The hetnet awakens: understanding complex diseases through data integration and open science

by

Daniel S. Himmelstein

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

Acknowledgments

I'd like to thank my PhD advisor, Dr. Sergio Baranzini, as well the other two members of my Thesis Committee, Dr. John Witte and Dr. Andrej Sali. In addition, I thank my Qualifying Exam Committee, which consisted of Dr. Patricia Babbitt as Chair, Dr. Sali, Dr. Katherine Pollard, and Dr. Jun Song. During the first year of my PhD, I rotated with Dr. Ryan Hernandez, Dr. Witte, and Dr. Baranzini. Prior to coming to the University of California, San Francisco, I was fortunate to research under Dr. Jason Moore, Dr. Olga Troyanskaya, and Dr. Casey Greene. To these individuals and to others who have mentored me during my existence, I offer my deep gratitude and appreciation. Finally, I'd like to thank my parents for rearing me and my friends for contributing to a joyful life.

Abstract

The hetnet awakens: understanding complex diseases through data integration and open science

By Daniel S. Himmelstein

Human disease is complex. However, the explosion of biomedical data is providing new opportunities to improve our understanding. My dissertation focused on how to harness the biodata revolution. Broadly, I addressed three questions: how to integrate data, how to extract insights from data, and how to make science more open.

To integrate data, we pioneered the hetnet — a network with multiple node and relationship types. After several preludes, we released Hetionet v1.0, which contains 2,250,197 relationships of 24 types. Hetionet encodes the collective knowledge produced by millions of studies over the last half century.

To extract insights from data, we developed a machine learning approach for hetnets. In order to predict the probability that an unknown relationship exists, our algorithm identifies influential network patterns. We used the approach to prioritize disease–gene associations and drug repurposing opportunities. By evaluating our predictions on withheld knowledge, we demonstrated the systematic success of our method.

After encountering friction that interfered with data integration and rapid communication, I began looking at how to make science more open. The quest led me to explore realtime open notebook science and expose publishing delays at journals as well as the problematic licensing

of publicly-funded research data.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The amount of available data is growing astronomically. This growth is especially visible in genomics [1,2], bioinformatics [3], and medicine [4]. Much of the data explosion consists of more data of the same type, such as genomic sequence. However, we're also seeing new types of data arise [5]. The challenges resulting from the data explosion fall into two categories: *too much* data and *too diverse* of data. Big data refers to when the quantity of data becomes large leading to the challenge of *too much*. Heterogeneous data refers to when datasets contain many different components or types and leads to the challenge of *too diverse*. Solutions for *too much* are in a constant arms race with the unbridled production of data [6,7]. My thesis focuses instead on the *too diverse* problem, where solutions, especially in biomedicine, have been late to the race.

One fundamental characteristic of the data explosion is that each unit of data becomes less and less informative. At any given time, storage and analysis capacity is filled with the most meaningful information available. For example, imagine if each individual got to choose 10 photos to commemorate their existence. Each photo would be highly commemorative. Now imagine if instead every photo an individual has ever taken is retained. Under this later regime, a random photo would rarely be highly commemorative. The same trend applies to the modern expansion in data availability: computational advances have allowed us to store and analyze more data but do not inherently create more valuable data. In addition, diminishing marginal informativeness emerges from the greater ease of generating less informative content. For example, new techniques in biology tend to generate more data but less information per unit. For example, an individual nucleotide uncovered by sequencing is generally of no value. A gene's transcriptional abundance, which can now be easily measured for all genes in a human, often has little phenotypic relevance.

Nonetheless, the advent of high-throughput technologies to assess genetic sequence and expression are incredibly valuable. Their value comes in providing a systematic exploration of a certain aspect of biology. By offering a comprehensive view of a previously inaccessible level of biology, data from these technologies offers novel insight [8]. However, a subset of diseases have been highly resilient to complete insights. These diseases are called complex and are caused by a multitude of genetic and environmental factors. For these diseases, signals in high-throughput datasets tend to be weak. While this should not come as a surprise given the complexity of biology, the result is that the data explosion has yet to crack complex disease.

