Factors Affecting Network-Based Gene Prediction Across Diverse Diseases

Alexander King,[1*] Ibrahim Youssef,[1,2] and Anna Ritz[1]

Abstract—**There are many current efforts to integrate biological interaction data with disease information in order to predict new genes associated with complex diseases. Network-based learning methods such as logistic regression can utilize this information to identify disease genes, and are typically applied to protein-protein interaction networks. However, little is reported about what factors influence the performance of these network-based methods. Here, we explore features that affect network-based disease gene prediction performance. We devise two cross-validation schemes to evaluate the impact of various parameters, settings and disease qualities across a wide range of diseases. We demonstrate that including gene regulatory interactions and including low-confidence disease genes improves disease gene prediction performance. Further, network connectivity among high-confidence disease genes is a strong indicator of prediction performance. We demonstrate that network and input features can have a dramatic effect on prediction performance, and these should be carefully considered when designing network-based algorithms to find new disease genes.**

I. Introduction

Predicting the genetic variants that underlie disease has been a long-standing area of research, and high-throughput Genome Wide Association studies (GWAS) provide tens of thousands to millions of such candidates for complex diseases [1]. However, many of these mutations are not causal to the disease of interest, and further, some genes affect the disease at the transcriptional or translational level [2]. Over the past decade, researchers have developed approaches that leverage the hypothesis mutated genes typically do not act alone in complex diseases, but rather work together within a network of interactions to alter a particular phenotype [3]–[8]. Network-based approaches have become incredibly useful extracting causal variants, particularly for complex diseases such as cancer [6], [9], [10] and neurological diseases [11]–[13].

Alongside the development of these network-based methods, there have been major efforts to catalog lists of known genetic variants for multiple diseases. Resources such as the Online Mendelian Inheritance in Man [14], [15], NCBI's dbGAP [16], [17], and the Comparative Toxicogenomics Database [18], [19] have become critical benchmarks for different types of diseases. More recently, DisGeNet [20], [21] aims to encompass mendelian, environmental, and complex diseases into a single resource with a consistent methodology for disease gene identification. DisGeNet has already been used for disease gene prediction with networks using random walks [22], [23]. It has also been used to benchmark dozens of different networks for disease gene prediction [7], [8] and as a means to evaluate functional similarity in human gene networks [24], illustrating DisGeNet's use as a comprehensive resource for such analyses.

Despite DisGeNet's promise for assessment in developing new disease gene association methods, the current approaches that use DisGeNet evaluate the prediction task on protein-protein interaction networks. Many methods include gene regulatory information in functional interactomes through co-expression networks or shared transcription-factor information in bayesian networks [25]–[27]. However, physical gene regulatory interactions have been underutilized in interaction networks, which can be important mechanisms for many known diseases. Further, established approaches have typically adopted a belief propagation or random-walk methodological approach [7], [22], [23], [27] (though there are notable exceptions [8], [12]). In this paper, we evaluate the effect of regulatory interactions on the disease gene association problem in multiple experimental settings using DisGeNet as an assessment framework. We make three contributions:

1) Network of protein and regulatory interactions: We construct an interaction network that consists of protein-protein interactions as well as gene regulatory interactions such as transcription factor targets, transcript-transcript interactions, and RNA-binding protein interactions. This network integrates protein-protein interactions from the STRING databases with regulatory interactions from four databases.

2) Pan-Disease assessment: We use logistic regression to predict novel disease genes given a set of high-confidence disease genes across 49 diseases in DisGeNet. We control for the number of disease genes in DisGeNet by capping the number of disease genes to 100. We show that, despite having the same number of positive examples, different disease types perform better than others. This performance is independent of using the DisGeNet score, and it is correlated with the connectivity of the positives within the network. We also show that including gene regulatory interactions generally benefits disease gene prediction performance.

3) Disease class assessment: We then consider classifying neoplasms using all available DisGeNet genes as positive examples. We find that using all DisGeNet genes achieves better performance across the board compared to using the top 100 positives. Strikingly, this performance is also independent of using the DisGeNet score, meaning that the high-confidence and low-confidence positives are weighted equally.

[1] Department of Biology, Reed College, Portland, OR, USA.
[2] Department of Biomedical Engineering, Cairo University, Giza, Egypt.
[*] Current Affiliation: Department of Neuroscience, University of California Riverside, Riverside, CA, USA.
Author Emails: aking035@ucr.edu; youssefi@reed.edu; aritz@reed.edu

978-1-7281-1867-3/19/$31.00 ©2019 IEEE

These highly generalizable findings can be used to optimize disease gene prediction in a variety of diseases.

II. Methods

A. Network Construction

We collated existing databases of curated data, high-throughput experiments, and computationally predicted interactions to create two weighted networks: a protein-protein interaction (PPI) network of physical interactions, and a Molecular Atlas of Phenomenon (MAP) that includes regulatory interactions and physical interactions (Fig. 1A). We first collected protein-protein interactions from the STRING database [25]. To ensure that only biochemical/genetic interaction information was used in gene prediction, we retained interactions from STRING's "experimental" channel (and their experimental confidence scores). This collection of STRING interactions formed a directed, weighted protein-protein interactome (PPI) consisting of 17,844 nodes and 4,441,009 edges.

