Reusing label functions to extract multiple types of relationships from biomedical abstracts at scale

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@e0278f8 on January 23, 2020.

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 bases support multiple research efforts such as providing contextual information for biomedical entities, constructing networks, and supporting the interpretation of high-throughput analyses. Some knowledge bases are automatically constructed, but most are populated via some form of manual curation. Manual curation is time consuming and difficult to scale in the context of an increasing publication rate. A recently described “data programming” paradigm seeks to circumvent this arduous process by combining distant supervision with simple rules and heuristics written as labeling functions that can be automatically applied to inputs. Unfortunately writing useful label functions requires substantial error analysis and is a nontrivial task: in early efforts to use data programming we found that producing each label function could take a few days. Producing a biomedical knowledge base with multiple node and edge types could take hundreds or possibly thousands of label functions. In this paper we sought to evaluate the extent to which label functions could be re-used across edge types. We used a subset of Hetionet v1 that centered on disease, compound, and gene nodes to evaluate this approach. We compared a baseline distant supervision model with the same distant supervision resources added to edge-type-specific label functions, edge-type-mismatch label functions, and all label functions. We confirmed that adding additional edge-type-specific label functions improves performance. We also found that adding one or a few edge-type-mismatch label functions nearly always improved performance. Adding a large number of edge-type-mismatch label functions produce variable performance that depends on the edge type being predicted and the label function’s edge type source. Lastly, we show that this approach, even on this subgraph of Hetionet, could add new edges to Hetionet v1 with high confidence. We expect that practical use of this strategy would include additional filtering and scoring methods which would further enhance precision.

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 [1,2,3] 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 [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 large corpora of well-labeled training data. These corpora are difficult to obtain, because they are constructed via particularly detailed manual curation. Distant supervision is a technique designed to sidestep the dependence on manual curation and quickly generate large training datasets. This technique makes the assumption 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 immense amount of false positives [10].

Ratner et al. [11] 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 conduct 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 time required to populate a knowledge base with a multitude of heterogeneous relationships.

Related Work

Relationship extraction is the process of detecting and classifying 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 the identification of key text patterns for relationship extraction, (2) the use of unsupervised methods via co-occurrence scores or clustering, and (3) supervised or semi-supervised machine learning using annotated datasets for the classification of documents or sentences. In this section, we discuss selected efforts for each type of edge that we include in this project.

Disease-Gene Associations

Efforts to extract Disease-associates-Gene (DaG) relationships have often used manually crafted rules or unsupervised methods. One study used hand crafted rules based on a sentence's grammatical structure, represented as dependency trees, to extract DaG relationships [12]. Some of these rules inspired certain DaG text pattern label functions in our work. Another study used co-occurrence frequencies within abstracts and sentences to score the likelihood of association between disease and gene pairs [13]. The results of this study were incorporated into Hetionet v1 [3], so this served as one of our distant supervision label functions. Another approach built off of the above work by incorporating a supervised classifier, trained via distant supervision, into a scoring scheme [14]. Each sentence containing a disease and gene mention is scored using a logistic regression model and combined using the same co-occurrence approach used in Pletscher-Frankild et al. [13]. We compared our results to this approach to measure how well our overall method performs relative to other methods. Besides the mentioned three studies, researchers have used co-occurrences for extraction alone [15,16,17] or in combination with other features to recover DaG relationships [18]. One recent effort relied on a bi-clustering approach to detect DaG-relevant sentences from Pubmed abstracts [19] with clustering of dependency paths grouping similar sentences together. The results of this work supply our domain heuristic label functions. These approaches do not rely on a well-annotated training performance and tend to provide excellent recall, though the precision is often worse than with supervised methods [20,21].

Hand-crafted high-quality datasets [22,23,24,25] often serve as a gold standard for training, tuning, and testing supervised machine learning methods in this setting. Support vector machines have been repeatedly used to detect DaG relationships [22,26,27]. These models perform well in large feature spaces, but are slow to train as the number of data points becomes large. Recently, some studies have used deep neural network models. One used a pre-trained recurrent neural network [28], and another used distant supervision [29]. Due to the success of these two models, we decided to use a deep neural network as our discriminative model.

Compound Treats Disease

The goal of extracting Compound-treats-Disease (CtD) edges is to identify sentences that mention current drug treatments or propose new uses for existing drugs. One study combined an inference model from previously established drug-gene and gene-disease relationships to infer novel drug-disease interactions via co-occurrences [30]. A similar approach has also been applied to CtD extraction [31]. Manually-curated rules have also been applied to PubMed abstracts to address this task [32]. The rules were based on identifying key phrases and wordings related to using drugs to treat a disease, and we used these patterns as inspirations for some of our CtD label functions. Lastly, one study used a bi-clustering approach to identify sentences relevant to CtD edges [19]. As with DaG edges, we use the results from this study to provide what we term as domain heuristic label functions.

Recent work with supervised machine learning methods has often focused on compounds that induce a disease: an important question for toxicology and the subject of the BioCreative V dataset [33]. We don't consider environmental toxicants in our work, as our source databases for distant supervision are primarily centered around FDA-approved therapies.

Compound Binds Gene

The BioCreative VI track 5 task focused on classifying compound-protein interactions and has led to a great deal of work on the topic [34]. The equivalent edge in our networks is Compound-binds-Gene (CbG). Curators manually annotated 2,432 PubMed abstracts for five different compound protein interactions (agonist, antagonist, inhibitor, activator and substrate/product production) as part of the BioCreative task. The best performers on this task achieved an F1 score of 64.10% [34]. Numerous additional groups have now used the publicly available dataset, that resulted from this competition, to train supervised machine learning methods [28,35,36,37,37,38,39,40,41] and semi-supervised machine learning methods [42]. These approaches depend on well-annotated training datasets, which creates a bottleneck. In addition to supervised and semi-supervised machine learning methods, hand crafted rules [43] and bi-clustering of dependency trees [19] have been used. We use the results from the bi-clustering study to provide a subset of the CbG label functions in this work.

Gene-Gene Interactions

Akin to the DaG edge type, many efforts to extract Gene-interacts-Gene (GiG) relationships used co-occurrence approaches. This edge type is more frequently referred to as a protein-protein interaction. Even approaches as simple as calculating Z-scores from PubMed abstract co-occurrences can be informative [44], and there are numerous studies using co-occurrences [17,45,46,47]. However, more sophisticated strategies such as distant supervision appear to improve performance [14]. Similarly to the other edge types, the bi-clustering approach over dependency trees has also been applied to this edge type [19]. This manuscript provides a set of label functions for our work.

Most supervised classifiers used publicly available datasets for evaluation [48,49,50,51,52]. These datasets are used equally among studies, but can generate noticeable differences in terms of performance [53]. Support vector machines were a common approach to extract GiG edges [54,55]. However, with the growing popularity of deep learning numerous deep neural network architectures have been applied [42,56,57,58]. Distant supervision has also been used in this domain [59], and in fact this effort was one of the motivating rationales for our work.

Materials and Methods

Hetionet

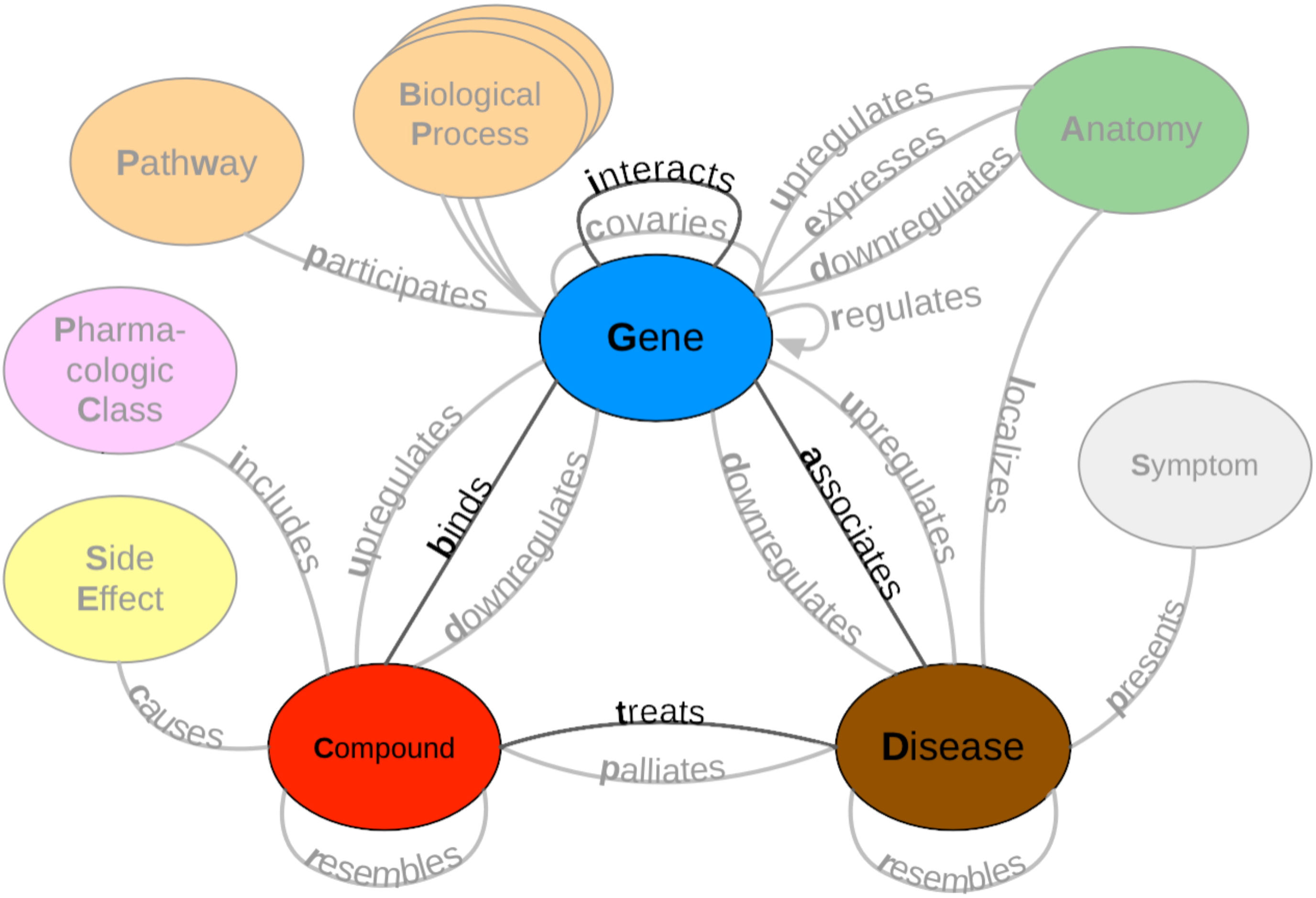

Figure 1: A metagraph (schema) of Hetionet where biomedical entities are represented as nodes and the relationships between them are represented as edges. We examined performance on the highlighted subgraph; however, the long-term vision is to capture edges for the entire graph.

