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@fe91de6 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 unstructed 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 manual curation process requires a significant amount of effort and time: in 2007 researchers estimated that filling in the missing annotations would require approximately 8.4 years [8]. The rate of publications has continued to increase exponentially [9]. This has been recognized as a considerable challenge, which can lead to gaps in knowledge bases [8].

Relationship extraction has been studied as a solution towards handling this problem [7]. This process consists of creating a machine learning system to automatically scan and extract relationships from textual sources. Machine learning methods often leverage a large corpus of well-labeled training data, which still requires manual curation. Distant supervision is one technique to sidestep the requirement of well-annotated sentences: with distant supervision one makes the assumption that all sentences containing an entity pair found in a selected database provide evidence for a relationship [4]. Distant supervision provides many labeled examples; however it is accompanied by a decrease in the quality of the labels.

Ratner et al. [10] recently introduced “data programming” as a solution. Data programming combines distant supervision with the automated labeling of text using hand-written label functions. The distant supervision sources and label functions are integrated using a noise aware generative model that is used to produce training labels. Combining distant supervision with label functions can dramatically reduce the time required to acquire sufficient training data. However, constructing a knowledge base of heterogeneous relationships through this framework still requires tens of hand-written label functions for each relationship type. Writing useful label functions requires significant error analysis, which can be a time-consuming process.

In this paper, we aim to address the question: to what extent can label functions be re-used across different relationship types? We hypothesized that sentences describing one relationship type may share information in the form of keywords or sentence structure with sentences that indicate other relationship types. We designed a series of experiments to determine the extent to which label function re-use enhanced performance over distant supervision alone. We examined relationships that indicated 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). The re-use of label functions could dramatically reduce the number required to generate and update a heterogeneous knowledge graph.

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 [11]. 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 [12]. 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 [13]. 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. [12]. 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 [14,15,16] or in combination with other features to recover DaG relationships [17]. One recent effort relied on a bi-clustering approach to detect DaG-relevant sentences from Pubmed abstracts [18] 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 [19,20].

Hand-crafted high-quality datasets [21,22,23,24] 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 [21,25,26]. 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 [27], and another used distant supervision [28]. 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 [29]. A similar approach has also been applied to CtD extraction [30]. Manually-curated rules have also been applied to PubMed abstracts to address this task [31]. 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 [18]. 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 [32]. 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 [33]. 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% [33]. Numerous additional groups have now used the publicly available dataset, that resulted from this competition, to train supervised machine learning methods [27,34,35,36,36,37,38,39,40] and semi-supervised machine learning methods [41]. 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 [42] and bi-clustering of dependency trees [18] 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 [43], and there are numerous studies using co-occurrences [16,44,45,46]. However, more sophisticated strategies such as distant supervision appear to improve performance [13]. Similarly to the other edge types, the bi-clustering approach over dependency trees has also been applied to this edge type [18]. This manuscript provides a set of label functions for our work.

Most supervised classifiers used publicly available datasets for evaluation [47,48,49,50,51]. These datasets are used equally among studies, but can generate noticeable differences in terms of performance [52]. Support vector machines were a common approach to extract GiG edges [53,54]. However, with the growing popularity of deep learning numerous deep neural network architectures have been applied [41,55,56,57]. Distant supervision has also been used in this domain [58], and in fact this effort was one of the motivating rationales for our work.

Materials and Methods

Hetionet

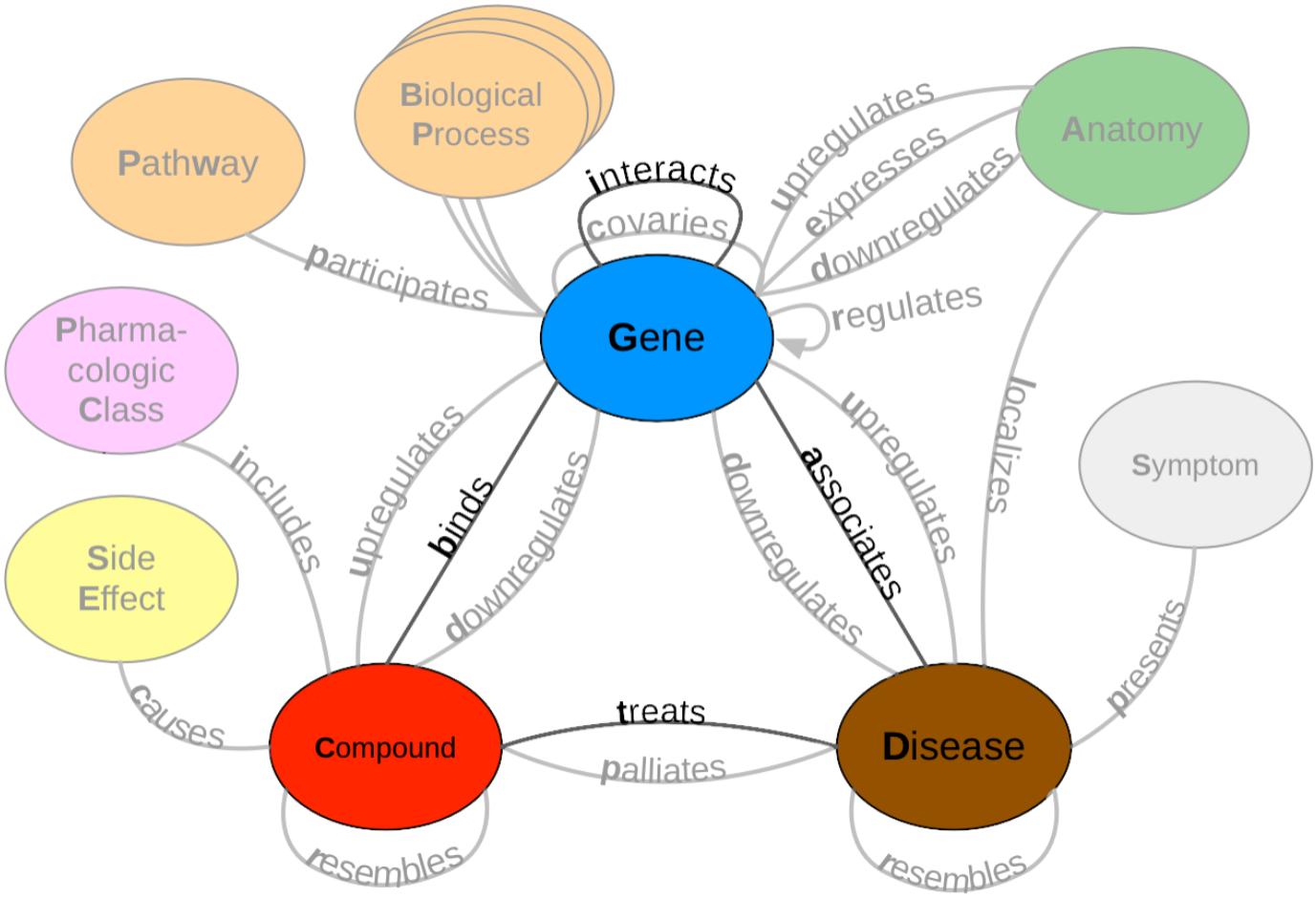

Figure 1: A metagraph (schema) of Hetionet where biomedical entities are represented as nodes and the relationships between them are represented as edges. We examined performance on the highlighted subgraph; however, the long-term vision is to capture edges for the entire graph.