Additional gene regulatory interactions were collected from various sources and combined with the PPI for the MAP (Table I). RNA associated interactions were collected from the RNA Association Interaction Database (RAID) [28]. Confidence scores for each interaction were taken from the RAID entries, which are categorized as "strong experimental evidence," "weak experimental evidence," and "computationally-predicted evidence.". Gene regulatory interactions were also collected from three other databases: the Transcriptional Regulatory Relationships Unraveled by Sentence-based Textmining (TRRUST) database [29], RegNetwork [30] and the Open-access Repository of Transcriptional Interactions (ORTI) [31]. For consistency with the RAID scoring methodology, each interaction from these data sources was labeled with a "strong experimental evidence" score under the RAID scoring method. We took the maximum score for any interaction that appeared in multiple regulatory databases. Regulatory interactions were added to the PPI by adding the regulatory score to an existing edge if one existed, and creating a new edge if one did not exist. In total, the MAP consists of 17,844 nodes and 5,178,891 edges. All data was gathered in November 2018.

B. Annotated Disease Gene Dataset

Disease-gene annotations were gathered from DisGeNet [21]. Diseases were classified according to Unified Medical Language System (UMLS) semantic types, which is described in more detail in the DisGeNet publication [21]. We will call these types *disease classes*. Only diseases with a disease class of "Mental or Behavioral Dysfunction", "Neoplastic Process", or "Diseases or Syndrome" were used. Further, diseases with at least 100 genes with a confidence score over 0.2 were considered. In total, we considered 7 "Mental or Behavioral Dysfunction" diseases, 17 "Neoplastic Process" diseases, and 25 diseases labeled as"Disease or Syndrome" (totaling 49 diseases).

Interaction Set	Source	Number of Edges
PPI		**4,441,009**
	STRING [25]	4,441,009
Gene Regulation		**737,882**
	RAID [28]	685,762
	ORTI [31]	36,510
	RegNetwork [30]	9,701
	TRRUST [29]	5,909
MAP		**5,178,891**

TABLE I
Sources of interactions contributed to the MAP. The bolded rows are the sums of the edges for each interaction set.

C. Machine Learning Methods

We are given a weighted, directed graph $G = (V, E, w)$, where the nodes V are genes and the edges E are interactions between genes with weights w. We also have a set of labeled nodes $L \subset V$ that contains genes associated with a disease (positives, L^+) and randomly selected genes not associated with the disease (negatives, L^-) such that $L = L^+ \cup L^-$ and $L^+ \cap L^- = \emptyset$. Recent work has found that supervised learning approaches outperform label propagation methods for gene classification [8]. Thus, we trained a logistic regression classifier using the labeled nodes as our training set. For a given node v, we generate a feature vector $\mathbf{x}^{(v)}$ of length $|L|$ that represents the incoming labeled neighbors. The features $x_u^{(v)}$ for each $u \in L$ correspond to the edge weight from v's incoming labeled nodes:

$$x_u^{(v)} = \begin{cases} w_{uv} & \text{if } (u, v) \in E \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Each labeled node $u \in L$ has a regression coefficient β_u, and these coefficients are fitted to predict the probability p_v that node v is a disease positive, such that

$$logit(p_v) = \beta_0 + \sum_{u \in L} \beta_u x_u^{(v)}. \tag{2}$$

We call this the *unweighted logistic regression classifier* because the labels L are not weighted.

We also apply weighted logistic regression, wherein we consider the score of the node labels L. The weights for each positive L^+ is the node's DisGeNet disease association score (which may change depending on the disease). We assigned a weight of 0.2 to each negative L^-, corresponding to the minimum score possible in the pan-disease evaluation. Logistic regression was implemented using Python's sci-kit learn package [32].

D. Evaluation

We describe two experimental settings which require different numbers of DisGeNet diseases, different number of positive nodes, and different validation frameworks (Fig. 1B). Each experimental setting aims to answer a different question.

1) Pan-Disease: In the pan-disease experimental setting, we compare disease gene prediction performance across diverse diseases and focus on whether regulatory interactions (e.g. the MAP) affect this performance. We control for the number of genes in the gold standards for each disease, using the top 100 annotated disease genes as positives and randomly selecting 100 genes not annotated to the particular disease as

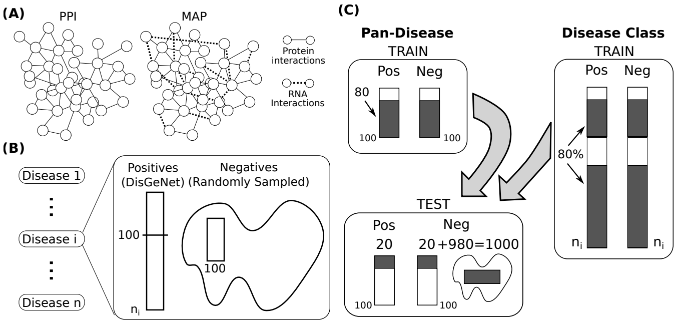

Fig. 1. Networks and Validation Frameworks. **(A)** Network Construction. **(B)** General validation framework: for each DisGeNet disease, we define two sets of positives: the 100 top-scoring DisGeNet genes, and all DisGeNet genes. We sample negatives randomly from the remaining genes not associated with the disease. **(C)** Experimental settings. In the pan-disease setting, we use 5-fold cross validation on the top 100 DisGeNet positives and 100 randomly-sampled negatives. In the disease class setting, we use all DisGeNet positives for training, but use stratified 5-fold cross validation to ensure that we test on the exact same positives as in the pan-disease setting.

negatives (Fig. 1C). We trained a weighted and an unweighted logistic regression classifier on the PPI and the MAP using these positive and negative sets, resulting in four different runs (weighted and unweighted; PPI and MAP).

