factor (m - csf) , and flt3 ligand .

CbG	D-Tyrosine	ERBB4	0.101	0.848	115	Novel	the efgr family be a group of four structurally similar [tyrosine].{compound_color} kinase (egfr , her2 / neu , erbb-3 [, and erbb-4].{gene_color}) that dimerize on bind with a number of ligand , include egf and transform growth factor alpha .
CbG	D-Tyrosine	EGFR	0.969	0.848	34223	Novel	the [epidermal growth factor receptor].{gene_color} be a member of type - -pron- growth factor receptor [family].{compound_color} with tyrosine kinase activity that be activate follow the binding of multiple cognate ligand .
CbG	D-Tyrosine	VAV1	0.601	0.842	187	Novel	stimulation of quiescent rodent fibroblast with either epidermal or platelet - derive growth factor induce an increase affinity of vav for cbl - b and result in the [subsequent].{gene_color} formation of a vav - [dependent].{compound_color} trimeric complex with the ligand - stimulate tyrosine kinase receptor .
CbG	Tretinoin	RORB	0.601	0.840	7	Novel	the retinoid z receptor beta ([rzr beta)].{gene_color} , an orphan receptor , be a member of the [retinoic acid].{compound_color} receptor (rar)/thyroid hormone receptor (tr) subfamily of nuclear receptor .
CbG	L-Tryptophan	TACR1	0.891	0.839	4	Novel	these result suggest that the [tryptophan].{compound_color} and quinuclidine series of nk-1 antagonist bind to similar bind site on the human [nk-1 receptor].{gene_color} .
GiG	CYSLTR2	CYSLTR2	0.967	0.564	37	Novel	the bind pocket of [cyslt2].{gene2_color} receptor and the proposition of the interaction mode between [cyslt2].{gene1_color} and hami3379 be identify .

GiG	RXRA	PPARA	1.000	0.563	143	Novel	after bind ligand , the [ppar].{gene2_color} - y receptor heterodimerize [with].{gene1_color} the rxr receptor .
GiG	RXRA	RXRA	0.824	0.551	1101	Existing	nuclear hormone receptor , for example , bind either as homodimer or as heterodimer with [retinoid x receptor].{gene1_color} ([rxr)].{gene2_color} to half - site repeat that be stabilize by protein - protein interaction mediate by residue within both the dna- and ligand - bind domain .
GiG	ADRBK1	ADRA2A	0.822	0.543	3	Novel	mutation of these residue within the [holo - alpha(2a)ar diminish grk2-promoted].{gene2_color} phosphorylation [of].{gene1_color} the receptor as well as the ability of the kinase to be activate by receptor binding .
GiG	ESRRA	ESRRA	0.001	0.531	308	Existing	the crystal structure of the ligand bind domain (lbd) of the estrogen - relate receptor [alpha].{gene2_color} ([erralpha ,].{gene1_color} nr3b1) complexe with a coactivator peptide from peroxisome proliferator - activate receptor coactivator-1alpha (pgc-1alpha) reveal a transcriptionally active conformation in the absence of a ligand .
GiG	GP1BA	VWF	0.518	0.527	144	Existing	these finding indicate the novel bind site require for [vwf].{gene2_color} binding of human [gpibalpha].{gene1_color} .
GiG	NR2C1	NR2C1	0.027	0.522	26	Novel	the human [testicular receptor 2].{gene1_color} ([tr2)].{gene2_color} , a member of the nuclear hormone receptor superfamily , have no identify ligand yet .
GiG	NCOA1	ESRRG	0.992	0.518	1	Novel	the crystal structure of the ligand bind domain (lbd) of the estrogen - relate receptor [3 ().{gene2_color} err3) complexe with a steroid receptor [coactivator-1 ().{gene1_color} src-1) peptide reveal a transcriptionally active conformation in absence of any ligand .
GiG	PPARG	PPARG	0.824	0.504	2497	Existing	although these agent can bind and activate an orphan nuclear receptor , [peroxisome proliferator - activate].{gene2_color} receptor [gamma ( ].{gene1_color} ppargamma) , there be no direct evidence to conclusively implicate this receptor in the regulation of mammalian glucose homeostasis .

GiG	ESR2	ESR1	0.995	0.503	1715	Novel	ligand bind experiment with purify [er alpha].{gene2_color} and [er beta].{gene1_color} confirm that the two phytoestrogen be er ligand .
GiG	FGFR2	FGFR2	1.000	0.501	584	Existing	receptor modeling of [kgfr].{gene1_color} be use to identify selective kgfr tyrosine kinase (tk) inhibitor molecule that have the potential to bind selectively to the [kgfr].{gene2_color} .