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

Authors

- **David N. Nicholson**
  [0000-0003-0002-5761](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](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](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 important resources that hold complex structured and unstructured information. These resources have been used in important tasks such as network analysis for drug repurposing discovery [1]–[3] or as a source of training labels for text mining systems [4]–[6]. Populating knowledge bases often requires highly trained scientists to read biomedical literature and summarize the results [7]. This time-consuming process is referred to as manual curation. In 2007, researchers estimated that filling a knowledge base via manual curation would require approximately 8.4 years to complete [8]. The rate of publications continues to exponentially increase [9], so using only manual curation to fully populate a knowledge base has become impractical.

Relationship extraction has been studied as a solution towards handling the challenge posed by an exponentially growing body of literature [7]. This process consists of creating an expert system to automatically scan, detect and extract relationships from textual sources. Typically, these systems utilize machine learning techniques that require extensive corpora of well-labeled training data. These corpora are difficult to obtain, because they are constructed via extensive manual curation pipelines.

Distant supervision is a technique also designed to sidestep the dependence on manual curation and quickly generate large training datasets. This technique assumes that positive examples established in selected databases can be applied to any sentence that contains them [4]. The central problem with this technique is that generated labels are often of low quality which results in an expansive amount of false positives **raw:Jiang2018RevisitingDS?**.

Ratner et al. [10] recently introduced "data programming" as a solution. Data programming is a paradigm that combines distant supervision with simple rules and heuristics written as small programs called label functions. These label functions are consolidated via a noise aware generative model that is designed to produce training labels for large datasets. Using this paradigm can dramatically reduce the time required to obtain sufficient training data; however, writing a useful label function requires a significant amount of time and error analysis. This dependency makes

constructing a knowledge base with a myriad of heterogenous relationships nearly impossible as tens or possibly hundreds of label functions are required per relationship type.

In this paper, we seek to accelerate the label function creation process by measuring the extent to which label functions can be re-used across different relationship types. We hypothesize that sentences describing one relationship type may share linguistic features such as keywords or sentence structure with sentences describing other relationship types. We conducted a series of experiments to determine the degree to which label function re-use enhanced performance over distant supervision alone. We focus on relationships that indicate similar types of physical interactions (i.e., gene-binds-gene and compound-binds-gene) as well as different types (i.e., disease-associates-gene and compound-treats-disease). Re-using label functions could dramatically reduce the time required to populate a knowledge base with a multitude of heterogeneous relationships.

Related Work

Relationship extraction is the process of detecting semantic relationships from a collection of text. This process can be broken down into three different categories: (1) the use of natural language processing techniques such as manually crafted rules and heuristics for relationship extraction (Rule Based Extractors), (2) the use of unsupervised methods such as co-occurrence scores or clustering to find patterns within sentences and documents (Unsupervised Extractors), and (3) the use of supervised or semi-supervised machine learning for classifying the presence of a relation within documents or sentences (Supervised Extractors). In this section, we briefly discuss selected efforts under each category.

Rule Based Extractors

Rule based extractors rely heavily on expert knowledge to perform extraction. Typically, these systems use linguistic rules and heuristics to identify key sentences or phrases. For example, a hypothetical extractor focused on protein phosphorylation events would identify sentences containing the phrase "gene X phosphorylates gene Y" [11]. This phrase is a straightforward indication that two genes have a fundamental role in protein phosphorylation. Other phrase extractors have been used to identify drug-disease treatments **tag:ctd_medline?**, pharmcogenomic events **tag:pharmpresso?** and protein-protein interactions **tag:ppinterfinder?**, **tag:hpiminer?**. These extractors provide a simple and effective way to extract sentences; however, they depend on extensive knowledge about the text to be properly constructed.

A sentence's grammatical structure can also support relationship extraction via dependency trees. Dependency trees are data structures that depict a sentence's grammatical relation structure in the form of nodes and edges. Nodes represent words and edges represent the dependency type each word shares between one another. For example, a possible extractor would classify sentences as a positive if a sentence contained the following dependency tree path: "gene X (subject)-> promotes (verb)<- cell death (direct object) <- in (preposition) <-tumors (object of preposition)" **tag:pkde4j?**. This approach provides extremely precise results, but the quantity of positive results remains modest as sentences appear in distinct forms and structure. Because of this limitation, recent approaches have incorporated methods on top of rule based extractors such as co-occurrence and machine learning systems [12], **tag:limtox?**. We discuss the pros and cons of added methods in a later section. For this project, we constructed our label functions without the aid of these works; however, approaches discussed in this section provide substantial inspiration for novel label functions in future endeavors.

Unsupervised Extractors

Unsupervised extractors detect relationships without the need of annotated text. Notable approaches exploit the fact that two entities can occur together in text. This event is referred to as co-occurrence. Extractors utilize these events by generating statistics on the frequency of entity pairs occurring in text. For example, a possible extractor would say gene X is associated with disease Y, because gene X and disease Y appear together more often than individually **tag:diseases?**. This approach has been used to establish the following relationship types: disease-gene relationships **tag:diseases?**, **tag:polysearch?**, **tag:dg_text_pubmed?**, **tag:lgscore?**, **tag:full_text_co_abstracts?**, **tag:copub_discovery?**, protein-protein interactions [13], **tag:protein_protein_co_network?**, **tag:full_text_co_abstracts?**, drug-disease treatments **tag:abc_drugs?**, and tissue-gene relations [14]. Extractors using the co-occurrence strategy provide exceptional recall results; however, these methods may fail to detect underreported relationships, because they depend on entity-pair frequency for detection. Junge et al. created a hybrid approach to account for this issue using distant supervision to train a classifier to learn the context of each sentence **tag:cocoscore?**. Once the classifier was trained, they scored every sentence within their corpus, and each sentence's score was incorporated into calculating co-occurrence frequencies to establish relationship existence **tag:cocoscore?**. Co-occurrence approaches are powerful in establishing edges on the global scale; however, they cannot identify individual sentences without the need for supervised methods.

Clustering is an unsupervised approach that extracts relationships from text by grouping similar sentences together. Percha et al. used this technique to group sentences based on their grammatical structure **tag:global_network?**. Using Stanford's Core NLP Parser [15], a dependency tree was generated for every sentence in each Pubmed abstract **tag:global_network?**. Each tree was clustered based on similarity and each cluster was manually annotated to determine which relationship each group represented **tag:global_network?**. For our project we incorporated the results of this work as domain heuristic label functions. Overall, unsupervised approaches are desirable since they do not require well-annotated training data. Such approaches provide excellent recall; however, performance can be limited in terms of precision when compared to supervised machine learning methods [16], [17].

Supervised Extractors

Supervised extractors consist of training a machine learning classifier to predict the existence of a relationship within text. These classifiers require access to well-annotated datasets, which are usually created via some form of manual curation. Previous work consists of research experts curating their own datasets to train classifiers **tag:befree?**, **tag:eu_adr?**, **tag:aimed?**, **tag:bioinfer?**, **tag:hprd50?**; however, there have been community-wide efforts to create datasets for shared tasks [18], [19], **tag:biocreative_v?**. Shared tasks are open challenges that aim to build the best classifier for natural language processing tasks such as named entity tagging or relationship extraction. A notable example is the BioCreative community that hosted a number of shared tasks such as predicting compound-protein interactions (BioCreative VI track 5) [18] and compound induced diseases [19]. Often these datasets are well annotated, but are modest in size (2,432 abstracts for BioCreative VI [18] and 1500 abstracts for BioCreative V [19]). As machine learning classifiers become increasingly complex, these small dataset sizes cannot suffice. Plus, these multitude of datasets are uniquely annotated which can generate noticeable differences in terms of classifier performance [19]. Overall, obtaining large well-annotated datasets still remains as an open non-trivial task.

Before the rise of deep learning, a classifier that was most frequently used was support vector machines. This classifier uses a projection function called a kernel to map data onto a high dimensional space so datapoints can be easily discerned between classes [20]. This method was used to extract disease-gene associations **tag:befree?**, **tag:dtminer?**, **tag:ensemble_svm?**, protein-protein interactions**tag:ppi_graph_kernels?**, **tag:limtox?**, **tag:lptk?** and protein docking information **tag:protein_docking?**. Generally, support vector machines perform well on small datasets with large feature spaces but are slow to train as the number of datapoints becomes asymptotically large.

Deep learning has been increasingly popular as these methods can outperform common machine learning methods [21]. Approaches in this field consist of using various neural network architectures, such as recurrent neural networks **tag:ppi_bilstm?**, **tag:cbg_ensemble_dl?**, **tag:cbg_neural_attention?**, **tag:recursive_nn?**, **tag:semi_supervised_vae?**, **tag:biobert?** and convolutional neural networks **tag:ppi_deep_conv?**, **tag:mcdepcnn?**, **tag:cbg_ensemble_dl?**, **tag:semi_supervised_vae?**, **tag:cbg_transfer_learning?**, to extract relationships from text. In fact approaches in this field were the winning model within the BioCreative VI shared task [18], **raw:chemprot_winner?**. Despite the substantial success of these models, they often require large amounts of data to perform well. Obtaining large datasets is a time-consuming task, which makes training these models a non-trivial challenge. Distant supervision has been used as a solution to fix the barren amount of large datasets [4]. Approaches have used this paradigm to extract chemical-gene interactions **tag:semi_supervised_vae?**, disease-gene associations **tag:cocoscore?** and protein-protein interactions **tag:deep_dive?**, **tag:cocoscore?**, **tag:semi_supervised_vae?**. In fact, efforts done in **tag:deep_dive?** served as one of the motivating rationales for our work.

Overall, deep learning has provided exceptional results in terms of relationships extraction. Thus, we decided to use a deep neural network as our discriminative model.