Hetionet [3] is a large heterogenous network that contains pharmacological and biological information. This network depicts information in the form of nodes and edges of different types: nodes that represent biological and pharmacological entities and edges which represent relationships between entities. Hetionet v1.0 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 were obtained from open databases, such as the GWAS Catalog [60] and DrugBank [61]. For this project, we analyzed performance over a subset of the Hetionet relationship types: disease associates with a gene (DaG), compound binds to a gene (CbG), gene interacts with gene (GiG) and compound treating a disease (CtD).

Dataset

We used PubTator [62] as input to our analysis. PubTator provides MEDLINE abstracts that have been annotated with well-established entity recognition tools including DNorm [63] for disease mentions, GeneTUKit [64] for gene mentions, Gnorm [65] for gene normalizations and a dictionary based search system for compound mentions [66]. We downloaded PubTator on June 30, 2017, at which point it contained 10,775,748 abstracts. Then we filtered out mention tags that were not contained in hetionet. We used the Stanford CoreNLP parser [67] to tag parts of speech and generate dependency trees. We extracted sentences with two or more mentions, termed candidate sentences. Each candidate sentence was stratified by co-mention pair to produce a training set, tuning set and a testing set (shown in Table 1). Each unique co-mention pair is sorted into four categories: (1) in hetionet and has sentences, (2) in hetionet and doesn't have sentences, (3) not in hetionet and does have sentences and (4) not in hetionet and doesn't have sentences. Within these four categories each pair is randomly assigned their own individual partition rank (continuous number between 0 and 1).

Any rank lower than 0.7 is sorted into the training set, while any rank greater than 0.7 and lower than 0.9 is assigned to the tuning set. The rest of the pairs with a rank greater than or equal to 0.9 is assigned to the test set. Sentences that contain more than one co-mention pair are treated as multiple individual candidates. We hand labeled five hundred to a thousand candidate sentences of each relationship type to obtain a ground truth set (Table 1)[1].

Table 1: Statistics of Candidate Sentences. We sorted each candidate sentence 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	2.35 M	31K (397+, 603-)	313K (351+, 649-)
Compound Binds Gene	1.7M	468K (37+, 463-)	227k (31+, 469-)
Compound Treats Disease	1.013M	96K (96+, 404-)	32K (112+, 388-)
Gene Interacts Gene	12.6M	1.056M (60+, 440-)	257K (76+, 424-)

Label Functions for Annotating Sentences

The challenge of having too few ground truth annotations is common to many natural language processing settings, even when unannotated text is abundant. Data programming circumvents this issue by quickly annotating large datasets by using multiple noisy signals emitted by label functions [11]. Label functions are simple pythonic functions that emit: a positive label (1), a negative label (-1) or abstain from emitting a label (0). We combine these functions using a generative model to output a single annotation, which 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 discriminator model that makes the final classification step.

Label Function Categories

Label functions can be constructed in a multitude of ways; however, many label functions share similar characteristics with one another.
We group these characteristics into the following categories: databases, text patterns and domain heuristics. Most of our label functions fall into the text pattern category, while the others were distributed across the database and domain heuristic categories (Table 2). We describe each category and provide an example using the 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. If the pair is present, our label function emits a positive label 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 that we encountered during development: emitting positives and negatives from the same label function causes downstream classifiers to generate almost exclusively negative predictions.

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

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

Domain Heuristics: These label functions used results from published text-based analyses to generate a signal. We used dependency path cluster themes generated by Percha et al. [19]. If a candidate sentence's dependency path belonged to a previously generated cluster, then the label function emitted a positive label and abstained otherwise.

$$\Lambda_{DH}(D, G) = \begin{cases} 1 & Candidate\ Sentence \in Cluster\ Theme \\ 0 & otherwise \end{cases}$$

Text Patterns: These label functions are designed to use keywords and sentence context to generate a signal. For example, a label function could focus on the number of words between two mentions or focus on the grammatical structure of a sentence. These functions emit a positive or negative label depending on the context.

$$\Lambda_{TP}(D, G) = \begin{cases} 1 & "\ target\ " \in Candidate\ Sentence \\ 0 & otherwise \end{cases}$$

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

Each text pattern label function was constructed by manual examination of sentences within the training set. For example, in the candidate sentence above one would extract the keywords "novel therapeutic target" and incorporate them in a text pattern label function. After initial construction, we tested and augmented the label function using sentences in the tune set. We repeated the above process for each label function in our repertoire.

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

Relationship	Databases (DB)	Text Patterns (TP)	Domain Heuristics (DH)
DaG	7	20	10
CtD	3	15	7
CbG	9	13	7
GiG	9	20	8

Training Models

Generative Model

The generative model is a core part of this automatic annotation framework. It integrates multiple signals emitted by label functions and assigns a training class to each candidate sentence. This model assigns training classes by estimating the joint probability distribution of the latent true class (Y) and label function signals (Λ), ($P_{\theta}(\Lambda, Y)$). Assuming each label function is conditionally independent, the joint distribution is defined as follows:

$$P_{\theta}(\Lambda, Y) = \frac{\exp(\sum_{i=1}^{m} \theta^{T} F_{i}(\Lambda, y))}{\sum_{\Lambda'} \sum_{y'} \exp(\sum_{i=1}^{m} \theta^{T} F_{i}(\Lambda', y'))}$$

where m is the number of candidate sentences, F is the vector of summary statistics and θ is a vector of weights for each summary statistic. The summary statistics used by the generative model are as follows:

$$F_{i,j}^{Lab}(\Lambda, Y) = \mathbb{1}\{\Lambda_{i,j} \neq 0\}$$
$$F_{i,j}^{Acc}(\Lambda, Y) = \mathbb{1}\{\Lambda_{i,j} = y_{i,j}\}$$

Lab is the label function's propensity (the frequency of a label function emitting a signal). *Acc* is the individual label function's accuracy given the training class. This model optimizes the weights (θ) by minimizing the negative log likelihood:

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

In the framework we used predictions from the generative model, $\hat{Y} = P_{\hat{\theta}}(Y \mid \Lambda)$, as training classes for our dataset [68,69].

Experimental Design

Being able to re-use 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. For example, in the Gene interacts Gene (GiG) edge type we used label functions that returned a **1** if the pair of genes were included in the Human Interaction database [70], the iRefIndex database [71] or in the Incomplete Interactome database [72]. Then we compared the baseline model with models that also included text and domain-heuristic label functions. Using a sampling with replacement approach, we sampled these text and domain-heuristic label functions separately 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. For each edge type we sampled a fixed number of label functions 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. We evaluated both generative and discriminative (training and downstream analyses are described in the supplemental methods section) models at each point, and report performance of each in terms of the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPR).

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 report results in terms of AUROC (Figures 2 and ??) and AUPR (Supplemental Figure 7 and 8).

Performance increases when edge-specific label functions are added to an edge-specific baseline model, while label function reusability shows modest results. The quintessential example of the 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 vise-versa. This suggests that sentences for GiG and CbG may share similar linguistic features or terminology that allows for label functions to be reused. 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 show significant signs of reusability which suggests label functions could be shared between the two edge types.