We use 5-fold cross validation, wherein 80 positive nodes and 80 negative nodes were used as training examples and each model was assessed on its ability to prioritize the remaining 20 positive genes. In practice, there are typically many more genes *not* associated with a disease than genes associated with a disease. For a more realistic scenario, we tested 1,000 negatives, including the 20 hidden negatives from training and 980 nodes not associated with the given disease that were randomly selected (Fig. 1C). We computed precision and recall for the ranked list of predictions for five validation iterations (corresponding to $5 \times 5 = 25$ runs), and consider the area under the precision recall curve (AUPRC) when comparing across diseases. For this and other experimental settings, we use precision and recall rather than ROC curves due to the imbalanced positive and negative sets used in testing [33].

2) Disease Class: In the disease class experimental setting, we compare disease gene prediction performance for neoplastic processes and focus on whether adding low-confidence disease genes affect this performance. Here, all DisGeNet genes annotated to each cancer were used for training. However, we wanted to assess the exact same 100 positives as in the pan-disease setting Thus, we stratified the training positives into two groups: the top 100 disease genes and the remaining disease genes. We then performed stratified cross-validation such that one-fifth of each group were hidden (Fig. 1C). We assessed the unweighted MAP classifier on the hidden genes from the top 100 positives and 1,000 negatives were constructed as in the pan-disease setting.

III. Results

A. Pan-Disease Assessment

We ran all four methods (weighted and unweighted logistic regression on the PPI and the MAP) on the 49 DisGeNet

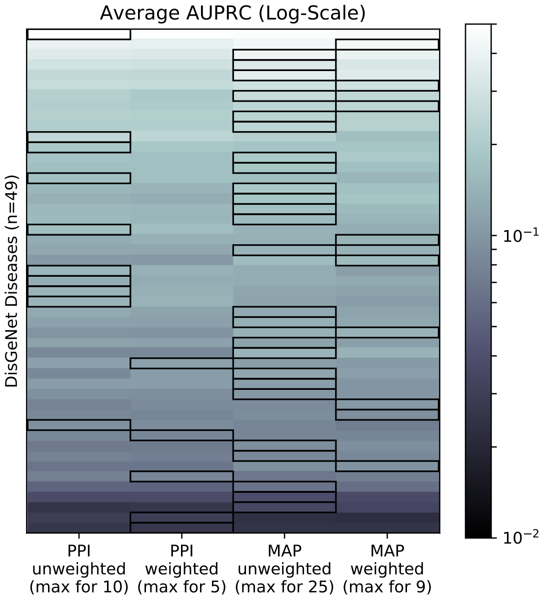

Fig. 2. Performance of the Four Classifiers (columns) for each Disease (rows). Area under the precision-recall curve (AUPRC) values were averaged across 25 iterations (5 runs of 5-fold cross validation) and plotted on a log-scale. Diseases are ordered by average AUPRC, and boxes denote the largest AUPRC for each row.

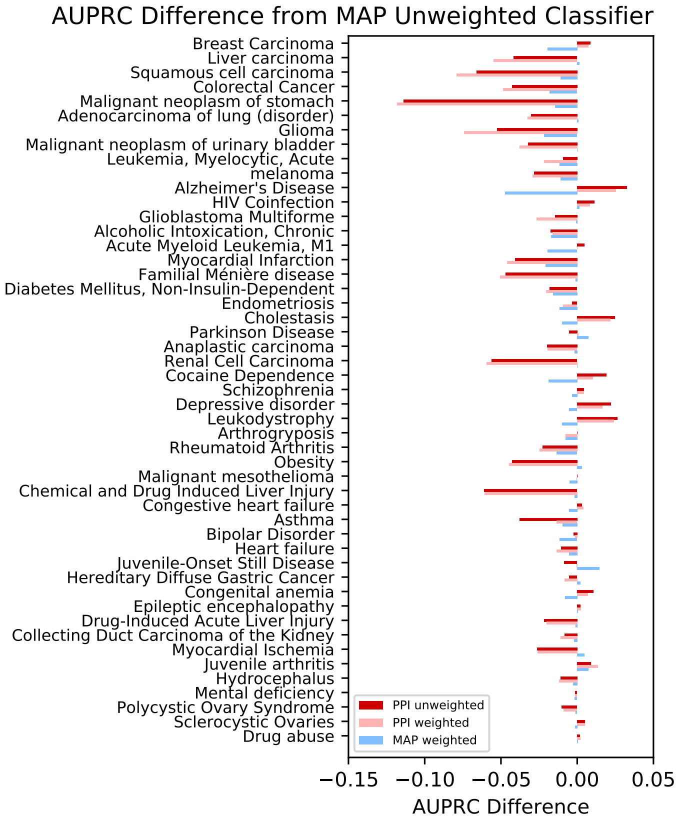

Fig. 3. Difference in Performance from the MAP Unweighted Classifier. The average AUPRC of the MAP unweighted classifier was subtracted from the average AUPRC of all other classifiers for each disease; bars to the left of 0 are worse than the MAP unweighted classifier. Diseases are ordered as in Fig. 2.

Feature Type	Feature	Estimate	S.E.	t	p
Intercept		7.438e-03	3.324e-03	2.237	2.530e-02
Network	PPI vs. MAP	1.354e-02	2.354e-03	5.754	9.252e-09
Node Labels	Weighted	-4.596e-03	2.354e-03	-1.953	5.092e-02
Disease Genes	Subnetwork Density	1.573	3.634e-02	43.281	<2e-16
Disease Genes	Regulatory Density	0.2669	3.601e-02	7.410	1.484e-13
Disease Class	MBD	-2.918e-02	3.573e-03	-8.167	3.983e-16
Disease Class	NP	3.431e-03	3.609e-03	0.9506	0.3418

TABLE II
LINEAR MODEL OF DISEASE GENE PREDICTION PERFORMANCE (PAN-DISEASE ASSESSMENT).