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 [22], Human Interaction database [23] and DrugBank [24]. 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).
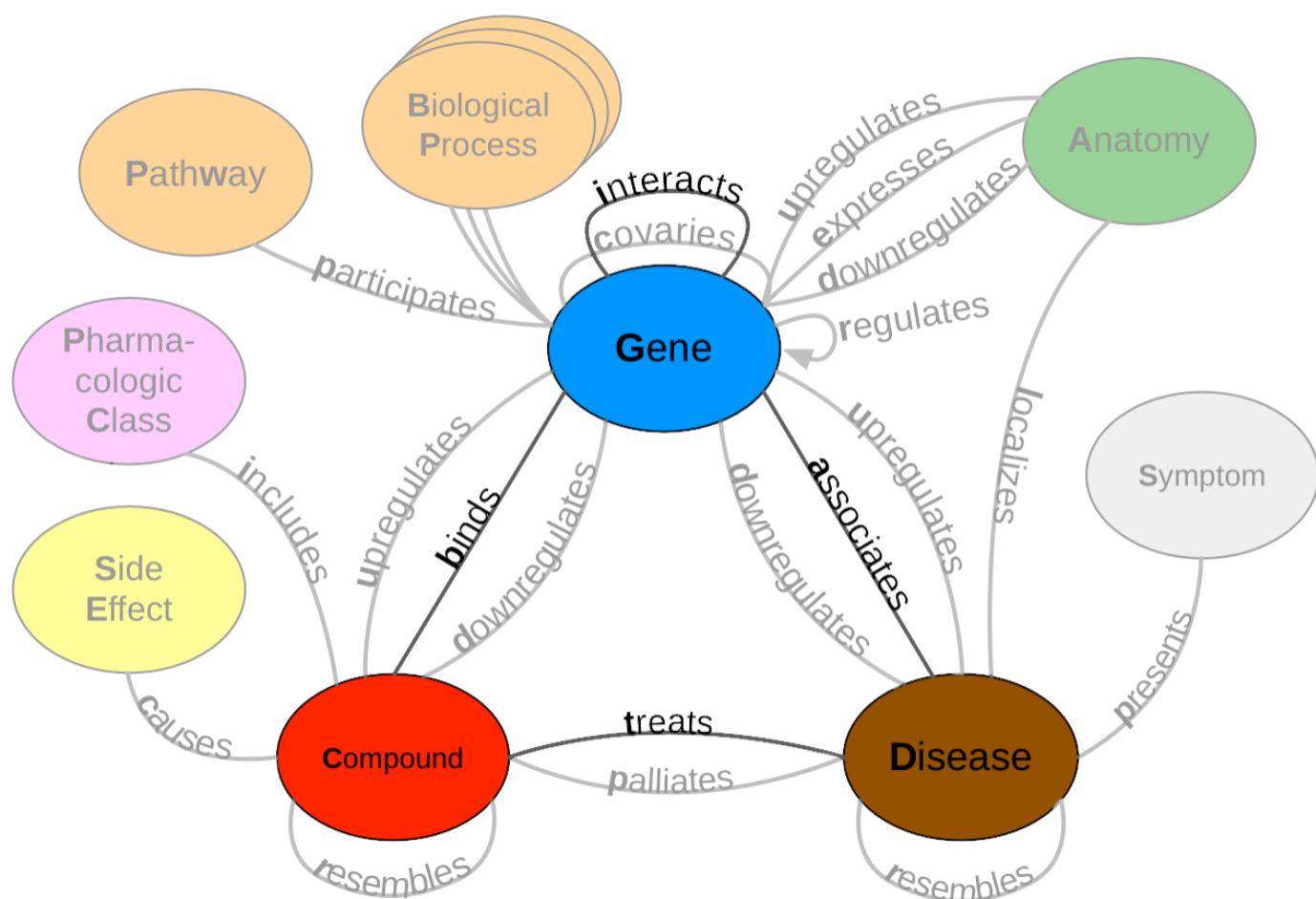
Figure 1: A metagraph (schema) of Hetionet v1 where biomedical entities are represented as nodes and the relationships between them are represented as edges. We examined performance on the highlighted subgraph; however, the long-term vision is to capture edges for the entire graph.

Dataset

We used PubTator Central [25] as input to our analysis. PubTator Central provides MEDLINE abstracts that have been annotated with well-established entity recognition tools including Tagger One [26] for disease, chemical and cell line entities, tmVar [27] for genetic variation tagging, GNormPlus [28] for gene entities and SR4GN [29] 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) [30] 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 [10]. 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?**.

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

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

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

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 [31].

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 [32], a BERT

[33] model trained on all papers and abstracts within Pubmed Central [34], 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 [35] 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 [36] 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 the extent to which label functions can be reused. Our hypothesis was that certain edge types share similar linguistic features such as keywords and/or sentence structure. This shared characteristic would make certain edge types amenable to label function reuse. We designed a set of experiments to test this hypothesis on an individual level (edge vs edge) as well as a global level (collective pool of sources). 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 4). We also evaluated randomly selecting label functions from among all sets and observed similar performance (AUROC Supplemental Figure 5, AUPR Supplemental Figure 6) The quintessential example of this overarching trend is the Compound treats Disease (CtD) edge type, where edge-specific label functions always outperformed transferred label functions. However, there are hints of label function transferability for selected edge types and label function sources. Performance increases as more CbG label functions are incorporated to the GiG baseline model and vice versa. This suggests that sentences for GiG and CbG may share similar linguistic features or terminology that allows for label functions to be reused. 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 which suggests label functions could be shared between the two edge types.
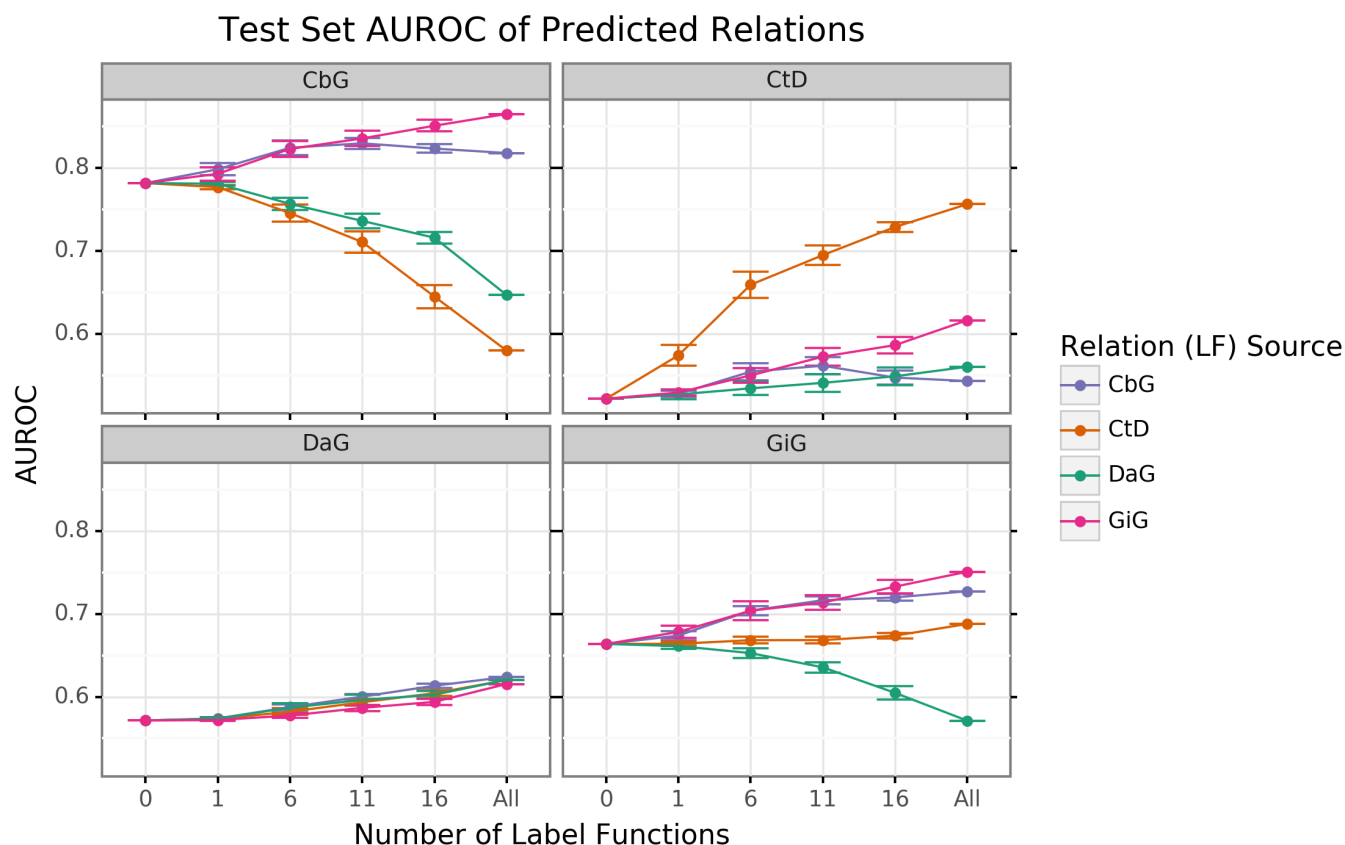
Test Set AUROC of Predicted Relations

Figure 2: Edge-specific label functions are better performing 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 being incorporated into the database-only baseline model (point at 0). The y axis shows 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 (Supplemental Figures 5 and 6). As more label functions were sampled, the gap between edge-specific sources and all sources widened. CbG is a prime example of this trend (Supplemental Figures 5 and 6), 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). The decreasing trend when pooling all label functions supports the notion that label functions cannot easily transfer between edge types (exception being CbG on GiG and vice versa).

Discriminative Model Performance

The discriminative model is designed to augment performance over the generative model by incorporating textual features along with estimated training labels. The discriminative model is a piecewise convolutional neural network trained over word embeddings (See Methods and Materials). We found that the discriminative model generally out-performed the generative model as more edge-specific label functions are incorporated (Figure 3 and Supplemental Figure 7). The discriminative model's performance is often poorest when very few edge-specific label functions are added to the baseline model (seen in Disease associates Gene (DaG), Compound binds Gene (CbG) and Gene interacts Gene (GiG)). This suggests that generative models trained with more label functions produce

outputs that are more suitable for training discriminative models. An exception to this trend is Compound treats Disease (CtD) where the discriminative model out-performs the generative model at all levels of sampling. We observed the opposite trend with the Compound-binds-Gene (CbG) edges: the discriminative model was always poorer 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 7). This suggests that the discriminative model might be predicting more false positives in this setting. Incorporating more edge-specific label functions usually improves performance for the discriminative model over the generative model.
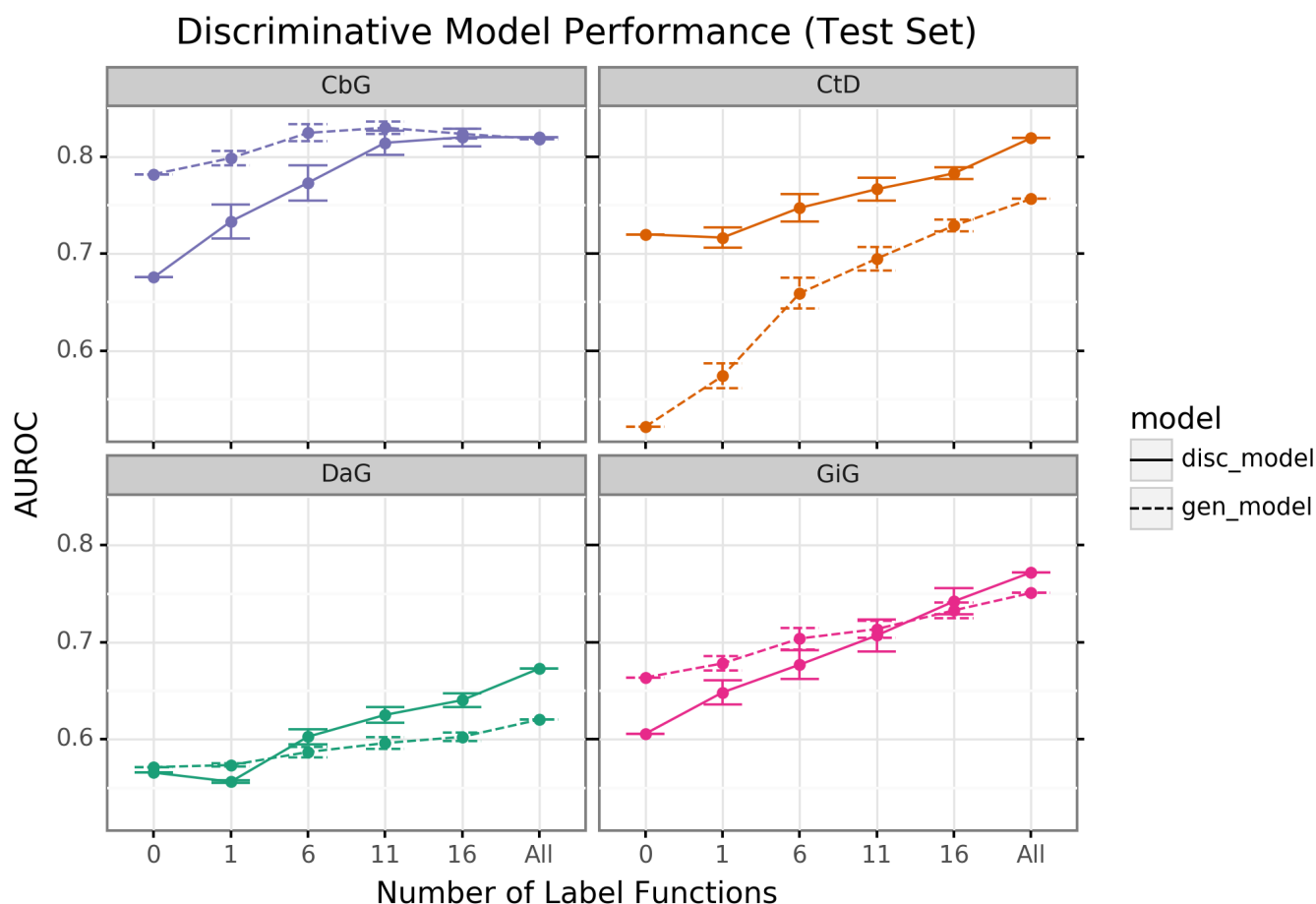