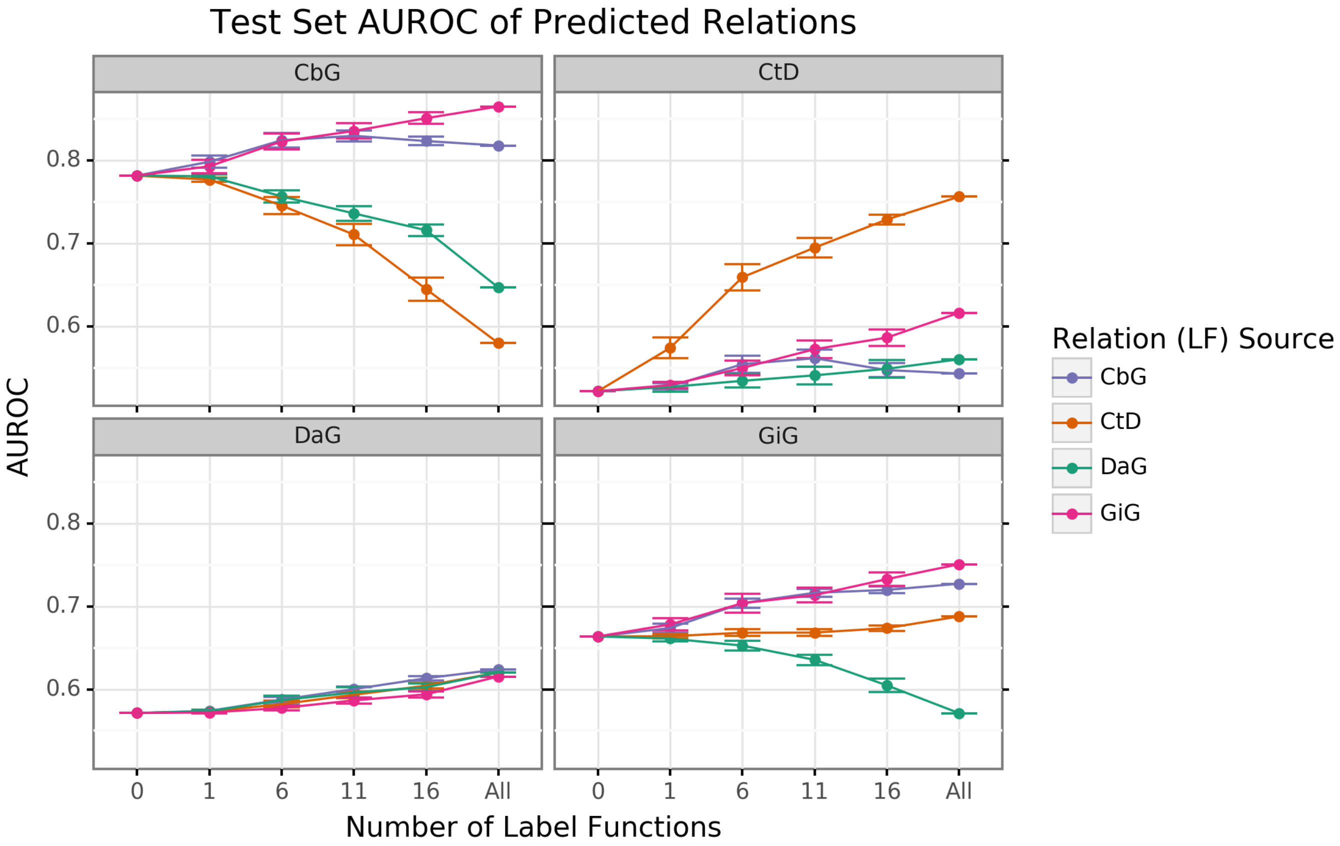

Figure 2: Edge-specific label functions are better performing than edge-mismatch label functions but certain mismatch situations show signs of successful transfer. Each line plot header depicts the edge type the generative model is trying to predict, while the colors represent the source of label functions. 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 onto 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 (Figure {#fig:auroc_grabbag_gen_model_test_set}). As more label functions were sampled, the gap between edge-specific sources and all sources widened. CbG is a prime example of this trend (Figure {#fig:auroc_grabbag_gen_model_test_set}), 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 {#fig:auroc_gen_model_test_set}). 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 vise versa).

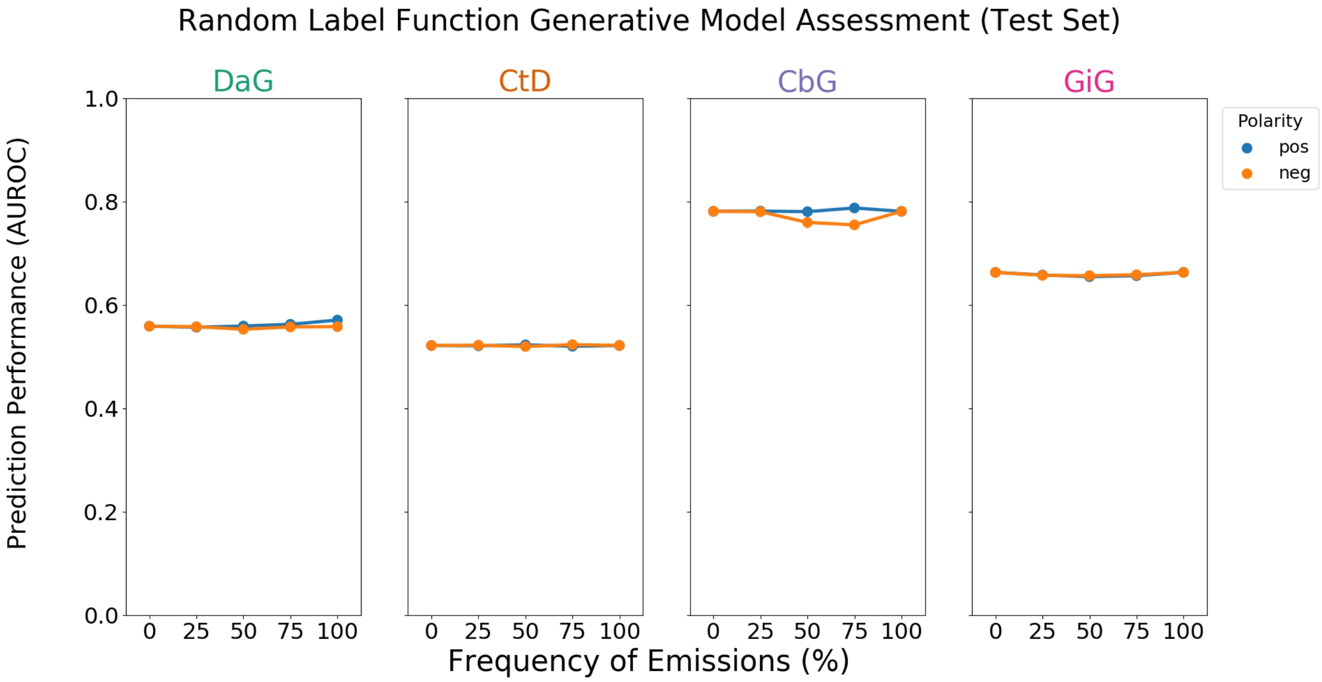

Figure 3: A grid of AUROC (A) scores for each edge type. Each plot consists of adding a single label function on top of the baseline model. This label function emits a positive (shown in blue) or negative (shown in orange) label at specified frequencies, and performance at zero is equivalent to not having a randomly emitting label function. The error bars represent 95% confidence intervals for AUROC or AUPR (y-axis) at each emission frequency.

We observed that including one label function of a mismatched type to distant supervision often improved performance, so we evaluated the effects of adding a random label function in the same setting. We found that usually adding random noise did not improve performance (Figure 3 and Supplemental Figure ??). For the CbG edge type we did observe slightly increased performance via AUPR (Supplemental Figure ??). However, performance changes in general were smaller than those observed with mismatched label types.

Discriminative Model Performance

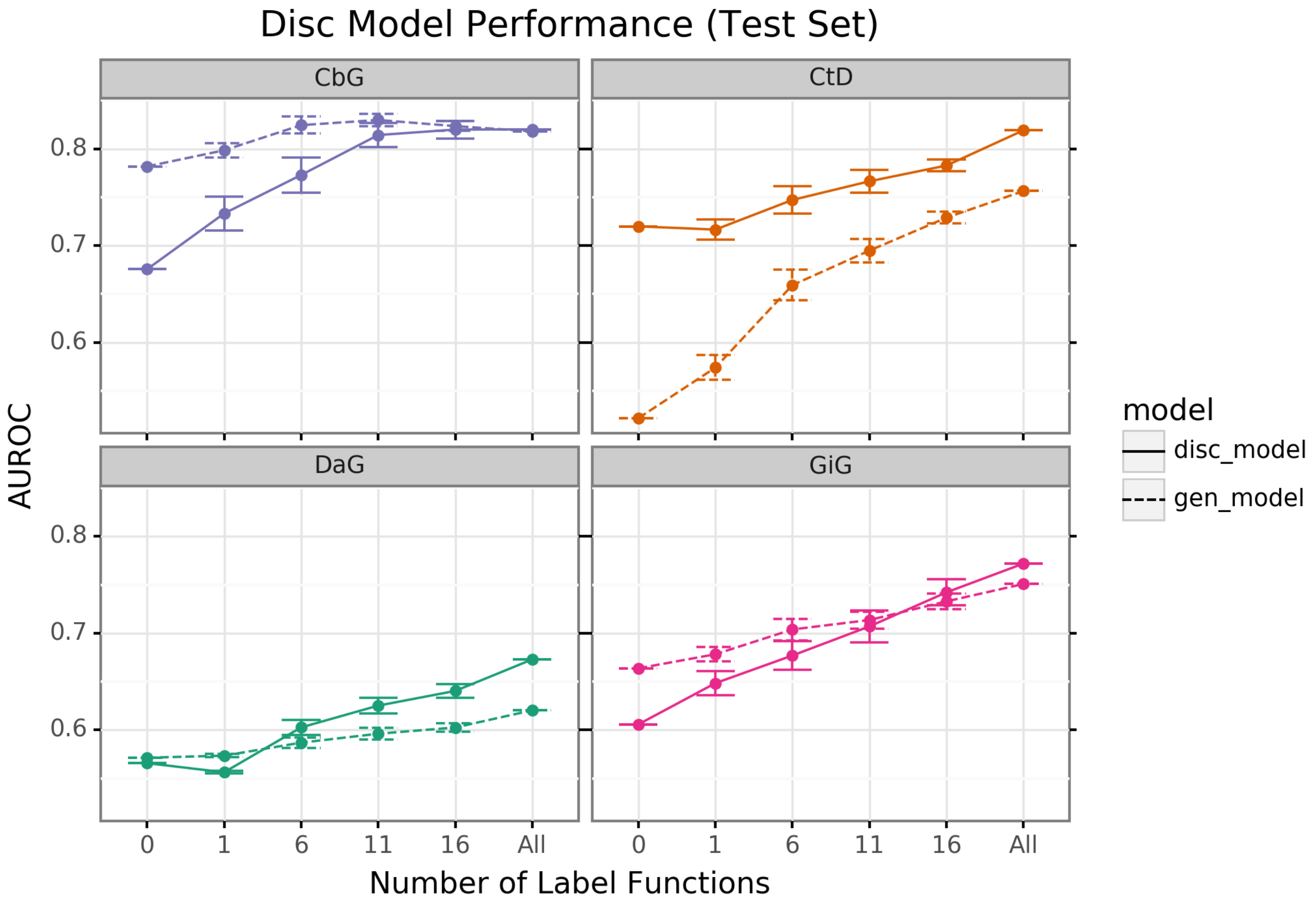

Figure 4: The discriminator model usually improves at a faster rate than the generative model as more edge-specific label function are included. The line plot headers represents the specific edge type the discriminator model is trying to predict. The x-axis shows the number of randomly sampled label functions that are incorporated on top of the baseline model (point at 0). The y axis shows the area under the receiver operating curve (AUROC). Each datapoint represents the average of each 50 sample run 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. This makes the error bars appear flat.