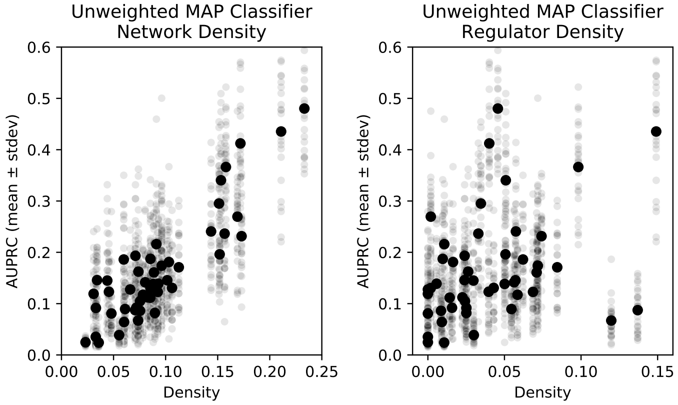

Fig. 4. DisGeNet gene prediction performance compared to subnetwork density (left) and regulatory density (right) for the unweighted MAP classifier. Each light gray point represents the AUPRC for one cross-validation run from one of the 49 diseases, and the black points represent average AUPRC across 25 runs.

diseases with at least 100 positives that had a score greater than 0.2. In terms of area under the precision recall curve (AUPRC), 34 of the 49 diseases had better performance using the MAP compared to the PPI (Fig. 2). Thus, including regulatory information helped in nearly 70% of the diseases. Further, 25 of the 34 diseases with improved MAP performance had the best average AUPRC using the unweighted logistic regression classifier, indicating that weighting the nodes by disease association score decreased performance (Fig. 2). The magnitude of the performance differences compared to the MAP unweighted classifier were generally larger for the diseases with the best overall performance, and when the MAP unweighted classifier was not the best, its average AUPRC was within 0.05 of the best-performing classifier (Fig. 3).

Next, we wanted to assess whether network features or disease classes were contributing to the performance differences among the classifiers. We calculated two statistics to describe the general network properties of the disease positives by considering the induced subgraph of the top 100 positives used for training. We calculated the *subnetwork density* (the number of edges divided by the number of possible edges.) and the *regulatory density* (the proportion of edges in induced subgraph that included a regulatory interaction between a positive node and a gene's transcript). Diseases may be classified as one "Mental or Behavioral Dysfunction" (abbreviated *MBD*) or "Neoplastic Process" (abbreviated *NP*).

We performed multiple linear regression to predict AUPRC based on five features related to the classifier, the network, and the disease (Table II). We found a significant regression equation ($F(6, 4893) = 737.5$, p $< 2.2 \times 10^{-16}$), with an adjusted R^2 of 0.4749. As expected, the network type (MAP or PPI) was a significant predictor, while weighting the labeled positives had a p-value of about 0.05. The disease classification of "Mental or Behavioral Dysfunction" was a more significant predictor than "Neoplastic Process," illustrating that the performance increase for cancers could be explained by the other terms in the equation. Both the subnetwork density and regulatory density were significant predictors of AUPRC (examples of this trend for the unweighted MAP classifier are shown in Fig. 4).

Since the MAP performed better overall than the PPI, and weighting was not a significant predictor of AUPRC, we used five-fold cross-validation on the MAP with unweighted logistic regression to compare all diseases. We found that gene-prediction performance varied across diseases, with neoplastic processes generally outperforming other diseases (Fig. 5). There was significant variation in AUPRC between disease classes for this classifier (Kruskal-Wallis rank sum, chi-squared = 1057.6, p $< 2.2 \times 10^{-16}$).

Finally, we measured the correlation between performance and the features in the linear model while removing the effect of possible confounders. We identified three additional DisGeNet features (number of disease publications, number of publications for all disease genes, and total number of disease genes) and two additional features of the disease gene induced subgraph (average degree and weighted average degree) as possible confounders. We calcuated the partial correlation of the PRAUC and each feature type while controlling for each one of these confounders individually (Table III). Here, the "Disease Class" feature is categorical (e.g., the colors in Fig. 5). All features remained significant when controlling for individual confounders except for the weighted labels. Surprisingly, when controlling for all confounders simultaneously, the regulatory density is no longer significant (last column in Table III).

Running time: We found that with only 100 disease positive labeled nodes and only 100 disease negative nodes, our method runs relatively quickly. With the MAP loaded in a NetworkX graph, it takes a late-2016 8 GB Macbook Pro one minute to assemble the input tables for the classifier, and it takes approximately seven seconds for it to complete one cross-validation run, which includes both the weighted and unweighted classifiers for the MAP.

B. *Disease Class Assessment*

We then evaluated the complete DisGeNet positive sets for diseases in the "Neoplastic Process" disease class. These