Figure 3: The discriminative model usually improves at a faster rate than the generative model as more edge-specific label function are included. The line plot headers represent the specific edge type the discriminative model is trying to predict. The x-axis shows the number of randomly sampled label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

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 4). Furthermore, we found that using all label functions at once generally hurts performance with the exception being the DaG edge type (Supplemental Figures 5 and 6). 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 [3](#) and Supplemental Figure [7](#)). 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 [3](#) and Supplemental Figure [7](#)). 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 [8](#)). 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 [9](#)). 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], [37]), 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 [38]–[40]. New multitask learning [41] 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/annotated_sentences. Source code for this work is available under open licenses at: https://github.com/greenelab/snorkeling/.

Acknowledgements

References

[1] R. Gramatica, T. Di Matteo, S. Giorgetti, M. Barbiani, D. Bevec, and T. Aste, "Graph Theory Enables Drug Repurposing – How a Mathematical Model Can Drive the Discovery of Hidden Mechanisms of Action," *PLoS ONE*, vol. 9, no. 1, p. e84912, Jan. 2014 [Online]. Available: https://doi.org/gf45zp. [Accessed: 29-Mar-2022]

[2] M. Alshahrani and R. Hoehndorf, "Drug repurposing through joint learning on knowledge graphs and literature," Bioinformatics, preprint, Aug. 2018 [Online]. Available: https://doi.org/gf45zk. [Accessed: 29-Mar-2022]

[3] D. S. Himmelstein *et al.*, "Systematic integration of biomedical knowledge prioritizes drugs for repurposing," *eLife*, vol. 6, p. e26726, Sep. 2017 [Online]. Available: https://doi.org/cdfk. [Accessed: 29-Mar-2022]

[4] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," in *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*, Suntec, Singapore, 2009, vol. 2, p. 1003 [Online]. Available: https://doi.org/fg9q43. [Accessed: 29-Mar-2022]

[5] A. Junge and L. J. Jensen, "CoCoScore: Context-aware co-occurrence scoring for text mining applications using distant supervision," Bioinformatics, preprint, Oct. 2018 [Online]. Available: https://doi.org/gf45zm. [Accessed: 29-Mar-2022]

[6] H. Zhou, C. Lang, Z. Liu, S. Ning, Y. Lin, and L. Du, "Knowledge-guided convolutional networks for chemical-disease relation extraction," *BMC Bioinformatics*, vol. 20, no. 1, p. 260, Dec. 2019 [Online]. Available: https://doi.org/gf45zn. [Accessed: 29-Mar-2022]

[7] R. Winnenburg, T. Wachter, C. Plake, A. Doms, and M. Schroeder, "Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies?" *Briefings in Bioinformatics*, vol. 9, no. 6, pp. 466–478, Jul. 2008 [Online]. Available: https://doi.org/bfsnwg. [Accessed: 29-Mar-2022]

[8] W. A. Baumgartner, K. B. Cohen, L. M. Fox, G. Acquaah-Mensah, and L. Hunter, "Manual curation is not sufficient for annotation of genomic databases," *Bioinformatics*, vol. 23, no. 13, pp. i41–i48, Jul. 2007 [Online]. Available: https://doi.org/dtck86. [Accessed: 29-Mar-2022]

[9] L. Bornmann and R. Mutz, "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," *Journal of the Association for Information Science and Technology*, vol. 66, no. 11, pp. 2215–2222, Nov. 2015 [Online]. Available: https://doi.org/gfj5zc. [Accessed: 29-Mar-2022]

[10] A. Ratner, C. De Sa, S. Wu, D. Selsam, and C. Ré, "Data Programming: Creating Large Training Sets, Quickly," arXiv, 1605.07723, Dec. 2018 [Online]. Available: https://arxiv.org/abs/1605.07723

[11] M. Torii, C. N. Arighi, G. Li, Q. Wang, C. H. Wu, and K. Vijay-Shanker, "RLIMS-P 2.0: A Generalizable Rule-Based Information Extraction System for Literature Mining of Protein Phosphorylation Information," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12, no. 1, pp. 17–29, Jan. 2015 [Online]. Available: https://doi.org/gf8fpv. [Accessed: 29-Mar-2022]

[12] H.-M. Müller, K. M. Van Auken, Y. Li, and P. W. Sternberg, "Textpresso Central: a customizable platform for searching, text mining, viewing, and curating biomedical literature," *BMC Bioinformatics*, vol. 19, no. 1, p. 94, Dec. 2018 [Online]. Available: https://doi.org/gf7rbz. [Accessed: 29-Mar-2022]

[13] J. X. Binder *et al.*, "COMPARTMENTS: unification and visualization of protein subcellular localization evidence," *Database*, vol. 2014, no. 0, pp. bau012–bau012, Feb. 2014 [Online]. Available: https://doi.org/btbm. [Accessed: 29-Mar-2022]

[14] A. Santos, K. Tsafou, C. Stolte, S. Pletscher-Frankild, S. I. O'Donoghue, and L. J. Jensen, "Comprehensive comparison of large-scale tissue expression datasets," *PeerJ*, vol. 3, p. e1054, Jun. 2015 [Online]. Available: https://doi.org/f3mn6p. [Accessed: 29-Mar-2022]

[15] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Baltimore, Maryland, 2014, pp. 55–60 [Online]. Available: https://doi.org/gf3xhp. [Accessed: 29-Mar-2022]

[16] L. J. Jensen, J. Saric, and P. Bork, "Literature mining for the biologist: from information retrieval to biological discovery," *Nature Reviews Genetics*, vol. 7, no. 2, pp. 119–129, Feb. 2006 [Online]. Available: https://doi.org/bgq7q9. [Accessed: 29-Mar-2022]

[17] W. W. M. Fleuren and W. Alkema, "Application of text mining in the biomedical domain," *Methods*, vol. 74, pp. 97–106, Mar. 2015 [Online]. Available: https://doi.org/f64p6n. [Accessed: 29-Mar-2022]

[18] M. Krallinger, O. Rabal, S. A. Akhondi, *et al.*, "Overview of the biocreative vi chemical-protein interaction track," in *Proceedings of the sixth biocreative challenge evaluation workshop*, 2017, vol. 1, pp. 141–146 [Online]. Available: https://www.semanticscholar.org/paper/Overview-of-the-BioCreative-VI-chemical-protein-Krallinger-Rabal/eed781f498b563df5a9e8a241c67d63dd1d92ad5

[19] S. Pyysalo, A. Airola, J. Heimonen, J. Björne, F. Ginter, and T. Salakoski, "Comparative analysis of five protein-protein interaction corpora," *BMC Bioinformatics*, vol. 9, no. S3, p. S6, Apr. 2008 [Online]. Available: https://doi.org/fh3df7. [Accessed: 29-Mar-2022]

[20] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, Jul. 1998 [Online]. Available: https://doi.org/fwgxrj. [Accessed: 29-Mar-2022]

[21] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, Jan. 2015 [Online]. Available: https://doi.org/f6v78n. [Accessed: 29-Mar-2022]

[22] J. MacArthur *et al.*, "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)," *Nucleic Acids Research*, vol. 45, no. D1, pp. D896–D901, Jan. 2017 [Online]. Available: https://doi.org/f9v7cp. [Accessed: 29-Mar-2022]

[23] T. Rolland *et al.*, "A Proteome-Scale Map of the Human Interactome Network," *Cell*, vol. 159, no. 5, pp. 1212–1226, Nov. 2014 [Online]. Available: https://doi.org/f3mn6x. [Accessed: 29-Mar-2022]

[24] D. S. Wishart *et al.*, "DrugBank 5.0: a major update to the DrugBank database for 2018," *Nucleic Acids Research*, vol. 46, no. D1, pp. D1074–D1082, Jan. 2018 [Online]. Available: https://doi.org/gcwtzk. [Accessed: 29-Mar-2022]

[25] C.-H. Wei, A. Allot, R. Leaman, and Z. Lu, "PubTator central: automated concept annotation for biomedical full text articles," *Nucleic Acids Research*, vol. 47, no. W1, pp. W587–W593, Jul. 2019 [Online]. Available: https://doi.org/ggzfsc. [Accessed: 29-Mar-2022]

[26] R. Leaman and Z. Lu, "TaggerOne: joint named entity recognition and normalization with semi-Markov Models," *Bioinformatics*, vol. 32, no. 18, pp. 2839–2846, Sep. 2016 [Online]. Available: https://doi.org/f855dg. [Accessed: 29-Mar-2022]

[27] C.-H. Wei, L. Phan, J. Feltz, R. Maiti, T. Hefferon, and Z. Lu, "tmVar 2.0: integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine," *Bioinformatics*, vol. 34, no. 1, pp. 80–87, Jan. 2018 [Online]. Available: https://doi.org/gbzsmc. [Accessed: 29-Mar-2022]

[28] C.-H. Wei, H.-Y. Kao, and Z. Lu, "GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains," *BioMed Research International*, vol. 2015, pp. 1–7, 2015 [Online]. Available: https://doi.org/gb85jb. [Accessed: 29-Mar-2022]

[29] C.-H. Wei, H.-Y. Kao, and Z. Lu, "SR4GN: A Species Recognition Software Tool for Gene Normalization," *PLoS ONE*, vol. 7, no. 6, p. e38460, Jun. 2012 [Online]. Available: https://doi.org/gpq498. [Accessed: 29-Mar-2022]

[30] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017.