The discriminator 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). We found that the discriminative model generally out-performed the generative model as more edge-specific label functions are incorporated (Figure 4 and Supplemental Figure 9). The discriminator 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 discriminator model out-performs the generative model at all levels of sampling. We observed the opposite trend with the Compound-binds-Gene (CbG) edges: the discriminator model was always poorer or indistinguishable from the generative model. Interestingly, the AUPR for CbG plateaus below the generative model and the decreases when all edge-specific label functions are used (Supplemental Figure 9). This suggests that the discriminator model might be predicting more false positives in this setting. Incorporating more edge-specific label functions usually improves performance for the discriminator model over the generator model.

Discriminative Model Calibration

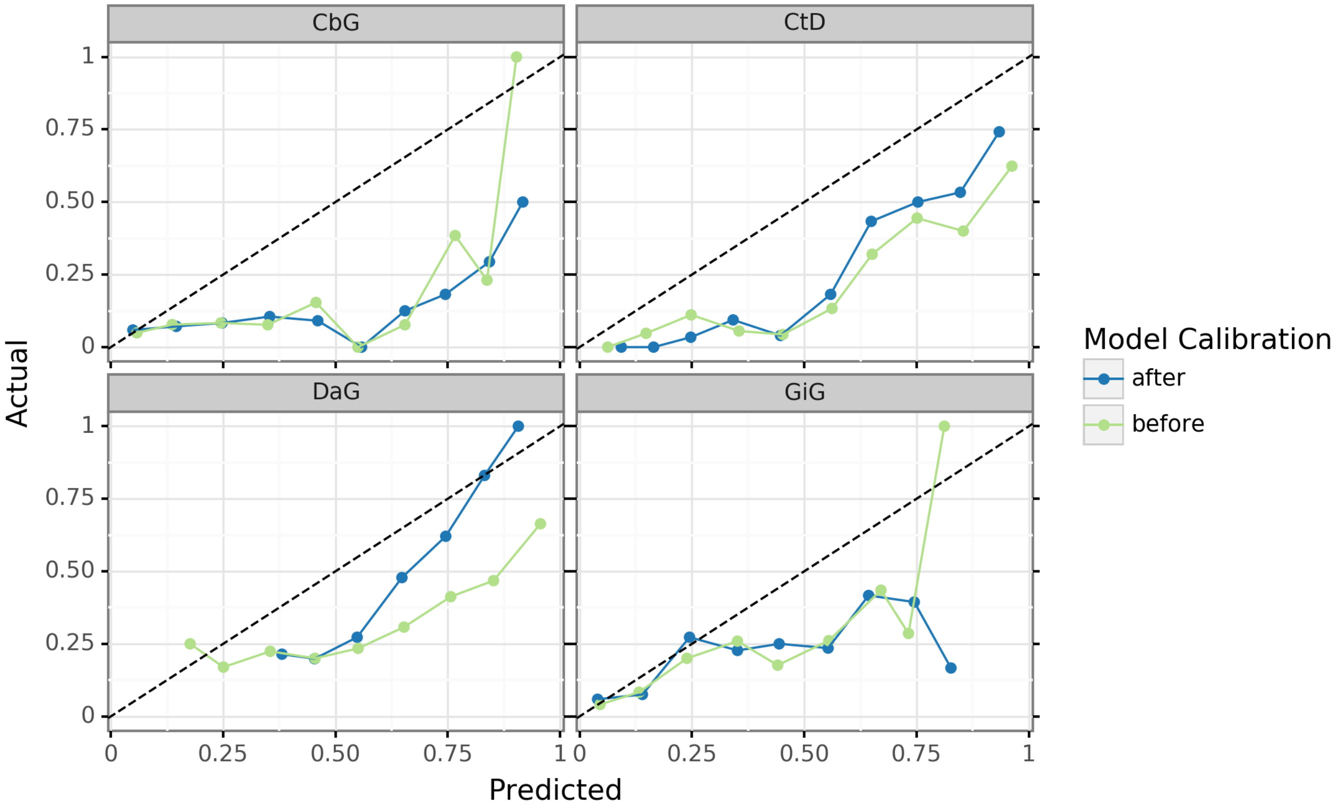

Figure 5: Deep learning models are overconfident in their predictions and need to be calibrated after training. These are calibration plots for the discrimintative model. The green line represents the predictions before calibration and the blue line shows predictions after calibration. Data points that lie closer to diagonal line show better model calibration, while data points far from the diagonal show poor performance. A perfectly calibrated model would align straight along the diagonal line.

Even deep learning models with good AUROC and AUPR statistics can be subject to poor calibration. Typically, these models are overconfident in their predictions [73,74]. We attempted to use temperature scaling to fix the calibration of the best performing discriminative models (Figure 5). 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.

Discussion

We tested the feasibility of re-using label functions to extract relationships from literature. Through our sampling experiment, we found that adding relevant label functions increases prediction performance (shown in the on-diagonals of Figures ?? and Supplemental Figure ??). We found that label functions designed from relatively related edge types can increase performance (seen when GiG label functions predicts CbG and vice versa). We noticed that one edge type (DaG) is agnostic to label function source (Figure ?? and Supplemental Figure ??). Performance routinely increases when adding a single mismatched label function to our baseline model (the generative model trained only on distant supervision label functions). These results led us to hypothesize that adding a small amount of noise aided the model, but our experiment with a random label function reveals that this was not the case (Figures 3 and ??). Based on these results one question still remains: why does performance drastically increase when adding a single label function to our distant supervision baseline?

The discriminative model didn't work as intended. The majority of the time the discriminative model underperformed the generative model (Supplemental Figures 4 and 9). Potential reasons for this are the discriminative model overfitting to the generative model's predictions and a negative class bias in some of our datasets (Table 1). The challenges with the discriminative model are likely to have led to issues in our downstream analyses: poor model calibration (Supplemental Figure 5) and poor recall in detecting existing Hetionet edges (Supplemental Figure 11). Despite the above complications, our model had similar performance with a published baseline model (Supplemental Figure 10). This implies that with better tuning the discriminative model has the potential to perform better than the baseline model.

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 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 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 ingest 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 where text-derived edges generally cannot be exactly attributed to excerpts from literature [3,75], 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 [76,77,78]. New multitask learning [69] strategies may make it even more practical to reuse label functions to construct continuously updating literature-derived knowledge graphs.

Supplemental Information

This manuscript and supplemental information are available at https://greenelab.github.io/text_mined_hetnet_manuscript/. 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 Univeristy, 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
Cold Spring Harbor Laboratory (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

5. **CoCoScore: Context-aware co-occurrence scoring for text mining applications using distant supervision**
Alexander Junge, Lars Juhl Jensen
Cold Spring Harbor Laboratory (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 (2019-05-21) 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 Jr, K. Bretonnel 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**
Lutz Bornmann, Rüdiger Mutz

Journal of the Association for Information Science and Technology (2015-04-29) https://doi.org/gfj5zc
DOI: 10.1002/asi.23329

10. **Revisiting distant supervision for relation extraction**
Tingsong Jiang, Jing Liu, Chin-Yew Lin, Zhifang Sui
LREC (2018)

11. **Data Programming: Creating Large Training Sets, Quickly**
Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, Christopher Ré
arXiv (2016-05-25) https://arxiv.org/abs/1605.07723v3

12. **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

13. **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

14. **CoCoScore: context-aware co-occurrence scoring for text mining applications using distant supervision**
Alexander Junge, Lars Juhl Jensen
Bioinformatics (2019-06-14) https://doi.org/gf4789
DOI: 10.1093/bioinformatics/btz490 · PMID: 31199464

15. **LGscore: A method to identify disease-related genes using biological literature and Google data**
Jeongwoo Kim, Hyunjin Kim, Youngmi Yoon, Sanghyun Park
Journal of Biomedical Informatics (2015-04) https://doi.org/f7bj9c
DOI: 10.1016/j.jbi.2015.01.003 · PMID: 25617670

16. **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 (2015-04-29) https://doi.org/f7nzn5
DOI: 10.1093/nar/gkv383 · PMID: 25925572 · PMCID: PMC4489268

17. **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

18. **The research on gene-disease association based on text-mining of PubMed**
Jie Zhou, Bo-quan Fu
BMC Bioinformatics (2018-02-07) https://doi.org/gf479k
DOI: 10.1186/s12859-018-2048-y · PMID: 29415654 · PMCID: PMC5804013

19. **A global network of biomedical relationships derived from text**
Bethany Percha, Russ B Altman

Bioinformatics (2018-02-27) https://doi.org/gc3ndk
DOI: 10.1093/bioinformatics/bty114 · PMID: 29490008 · PMCID: PMC6061699