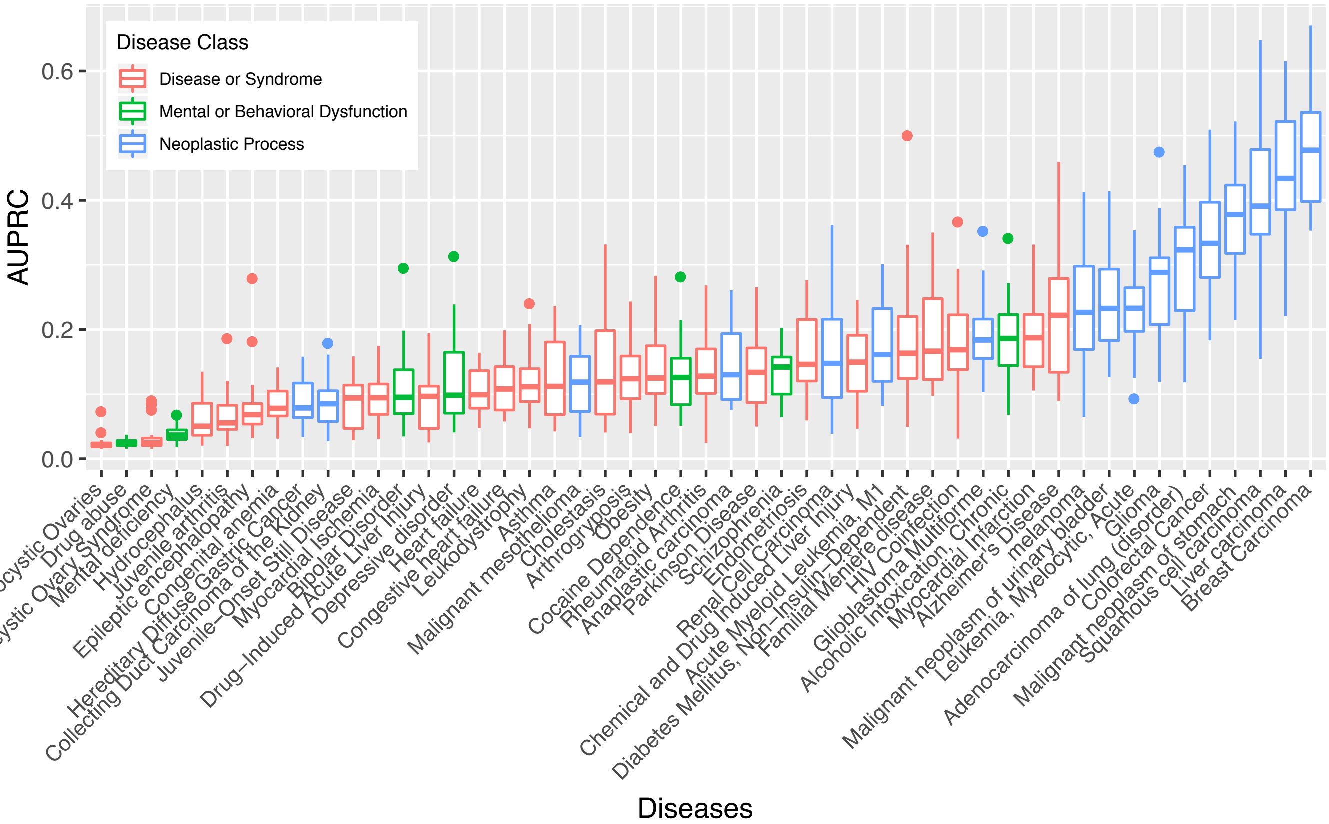

Fig. 5. DisGeNet gene prediction performance for the pan-disease assessment with the unweighted MAP classifier, ordered by median performance.

Feature	Potential Confounders # Disease Publications	# Disease Gene Publications	Total # Disease Genes	Subnetwork Average Degree	Subnetwork Weighted Average Degree	All Confounders
PPI vs. MAP	2.72e-05	1.873e-05	2.620e-08	1.056e-07	1.276e-06	2.442e-09
Weighted	0.155	0.147	0.059	0.072	0.101	0.0225
Subnetwork Density	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16	1.624e-07
Regulatory Density	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16	0.5232
Disease Class	<2e-16	<2e-16	<2e-16	0.006	<2e-16	7.626e-10

TABLE III

PARTIAL CORRELATION p-VALUES FOR DISEASE GENE PREDICTION (PAN-DISEASE ASSESSMENT). ENTRIES REPRESENT THE PARTIAL CORRELATION p-VALUE BETWEEN AUPRC AND THE FEATURE (ROW) WHEN CONTROLLING FOR THE POTENTIAL CONFOUNDER (COLUMN).

diseases had the largest AUPRC values in the pan-disease assessment (Fig. 5). To compare the disease class assessment to the pan-disease assessment, we tested on the same positives using stratified 5-fold cross validation (See Methods). When we consider all disease genes instead of the top 100, we see a marked improvement in performance (Fig. 6, Wilcoxon rank sum test, W = 997,420, p $< 2.2 \times 10^{-16}$). Thirteen of the 17 cancers have better performance with Wilcoxon rank sum test $p < 0.01$. Fig. 7 shows the disease gene performance of four classifiers on the two validation settings for three example cancers: lung, liver, and skin. In these examples, there is a visual effect of including all disease genes, improving the MAP classifiers in particular (yellow and green lines.)

As before, we used multiple linear regression to predict AUPRC based on network features (Table IV). A significant regression equation was found ($F(4, 1663) = 476.3$, p $< 2.2\times 10^{-16}$), with an adjusted R^2 of 0.5328. Similar to the pan-disease setting, the network type (PPI vs. MAP), subnetwork

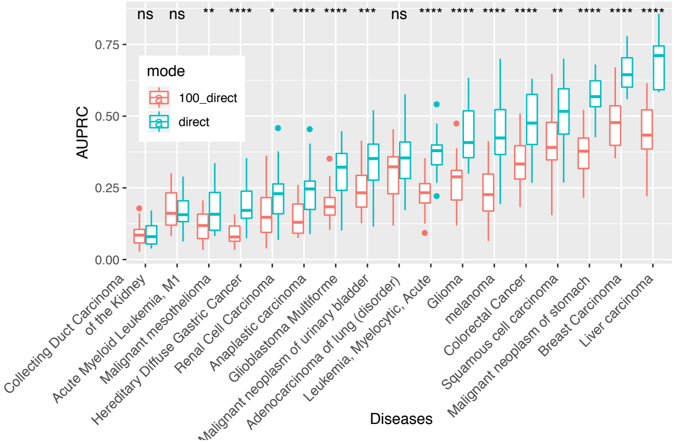

Fig. 6. Five-fold cross-validation performance (unweighted MAP classifier) of DisGeNet gene prediction for neoplastic processes. "100_direct": classifier trained on 100 positives; "direct": classifier trained on all positives. (Wilcoxon rank sum test, ns: $p > 0.05$, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$).