[31] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, "Snorkel: rapid training data creation with weak supervision," *The VLDB Journal*, vol. 29, no. 2–3, pp. 709–730, May 2020 [Online]. Available: https://doi.org/ghbw5f. [Accessed: 29-Mar-2022]

[32] J. Lee *et al.*, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," arXiv, 1901.08746, Oct. 2019 [Online]. Available: https://arxiv.org/abs/1901.08746

[33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "[No title found]," in *Proceedings of the 2019 Conference of the North*, Minneapolis, Minnesota, 2019, pp. 4171–4186 [Online]. Available: https://doi.org/ggbwf6. [Accessed: 29-Mar-2022]

[34] R. J. Roberts, "PubMed Central: The GenBank of the published literature," *Proceedings of the National Academy of Sciences*, vol. 98, no. 2, pp. 381–382, Jan. 2001 [Online]. Available: https://doi.org/bbn9k8. [Accessed: 29-Mar-2022]

[35] T. Wolf *et al.*, "Transformers: State-of-the-Art Natural Language Processing," 2020, pp. 38–45 [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

[36] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," arXiv, 1412.6980, Jan. 2017 [Online]. Available: https://arxiv.org/abs/1412.6980

[37] "[No title found]." [Online]. Available: https://doi.org/f3mqwp

[38] "[No title found]." [Online]. Available: https://doi.org/bfmk

[39] S. Oxenham, "Legal confusion threatens to slow data science," *Nature*, vol. 536, no. 7614, pp. 16–17, Aug. 2016 [Online]. Available: https://doi.org/bndt. [Accessed: 29-Mar-2022]

[40] S. Carbon, R. Champieux, J. A. McMurry, L. Winfree, L. R. Wyatt, and M. A. Haendel, "An analysis and metric of reusable data licensing practices for biomedical resources," *PLOS ONE*, vol. 14, no. 3, p. e0213090, Mar. 2019 [Online]. Available: https://doi.org/gf5m8v. [Accessed: 29-Mar-2022]

[41] A. Ratner, B. Hancock, J. Dunnmon, R. Goldman, and C. Ré, "Snorkel MeTaL: Weak Supervision for Multi-Task Learning," in *Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning*, Houston TX USA, 2018, pp. 1–4 [Online]. Available: https://doi.org/gf3xk7. [Accessed: 29-Mar-2022]

[42] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," arXiv, 1706.04599, Aug. 2017 [Online]. Available: https://arxiv.org/abs/1706.04599

[43] V. Kuleshov, N. Fenner, and S. Ermon, "Accurate Uncertainties for Deep Learning Using Calibrated Regression," arXiv, 1807.00263, Jul. 2018 [Online]. Available: https://arxiv.org/abs/1807.00263

Supplemental Figures

Generative Model Using Randomly Sampled Label Functions

Individual Sources
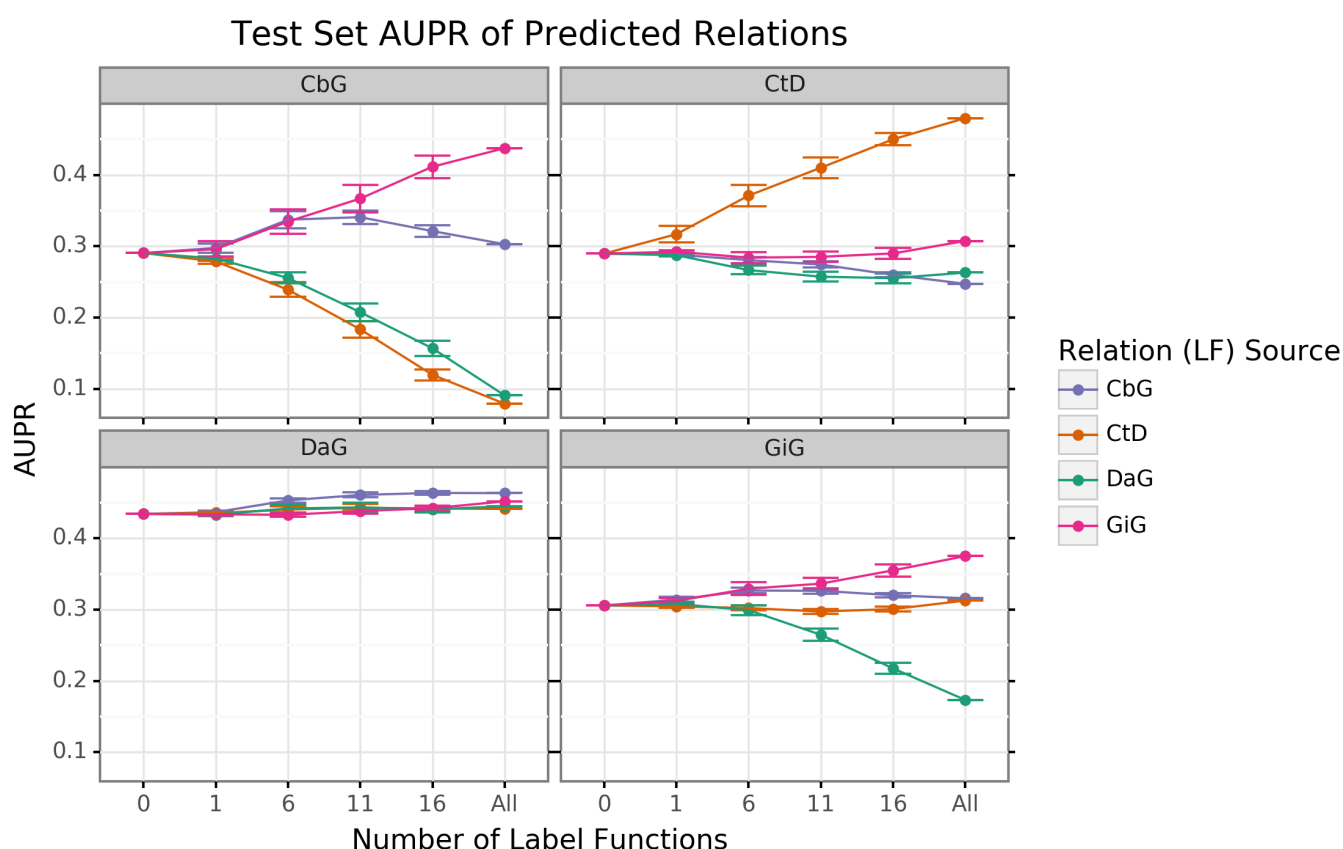
Figure 4: Edge-specific label functions improves performance over edge-mismatch label functions. Each line plot header depicts the edge type the generative model is trying to predict, while the colors represent the source of label functions. For example orange represents sampling label functions designed to predict the Compound treats Disease (CtD) edge type. The x axis shows the number of randomly sampled label functions being incorporated into the database-only baseline model (point at 0). The y axis shows 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 5: 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 being incorporated into the database-only baseline model (point at 0). The y axis shows 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.
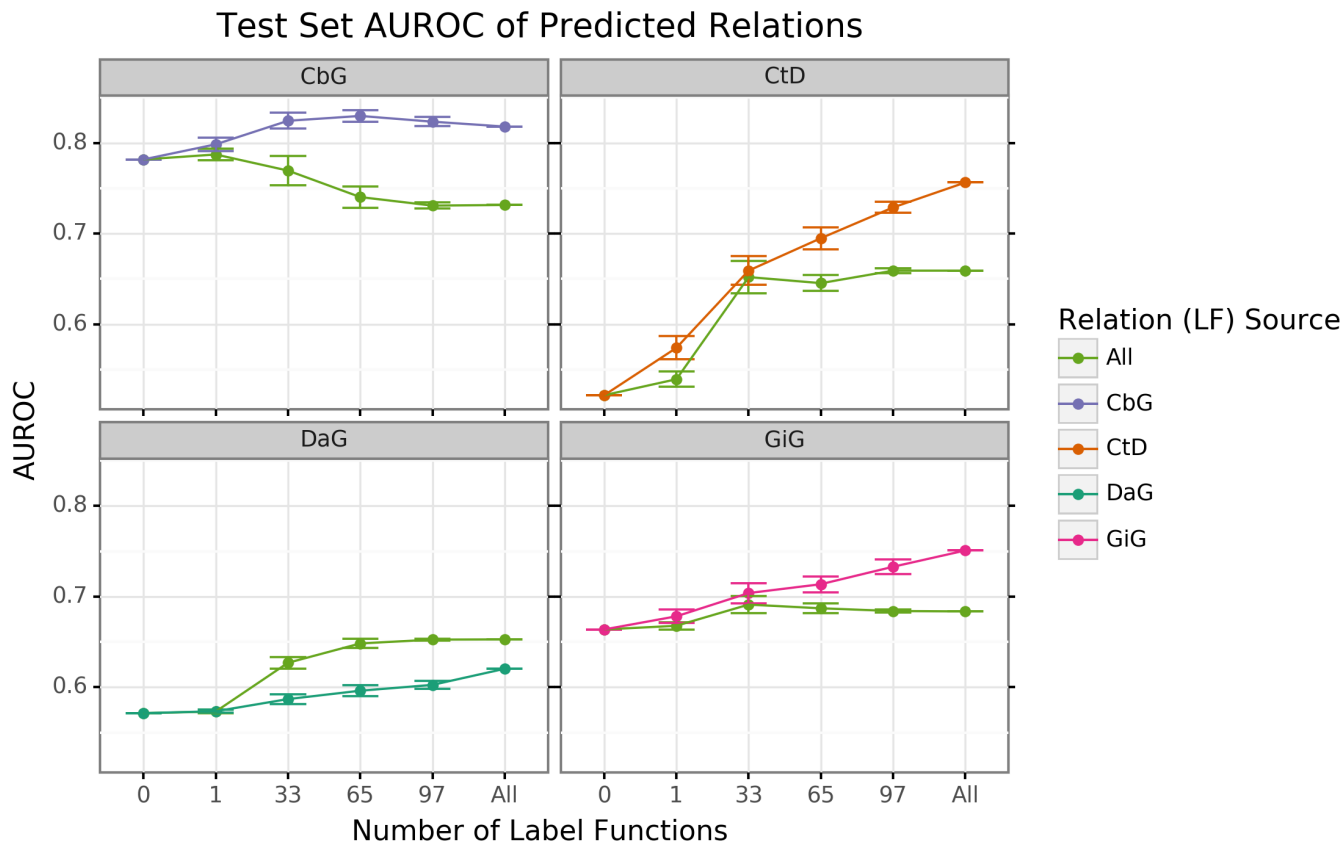
Figure 6: Using all label functions generally hinders generative model performance. Each line plot header depicts the edge type the generative model is trying to predict, while the colors represent the source of label functions. For example, orange represents sampling label functions designed to predict the Compound treats Disease (CtD) edge type. The x axis shows the number of randomly sampled label functions being incorporated into the database-only baseline model (point at 0). The y axis shows 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 7: The discriminator model improves performance as the number of edge-specific label functions is added to the baseline model. The line plot headers represents the specific edge type the discriminator model is trying to predict. The x-axis shows the number of randomly sampled label functions incorporated on top of the baseline model (point at 0). The y axis shows the area under the precision recall curve (AUPR). Each datapoint shows the average of 50 sample runs, while the error bars represents the 95% confidence interval at each point. The baseline and "All" data points consist of sampling from the entire fixed set of label functions.
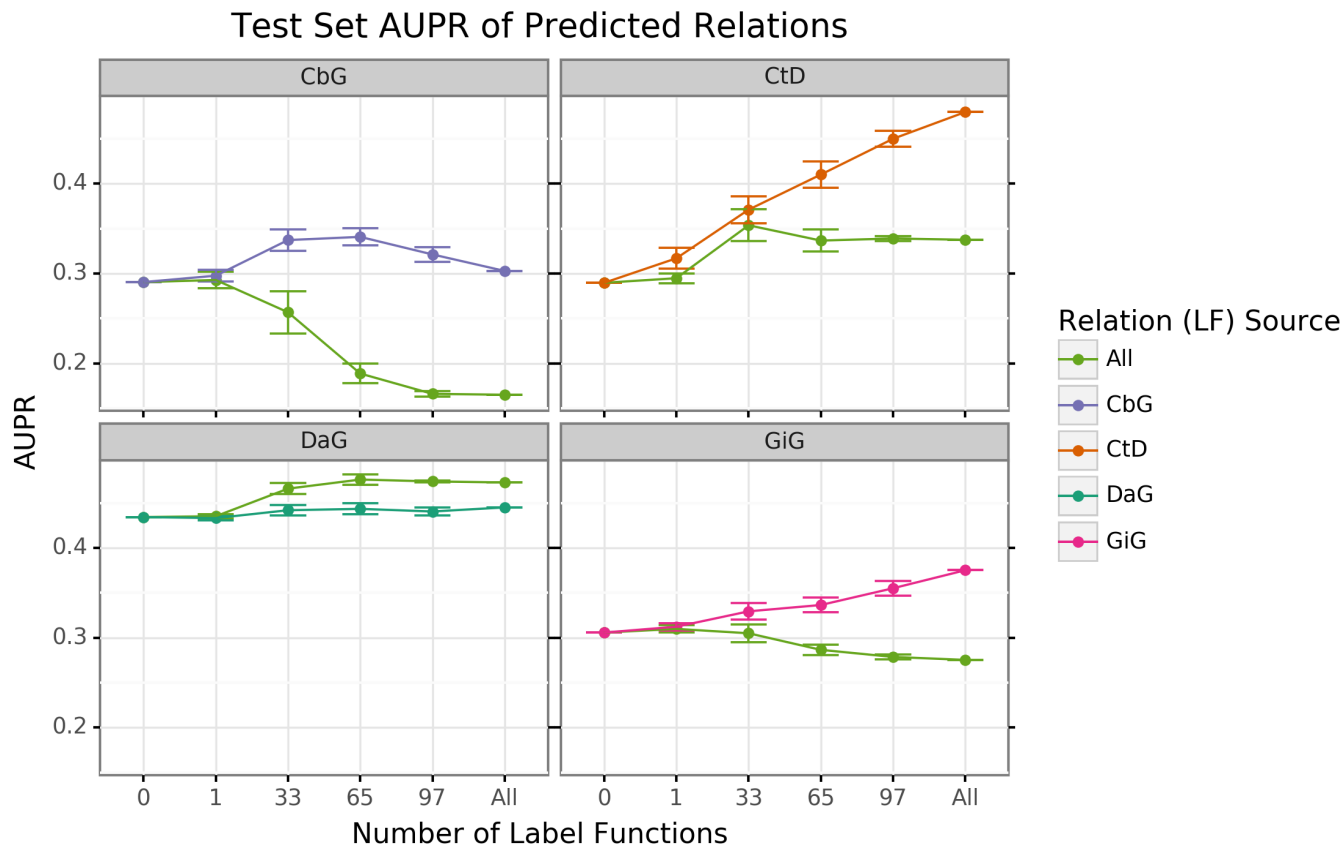
Discriminative Model Calibration
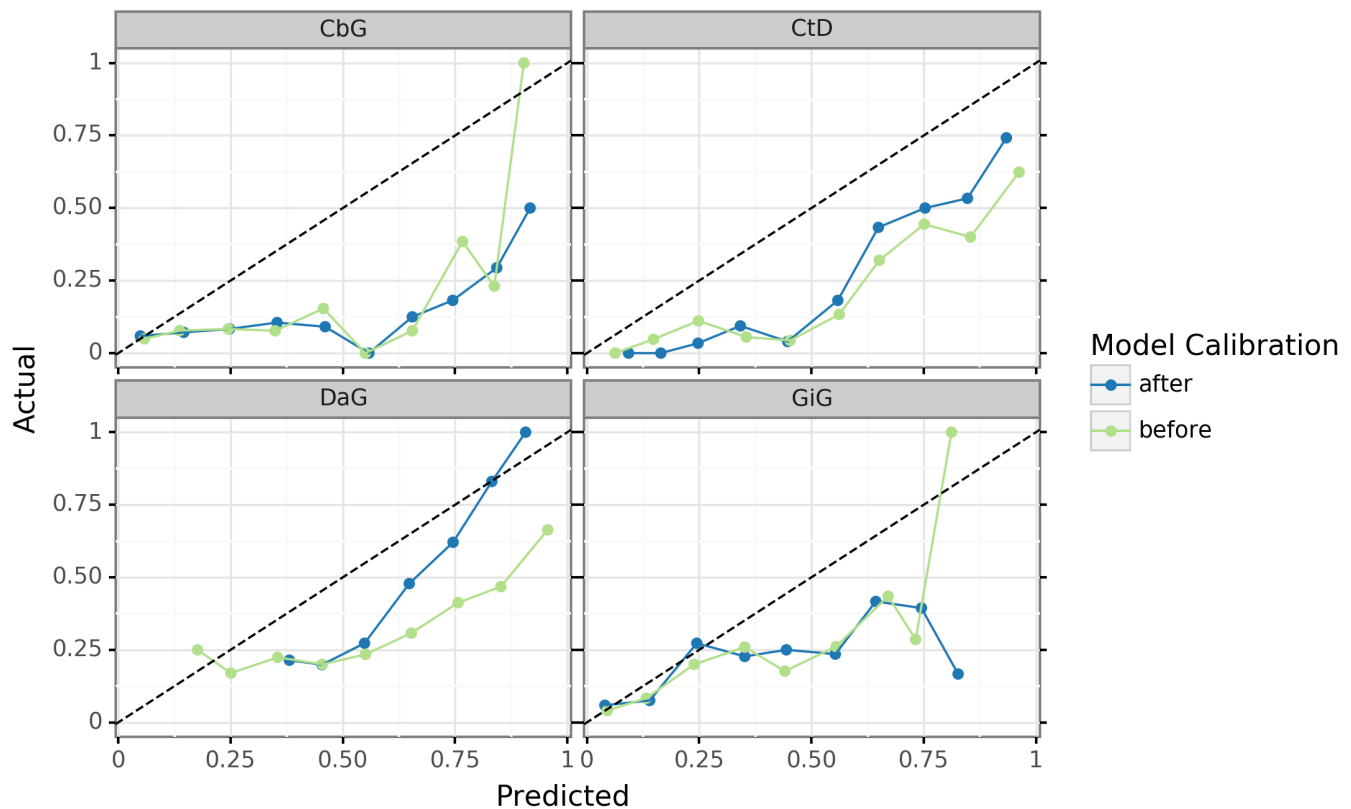
Figure 8: Deep learning models are overconfident in their predictions and need to be calibrated after training. These are calibration plots for the discriminative model, where the green line represents the predictions before calibration and the blue line shows predictions after calibration. Data points that lie closer to the diagonal line show better model calibration, while data points far from the diagonal line show poor performance. A perfectly calibrated model would align straight along the diagonal line.