20. **Literature mining for the biologist: from information retrieval to biological discovery**
Lars Juhl Jensen, Jasmin Saric, Peer Bork
Nature Reviews Genetics (2006-02) https://doi.org/bgq7q9
DOI: 10.1038/nrg1768 · PMID: 16418747

21. **Application of text mining in the biomedical domain**
Wilco W. M. Fleuren, Wynand Alkema
Methods (2015-03) https://doi.org/f64p6n
DOI: 10.1016/j.ymeth.2015.01.015 · PMID: 25641519

22. **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 (2015-02-21) https://doi.org/f7kn8s
DOI: 10.1186/s12859-015-0472-9 · PMID: 25886734 · PMCID: PMC4466840

23. **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

24. **CoMAGC: a corpus with multi-faceted annotations of gene-cancer relations**
Hee-Jin Lee, Sang-Hyung Shim, Mi-Ryoung Song, Hyunju Lee, Jong C Park
BMC Bioinformatics (2013) https://doi.org/gb8v5s
DOI: 10.1186/1471-2105-14-323 · PMID: 24225062 · PMCID: PMC3833657

25. **Concept annotation in the CRAFT corpus**
Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A Baumgartner, K Bretonnel Cohen, Karin Verspoor, Judith A Blake, Lawrence E Hunter
BMC Bioinformatics (2012-07-09) https://doi.org/gb8vdr
DOI: 10.1186/1471-2105-13-161 · PMID: 22776079 · PMCID: PMC3476437

26. **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

27. **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

28. **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 (2019-01-25) https://arxiv.org/abs/1901.08746v4
DOI: 10.1093/bioinformatics/btz682

29. **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 (2018-11-20)
https://doi.org/gf5hfv
DOI: 10.1007/978-981-13-2354-6_39

30. **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

31. **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

32. **Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing**
Rong Xu, QuanQiu Wang
BMC Bioinformatics (2013-06-06) https://doi.org/gb8v3k
DOI: 10.1186/1471-2105-14-181 · PMID: 23742147 · PMCID: PMC3702428

33. **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

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

35. **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 (2017-05-22) https://doi.org/gf479h
DOI: 10.1093/nar/gkx462 · PMID: 28531339 · PMCID: PMC5570141

36. **LPTK: a linguistic pattern-aware dependency tree kernel approach for the BioCreative VI CHEMPROT task**
Neha Warikoo, Yung-Chun Chang, Wen-Lian Hsu
Database (2018-01-01) https://doi.org/gfhjr6
DOI: 10.1093/database/bay108 · PMID: 30346607 · PMCID: PMC6196310

37. **Extracting chemical–protein relations with ensembles of SVM and deep learning models**
Yifan Peng, Anthony Rios, Ramakanth Kavuluru, Zhiyong Lu

Database (2018-01-01) https://doi.org/gf479f
DOI: 10.1093/database/bay073 · PMID: 30020437 · PMCID: PMC6051439

38. **Extracting chemical–protein interactions from literature using sentence structure analysis and feature engineering**
Pei-Yau Lung, Zhe He, Tingting Zhao, Disa Yu, Jinfeng Zhang
Database (2019-01-01) https://doi.org/gf479g
DOI: 10.1093/database/bay138 · PMID: 30624652 · PMCID: PMC6323317

39. **Improving the learning of chemical-protein interactions from literature using transfer learning and specialized word embeddings**
P Corbett, J Boyle
Database (2018-01-01) https://doi.org/gf479d
DOI: 10.1093/database/bay066 · PMID: 30010749 · PMCID: PMC6044291

40. **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: 10.1093/database/bay102 · PMID: 30295724 · PMCID: PMC6174551

41. **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

42. **Exploring Semi-supervised Variational Autoencoders for Biomedical Relation Extraction**
Yijia Zhang, Zhiyong Lu
arXiv (2019-01-18) https://arxiv.org/abs/1901.06103v1

43. **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

44. **STRING v10: protein–protein interaction networks, integrated over the tree of life**
Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P. Tsafou, … Christian von Mering
Nucleic Acids Research (2014-10-28) https://doi.org/f64rfn
DOI: 10.1093/nar/gku1003 · PMID: 25352553 · PMCID: PMC4383874

45. **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

46. **HPIminer: A text mining system for building and visualizing human protein interaction networks and pathways**
Suresh Subramani, Raja Kalpana, Pankaj Moses Monickaraj, Jeyakumar Natarajan

Journal of Biomedical Informatics (2015-04) https://doi.org/f7bgnr
DOI: 10.1016/j.jbi.2015.01.006 · PMID: 25659452

47. **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 (2019-02-08) https://doi.org/gf49nm
DOI: 10.1186/s12859-019-2634-7 · PMID: 30736752 · PMCID: PMC6368766

48. **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

49. **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 (2007-02-09) https://doi.org/b7bhhc
DOI: 10.1186/1471-2105-8-50 · PMID: 17291334 · PMCID: PMC1808065

50. **Learning language in logic - genic interaction extraction challenge**
C. Nédellec
Proceedings of the learning language in logic 2005 workshop at the international conference on machine learning (2005)

51. **RelEx–Relation extraction using dependency parse trees**
K. Fundel, R. Kuffner, R. Zimmer
Bioinformatics (2006-12-01) https://doi.org/cz7q4d
DOI: 10.1093/bioinformatics/btl616 · PMID: 17142812

52. **Mining medline: Abstracts, sentences, or phrases?**
Jing Ding, Daniel Berleant, Dan Nettleton, Eve Syrkin Wurtele
Pacific symposium on biocomputing (2002) http://helix-web.stanford.edu/psb02/ding.pdf

53. **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

54. **Exploiting graph kernels for high performance biomedical relation extraction**
Nagesh C. Panyam, Karin Verspoor, Trevor Cohn, Kotagiri Ramamohanarao
Journal of Biomedical Semantics (2018-01-30) https://doi.org/gf49nn
DOI: 10.1186/s13326-017-0168-3 · PMID: 29382397 · PMCID: PMC5791373

55. **Text Mining for Protein Docking**
Varsha D. Badal, Petras J. Kundrotas, Ilya A. Vakser
PLOS Computational Biology (2015-12-09) https://doi.org/gcvj3b
DOI: 10.1371/journal.pcbi.1004630 · PMID: 26650466 · PMCID: PMC4674139

56. **Feature assisted stacked attentive shortest dependency path based Bi-LSTM model for protein–protein interaction**
Shweta Yadav, Asif Ekbal, Sriparna Saha, Ankit Kumar, Pushpak Bhattacharyya

Knowledge-Based Systems (2019-02) https://doi.org/gf4788
DOI: 10.1016/j.knosys.2018.11.020

57. **Extraction of protein–protein interactions (PPIs) from the literature by deep convolutional neural networks with various feature embeddings**
Sung-Pil Choi
Journal of Information Science (2016-11-01) https://doi.org/gcv8bn
DOI: 10.1177/0165551516673485

58. **Deep learning for extracting protein-protein interactions from biomedical literature**
Yifan Peng, Zhiyong Lu
arXiv (2017-06-05) https://arxiv.org/abs/1706.01556v2

59. **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

60. **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 (2016-11-29) https://doi.org/f9v7cp
DOI: 10.1093/nar/gkw1133 · PMID: 27899670 · PMCID: PMC5210590

61. **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 (2017-11-08) https://doi.org/gcwtzk
DOI: 10.1093/nar/gkx1037 · PMID: 29126136 · PMCID: PMC5753335

62. **PubTator: a web-based text mining tool for assisting biocuration**
Chih-Hsuan Wei, Hung-Yu Kao, Zhiyong Lu
Nucleic Acids Research (2013-05-22) https://doi.org/f475th
DOI: 10.1093/nar/gkt441 · PMID: 23703206 · PMCID: PMC3692066

63. **DNorm: disease name normalization with pairwise learning to rank**
R. Leaman, R. Islamaj Dogan, Z. Lu
Bioinformatics (2013-08-21) https://doi.org/f5gj9n
DOI: 10.1093/bioinformatics/btt474 · PMID: 23969135 · PMCID: PMC3810844

64. **GeneTUKit: a software for document-level gene normalization**
M. Huang, J. Liu, X. Zhu
Bioinformatics (2011-02-08) https://doi.org/dng2cb
DOI: 10.1093/bioinformatics/btr042 · PMID: 21303863 · PMCID: PMC3065680

65. **Cross-species gene normalization by species inference**
Chih-Hsuan Wei, Hung-Yu Kao
BMC Bioinformatics (2011-10-03) https://doi.org/dnmvds
DOI: 10.1186/1471-2105-12-s8-s5 · PMID: 22151999 · PMCID: PMC3269940

66. **Collaborative biocuration–text-mining development task for document prioritization for curation**
T. C. Wiegers, A. P. Davis, C. J. Mattingly

Database (2012-11-22) https://doi.org/gbb3zw
DOI: 10.1093/database/bas037 · PMID: 23180769 · PMCID: PMC3504477

67. **The Stanford CoreNLP Natural Language Processing Toolkit**
Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, David McClosky
Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations (2014) https://doi.org/gf3xhp
DOI: 10.3115/v1/p14-5010

68. **Snorkel**
Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, Christopher Ré
Proceedings of the VLDB Endowment (2017-11-01) https://doi.org/ch44
DOI: 10.14778/3157794.3157797 · PMID: 29770249 · PMCID: PMC5951191

69. **Snorkel MeTaL**
Alex Ratner, Braden Hancock, Jared Dunnmon, Roger Goldman, Christopher Ré
Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning - DEEM'18 (2018) https://doi.org/gf3xk7
DOI: 10.1145/3209889.3209898 · PMID: 30931438 · PMCID: PMC6436830