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 [59] and DrugBank [60]. 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 [61] as input to our analysis. PubTator provides MEDLINE abstracts that have been annotated with well-established entity recognition tools including DNorm [62] for disease mentions, GeneTUKit [63] for gene mentions, Gnorm [64] for gene normalizations and a dictionary based search system for compound mentions [65]. 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 [66] 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 [10]. 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. Our label functions fall into three categories: databases, text patterns and domain heuristics. We provide examples for each category in our supplemental methods section.

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 [67,68].

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 [69], the iRefIndex database [70] or in the Incomplete Interactome database [71]. 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

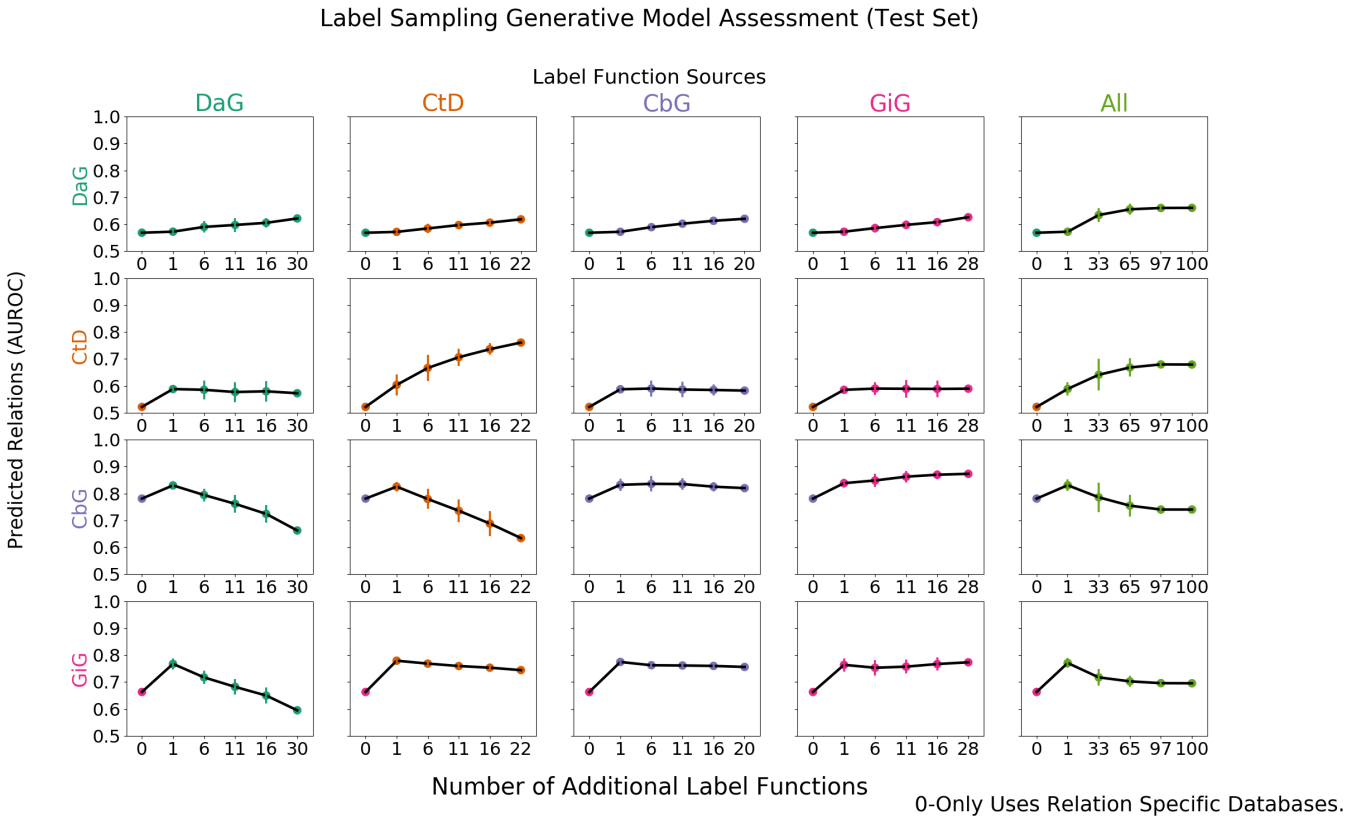

Figure 2: Grid of AUROC scores for each generative model trained on randomly sampled label functions. The rows depict the relationship each model is trying to predict and the columns are the edge type specific sources from which each label function is sampled. The right most column consists of pooling every relationship specific label function and proceeding as above.

We added randomly sampled label functions to a baseline for each edge type to evaluate the feasibility of label function re-use. Our baseline model consisted of a generative model trained with only edge-specific distant supervision label functions. We reported the results in AUROC and AUPR (Figure 2 and Supplemental Figure 6).
The on-diagonal plots of figure 2 and supplemental figure 6 show increasing performance when edge-specific label functions are added on top of the edge-specific baselines. The CtD edge type is a quintessential example of this trend. The baseline model starts off with an AUROC score of 52% and an AUPRC of 28%, which increase to 76% and 49% respectively as more CtD label functions are included. DaG edges have a similar trend: performance starting off with an AUROC of 56% and AUPR of 41% then increases to 62% and 45% respectively. Both the CbG and GiG edges have an increasing trend but plateau after a few label functions are added.

The off-diagonals in figure 2 and supplemental figure 6 show how performance varies when label functions from one edge type are added to a different edge type's baseline. In certain cases (apparent for DaG), performance increases regardless of the edge type used for label functions. In other cases (apparent with CtD), one label function appears to improve performance; however, adding more label functions does not improve performance (AUROC) or decreases it (AUPR). In certain cases, the source of the label functions appears to be important: the performance of CbG edges decrease when using label functions from the DaG and CtD categories.

Our initial hypothesis was based on the idea that certain edge types capture similar physical relationships and that these cases would be particularly amenable for label function transfer. For example, CbG and GiG both describe physical interactions. We observed that performance increased as assessed by both AUROC and AUPR when using label functions from the GiG edge type to predict CbG edges. A similar trend was observed when predicting the GiG edge; however, the performance differences were small for this edge type making the importance difficult to assess.

The last column shows increasing performance (AUROC and AUPR) for both DaG and CtD when sampling from all label functions. CbG and GiG also had increased performance when one random label function was sampled, but performance decreased drastically as more label functions were added. It is possible that a small number of irrelevant label functions are able to overwhelm the distant supervision label functions in these cases (see Figure 3 and Supplemental Figure 7).