So we face an onslaught of heterogeneous data where each individual type tends to be only weakly informative. Analyzing each dataset in isolation yields weak findings. For example, genotyping data predicts complex disease susceptibility but only weakly informs a patient of their risk compared to the general population [9]. The motivation behind my dissertation is that combining many weak datasets of different types can yield strong findings. As stated by a 2015 review [10], "a comprehensive understanding of a biological system can come only from a joint analysis of all omics layers."

Our journey to understand human disease using heterogeneous data led us to five fields: data integration, hetnets, machine learning, data visualization, and open science. Data integration means combining data from multiple sources and is thus a prerequisite to analyzing heterogeneous data [11]. Hetnets, short for heterogeneous networks, are networks with multiple types of nodes or relationships. We adopted hetnets as a data structure to encode heterogeous information. Machine learning allows us to extract insights on data that is poorly suited for human cognitive understanding. For example, we use machine learning to identify patterns and make predictions from datasets that are too diverse and large for human learning. We use data visualization to understand our data and express our findings. Open science is a movement to make all aspects of research more accessible. Specifically, open data has been the bedrock of my dissertation by providing compendia of information ripe for the integration.

Chapter 2, describes our study to integrate genetic studies into a single network of disease similarity. This study leverages genome-wide data to assess disease similarity free from the biases of existing knowledge. However, this disease network contains only a single type of relationship. Chapter 3 details our adoption of hetnets and lessons along the way. Chapter 4 describes our study to prioritize gene–disease associations using a hetnet. In Chapter 5, we use a similar approach to predict repurposed drug therapies. Chapter 6 discusses our use of and contributions to open science. Finally, Chapter 7 concludes my dissertation.

To close my introduction, I will show my Facebook friendship network as it existed in 2014. The network visualization below was produced using only friendship (Figure 1.1). In other words, the graph doesn't contain any identifying information about my friends. However, both the layout and community detection algorithms were able to finely characterize my past social experience. Networks make grand simplifications. Here all friendships are represented equally as a single unweighted edge. There is no attribute for how long two users have been friends, how many times they've posted on each other's wall, or whether they have even met in person. Successful applications of network science require the right simplification for the problem. My Facebook network could be algorithmically understood using just mutual friendship information. However, for disease biology where the problems are more complex and the datasets more imperfect, my dissertation will argue that a powerful simplification is combining heterogeneous data into a single network while retaining type.