70. **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

71. **iRefIndex: A consolidated protein interaction database with provenance**
Sabry Razick, George Magklaras, Ian M Donaldson
BMC Bioinformatics (2008) https://doi.org/b99bjj
DOI: 10.1186/1471-2105-9-405 · PMID: 18823568 · PMCID: PMC2573892

72. **Uncovering disease-disease relationships through the incomplete interactome**
J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, A.-L. Barabasi
Science (2015-02-19) https://doi.org/f3mn6z
DOI: 10.1126/science.1257601 · PMID: 25700523 · PMCID: PMC4435741

73. **On Calibration of Modern Neural Networks**
Chuan Guo, Geoff Pleiss, Yu Sun, Kilian Q. Weinberger
arXiv (2017-06-14) https://arxiv.org/abs/1706.04599v2

74. **Accurate Uncertainties for Deep Learning Using Calibrated Regression**
Volodymyr Kuleshov, Nathan Fenner, Stefano Ermon
arXiv (2018-07-01) https://arxiv.org/abs/1807.00263v1

75. **Mining knowledge from MEDLINE articles and their indexed MeSH terms**
Daniel Himmelstein, Alex Pankov
ThinkLab (2015-05-10) https://doi.org/f3mqwp
DOI: 10.15363/thinklab.d67

76. **Integrating resources with disparate licensing into an open network**
Daniel Himmelstein, Lars Juhl Jensen, MacKenzie Smith, Katie Fortney, Caty Chung
ThinkLab (2015-08-28) https://doi.org/bfmk
DOI: 10.15363/thinklab.d107

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

78. **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

79. **A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification**
Ye Zhang, Byron Wallace
arXiv (2015-10-13) https://arxiv.org/abs/1510.03820v4

80. **Adam: A Method for Stochastic Optimization**
Diederik P. Kingma, Jimmy Ba
arXiv (2014-12-22) https://arxiv.org/abs/1412.6980v9

81. **Distributed Representations of Words and Phrases and their Compositionality**
Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean
arXiv (2013-10-16) https://arxiv.org/abs/1310.4546v1

82. **Enriching Word Vectors with Subword Information**
Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov
arXiv (2016-07-15) https://arxiv.org/abs/1607.04606v2

83. **Efficient Estimation of Word Representations in Vector Space**
Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean
arXiv (2013-01-16) https://arxiv.org/abs/1301.3781v3

Supplemental Methods

Adding Random Noise to Generative Model

We discovered in the course of this work that adding a single label function from a mismatched type would often improve the performance of the generative model (see Results). We designed an experiment to test whether adding a noisy label function also increased performance. This label function emitted a positive or negative label at varying frequencies, which were evenly spaced from zero to one. Zero was the same as distant supervision and one meant that all sentences were randomly labeled. We trained the generative model with these label functions added and reported results in terms of AUROC and AUPR.

Discriminative Model

The discriminative model is a neural network, which we train to predict labels from the generative model. The expectation is that the discriminative model can learn more complete features of the text than the label functions used in the generative model. We used a convolutional neural network with multiple filters as our discriminative model. This network uses multiple filters with fixed widths of 300 dimensions and a fixed height of 7 (Figure 6), because this height provided the best performance in terms of relationship classification [79]. We trained this model for 20 epochs using the adam optimizer [80] with pytorch's default parameter settings and a learning rate of 0.001. We added a L2 penalty on the network weights to prevent overfitting. Lastly, we added a dropout layer (p=0.25) between the fully connected layer and the softmax layer.

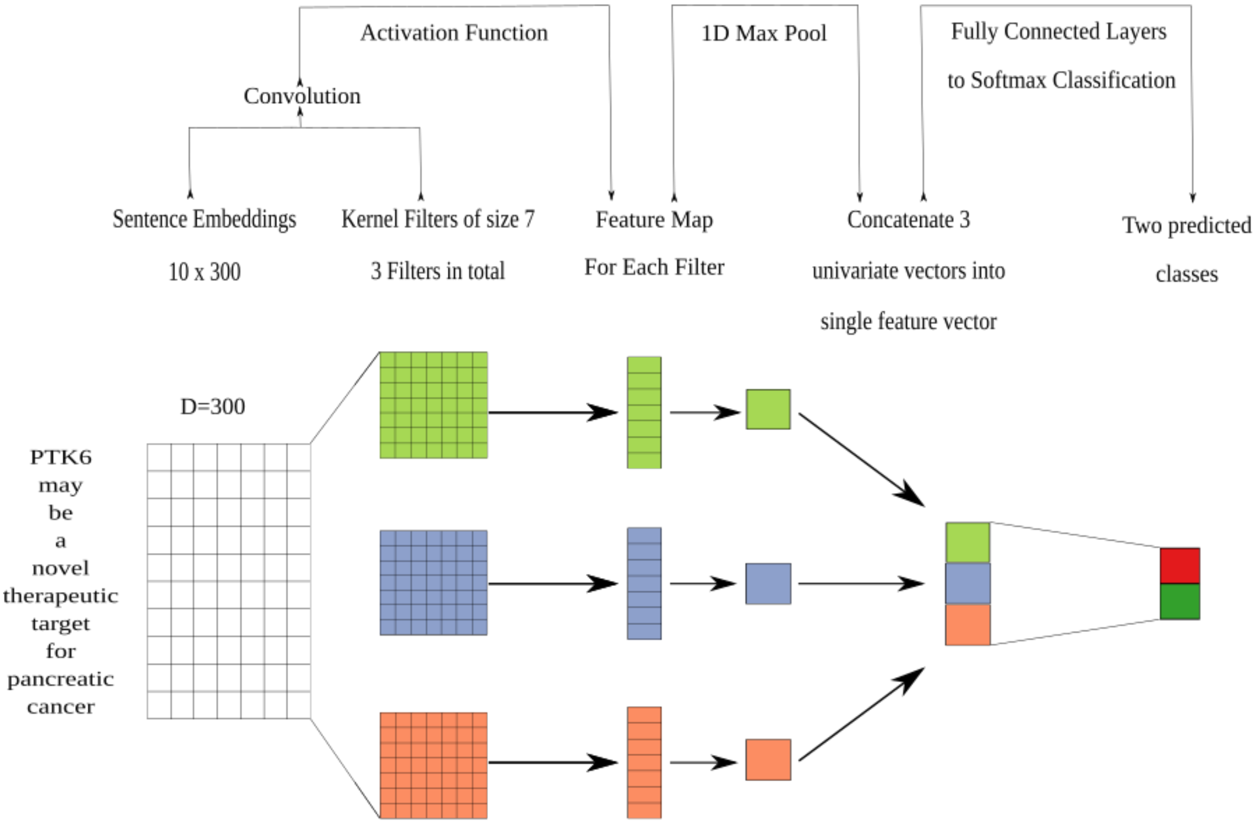

Figure 6: The architecture of the discriminative model was a convolutional neural network. We performed a convolution step using multiple filters. The filters generated a feature map that was sent into a maximum pooling layer that was designed to extract the largest feature in each map. The extracted features were concatenated into a singular vector that was passed into a fully connected network. The fully connected network had 300 neurons for the first layer, 100 neurons for the second layer and 50 neurons for the last layer. The last step from the fully connected network was to generate predictions using a softmax layer.

Word Embeddings

Word embeddings are representations that map individual words to real valued vectors of user-specified dimensions. These embeddings have been shown to capture the semantic and syntactic information between words [81]. We trained Facebook's fastText [82] using all candidate sentences for each individual relationship pair to generate word embeddings. fastText uses a skipgram model [83] that aims to predict the surrounding context for a candidate word and pairs the model with a novel scoring function that treats each word as a bag of character n-grams. We trained this model for 20 epochs using a window size of 2 and generated 300-dimensional word embeddings. We use the optimized word embeddings to train a discriminative model.

Calibration of the Discriminative Model

Often many tasks require a machine learning model to output reliable probability predictions. A model is well calibrated if the probabilities emitted from the model match the observed probabilities: a well-calibrated model that assigns a class label with 80% probability should have that class appear 80% of the time. Deep neural network models can often be poorly calibrated [73,74]. These models are usually over-confident in their predictions. As a result, we calibrated our convolutional neural network using temperature scaling. Temperature scaling uses a parameter T to scale each value of the logit vector (z) before being passed into the softmax (SM) function.

$$\sigma_{SM}(\frac{z_i}{T}) = \frac{\exp(\frac{z_i}{T})}{\sum_i \exp(\frac{z_i}{T})}$$

We found the optimal T by minimizing the negative log likelihood (NLL) of a held out validation set. The benefit of using this method is that the model becomes more reliable and the accuracy of the model doesn't change [73].