Random Label Function Generative Model Analysis

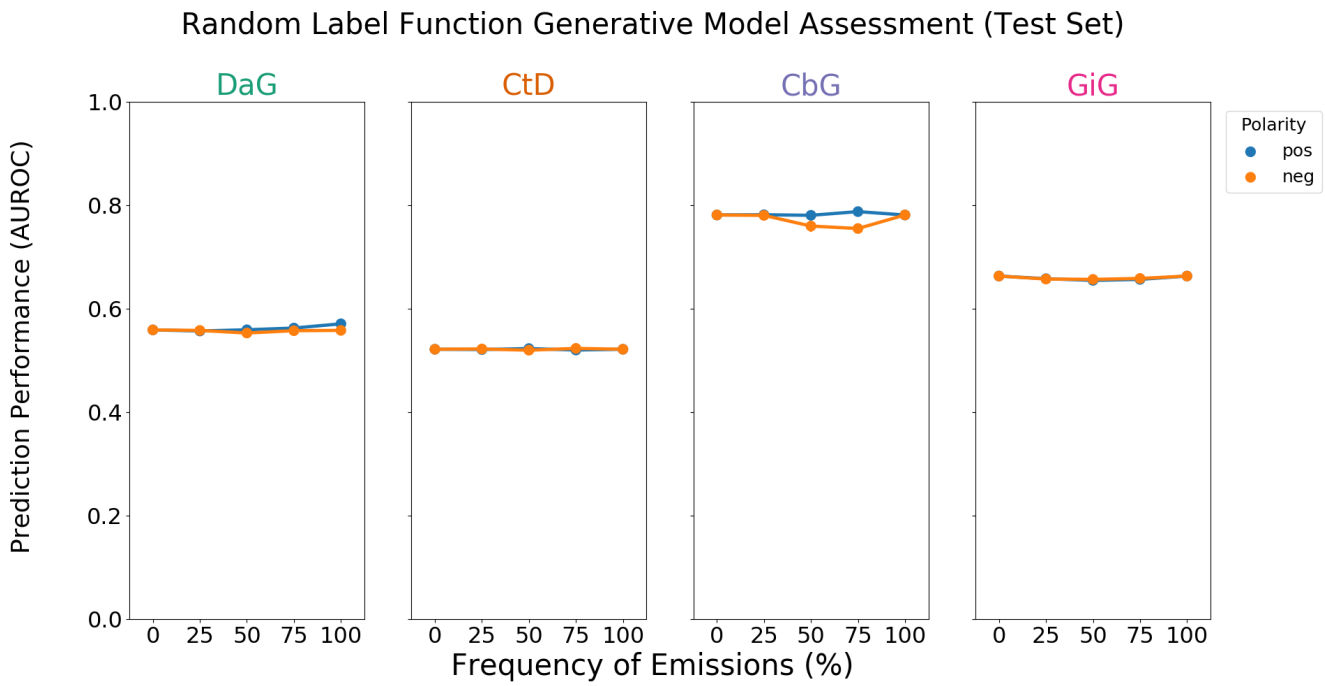

Figure 3: A grid of AUROC (A) scores for each edge type. Each plot consists of adding a single label function on top of the baseline model. This label function emits a positive (shown in blue) or negative (shown in orange) label at specified frequencies, and performance at zero is equivalent to not having 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 7). For the CbG edge type we did observe slightly increased performance via AUPR (Supplemental Figure 7). However, performance changes in general were smaller than those observed with mismatched label types.

Discriminative Model Calibration

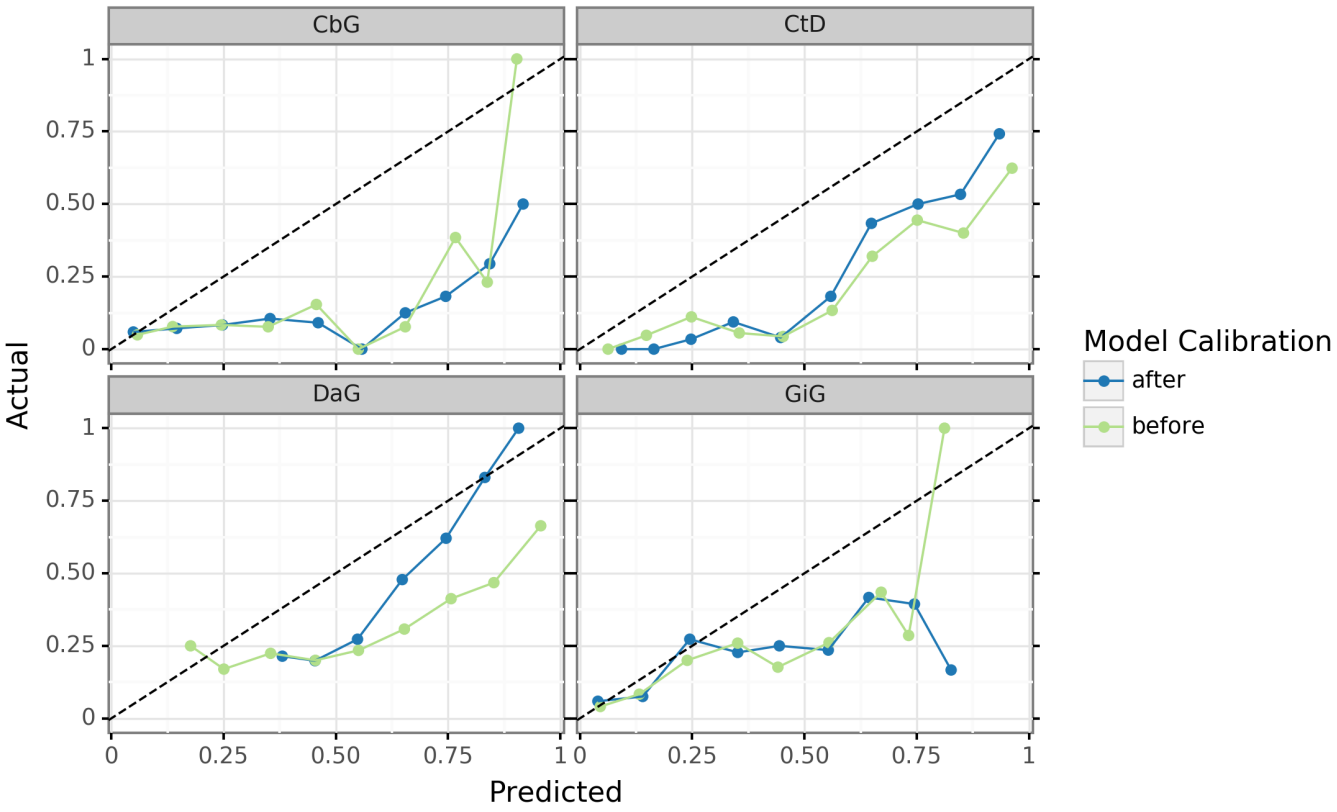

Figure 4: Deep learning models are overconfident in their predictions and need to be calibrated after training. These are calibration plots for the discrimintative model. The green line represents the predictions before calibration and the blue line shows predictions after calibration. Data points that 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 [72,73]. We attempted to use temperature scaling to fix the calibration of the best performing discriminative models (Figure 4). 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 2 and Supplemental Figure 6). 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 2 and Supplemental Figure 6). 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 7). 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 8 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 4) 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 manual curation becomes an infeasible approach. 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. A problem with this paradigm is that creating a useful label function takes a significant amount of time. We tested the feasibility of reusing label functions as a way to speed up the label function creation process. We conclude that label function re-use across edge types can increase performance when there are certain constraints on the number of functions re-used. More sophisticated methods of reuse may be able to capture many of the advantages and avoid many of the drawbacks. Adding more relevant label functions can increase overall performance. The discriminative model, under this paradigm, has a tendency to overfit to predictions of the generative model. We recommend implementing regularization techniques such as drop out and weight decay to combat this issue.

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 by consolidating sentence scores. Different from existing hetnets like Hetionet where text-derived edges generally cannot be exactly attributed to excerpts from literature [3,74], our approach would annotate each edge with 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 [75,76,77]. Accordingly, we plan to use this framework to create a robust multi-edge extractor via multitask learning [68] to construct continuously updating literature-derived hetnets.