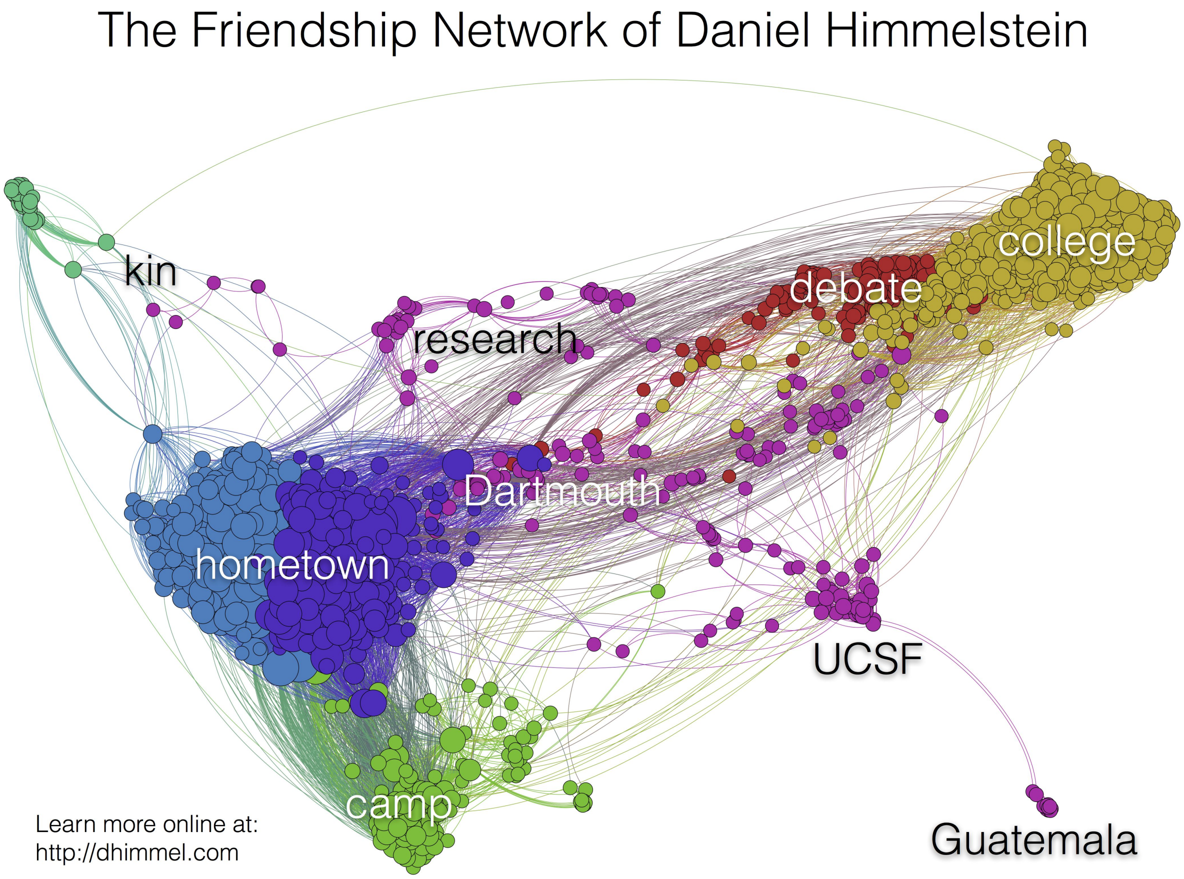

Figure 1.1. My network of friends on Facebook. Each of my 1,278 Facebook friends are nodes. The edges represent the 40,255 friendships amongst my friends. Node size indicates degree (number of mutual friends). I applied a force-directed layout to organize the network: highly connected nodes are pulled together, while distantly connected nodes are repulsed [12]. Ideally, I would have manually categorized my friends into communities. However, to save time, I classified friends using a community detection algorithm [13]. The algorithm succeeded in partitioning my friends into communities (denoted by color) that coincided with my past social experiences.

Chapter 2

GWAS gives rise to a new breed of homonet

This chapter describes a disease similarity network I created from published GWAS findings. The analysis was part of a larger study which assessed the genetic overlap of Hodgkin lymphoma and multiple sclerosis [14]:

> Khankhanian P, Cozen W, **Himmelstein DS**, Madireddy L, Din L, et al. (2016) **Meta-analysis of genome-wide association studies reveals genetic overlap between Hodgkin lymphoma and multiple sclerosis**. *Int J Epidemiol.* DOI: 10.1093/ije/dyv364

Rather than include this study as a chapter, I'm including a post I wrote for the *International Journal of Epidemiology* blog. The post focuses on my portion of the larger study and describes our method in the context of alternative approaches.

2.1 A puzzling similarity

Researchers have long noted puzzling similarities between Hodgkin lymphoma and multiple sclerosis. Although the first is a cancer and the second is an autoimmune disease, risk for both diseases appears to increase due to the Epstein–Barr virus and a lack of sunlight. In fact having a family member with multiple sclerosis may place you at increased risk for Hodgkin lymphoma and vice versa. Now, a recent study, on which I am a co-author, has identified genetic similarities.

Our analysis compared two studies designed to pinpoint the genetic variants behind disease susceptibility. Together these studies, referred to as genome-wide association studies or GWAS, analyzed 1,816 Hodgkin lymphoma patients, 9,772 multiple sclerosis patients, and 25,255 healthy individuals. The prevalence of 404,069 genetic variants were compared between patients and healthy individuals for each disease. We found a large number of variants that appeared to affect susceptibility in both diseases. Additionally, a genetic risk model designed for multiple sclerosis also predicted Hodgkin lymphoma.