Even deep learning models with impressive AUROC and AUPR statistics can be subject to poor calibration. Typically, these models are overconfident in their predictions [42], [43]. We attempted to use temperature scaling to fix the calibration of the best performing discriminative models (Figure 8). Before calibration (green lines), our models were aligned with the ideal calibration only when predicting low probability scores (close to 0.25). Applying the temperature scaling calibration algorithm (blue lines) did not substantially improve the calibration of the model in most cases. The exception to this pattern is the Disease associates Gene (DaG) model where high confidence scores are shown to be better calibrated. Overall, calbrating deep learning models is a nontrivial task that requires more complex approaches to accomplish.

Text Mined Edges Can Expand a Database-derived Knowledge Graph
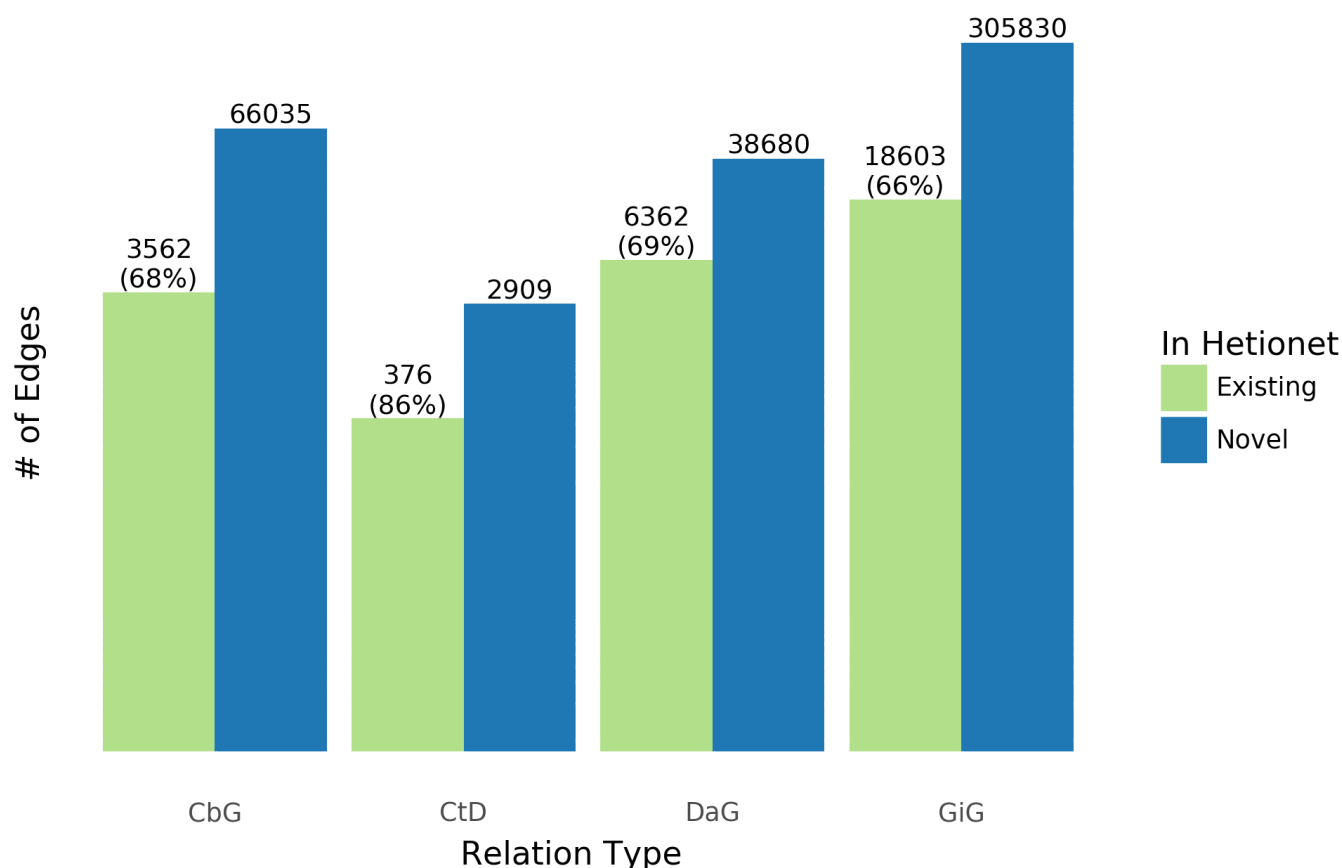
Figure 9: Text-mined edges recreate a substantial fraction of an existing knowledge graph and include new predictions. This bar chart shows the number of edges we can successfully recall in green and shows the number of new edges that can be added in blue.
The recall for the Hetionet v1 knowledge graph is shown as a percentage in parentheses. For example, for the Compound treats Disease (CtD) edge our method recalls 85% of existing edges and adds 6,088 new edges.