Supplemental Tables and Figures

Generative Model Using Randomly Sampled Label Functions

Individual Sources

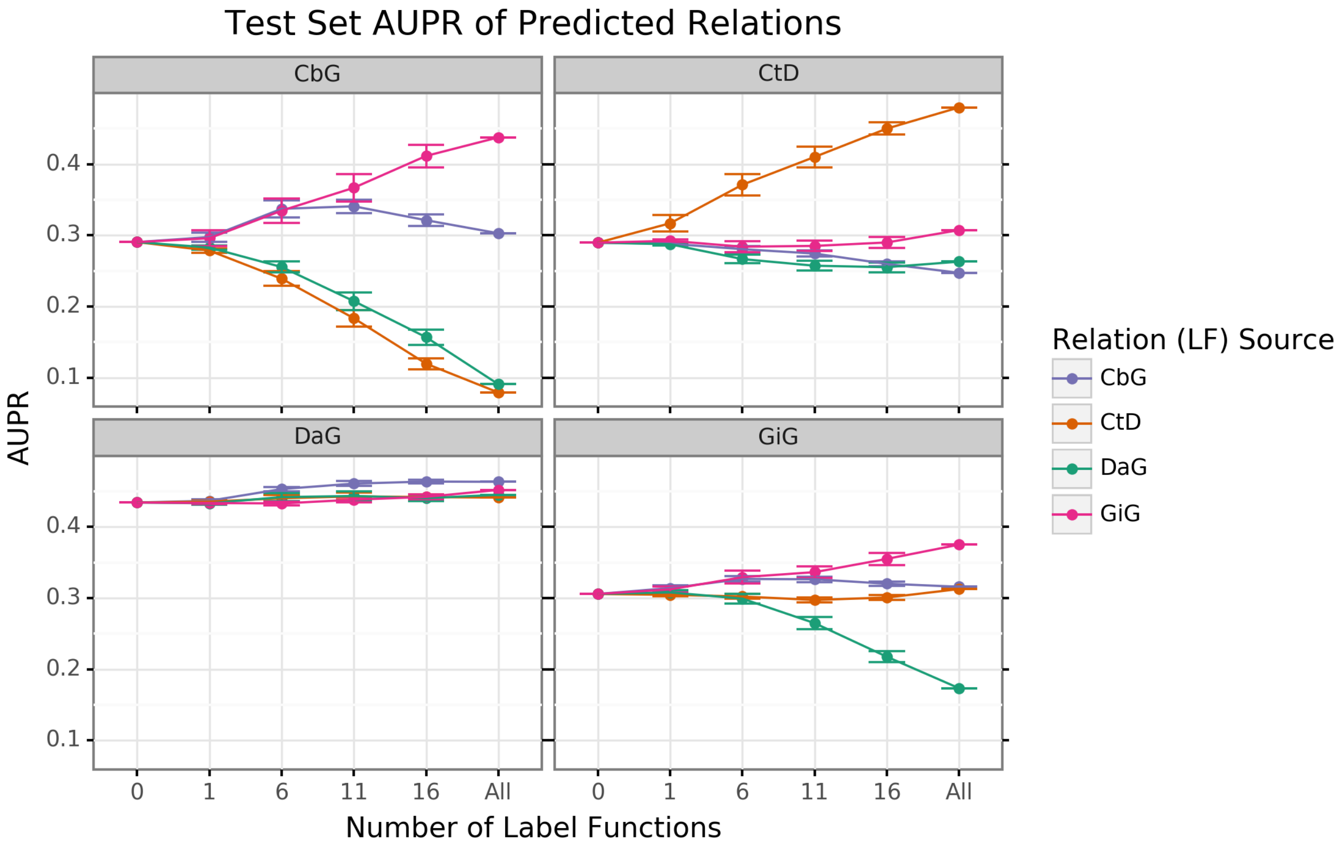

Figure 7: Edge-specific label functions improves performance over edge-mismatch label functions. Each line plot header depicts the edge type the generative model is trying to predict, while the colors represent the source of label functions. For example orange represents sampling label functions designed to predict the Compound treats Disease (CtD) edge type. The x axis shows the number of randomly sampled label functions being incorporated onto 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

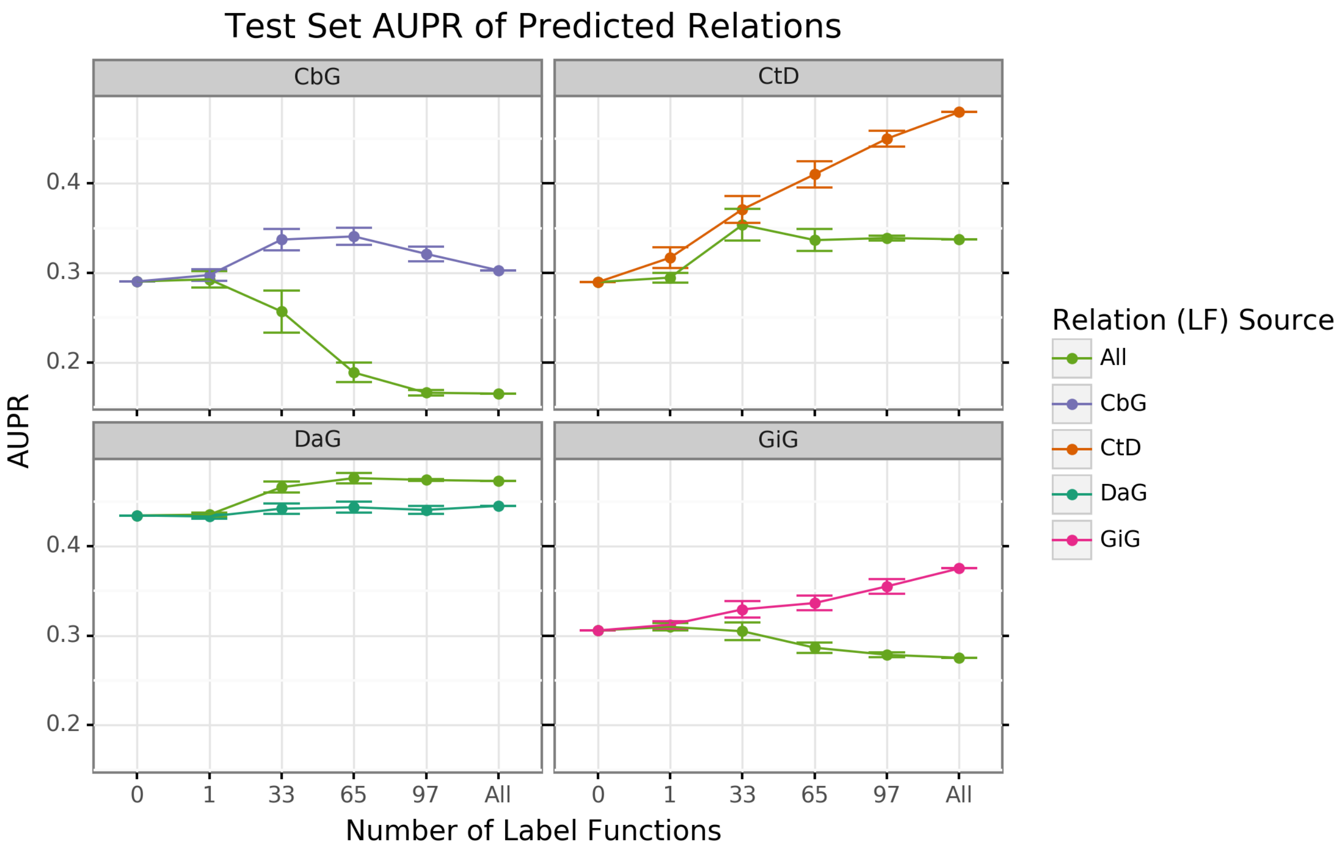

Figure 8: Using all label functions generally hinders generative model performance. Each line plot header depicts the edge type the generative model is trying to predict, while the colors represent the source of label functions. For example, orange represents sampling label functions designed to predict the Compound treats Disease (CtD) edge type. The x axis shows the number of randomly sampled label functions being incorporated onto 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

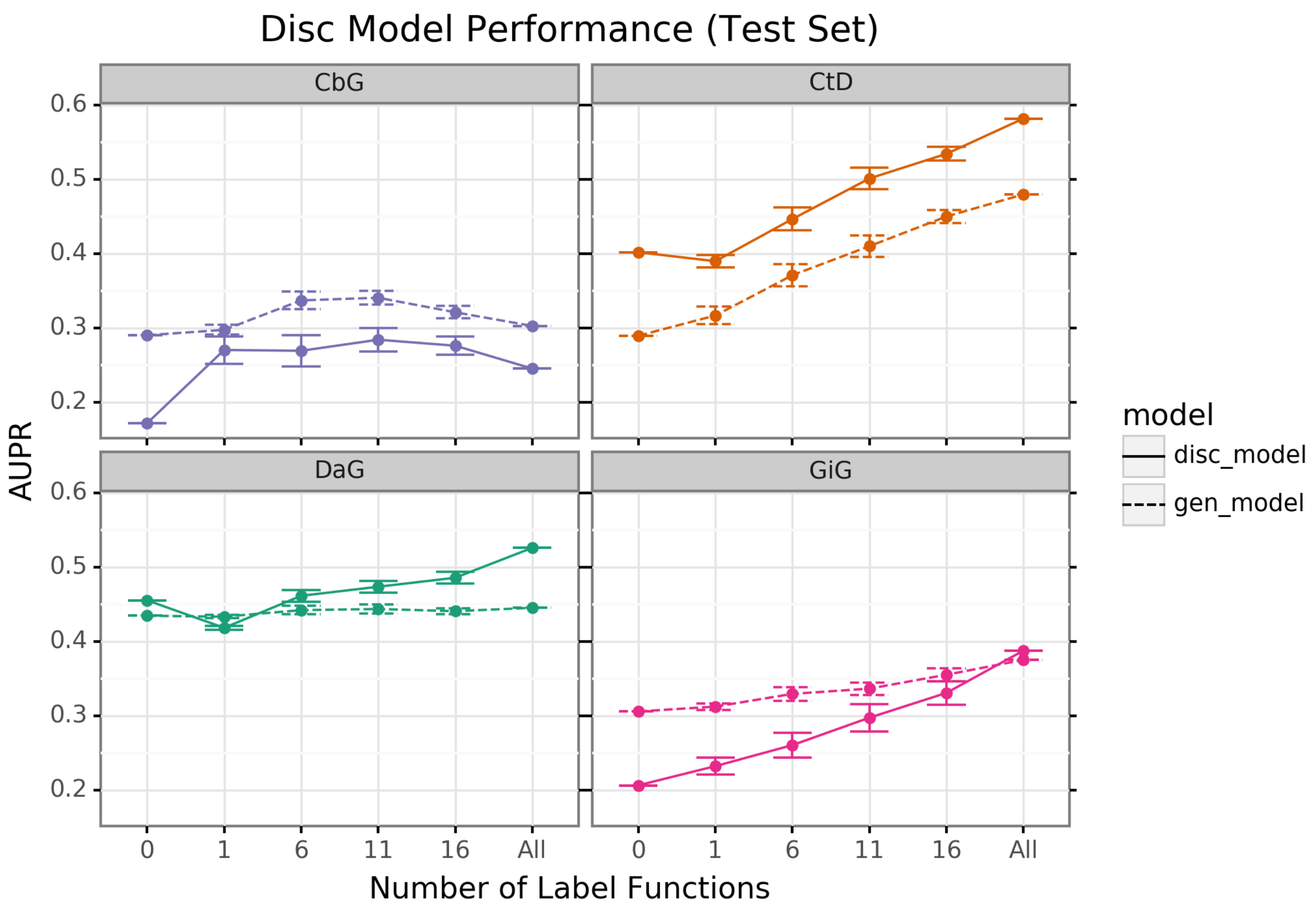

Figure 9: The discriminator model improves performance as the number of edge-specific label functions is added to the baseline model. The line plot headers represents the specific edge type the discriminator model is trying to predict. The x-axis shows the number of randomly sampled label functions incorporated 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 each 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. This makes the error bars appear flat.