2.2 Building a network

We were excited to find common genetic signals, but the similarity lacked context. For example, is Hodgkin lymphoma more similar to other cancers than to multiple sclerosis? To answer these questions, we needed GWAS results for many diseases. We turned to the GWAS Catalog, whose team of curators reads through GWAS publications and extracts the associated variants into a public database.

For 82 diseases, we identified associated regions of the genome, called loci. Then for each disease pair, we calculated a similarity score based on the number of overlapping loci. The score adjusts for the number of loci per disease which can vary widely — multiple sclerosis has 55 loci, whereas Hodgkin lymphoma has 7. Of the 3,321 possible disease pairs, 433 had at least one overlapping locus.

Next, we calculated the network proximity of disease pairs. Proximity is calculated by transmitting similarity scores between related diseases. By leveraging the same insight as PageRank — the founding algorithm behind Google's web search — the approach helps improve the robustness and connectedness of our network.

Below we display our network of disease proximities (Figure 2.1). See these tables for disease abbreviations and specific proximity scores (plotted with edge thickness). We applied a layout that pushes proximal diseases together and distant diseases apart.

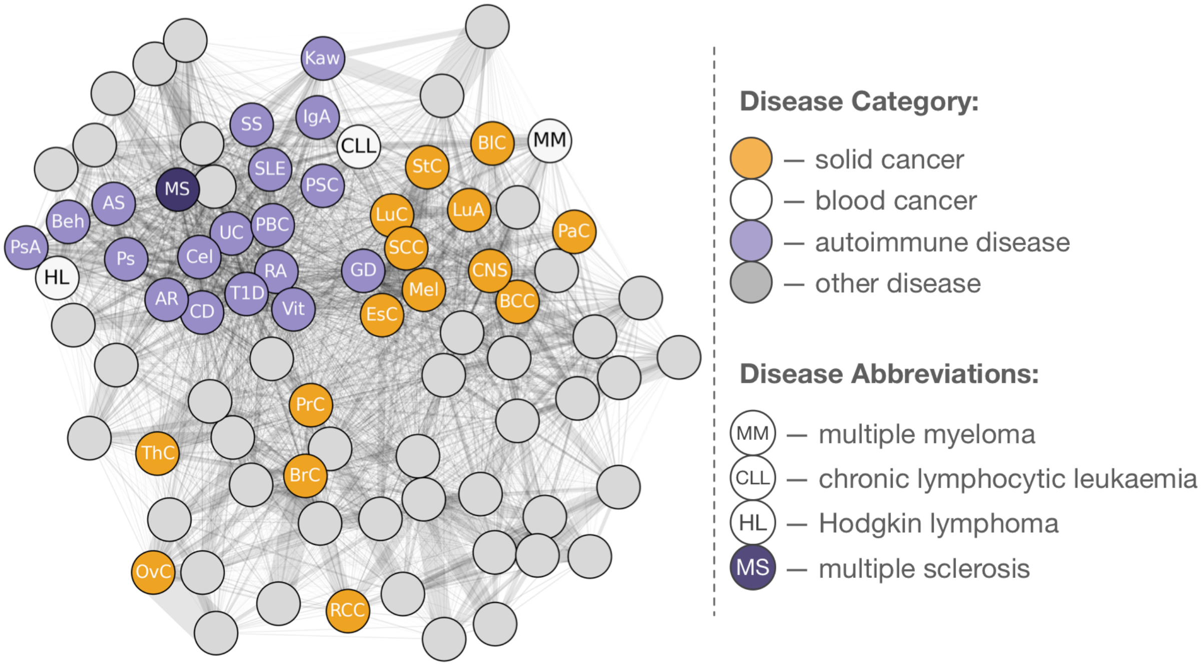

Figure 2.1. Disease proximity network from GWAS loci.