One of the goals in 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 how many new edges could be added with high confidence. Overall, we were able to recall more than half of preexisting edges for all edge types (Figure 9) and report our top ten scoring sentences for each edge type in Supplemental Table 11. Our best recall is with the Compound treats Disease (CtD) edge type, where we retain 85% of preexisting edges. Plus, we can add over 6,000 new edges to that category. In contrast, we could only recall close to 70% of existing edges for the other categories; however, we can add over 40,000 novel edges to each category. This highlights the fact that Hetionet v1 is missing a compelling amount of biomedical information and this framework is a viable way to close the information gap.

Comparison with CoCoScore using Hetionet v1 as an Evaluation Set
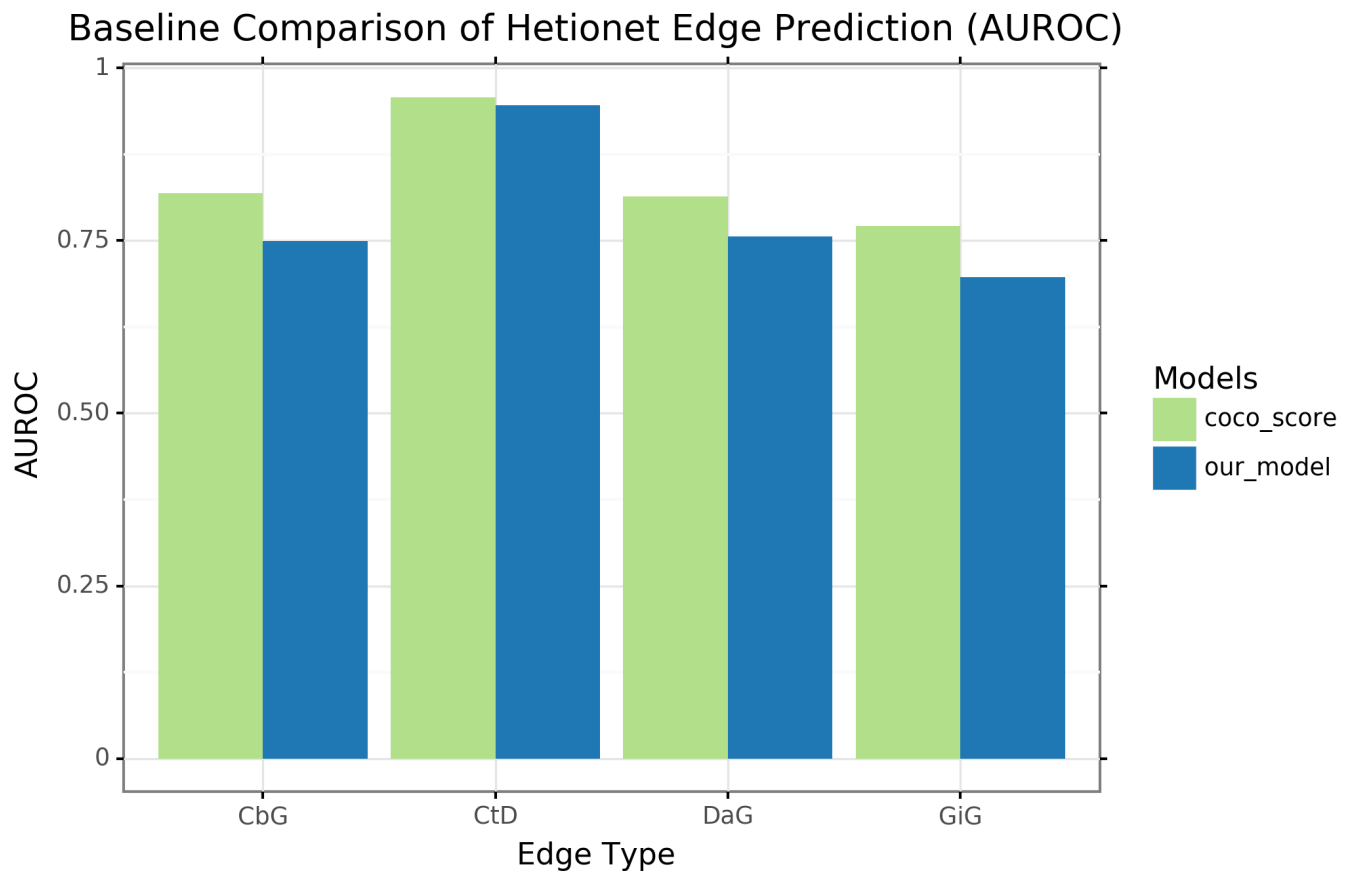
Figure 10: Our extractor shows similar performance to a previously published method when using Hetionet v1 as an evaluation set. We compared our model (blue) with the CoCoScore model **tag:cocoscore?** (green). The y axis represents AUROC and the x axis represents the edge type both models are trying to predict.

Our model showed promising performance in terms of recalling edges in Hetionet v1. We assessed our model's performance relative to a recently published method **tag:cocoscore?**. Though our method is primarily designed to predict assertions, not edges, we compared performance at an edge level because this was available for CoCoScore. We found that a simple summary approach, max sentence score, provided comparable performance to the CoCoScore for the compound treats disease (CtD) edge type and slightly poorer performance for other edge types (Supplemental Figure 10). Sentence-level scores can be integrated in multiple ways, and approaches that consider more complexity (e.g., the number of sentences with high-probability) should be evaluated in future work.

Supplemental Tables

Discriminative Model Calibration Tables

Table 3: Contains the top ten Disease-associates-Gene confidence scores before and after model calbration. Disease mentions are highlighted in brown and Gene mentions are highlighted in blue.

Disease Name	Gene Symbol	Text	Before Calibration	After Calibration
prostate cancer	DKK1	conclusion : high dkk-1 serum levels are associated with a poor survival in patients with prostate cancer .	0.999	0.916
breast cancer	ERBB2	conclusion : her-2 / neu overexpression in primary breast carcinoma is correlated with patients ' age (under age 50) and calcifications at mammography .	0.998	0.906

Disease Name	Gene Symbol	Text	Before Calibration	After Calibration
breast cancer	ERBB2	the results of multiple linear regression analysis , with her2 as the dependent variable , showed that family history of breast cancer was significantly associated with elevated her2 levels in the tumors (p = 0.0038) , after controlling for the effects of age , tumor estrogen receptor , and dna index .	0.998	0.904
colon cancer	SP3	ba also decreased expression of sp1 , sp3 and sp4 transcription factors which are overexpressed in colon cancer cells and decreased levels of several sp-regulated genes including survivin , vascular endothelial growth factor , p65 sub-unit of nfkb , epidermal growth factor receptor , cyclin d1 , and pituitary tumor transforming gene-1 .	0.998	0.902
breast cancer	ERBB2	in breast cancer , overexpression of her2 is associated with an aggressive tumor phenotype and poor prognosis .	0.998	0.898
breast cancer	BCL2	in clinical breast cancer samples , high bcl2 expression was associated with poor prognosis .	0.997	0.886
adrenal gland cancer	TP53	the mechanisms of adrenal tumorigenesis remain poorly established ; the r337h germline mutation in the p53 gene has previously been associated with acts in brazilian children .	0.996	0.883
prostate cancer	AR	the androgen receptor was expressed in all primary and metastatic prostate cancer tissues and no mutations were identified .	0.996	0.881
urinary bladder cancer	PIK3CA	conclusions : increased levels of fgfr3 and pik3ca mutated dna in urine and plasma are indicative of later progression and metastasis in bladder cancer .	0.995	0.866
ovarian cancer	EPAS1	the log-rank test showed that nuclear positive immunostaining for hif-1alpha (p = .002) and cytoplasmic positive immunostaining for hif-2alpha (p = .0112) in tumor cells are associated with poor prognosis of patients with ovarian carcinoma .	0.994	0.86

Table 4: Contains the bottom ten Disease-associates-Gene confidence scores before and after model calbration. Disease mentions are highlighted in brown and Gene mentions are highlighted in blue.

Disease Name	Gene Symbol	Text	Before Calibration	After Calibration
endogenous depression	EP300	from a clinical point of view , p300 amplitude should be considered as a psychophysiological index of suicidal risk in major depressive disorder .	0.202	0.379
Alzheimer's disease	PDK1	from prion diseases to alzheimer 's disease : a common therapeutic target , [pdk1] .	0.2	0.378

Disease Name	Gene Symbol	Text	Before Calibration	After Calibration
endogenous depression	HTR1A	gepirone , a selective serotonin (5ht1a) partial agonist in the treatment of major depression .	0.199	0.378
Gilles de la Tourette syndrome	FGF9	there were no differences in gender distribution , age at tic onset or td diagnosis , tic severity , proportion with current diagnoses of ocd/oc behavior or attention deficit hyperactivity disorder (adhd) , cbcl internalizing , externalizing , or total problems scores , ygtss scores , or gaf scores .	0.185	0.37
hematologic cancer	MLANA	methods : the sln sections (n = 214) were assessed by qrt assay for 4 established messenger rna biomarkers : mart-1 , mage-a3 , galnac-t , and pax3 .	0.18	0.368
endogenous depression	MAOA	alpha 2-adrenoceptor responsivity in depression : effect of chronic treatment with moclobemide , a selective mao-a-inhibitor , versus maprotiline .	0.179	0.367
chronic kidney failure	B2M	to evaluate comparative beta 2-m removal we studied six stable end-stage renal failure patients during high-flux 3-h haemodialysis , haemodia-filtration , and haemofiltration , using acrylonitrile , cellulose triacetate , polyamide and polysulphone capillary devices .	0.178	0.366
hematologic cancer	C7	serum antibody responses to four haemophilus influenzae type b capsular polysaccharide-protein conjugate vaccines (prp-d , hboc , c7p , and prp-t) were studied and compared in 175 infants , 85 adults and 140 2-year-old children .	0.174	0.364
hypertension	AVP	portohepatic pressures , hepatic function , and blood gases in the combination of nitroglycerin and vasopressin : search for additive effects in cirrhotic portal hypertension .	0.168	0.361
endogenous depression	GAD1	within-individual deflections in gad , physical , and social symptoms predicted later deflections in depressive symptoms , and deflections in depressive symptoms predicted later deflections in gad and separation anxiety symptoms .	0.149	0.349

Table 5: Contains the top ten Compound-treats-Disease confidence scores after model calbration. Disease mentions are highlighted in brown and Compound mentions are highlighted in red.