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. **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

11. **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

12. **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

13. **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

14. **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

15. **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

16. **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

17. **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

18. **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

19. **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

20. **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

21. **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

22. **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

23. **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

24. **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

25. **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

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

27. **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

28. **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

29. **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

30. **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

31. **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

32. **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

33. **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

34. **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

35. **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

36. **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

37. **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

38. **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

39. **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

40. **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

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

42. **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

43. **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

44. **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

45. **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

46. **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

47. **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

48. **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

49. **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)

50. **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

51. **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

52. **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

53. **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

54. **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

55. **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

56. **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

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

58. **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

59. **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

60. **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

61. **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

62. **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

63. **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

64. **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

65. **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

66. **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

67. **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

68. **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

69. **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

70. **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

71. **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

72. **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

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

74. **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

75. **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

76. **Legal confusion threatens to slow data science**
Simon Oxenham

Nature (2016-08) https://doi.org/bndt
DOI: 10.1038/536016a · PMID: 27488781

77. **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

78. **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

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

80. **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

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

82. **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

Label Function Categories

We provide examples of label function categories below. Each example regards the following candidate sentence: "PTK6 may be a novel therapeutic target for pancreatic cancer."

Databases: These label functions incorporate existing databases to generate a signal, as seen in distant supervision [4]. These functions detect if a candidate sentence's co-mention pair is present in a given database. If the candidate pair is present, our label function emitted a positive label and abstained otherwise. If the candidate pair wasn't present in any existing database, a separate label function emitted a negative label. We used a separate label function to prevent a label imbalance problem that we encountered during development: emitting positive and negatives from the same label functions appeared to result in classifiers that predict almost exclusively negative predictions.

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

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

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 situation.

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

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

Domain Heuristics: These label functions use the other experiment results to generate a signal. For this category, we used dependency path cluster themes generated by Percha et al. [18]. If a candidate sentence's dependency path belongs to a previously generated cluster, then the label function will emit a positive label and abstain otherwise.

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

Roughly half of our label functions are based on text patterns, while the others are distributed across the databases and domain heuristics (Table 2).

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

Relationship	Databases (DB)	Text Patterns (TP)	Domain Heuristics (DH)
GiG	9	20	8

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 5), because this height provided the best performance in terms of relationship classification [78]. We trained this model for 20 epochs using the adam optimizer [79] 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.

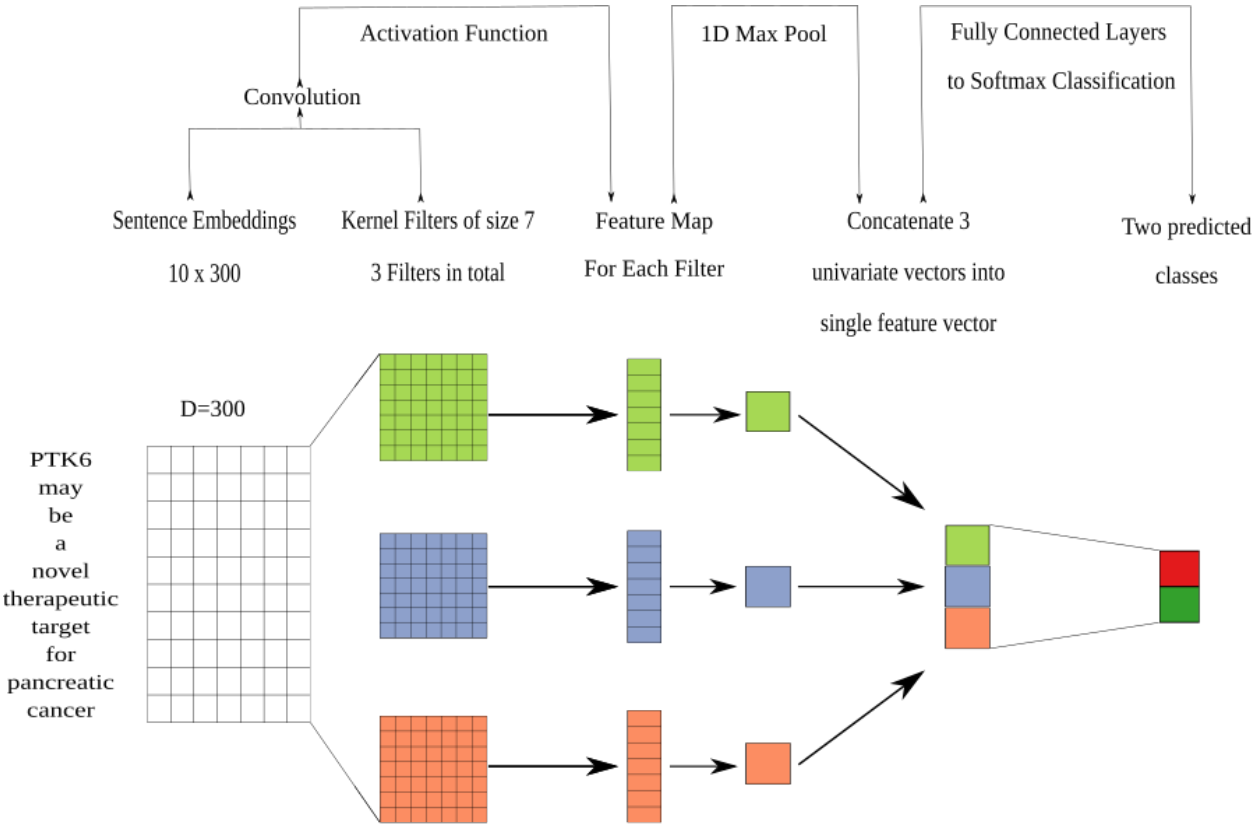

Figure 5: The architecture of the discriminative model was a convolutional neural network. We performed a convolution step using multiple filters. The filters generated a feature map that was sent into a maximum pooling layer that was designed to extract the largest feature in each map. The extracted 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 [80]. We trained Facebook's fastText [81] using all candidate sentences for each individual relationship pair to generate word embeddings. fastText uses a skipgram model [82] 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 [72,73]. 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 [72].