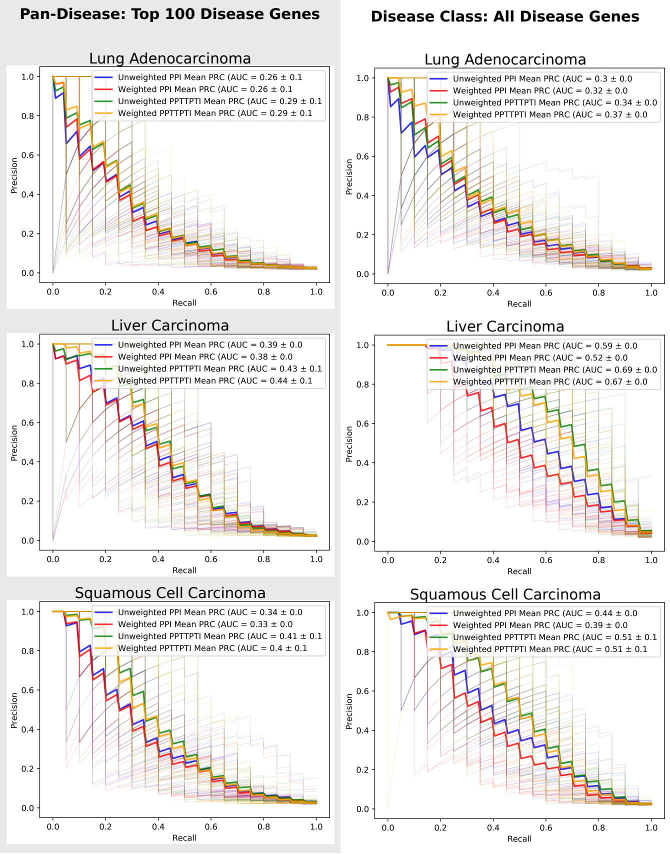

Fig. 7. Performance of all four classifiers trained on the top 100 genes in the pan-disease setting (left) or all genes in the disease class setting (right) for Lung Adenocarcinoma (top), Liver Carcinoma (middle) and Squamous Cell Carcinoma (bottom). Thin lines represent each of the 25 cross-validation runs for each classifier and the thick line is the median precision-recall curve by AUPRC value. "PPTTPTI" in the legend indicates the MAP network.

density, and regulatory density were all significant predictors for performance, whereas the weighted gene labels were not as significant (Table IV). However, when we controlled for the number of DisGeNet genes, the correlation between AUPRC and subnetwork density disappears (Table V). Further, the correlation between AUPRC and regulatory density is not significant when controlling for all confounders.

Running time: We noticed that runtime increased proportionally with the increase in disease genes. With the PPI loaded in a NetworkX graph, it takes a late-2016 8 GB Macbook Pro 17 minutes to assemble the input tables for the classifier for Alzheimer's Disease, which has 1826 disease genes. It takes approximately 75 seconds for it to complete one cross-validation run, not including the weighted classifier. For Liver carcinoma, which has 3226 genes, it takes 44 minutes to assemble the input tables and about 11 minutes to complete one cross-validation run.

IV. Conclusion

In this study, we demonstrated that including gene regulatory interaction data in network-based disease gene prediction significantly improves prediction performance over not including it (Fig. 2 and Fig. 3). While a handful of the diseases performed better under the PPI, most performed better under the MAP. These results underscore the importance of creating holistic models of molecular biology in computational inference of diseases.

In the pan-disease assessment, using the top 100 disease positives controlled for the difference in gene set size between diseases. In this evaluation, the subnetwork density of the top 100 disease genes was the best predictor of prediction performance in a multiple linear regression model (Table II). Recent work has found that, for the disease gene prediction problem, network density is correlated with PRAUC performance for multiple interaction networks, supporting our results that density is associated with classifier performance [8]. Partial correlation also showed that this relationship is not the result of how well the disease or the disease genes have been studied (Table III). The effect of density is significant even when accounting for the average disease gene degree, meaning that this feature is not accounted for by the general connectivity to the network at large. Because density was calculated based on the training nodes for each cross-validation run, density is reflective of the training set's connectivity to itself, rather than its connectivity to the test set. With these possible confounders accounted for, these results suggest that the density of the top 100 genes associated with a disease is indicative of how well network connectivity learning methods can inform us about whether or not a gene is a disease gene.

We also showed that including more disease data significantly improves disease gene prediction (Fig. 6). Surprisingly, we also show that weighting disease positives does not provide any observable benefit to prediction performance, despite the positive sets sometimes containing thousands of genes (Table IV). This might be due to the fitting process of logistic regression, wherein the regression coefficients change in order to optimize performance on the supervised set, effectively weighting the positives and negatives without additional data. However, established supervised learning weighting mechanisms involve weighting how much a gene's status as a positive or a negative affects the fitting process, rather than weighting its effect on other genes [32]. It is possible that including the additional positives simply leads to more ways that the network can train itself to recognize positives, rather than actually providing additional information. If this were the case, then any particular nodes could be added to the positive set and they would improve performance. However, Krishnan et. al. demonstrated that adding random positives does not improve performance when predicting autism disease genes [12].