Compound Name	Disease Name	Text	Before Calibration	After Calibration
Prazosin	hypertension	experience with prazosin in the treatment of hypertension .	0.997	0.961
Methyldopa	hypertension	oxprenolol plus cyclopenthiazide-kcl versus methyldopa in the treatment of hypertension .	0.997	0.961
Methyldopa	hypertension	atenolol and methyldopa in the treatment of hypertension .	0.996	0.957
Prednisone	asthma	prednisone and beclomethasone for treatment of asthma .	0.995	0.953
Sulfasalazine	ulcerative colitis	sulphasalazine , used in the treatment of ulcerative colitis , is cleaved in the colon by the metabolic action of colonic bacteria on the diazo bond to release 5-aminosalicylic acid (5-asa) and sulpharidine .	0.994	0.949
Prazosin	hypertension	letter : prazosin in treatment of hypertension .	0.994	0.949
Methylprednisolone	asthma	use of tao without methylprednisolone in the treatment of severe asthma .	0.994	0.948
Budesonide	asthma	thus , a regimen of budesonide treatment that consistently attenuates bronchial responsiveness in asthmatic subjects had no effect in these men ; larger and longer trials will be required to establish whether a subgroup of smokers shows a favorable response .	0.994	0.946
Methyldopa	hypertension	pressor and chronotropic responses to bilateral carotid occlusion (bco) and tyramine were also markedly reduced following treatment with methyldopa , which is consistent with the clinical findings that chronic methyldopa treatment in hypertensive patients impairs cardiovascular reflexes .	0.994	0.946
Fluphenazine	schizophrenia	low dose fluphenazine decanoate in maintenance treatment of schizophrenia .	0.994	0.946

Table 6: Contains the bottom ten Compound-treats-Disease confidence scores before and after model calbration. Disease mentions are highlighted in brown and Compound mentions are highlighted in red.

Compound Name	Disease Name	Text	Before Calibration	After Calibration
Indomethacin	hypertension	effects of indomethacin in rabbit renovascular hypertension .	0.033	0.13
Alprazolam	panic disorder	according to logistic regression analysis , the relationships between plasma alprazolam concentration and response , as reflected by number of panic attacks reported , phobia ratings , physicians ' and patients ' ratings of global improvement , and the emergence of side effects , were significant .	0.03	0.124

Compound Name	Disease Name	Text	Before Calibration	After Calibration
Mestranol	polycystic ovary syndrome	the binding capacity of plasma testosterone-estradiol-binding globulin (tebg) and testosterone (t) levels were measured in four women with proved polycystic ovaries and three women with a clinical diagnosis of polycystic ovarian disease before , during , and after administration of norethindrone , 2 mg. , and mestranol , 0.1 mg .	0.03	0.123
Creatine	coronary artery disease	during successful and uncomplicated angioplasty (ptca) , we studied the effect of a short lasting myocardial ischemia on plasma creatine kinase , creatine kinase mb-activity , and creatine kinase mm-isoforms (mm1 , mm2 , mm3) in 23 patients .	0.028	0.12
Creatine	coronary artery disease	in 141 patients with acute myocardial infarction , creatine phosphokinase isoenzyme (cpk-mb) was determined by the activation method with dithiothreitol (rao et al. : clin .	0.027	0.117
Morphine	brain cancer	the tissue to serum ratio of morphine in the hypothalamus , hippocampus , striatum , midbrain and cortex were also smaller in morphine tolerant than in non-tolerant rats .	0.026	0.115
Glutathione	anemia	our results suggest that an association between gsh px deficiency and hemolytic anemia need not represent a cause-and-effect relationship .	0.026	0.114
Dinoprostone	stomach cancer	prostaglandin e2 (pge2) - and 6-keto-pgf1 alpha-like immunoactivity was measured in incubates of forestomach and gastric corpus mucosa in (a) unoperated rats , (b) rats with sham-operation of the kidneys and (c) rats with bilateral nephrectomy .	0.023	0.107
Creatine	coronary artery disease	the value of the electrocardiogram in assessing infarct size was studied using serial estimates of the mb isomer of creatine kinase (ck mb) in plasma , serial 35 lead praecordial maps in 28 patients with anterior myocardial infarction , and serial 12 lead electrocardiograms in 17 patients with inferior myocardial infarction .	0.022	0.105
Sulfamethazine	multiple sclerosis	quantitation and confirmation of sulfamethazine residues in swine muscle and liver by lc and gc/ms .	0.017	0.093

Table 7: Contains the top ten Compound-binds-Gene confidence scores before and after model calbration. Gene mentions are highlighted in blue and Compound mentions are highlighted in red.

Compound Name	Gene Symbol	Text	Before Calibration	After Calibration

Compound Name	Gene Symbol	Text	Before Calibration	After Calibration
Cyclic Adenosine Monophosphate	B3GNT2	in sk-n-mc human neuroblastoma cells , the camp response to 10 nm isoproterenol (iso) is mediated primarily by beta 1-adrenergic receptors .	0.903	0.93
Indomethacin	AGT	indomethacin , a potent inhibitor of prostaglandin synthesis , is known to increase the maternal blood pressure response to angiotensin ii infusion .	0.894	0.922
Tretinoin	RXRA	the vitamin a derivative retinoic acid exerts its effects on transcription through two distinct classes of nuclear receptors , the retinoic acid receptor (rar) and the retinoid x receptor (rxr) .	0.882	0.912
Tretinoin	RXRA	the vitamin a derivative retinoic acid exerts its effects on transcription through two distinct classes of nuclear receptors , the retinoic acid receptor (rar) and the retinoid x receptor (rxr) .	0.872	0.903
D-Tyrosine	CSF1	however , the extent of gap tyrosine phosphorylation induced by csf-1 was approximately 10 % of that induced by pdgf-bb in the nih3t3 fibroblasts .	0.851	0.883
D-Glutamic Acid	GLB1	thus , the negatively charged side chain of glu-461 is important for divalent cation binding to beta-galactosidase .	0.849	0.882
D-Tyrosine	CD4	second , we use the same system to provide evidence that the physical association of cd4 with the tcr is required for effective tyrosine phosphorylation of the tcr zeta-chain subunit , presumably reflecting delivery of p56lck (lck) to the tcr .	0.825	0.859

Compound Name	Gene Symbol	Text	Before Calibration	After Calibration
Calcium Chloride	TNC	the possibility that the enhanced length dependence of ca2 + sensitivity after cardiac tnc reconstitution was attributable to reduced tnc binding was excluded when the length dependence of partially extracted fast fibres was reduced to one-half the normal value after a 50 % deletion of the native tnc .	0.821	0.855
Metoprolol	KCNMB2	studies in difi cells of the displacement of specific 125i-cyp binding by nonselective (propranolol) , beta 1-selective (metoprolol and atenolol) , and beta 2-selective (ici 118-551) antagonists revealed only a single class of beta 2-adrenergic receptors .	0.82	0.854
D-Tyrosine	PLCG1	epidermal growth factor (egf) or platelet-derived growth factor binding to their receptor on fibroblasts induces tyrosine phosphorylation of plc gamma 1 and stable association of plc gamma 1 with the receptor protein tyrosine kinase .	0.818	0.851

Table 8: Contains the bottom ten Compound-binds-Gene confidence scores before and after model calbration. Gene mentions are highlighted in blue and Compound mentions are highlighted in red.

Compound Name	Gene Symbol	Text	Before Calibration	After Calibration
Deferoxamine	TF	the mechanisms of fe uptake have been characterised using 59fe complexes of citrate , nitrilotriacetate , desferrioxamine , and 59fe added to eagle 's minimum essential medium (mem) and compared with human transferrin (tf) labelled with 59fe and iodine-125 .	0.02	0.011
Hydrocortisone	GH1	group iv patients had normal basal levels of lh and normal lh , gh and cortisol responses .	0.02	0.011
Carbachol	INS	at the same concentration , however , iapp significantly (p less than 0.05) inhibited carbachol-stimulated (10 (-7) m) release of insulin by 30 % , and cgrp significantly inhibited carbachol-stimulated release of insulin by 33 % when compared with the control group .	0.02	0.011

Compound Name	Gene Symbol	Text	Before Calibration	After Calibration
Adenosine	ME2	at physiological concentrations , atp , adp , and amp all inhibit the enzyme from atriplex spongiosa and panicum miliaceum (nad-me-type plants) , with atp the most inhibitory species .	0.019	0.01
Naloxone	POMC	specifically , opioids , including 2-n-pentyloxy-2-phenyl-4-methyl-morpholine , naloxone , and beta-endorphin , have been shown to interact with il-2 receptors (134) and regulate production of il-1 and il-2 (48-50 , 135) .	0.018	0.01
Cortisone acetate	POMC	sarcoidosis therapy with cortisone and acth – the role of acth therapy .	0.017	0.009
Epinephrine	INS	thermogenic effect of thyroid hormones : interactions with epinephrine and insulin .	0.017	0.009
Aldosterone	KNG1	important vasoconstrictor , fluid - and sodium-retaining factors are the renin-angiotensin-aldosterone system , sympathetic nerve activity , and vasopressin ; vasodilator , volume , and sodium-eliminating factors are atrial natriuretic peptide , vasodilator prostaglandins like prostacyclin and prostaglandin e2 , dopamine , bradykinin , and possibly , endothelial derived relaxing factor (edrf) .	0.016	0.008
D-Leucine	POMC	cross-reactivities of leucine-enkephalin and beta-endorphin with the eia were less than 0.1 % , while that with gly-gly-phe-met and oxidized gly-gly-phe-met were 2.5 % and 10.2 % , respectively .	0.011	0.005
Estriol	LGALS1	[diagnostic value of serial determination of estriol and hpl in plasma and of total estrogens in 24-h-urine compared to single values for diagnosis of fetal danger] .	0.01	0.005

Table 9: Contains the top ten Gene-interacts-Gene confidence scores before and after model calbration. Both gene mentions highlighted in blue.

Gene1 Symbol	Gene2 Symbol	Text	Before Calibration	After Calibration
ESR1	HSP90AA1	previous studies have suggested that the 90-kda heat shock protein (hsp90) interacts with the er , thus stabilizing the receptor in an inactive state .	0.812	0.864

Gene1 Symbol	Gene2 Symbol	Text	Before Calibration	After Calibration
TP53	TP73	cyclin g interacts with p53 as well as p73 , and its binding to p53 or p73 presumably mediates downregulation of p53 and p73 .	0.785	0.837
TP53	AKT1	treatment of c81 cells with ly294002 resulted in an increase in the p53-responsive gene mdm2 , suggesting a role for akt in the tax-mediated regulation of p53 transcriptional activity .	0.773	0.825
ABCB1	NR1I3	valproic acid induces cyp3a4 and mdr1 gene expression by activation of constitutive androstane receptor and pregnane x receptor pathways .	0.762	0.813
PTH2R	PTH2	thus , the juxtamembrane receptor domain specifies the signaling and binding selectivity of tip39 for the pth2 receptor over the pth1 receptor .	0.761	0.812

Gene1 Symbol	Gene2 Symbol	Text	Before Calibration	After Calibration
CCND1	ABL1	synergy with v-abl depended on a motif in cyclin d1 that mediates its binding to the retinoblastoma protein , suggesting that abl oncogenes in part mediate their mitogenic effects via a retinoblastoma protein-dependent pathway .	0.757	0.808
CTNND1	CDH1	these complexes are formed independently of ddr1 activation and of beta-catenin and p120-catenin binding to e-cadherin ; they are ubiquitous in epithelial cells .	0.748	0.798
CSF1	CSF1R	this is in agreement with current thought that the c-fms proto-oncogene product functions as the csf-1 receptor specific to this pathway .	0.745	0.795

Gene1 Symbol	Gene2 Symbol	Text	Before Calibration	After Calibration
EZR	CFTR	without ezrin binding , the cytoplasmic tail of cftr only interacts strongly with the first amino-terminal pdz domain to form a 1:1 c-cftr .	0.732	0.78
SRC	PIK3CG	we have demonstrated that the sh2 (src homology 2) domains of the 85 kda subunit of pi-3k are sufficient to mediate binding of the pi-3k complex to tyrosine phosphorylated , but not non-phosphorylated il-2r beta , suggesting that tyrosine phosphorylation is an integral component of the activation of pi-3k by the il-2r .	0.731	0.78

Table 10: Contains the bottom ten Gene-interacts-Gene confidence scores before and after model calbration. Both gene mentions highlighted in blue.