Supplemental Tables and Figures

Generative Model AUPR Performance

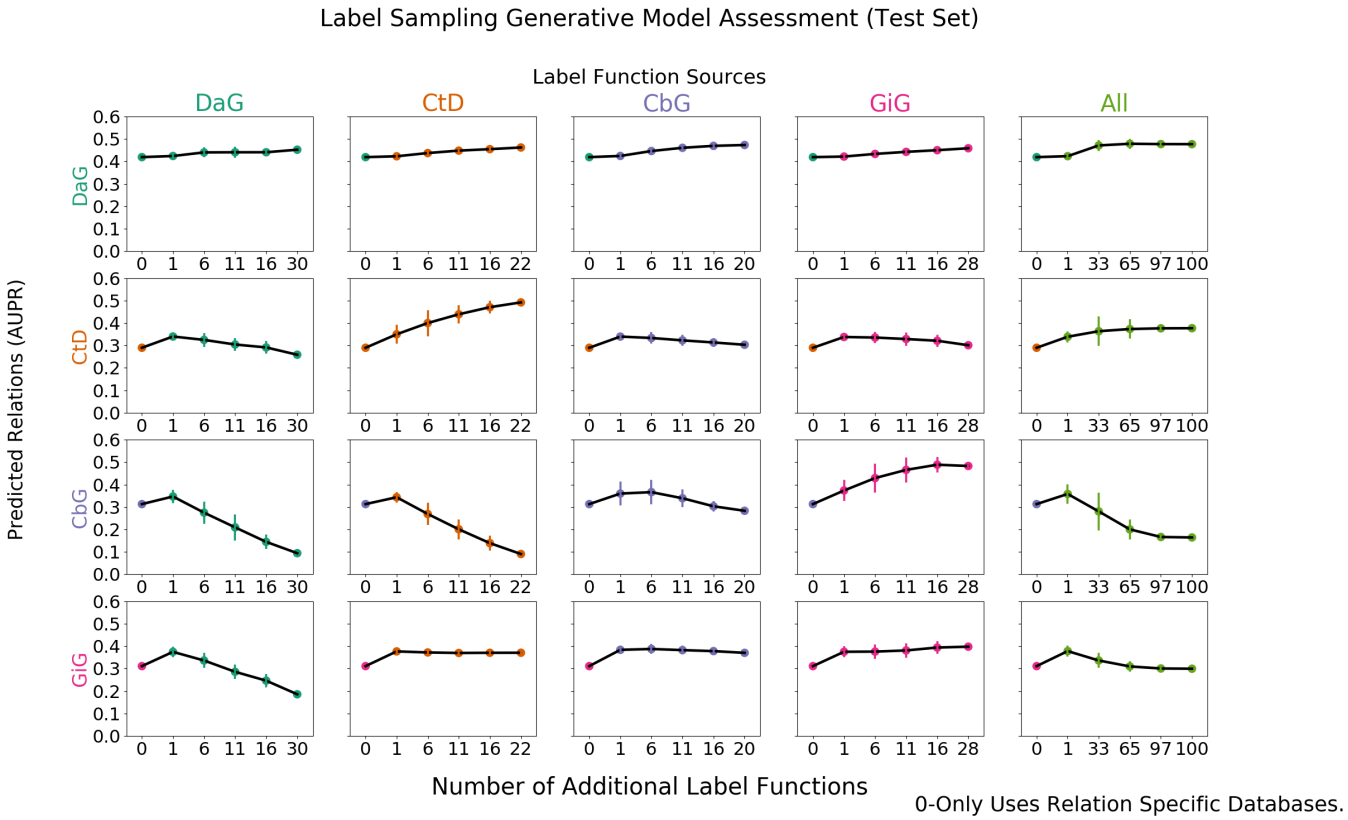

Figure 6: Grid of AUPR scores for each generative model trained on randomly sampled label functions. The rows depict the relationship each model is trying to predict and the columns are the edge type specific sources from which each label function is sampled. For example, the top-left most square depicts the generative model predicting DaG sentences, while randomly sampling label functions designed to predict the DaG relationship. The square towards the right depicts the generative model predicting DaG sentences, while randomly sampling label functions designed to predict the CtD relationship. This pattern continues filling out the rest of the grid. The right most column consists of pooling every relationship specific label function and proceeding as above.

Random Label Function Generative Model Analysis

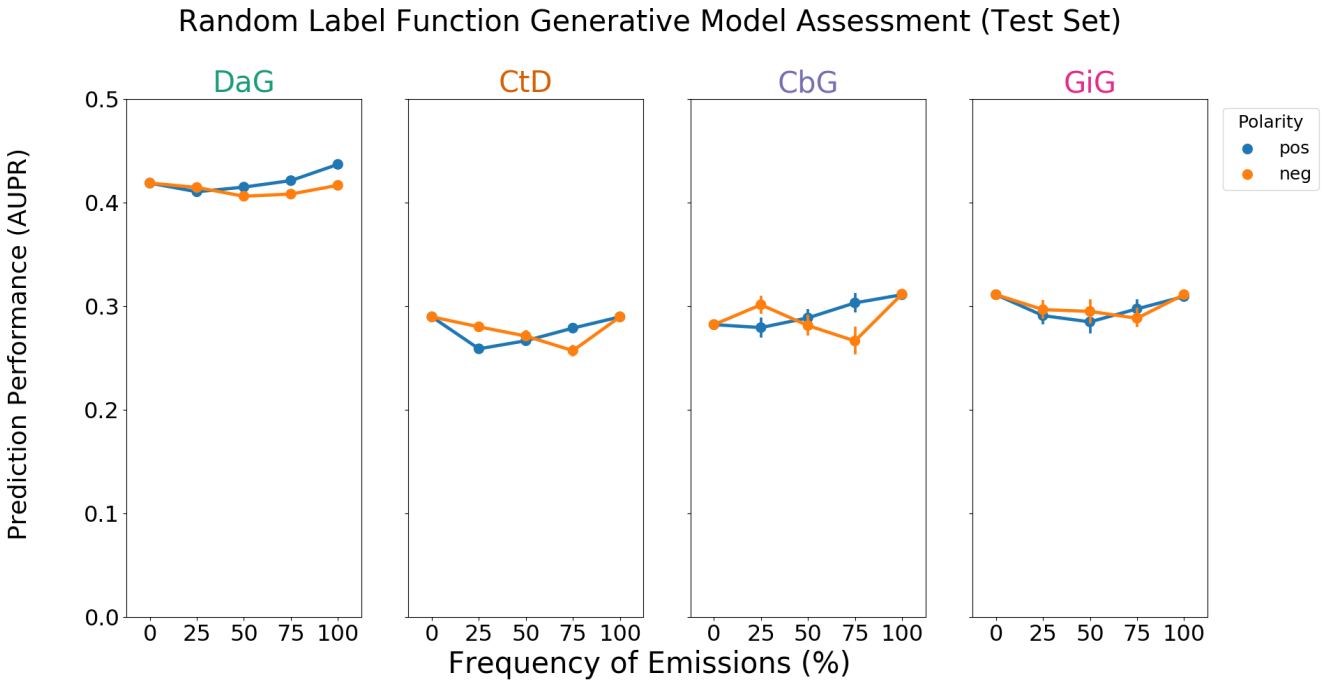

Figure 7: A grid of AUROC (A) scores for each edge type. Each plot consists of adding a single label function on top of the baseline model. This label function emits a positive (shown in blue) or negative (shown in orange) label at specified frequencies, and performance at zero is equivalent to not having a randomly emitting label function. The error bars represent 95% confidence intervals for AUROC or AUPR (y-axis) at each emission frequency.