We emphasize that we are not striving to create an optimal classifier for disease gene prediction. Instead, we are exploring how network features correlate with performance of logistic regression, focusing on different characteristics of the input data (the network and disease gene positives). We and others have worked on designing network-based algorithms to optimally predict disease genes [6]–[13], and there are many approaches beyond unregularized logistic regression that perform well. However, we felt that keeping the classifier simple helped us explore network features that may contribute to performance variation across diverse diseases. When predicting disease genes for a particular disease of interest, it is also important

Feature Type	Feature	Estimate	S.E.	t	p
Intercept		0.7156	0.0113	63.330	<2e-16
Network	PPI vs. MAP	0.0619	0.0060	10.249	<2e-16
Node Labels	Weighted	-0.0136	0.0060	-2.266	0.0236
Disease Genes	Subnetwork Density	-8.9776	0.2487	-36.090	<2e-16
Disease Genes	Regulatory Density	2.7616	0.2624	-1.953	<2e-16

TABLE IV

LINEAR MODEL OF DISEASE GENE PREDICTION PERFORMANCE (DISEASE CLASS ASSESSMENT).

	Potential Confounders					
Feature	# Disease Publications	# Disease Gene Publications	Total # Disease Genes	Subnetwork Average Degree	Subnetwork Weighted Average Degree	**All Confounders**
PPI vs. MAP	1.553e-12	4.381e-18	1.557e-34	4.511e-21	2.376e-16	5.629e-39
Weighted Labels	0.1207	0.0579	0.0080	0.0397	0.0725	0.0048
Subnetwork Density	5.221e-261	8.21e-134	0.5365	3.559e-178	1.242e-223	2.085e-26
Regulatory Density	1.105e-63	3.739e-13	0.0002	6.597e-86	1.003e-100	0.9101

TABLE V

PARTIAL CORRELATION p-VALUES FOR DISEASE GENE PREDICTION (DISEASE CLASS ASSESSMENT). ENTRIES REPRESENT THE PARTIAL CORRELATION p-VALUE BETWEEN AUPRC AND THE FEATURE (ROW) WHEN CONTROLLING FOR THE POTENTIAL CONFOUNDER (COLUMN).

to generate a null distribution of performance values (e.g. AUPRCs) from randomized data to have confidence in the gene classifications [12].

As we collect and aggregate information about diverse disease for disease gene prediction, it is important to consider the associations between the input features and a classifier's performance. Using logistic regression, we have shown that network-based features have the ability to distinguish performance among diverse disease types, and may offer a first step to indicate which classifiers are well-suited for different diseases.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation (NSF-BIO-ABI #1750981, to AR) and a Reed Neuroscience Fellowship Fund (to AK).

REFERENCES

[1] L. R. Cardon and J. I. Bell, "Association study designs for complex diseases," *Nature Reviews Genetics*, vol. 2, no. 2, p. 91, 2001.

[2] P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang, "Five years of GWAS discovery," *Am. J. Hum. Genet.*, vol. 90, no. 1, pp. 7–24, Jan 2012.

[3] M. Vidal, M. E. Cusick, and A.-L. Barabási, "Interactome networks and human disease," *Cell*, vol. 144, no. 6, pp. 986–998, 2011.

[4] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature reviews genetics*, vol. 12, no. 1, p. 56, 2011.

[5] S. Karni, H. Soreq, and R. Sharan, "A network-based method for predicting disease-causing genes," *Journal of Computational Biology*, vol. 16, no. 2, pp. 181–189, 2009.

[6] A. Califano, A. J. Butte, S. Friend, T. Ideker, and E. Schadt, "Leveraging models of cell regulation and gwas data in integrative network-based association studies," *Nature genetics*, vol. 44, no. 8, p. 841, 2012.

[7] J. K. Huang, D. E. Carlin, M. K. Yu, W. Zhang, J. F. Kreisberg, P. Tamayo, and T. Ideker, "Systematic evaluation of molecular networks for discovery of disease genes," *Cell systems*, vol. 6, no. 4, pp. 484–495, 2018.

[8] R. Liu, C. A. Mancuso, A. Yannakopoulos, K. A. Johnson, and A. Krishnan, "Supervised-learning is an accurate method for network-based gene classification," *bioRxiv*, 2019.

[9] P. Creixell, J. Reimand, S. Haider, G. Wu, T. Shibata, M. Vazquez, V. Mustonen, A. Gonzalez-Perez, J. Pearson, C. Sander *et al.*, "Pathway and network analysis of cancer genomes," *Nature methods*, vol. 12, no. 7, p. 615, 2015.

[10] N. J. Krogan, S. Lippman, D. A. Agard, A. Ashworth, and T. Ideker, "The cancer cell map initiative: defining the hallmark networks of cancer," *Molecular cell*, vol. 58, no. 4, pp. 690–698, 2015.

[11] V. K. Ramanan and A. J. Saykin, "Pathways to neurodegeneration: mechanistic insights from gwas in alzheimers disease, parkinsons disease, and related disorders," *American journal of neurodegenerative disease*, vol. 2, no. 3, p. 145, 2013.

[12] A. Krishnan, R. Zhang, V. Yao, C. L. Theesfeld, A. K. Wong, A. Tadych, N. Volfovsky, A. Packer, A. Lash, and O. G. Troyanskaya, "Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder," *Nature neuroscience*, vol. 19, no. 11, p. 1454, 2016.

[13] M. Bern, A. King, D. A. Applewhite, and A. Ritz, "Network-based prediction of polygenic disease genes involved in cell motility," *BMC bioinformatics*, vol. 20, no. 12, p. 313, 2019.

[14] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, "Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders," *Nucleic acids research*, vol. 33, no. suppl_1, pp. D514–D517, 2005.