Gene1 Symbol	Gene2 Symbol	Text	Before Calibration	After Calibration

Gene1 Symbol	Gene2 Symbol	Text	Before Calibration	After Calibration
AGTR1	ACE	result (s) : the luteal tissue is the major site of ang ii , ace , at1r , and vegf , with highest staining intensity found during the midluteal phase and at pregnancy .	0.009	0.003
ABCE1	ABCF2	in relation to normal melanocytes , abcb3 , abcb6 , abcc2 , abcc4 , abce1 and abcf2 were significantly increased in melanoma cell lines , whereas abca7 , abca12 , abcb2 , abcb4 , abcb5 and abcd1 showed lower expression levels .	0.008	0.002
IL4	IFNG	in contrast , il-13ralpha2 mrna expression was up-regulated by ifn-gamma plus il-4 .	0.007	0.002
FCAR	CD79A	we report here the presence of circulating soluble fcalphar (cd89) - iga complexes in patients with igan .	0.007	0.002

Gene1 Symbol	Gene2 Symbol	Text	Before Calibration	After Calibration
IL4	VCAM1	similarly , il-4 induced vcam-1 expression and augmented tnf-alpha-induced expression on huvec but did not affect vcam-1 expression on hdmec .	0.007	0.002
IL2	IFNG	prostaglandin e2 at priming of naive cd4 + t cells inhibits acquisition of ability to produce ifn-gamma and il-2 , but not il-4 and il-5 .	0.006	0.002
IL2	FOXP3	il-1b promotes tgf-b1 and il-2 dependent foxp3 expression in regulatory t cells .	0.006	0.002
IL2	IFNG	the detailed distribution of lymphokine-producing cells showed that il-2 and ifn-gamma-producing cells were located mainly in the follicular areas .	0.005	0.001

Gene1 Symbol	Gene2 Symbol	Text	Before Calibration	After Calibration
IFNG	IL10	results : we found weak mrna expression of interleukin-4 (il-4) and il-5 , and strong expression of il-6 , il-10 and ifn-gamma before therapy .	0.005	0.001
PIK3R1	PTEN	both pten (pi3k antagonist) and pp2 (unspecific phosphatase) were down-regulated .	0.005	0.001

Top Ten Sentences for Each Edge Type

Table 11: 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

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
DaG	urinary bladder cancer	TP53	1	0.945	2112	Existing	conclusion : our findings indicate that the dsp53-285 can upregulate wild-type p53 expression in human bladder cancer cells through rna activation , and suppresses cells proliferation and metastasis in vitro and in vivo .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
DaG	ovarian cancer	EGFR	1	0.937	1330	Existing	conclusion : our data showed that increased expression of egfr is associated with poor prognosis of patients with eoc and dacomitinib may act as a novel , useful chemotherapy drug .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
DaG	stomach cancer	TP53	1	0.937	2679	Existing	conclusion : this meta - analysis suggests that p53 arg72 pro polymorphism is associated with increased risk of gastric cancer in asians .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
DaG	lung cancer	TP53	1	0.936	6813	Existing	conclusion : these results suggest that high expression of the p53 onco protein is a favorable prognostic factor in a subset of patients with nsclc .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
DaG	breast cancer	TCF7L2	1	0.936	56	Existing	this meta-analysis demonstrated that tcf7l2 gene polymorphisms (rs12255372 and rs7903146) are associated with an increased susceptibility to breast cancer .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
DaG	skin cancer	COX2	1	0.935	73	Novel	elevated expression of cox-2 has been associated with tumor progression in skin cancer through multiple mechanisms .
DaG	thyroid cancer	VEGFA	1	0.933	592	Novel	as a conclusion , we suggest that vegf g +405 c polymorphism is associated with increased risk of ptc .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
DaG	stomach cancer	EGFR	1	0.933	1237	Existing	recently , high lymph node ratio is closely associated with egfr expression in advanced gastric cancer .
DaG	liver cancer	GPC3	1	0.933	1944	Novel	conclusions serum gpc3 was overexpressed in hcc patients .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
DaG	stomach cancer	CCR6	1	0.931	24	Novel	the cox regression analysis showed that high expression of ccr6 was an independent prognostic factor for gc patients .
CtD	Sorafenib	liver cancer	1	0.99	6672	Existing	tace plus sorafenib for the treatment of hepatocellular carcinoma : final results of the multicenter socrates trial .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
CtD	Methotrexate	rheumatoid arthritis	1	0.989	14546	Existing	comparison of low-dose oral pulse methotrexate and placebo in the treatment of rheumatoid arthritis .
CtD	Auranofin	rheumatoid arthritis	1	0.988	419	Existing	auranofin versus placebo in the treatment of rheumatoid arthritis .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
CtD	Lamivudine	hepatitis B	1	0.988	6709	Existing	randomized controlled trials (rcts) comparing etv with lam for the treatment of hepatitis b decompensated cirrhosis were included .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
CtD	Doxorubicin	urinary bladder cancer	1	0.988	930	Existing	17-year follow-up of a randomized prospective controlled trial of adjuvant intravesical doxorubicin in the treatment of superficial bladder cancer .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
CtD	Docetaxel	breast cancer	1	0.987	5206	Existing	currently , randomized phase iii trials have demonstrated that docetaxel is an effective strategy in the adjuvant treatment of breast cancer .
CtD	Cimetidine	psoriasis	0.999	0.987	12	Novel	cimetidine versus placebo in the treatment of psoriasis .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
CtD	Olanzapine	schizophrenia	1	0.987	3324	Novel	a double-blind , randomised comparative trial of amisulpride versus olanzapine in the treatment of schizophrenia : short-term results at two months .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
CtD	Fulvestrant	breast cancer	1	0.987	826	Existing	phase iii clinical trials have demonstrated the clinical benefit of fulvestrant in the endocrine treatment of breast cancer .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
CtD	Pimecrolimus	atopic dermatitis	1	0.987	531	Existing	introduction : although several controlled clinical trials have demonstrated the efficacy and good tolerability of 1 % pimecrolimus cream for the treatment of atopic dermatitis , the results of these trials may not apply to real-life usage .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
CbG	Gefitinib	EGFR	1	0.99	8746	Existing	morphologic features of adenocarcinoma of the lung predictive of response to the epidermal growth factor receptor kinase inhibitors erlotinib and gefitinib .
CbG	Adenosine	EGFR	1	0.987	644	Novel	it is well established that inhibiting atp binding within the egfr kinase domain regulates its function .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
CbG	Rosiglitazone	PPARG	1	0.987	1498	Existing	rosiglitazone is a potent peroxisome proliferator-activated receptor gamma agonist that decreases hyperglycemia by reducing insulin resistance in patients with type 2 diabetes mellitus .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
CbG	D-Tyrosine	INSR	0.998	0.987	1713	Novel	this result suggests that tyrosine phosphorylation of phosphatidylinositol 3-kinase by the insulin receptor kinase may increase the specific activity of the former enzyme in vivo .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
CbG	D-Tyrosine	IGF1	0.998	0.983	819	Novel	affinity-purified insulin-like growth factor i receptor kinase is activated by tyrosine phosphorylation of its beta subunit .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
CbG	Pindolol	HTR1A	1	0.983	175	Existing	pindolol , a beta blocker with weak partial 5-ht1a receptor agonist activity has been shown to produce a more rapid onset of antidepressant action of ssris .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
CbG	Progesterone	SHBG	1	0.981	492	Existing	however , dng also elicits properties of progesterone derivatives like neutrality in metabolic and cardiovascular system and considerable antiandrogenic activity , the latter increased by lack of binding to shbg as specific property of dng .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
CbG	Mifepristone	AR	1	0.98	78	Existing	ru486 bound to the androgen receptor .
CbG	Alfentanil	OPRM1	1	0.979	10	Existing	purpose : alfentanil is a high potency mu opiate receptor agonist commonly used during presurgical induction of anesthesia .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
CbG	Candesartan	AGTR1	1	0.979	36	Existing	tcv-116 is a new , nonpeptide , angiotensin ii type-1 receptor antagonist that acts as a specific inhibitor of the renin - angiotensin system .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
GiG	BRCA2	BRCA1	0.972	0.984	12257	Novel	a total of 9 families (16 %) showed mutations in the brca1 gene , including the one new mutation identified in this study (5382insc) , and 12 families (21 %) presented mutations in the brca2 gene .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
GiG	MDM2	TP53	0.938	0.978	17128	Existing	no mutations in the tp53 gene have been found in samples with amplification of mdm2 .
GiG	BRCA1	BRCA2	1	0.978	12257	Existing	pathogenic truncating mutations in the brca1 gene were found in two tumor samples with allelic losses , whereas no mutations were identified in the brca2 gene .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
GiG	KRAS	TP53	0.992	0.971	4106	Novel	mutations in the p53 gene did not correlate with mutations in the c-k-ras gene , indicating that colorectal cancer can develop through pathways independent not only of the presence of mutations in any of these genes but also of their cooperation .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
GiG	TP53	HRAS	0.992	0.969	451	Novel	pathologic examination of the uc specimens from aa-exposed patients identified heterozygous hras changes in 3 cases , and deletion or replacement mutations in the tp53 gene in 4 .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
GiG	REN	NR1H3	0.998	0.966	8	Novel	nuclear receptor lxralpha is involved in camp - mediated human renin gene expression .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
GiG	ESR2	CYP19A1	0.999	0.96	159	Novel	dna methylation , histone modifications , and binding of estrogen receptor , erb to regulatory dna sequences of cyp19a1 gene were evaluated by chromatin immunoprecipitation (chip) assay .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
GiG	RET	EDNRB	0.816	0.96	136	Novel	mutations in the ret gene , which codes for a receptor tyrosine kinase , and in ednrb which codes for the endothelin -b receptor , have been shown to be associated with hscr in humans .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
GiG	PKD1	PKD2	1	0.959	1614	Existing	approximately 85 % of adpkd cases are caused by mutations in the pkd1 gene , while mutations in the pkd2 gene account for the remaining 15 % of cases .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	Number of Sentences	In Hetionet	Text
GiG	LYZ	CTCF	0.999	0.959	2	Novel	in conjunction with the thyroid receptor (tr) , ctcf binding to the lysozyme gene transcriptional silencer mediates the thyroid hormone response element (tre) - dependent transcriptional repression .