Discriminative Model Performance

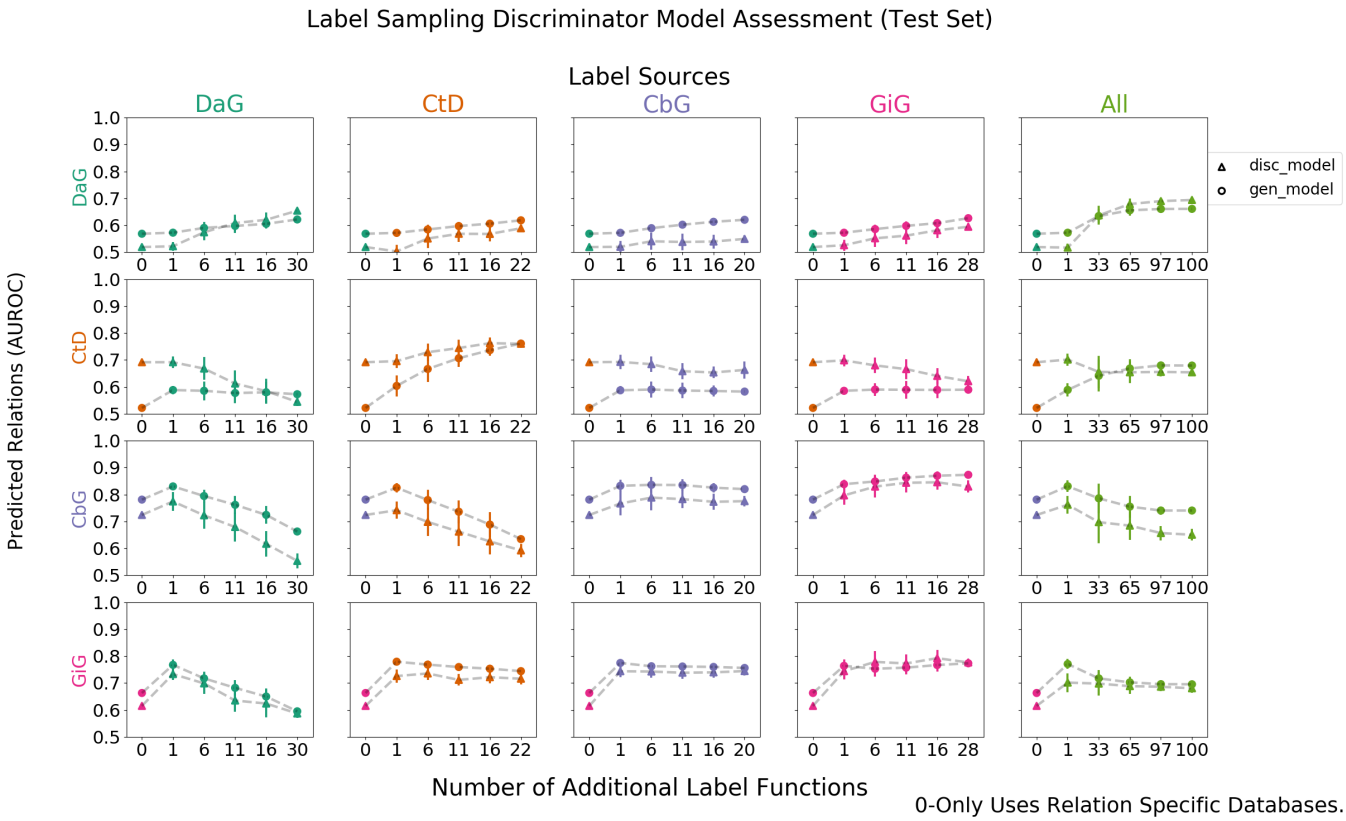

Figure 8: Grid of AUROC scores for each discriminative model trained using generated labels from the generative models. The rows depict the edge type each model is trying to predict and the columns are the edge type specific sources from which each label function was sampled. For example, the top-left most square depicts the discriminator model predicting DaG sentences, while randomly sampling label functions designed to predict the DaG relationship. The error bars over the points represents the standard deviation between sampled runs. The square towards the right depicts the discriminative model predicting DaG sentences, while randomly sampling label functions designed to predict the CtD relationship. This pattern continues filling out the rest of the grid. The right most column consists of pooling every relationship specific label function and proceeding as above.

In this framework we used a generative model trained over label functions to produce probabilistic training labels for each sentence. Then we trained a discriminative model, which has full access to a representation of the text of the sentence, to predict the generated labels. The discriminative model is a convolutional neural network trained over word embeddings (See Methods). We report the results of the discriminative model using AUROC and AUPR (Figures 8 and 9).

We found that the discriminative model under-performed the generative model in most cases. Only for the CtD edge does the discriminative model appear to provide performance above the generative model and that increased performance is only with a modest number of label functions. With the full set of label functions, performance of both models remain similar. The one or a few mismatched label functions (off-diagonal) improving generative model performance trend is retained despite the limited performance of the discriminative model.

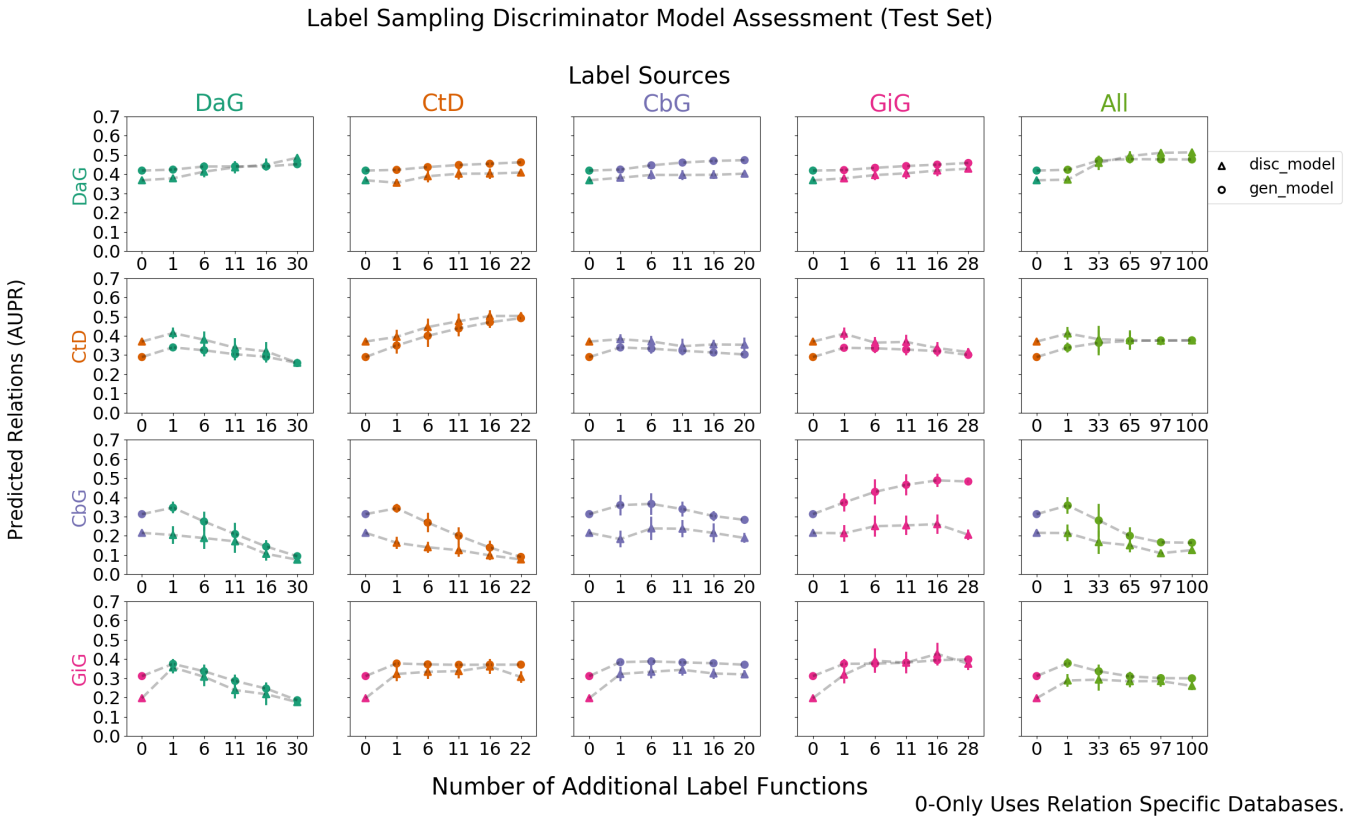

Figure 9: Grid of AUPR scores for each discriminative model trained using generated labels from the generative models. The rows depict the edge type each model is trying to predict and the columns are the edge type specific sources from which each label function was sampled. For example, the top-left most square depicts the discriminator model predicting DaG sentences, while randomly sampling label functions designed to predict the DaG relationship. The error bars over the points represents the standard deviation between sampled runs. The square towards the right depicts the discriminative model predicting DaG sentences, while randomly sampling label functions designed to predict the CtD relationship. This pattern continues filling out the rest of the grid. The right most column consists of pooling every relationship specific label function and proceeding as above.