[15] J. S. Amberger, C. A. Bocchini, F. Schiettecatte, A. F. Scott, and A. Hamosh, "Omim. org: Online mendelian inheritance in man (omim®), an online catalog of human genes and genetic disorders," *Nucleic acids research*, vol. 43, no. D1, pp. D789–D798, 2014.

[16] M. D. Mailman, M. Feolo, Y. Jin, M. Kimura, K. Tryka, R. Bagoutdinov, L. Hao, A. Kiang, J. Paschall, L. Phan *et al.*, "The ncbi dbgap database of genotypes and phenotypes," *Nature genetics*, vol. 39, no. 10, p. 1181, 2007.

[17] K. A. Tryka, L. Hao, A. Sturcke, Y. Jin, Z. Y. Wang, L. Ziyabari, M. Lee, N. Popova, N. Sharopova, M. Kimura *et al.*, "Ncbis database of genotypes and phenotypes: dbgap," *Nucleic acids research*, vol. 42, no. D1, pp. D975–D979, 2013.

[18] C. J. Mattingly, G. T. Colby, J. N. Forrest, and J. L. Boyer, "The comparative toxicogenomics database (ctd)." *Environmental health perspectives*, vol. 111, no. 6, pp. 793–795, 2003.

[19] A. P. Davis, C. J. Grondin, R. J. Johnson, D. Sciaky, B. L. King, R. McMorran, J. Wiegers, T. C. Wiegers, and C. J. Mattingly, "The comparative toxicogenomics database: update 2017," *Nucleic acids research*, vol. 45, no. D1, pp. D972–D978, 2016.

[20] J. Pinero, N. Queralt-Rosinach, A. Bravo, J. Deu-Pons, A. Bauer-Mehren, M. Baron, F. Sanz, and L. I. FOPTURLong, "Disgenet: a discovery platform for the dynamical exploration of human diseases and their genes," *Database*, vol. 2015, 2015.

[21] J. Piero, . Bravo, N. Queralt-Rosinach, A. Gutirrez-Sacristn, J. Deu-Pons, E. Centeno, J. Garca-Garca, F. Sanz, and L. I. FOPTURLong, "DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants," *Nucleic Acids Research*, vol. 45, no. D1, pp. D833–D839, Jan. 2017.

[22] A. Valdeolivas, L. Tichit, C. Navarro, S. Perrin, G. Odelin, N. Levy, P. Cau, E. Remy, and A. Baudot, "Random walk with restart on multiplex and heterogeneous biological networks," *Bioinformatics*, vol. 35, no. 3, pp. 497–505, 2018.

[23] D. E. Carlin, S. H. Fong, Y. Qin, T. Jia, J. K. Huang, B. Bao, C. Zhang, and T. Ideker, "A fast and flexible framework for network-assisted genomic association," *iScience*, vol. 16, pp. 155–161, 2019.

[24] S. Hwang, C. Y. Kim, S. Yang, E. Kim, T. Hart, E. M. Marcotte, and I. Lee, "Humannet v2: human gene networks for disease research," *Nucleic acids research*, vol. 47, no. D1, pp. D573–D580, 2018.

[25] D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, L. J. Jensen, and C. Mering, “STRING v11: proteinprotein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets,” *Nucleic Acids Research*, vol. 47, no. D1, pp. D607–D613, Jan. 2019.

[26] C. S. Greene, A. Krishnan, A. K. Wong, E. Ricciotti, R. A. Zelaya, D. S. Himmelstein, R. Zhang, B. M. Hartmann, E. Zaslavsky, S. C. Sealfon *et al.*, “Understanding multicellular function and disease with human tissue-specific networks,” *Nature genetics*, vol. 47, no. 6, p. 569, 2015.

[27] C. L. Poirel, A. Rahman, R. R. Rodrigues, A. Krishnan, J. R. Addesa, and T. Murali, “Reconciling differential gene expression data with molecular interaction networks,” *Bioinformatics*, vol. 29, no. 5, pp. 622–629, 2013.

[28] Y. Yi, Y. Zhao, C. Li, L. Zhang, H. Huang, Y. Li, L. Liu, P. Hou, T. Cui, P. Tan, Y. Hu, T. Zhang, Y. Huang, X. Li, J. Yu, and D. Wang, “RAID v2.0: an updated resource of RNA-associated interactions across organisms,” *Nucleic Acids Research*, vol. 45, no. D1, pp. D115–D118, Jan. 2017.

[29] H. Han, J.-W. Cho, S. Lee, A. Yun, H. Kim, D. Bae, S. Yang, C. Y. Kim, M. Lee, E. Kim, S. Lee, B. Kang, D. Jeong, Y. Kim, H.-N. Jeon, H. Jung, S. Nam, M. Chung, J.-H. Kim, and I. Lee, “TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions,” *Nucleic Acids Research*, vol. 46, no. D1, pp. D380–D386, Jan. 2018.

[30] Z.-P. Liu, C. Wu, H. Miao, and H. Wu, “RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse,” *Database*, vol. 2015, p. bav095, 2015.

[31] F. Vafaee, J. R. Krycer, X. Ma, T. Burykin, D. E. James, and Z. Kuncic, “ORTI: An Open-Access Repository of Transcriptional Interactions for Interrogating Mammalian Gene Expression Data,” *PLOS ONE*, vol. 11, no. 10, p. e0164535, Oct. 2016.

[32] Fabian Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, “Scikit-learn: Machine learning in Python,” *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[33] T. Saito and M. Rehmsmeier, “The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets,” *PLOS ONE*, vol. 10, no. 3, pp. 1–21, 03 2015.