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
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

Disease Name	Gene Symbol	Text	Before Calibration	After Calibration
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
endogenous depression	HTR1A	gepirone , a selective serotonin (5ht1a) partial agonist in the treatment of major depression .	0.199	0.378

Disease Name	Gene Symbol	Text	Before Calibration	After Calibration
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
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

Gene1 Symbol	Gene2 Symbol	Text	Before Calibration	After Calibration
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
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

Gene1 Symbol	Gene2 Symbol	Text	Before Calibration	After Calibration
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
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
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
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

Gene1 Symbol	Gene2 Symbol	Text	Before Calibration	After Calibration
PIK3R1	PTEN	both pten (pi3k antagonist) and pp2 (unspecific phosphatase) were down-regulated .	0.005	0.001

Baseline Comparison

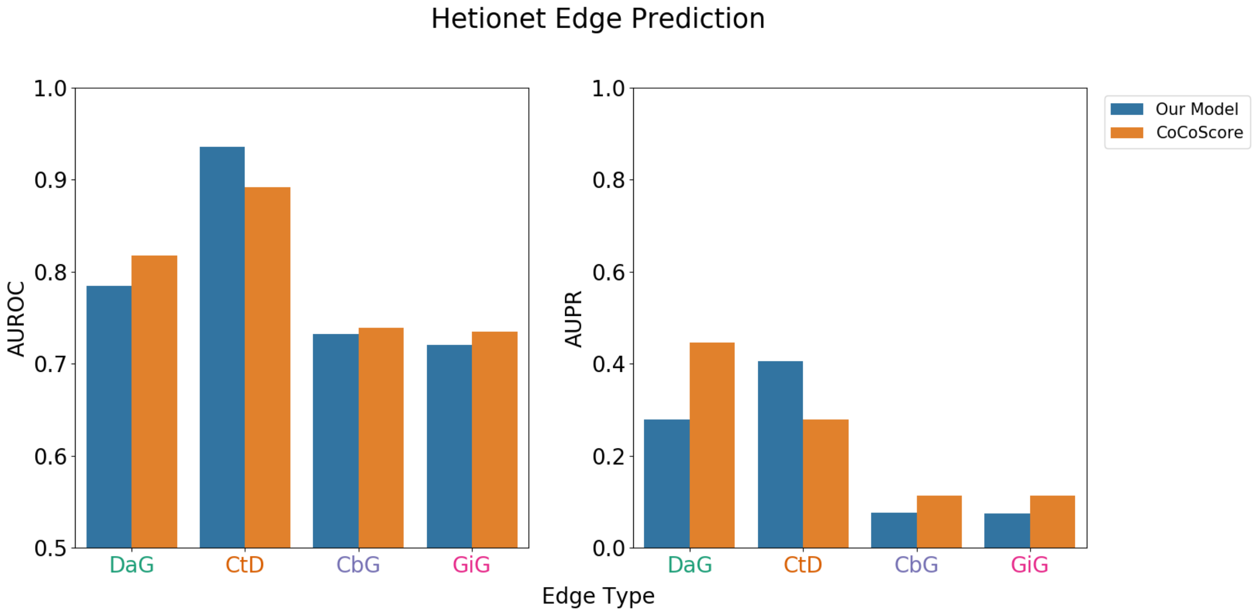

Figure 10: Comparion between our model and CoCoScore model [14]. We report both model's performance in terms of AUROC and AUPR. Our model achieves comparable performance against CoCoScore in terms of AUROC. As for AUPR, CoCoScore consistently outperforms our model except for CtD.

Once our discriminator model is calibrated, we grouped sentences based on mention pair (edges). We assigned each edge the maximum score over all grouped sentences and compared our model's ability to predict pairs in our test set to a previously published baseline model [14]. Performance is reported in terms of AUROC and AUPR (Figure 10). Across edge types our model shows comparable performance against the baseline in terms of AUROC. Regarding AUPR, our model shows hindered performance against the baseline. The exception for both cases is CtD where our model performs better than the baseline.

Reconstructing Hetionet

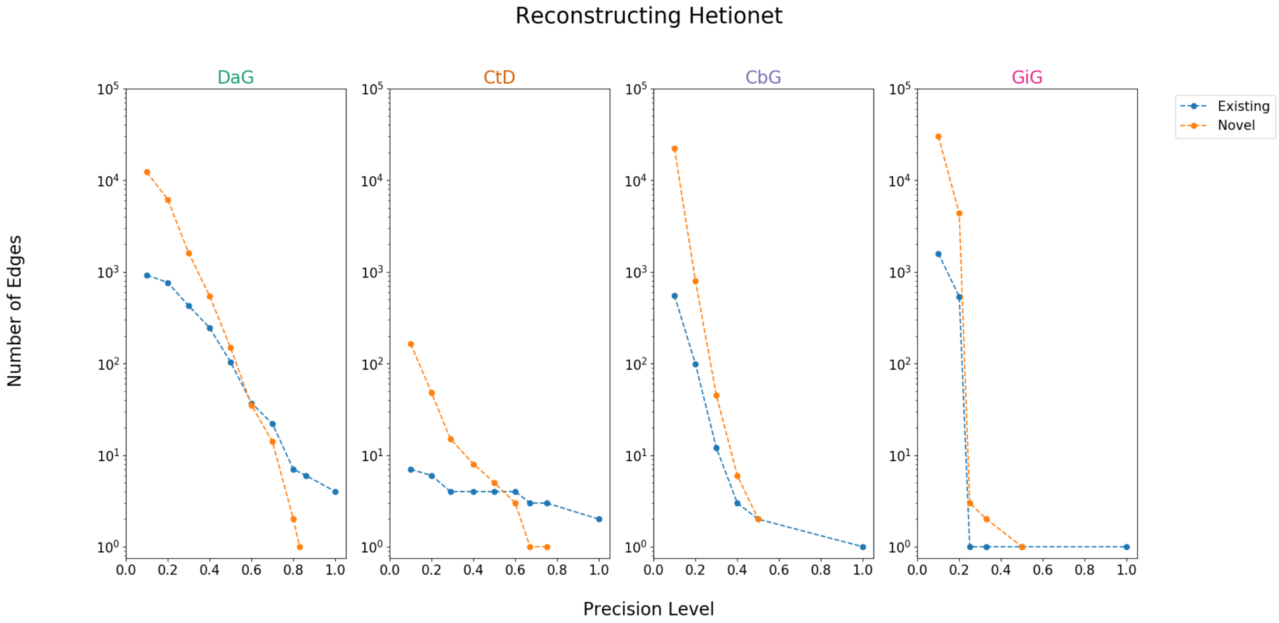

Figure 11: A scatter plot showing the number of edges (log scale) we can add or recall at specified precision levels. The blue depicts edges existing in hetionet and the orange depicts how many novel edges can be added.

We evaluated how many edges we can recall/add to Hetionet v1 (Supplemental Figure 11 and Table 11). In our evaluation we used edges assigned to our test set. Overall, we can recall a small amount of edges at high precision thresholds. A key example is CbG and GiG where we recalled only one exisiting edge at 100% precision. Despite the low recall, we are still able to add novel edges to DaG and CtD while retaining modest precision.

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	In Hetionet?	Number of Sentences	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 .
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	In Hetionet?	Number of Sentences	Text
DaG	stomach cancer	TP53	1	0.937	2679	Existing	conclusion : this meta-analysis suggests that p53 arg72pro polymorphism is associated with increased risk of gastric cancer in asians .
DaG	lung cancer	TP53	1	0.936	6813	Existing	conclusion : these results suggest that high expression of the p53 oncoprotein is a favorable prognostic factor in a subset of patients with nsclc .
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 .
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 .
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	In Hetionet?	Number of Sentences	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 .
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 .
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 .
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 .
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 .

Edge Type	Source Node	Target Node	Generative Model Prediction	Discriminative Model Prediction	In Hetionet?	Number of Sentences	Text
CtD	Cimetidine	psoriasis	0.999	0.987	12	Novel	cimetidine versus placebo in the treatment of psoriasis .
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 .
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 .
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 .
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	In Hetionet?	Number of Sentences	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 .
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 .
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 .
CbG	Pindolol	HTR1A	1	0.983	175	Existing	pindolol , a betablocker with weak partial 5-ht1a receptor agonist activity has been shown to produce a more rapid onset of antidepressant action of ssris .
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 .
CbG	Mifepristone	AR	1	0.98	78	Existing	ru486 bound to the androgen receptor .

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

---

1. Labeled sentences are available here.↩