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

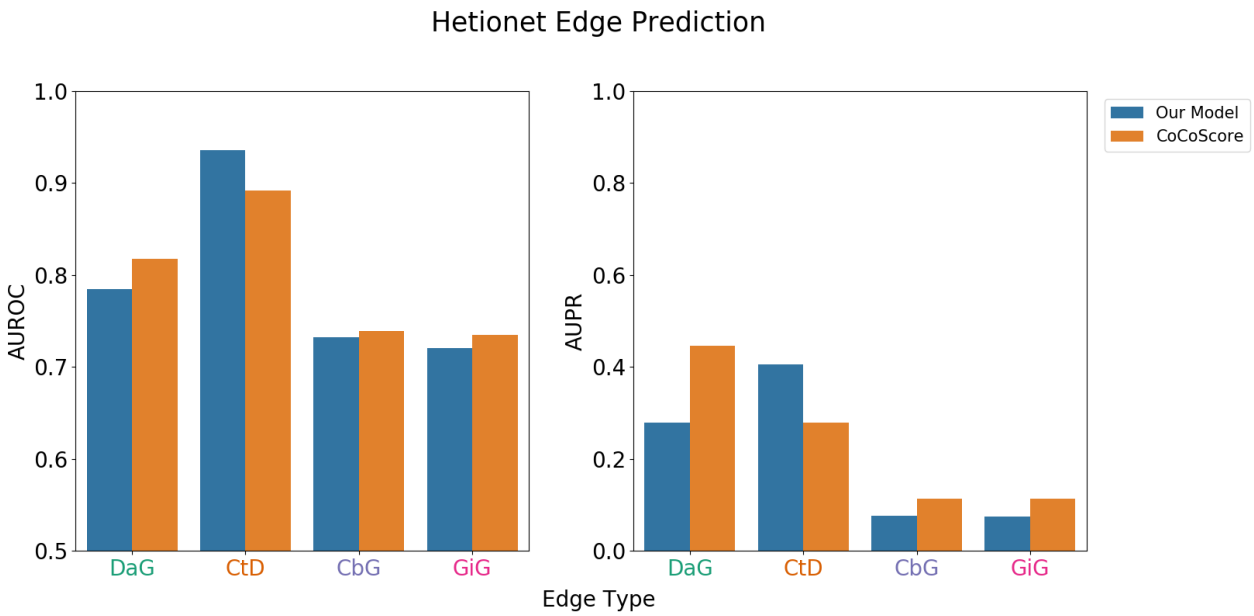

Figure 10: Comparion between our model and CoCoScore model [13]. We report both model's performance in terms of AUROC and AUPR. Our model achieves comparable performance against CoCoScore in terms of AUROC. As for AUPR, CoCoScore consistently outperforms our model except for CtD.

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 [13]. 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

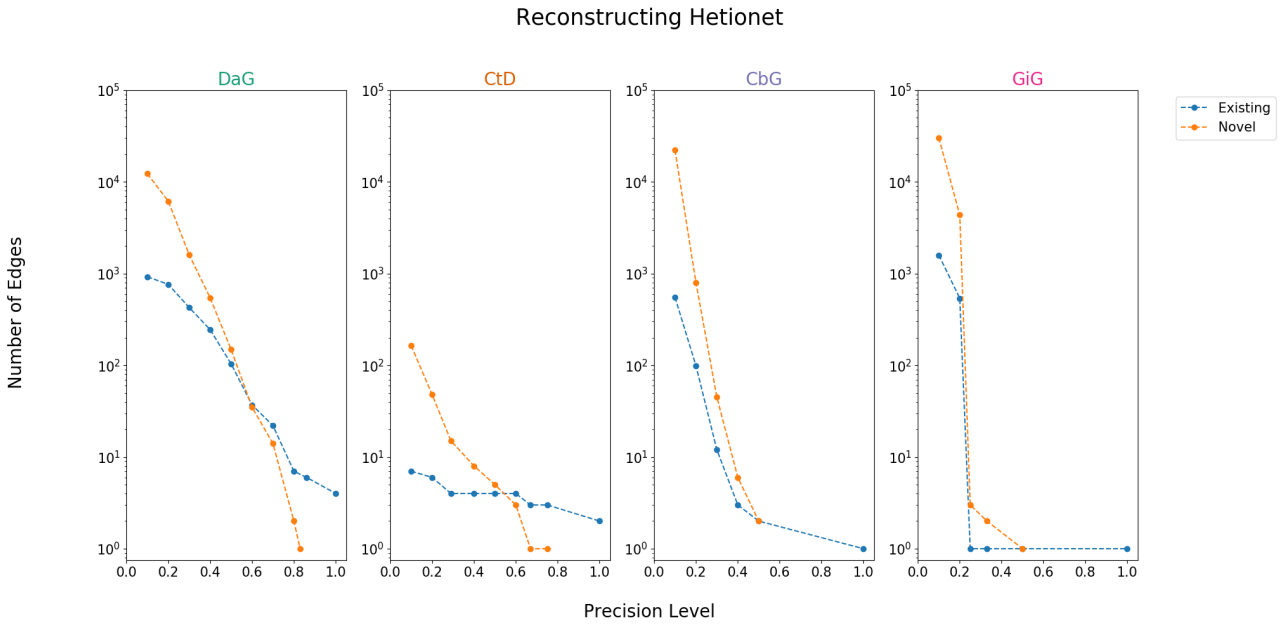

Figure 11: A scatter plot showing the number of edges (log scale) we can add or recall at specified precision levels. The blue depicts edges existing in hetionet and the orange depicts how many novel edges can be added.

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.

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	Gen Model Prediction	Disc Model Prediction	Number of Sentences	Text
DaG	lung cancer	VEGFA	1.000	0.912	3293	conclusion : the plasma vegf level is increased in nsclc patients with approximate1y one fourth to have cancer cells in the peripheral blood.
DaG	hematologic cancer	TP53	1.000	0.905	8660	mutations of the p53 gene were found in four cases of cml in blastic crisis (bc).
DaG	obesity	MC4R	1.000	0.901	1493	several mutations in the melanocortin 4 receptor gene are associated with obesity in chinese children and adolescents.
DaG	Alzheimer's disease	VLDLR	1.000	0.886	86	the 5-repeat allele in the very-low-density lipoprotein receptor gene polymorphism is not increased in sporadic alzheimer 's disease in japanese.

Edge Type	Source Node	Target Node	Gen Model Prediction	Disc Model Prediction	Number of Sentences	Text
DaG	lung cancer	XRCC1	1.000	0.885	662	results : xrcc1 gene polymorphism is associated with increased risk of lung cancer when the arg/arg genotype was used as the reference group.
DaG	prostate cancer	ESR1	1.000	0.883	500	conclusion : these results suggest that variants of the ggga polymorphism from the estrogen receptor alpha gene may be associated with an increased risk of developing prostate cancer.
DaG	breast cancer	REG1A	1.000	0.878	37	conclusion : high levels of reg1a expression within tumors are an independent predictor of poor prognosis in patients with breast cancer.
DaG	breast cancer	INSR	1.000	0.877	200	we have previously reported that insulin receptor expression is increased in human breast cancer specimens (v. papa et al. , j. clin.
DaG	rheumatoid arthritis	AR	1.000	0.877	53	conclusion : our results suggest no correlation between cag repeat polymorphism in the ar gene and response to treatment with lef in women with ra.
DaG	coronary artery disease	CTLA4	1.000	0.875	12	conclusion : the g/g genotype polymorphism of the ctla-4 gene is associated with increased risk of ami.
CtD	Zonisamide	epilepsy syndrome	1.000	0.943	1011	adjunctive zonisamide therapy in the long-term treatment of children with partial epilepsy : results of an open-label extension study of a phase iii , randomized , double-blind , placebo-controlled trial.

Edge Type	Source Node	Target Node	Gen Model Prediction	Disc Model Prediction	Number of Sentences	Text
CtD	Metformin	polycystic ovary syndrome	1.000	0.942	3217	in the present study , 23 pcos subjects [mean (+ / - se) body mass index 30.0 + / -1.1 kg/m2] were randomly assigned to double-blind treatment with metformin (500 mg tid) or placebo for 6 months , while maintaining their usual eating habits.
CtD	Piroxicam	rheumatoid arthritis	1.000	0.928	184	methods : a double-blind , randomized , crossover trial in 49 patients with active ra compared 6 weeks of treatment with tenidap (120 mg/day) versus 6 weeks of treatment with piroxicam (20 mg/day).
CtD	Irinotecan	stomach cancer	1.000	0.918	968	randomized phase ii trial of first-line treatment with tailored irinotecan and s-1 therapy versus s-1 monotherapy for advanced or recurrent gastric carcinoma (jfmc31-0301).
CtD	Treprostinil	hypertension	1.000	0.913	536	oral treprostinil for the treatment of pulmonary arterial hypertension in patients receiving background endothelin receptor antagonist and phosphodiesterase type 5 inhibitor therapy (the freedom-c2 study) : a randomized controlled trial.
CtD	Colchicine	gout	1.000	0.911	78	this is the first in vivo data to provide a biological rationale that supports the implementation of low dose , non-toxic , colchicine therapy for the treatment of gouty arthritis.
CtD	Propranolol	stomach cancer	1.000	0.898	45	74 cirrhotic patients with a history of variceal or gastric bleeding were randomly assigned to treatment with propranolol (40 to 360 mg/day) or placebo.

Edge Type	Source Node	Target Node	Gen Model Prediction	Disc Model Prediction	Number of Sentences	Text
CtD	Reboxetine	endogenous depression	1.000	0.894	439	data were obtained from four short-term (4-8-week) , randomized , placebo-controlled trials of reboxetine for the treatment of mdd.
CtD	Diclofenac	ankylosing spondylitis	1.000	0.892	61	comparison of two different dosages of celecoxib with diclofenac for the treatment of active ankylosing spondylitis : results of a 12-week randomised , double-blind , controlled study.
CtD	Tapentadol	osteoarthritis	1.000	0.880	29	driving ability in patients with severe chronic low back or osteoarthritis knee pain on stable treatment with tapentadol prolonged release : a multicenter , open-label , phase 3b trial.
CbG	Dexamethasone	NR3C1	1.000	0.850	1119	submicromolar free calcium modulates dexamethasone binding to the glucocorticoid receptor.
CbG	Vitamin A	RBP4	1.000	0.807	5512	the authors give serum retinol binding protein (rbp) normal values , established by immunonephelometry , for two healthy populations in their hospital laboratory.
CbG	D-Proline	IGFBP4	1.000	0.790	1	the insulin-like growth factor-i-stimulated l-proline uptake was inhibited by one of its binding protein , insulin-like growth factor binding protein-4 , in a concentration-dependent manner.
CbG	Sucrose	AR	0.996	0.789	37	the amount (maximal binding capacity of 24 to 30 fmol/mg protein) and hormone binding affinity (half-maximal saturation of 0.2 nm) of the androgen receptor in cultured skin fibroblasts was normal , but the receptor was qualitatively abnormal as evidenced by instability on sucrose density gradient centrifugation.

Edge Type	Source Node	Target Node	Gen Model Prediction	Disc Model Prediction	Number of Sentences	Text
CbG	D-Lysine	PLG	1.000	0.787	403	in both elisa and rocket immunoelectrophoresis systems , complex formation was inhibited by 10 mm epsilon-amino-n-caproic acid , implying that there is a role for the lysine binding sites of plg in mediating the interaction.
CbG	Adenosine	INSR	1.000	0.785	129	these findings demonstrate basal state binding of atp to the ckd leading to cis-autophosphorylation and novel basal state regulatory interactions among the subdomains of the insulin receptor kinase.
CbG	Adenosine	PLK1	1.000	0.783	104	most kinase inhibitors interact with the atp binding site on plk1 , which is highly conserved.
CbG	Calcium Chloride	ITPR3	0.995	0.777	1954	control of ca2 + influx in human neutrophils by inositol 1,4,5-trisphosphate (ip3) binding : differential effects of micro-injected ip3 receptor antagonists.
CbG	D-Arginine	C5AR1	1.000	0.775	808	thus , selected out of a multiplicity of possibilities by the natural binding partner , arg37 as well as arg40 appear to be anchor residues in binding to the c5a receptor.
CbG	Ticagrelor	P2RY12	1.000	0.773	322	purpose : ticagrelor is a reversibly binding p2y12 receptor antagonist used clinically for the prevention of atherothrombotic events in patients with acute coronary syndromes (acs).
GiG	ABL1	ABL1	0.999	0.600	9572	the acquired resistance in patients who failed to respond to imatinib seemed to be induced by several point mutations in the bcr-abl gene , which were likely to affect the binding of imatinib with bcr-abl.

Edge Type	Source Node	Target Node	Gen Model Prediction	Disc Model Prediction	Number of Sentences	Text
GiG	TP63	TP53	1.000	0.595	2557	tp63 , a member of the p53 gene family gene , encodes the np63 protein and is one of the most frequently amplified genes in squamous cell carcinomas (scc) of the head and neck (hnscc) and lungs (lusc).
GiG	FERMT1	FERMT1	0.004	0.590	194	ks is caused by mutations in the fermt1 gene encoding kindlin-1.
GiG	GRN	GRN	1.000	0.590	3842	background : mutations in the progranulin gene (pgrn) have recently been identified as a cause of frontotemporal lobar degeneration with ubiquitin-positive inclusions (ftld-u) in some families.
GiG	FASN	EP300	0.999	0.589	6	here , we demonstrated that p300 binds to and increases histone h3 lysine 27 acetylation (h3k27ac) in the fasn gene promoter.
GiG	SETBP1	SETBP1	1.000	0.588	354	the critical deleted region contains setbp1 gene (set binding protein 1).
GiG	BCL2	BAK1	0.118	0.587	1220	different expression patterns of bcl-2 family genes in breast cancer by estrogen receptor status with special reference to pro-apoptotic bak gene.
GiG	SP1	INSR	0.948	0.587	23	thus , the efficient expression of the human insulin receptor gene possibly requires the binding of transcriptional factor sp1 to four g-c boxes located -593 to -618 base pairs upstream of the atg translation initiation codon.
GiG	ABCD1	ABCD1	1.000	0.586	410	x-linked adrenoleukodystrophy (x-ald) is caused by mutations in the abcd1 gene encoding the peroxisomal abc transporter adrenoleukodystrophy protein (aldp).

Edge Type	Source Node	Target Node	Gen Model Prediction	Disc Model Prediction	Number of Sentences	Text
GiG	CYP1A1	AHR	0.996	0.586	1940	the liganded ah receptor activates transcription by binding to a specific dna-recognition motif within a dioxin-responsive enhancer upstream of the cyp1a1 gene.

---

1. Labeled sentences are available here.↩