Autoimmune diseases form a distinct cluster. Solid cancers cluster as well but less cohesively. And the three blood cancers span from the solid cancer to autoimmune extremes. Multiple myeloma sits in solid cancer territory; chronic lymphocytic leukemia rests in between; and Hodgkin lymphoma resides with the autoimmune. We interpret these findings as evidence that Hodgkin lymphoma is special amongst cancers in that its genetics align primarily with autoimmune disease.

Why is this important? Complex human diseases, such as Hodgkin lymphoma and multiple sclerosis, are often poorly understood, complicating prevention and treatment efforts. We hypothesize that genetically similar diseases will share more than just genetics. As lead author Dr. Khankhanian explains, “genetic similarity between diseases may have clinical implications. Drugs that treat one disease may be repurposed to treat a genetically similar disease.” Likewise, two diseases with similar genetics may also share risk factors, as we see with Hodgkin lymphoma and multiple sclerosis.

2.3 Three examples of the new breed

Why do we consider our approach a *new breed* of disease network? Many early approaches, such as this prominent example [15], were exposed to two biases. First, the genetic profiles used to describe each disease relied on targeted studies of disease association. Such studies are biased by researchers' existing knowledge. GWAS offers a systematic, comprehensive, and hypothesis-free alternative. However, the early GWAS-based approaches suffered from a second bias. As Dr. Ben Voight — Assistant Professor of Systems Pharmacology and Translational Therapeutics at the University of Pennsylvania — explains, "for many loci we just don't know what the causal variant(s) are, and we certainly don't know the causal gene(s) linked to these variants." Approaches which require converting GWAS variants to genes introduce bias and potentially obscure signals.

Here, we investigate the new breed of approaches that avoid the two biases. The disease networks we mention below use only GWAS data and do not operate in gene space. Our method operates on loci rather than genes. To define loci, we identify a region around each lead variant uncovered by GWAS. The region boundaries are calculated by looking at the patterns of variation across the human genome. Farh et al. 2015 took a similar approach [16], which also used the GWAS Catalog loci (see their Figure 1a and the "Shared genetic loci" section).

Both Farh et al. and our approach faced the same hurdles. We both applied p-value filters to remove low-confidence associations. We both condensed multiple studies on the same disease. Additionally, some GWAS studies lack statistical power and should thus be discarded. Accordingly, Farh et al. excluded studies with fewer than 6 significant variants, while we excluded studies on fewer than 1000 individuals. Consequently, the Farh network is considerably smaller with just 39 diseases. However, both networks offer a genome-wide glimpse into the genetic similarities between complex disease.

Nonetheless, these methods are not without limitations. As Dr. Voight notes, oftentimes there may be “multiple associations at the same locus arising from different variants.” Our approaches interpret variant co-localization as shared genetic architecture, which is not always the case. Dr. Voight continues that even if two diseases associate with the same variant, “the risk allele for one disease may be protective for the other.” Bulik-Sullivan et al. 2015 sidestepped these concerns by analyzing trends in summary statistics across all variants [17]. The drawback is that genome-wide summary statistics are lamentably not always available. Hence, the Bulik-Sullivan analysis focused on only 24 traits with poor disease coverage. The study uncovered several cases where the genetic profiles of two diseases were anti-correlated (see red in Figure 2). These cases are particularly interesting as our method would overlook the opposing genetic nature of the two diseases.

In closing, GWAS has given rise to a new breed of disease similarity network. These networks offer unbiased insights into commonalities between diseases. Here we explored three approaches and their trade-offs. While we initially constructed the disease network to contextualize the similarity between Hodgkin lymphoma and multiple sclerosis, we created a general resource covering 82 diseases. And we’ve dedicated the code and data for our network, available on GitHub [18], to the public domain.

Chapter 3

The rise of the hetnet

The previous chapter discussed creating a network of disease similarity. The network was homogeneous: it had only one node type (diseases) and one relationship type (genetic similarity). However, the data extracted from the GWAS Catalogue contained several types of entities including SNPs, genes, diseases, and studies. For a simple visualization of genetic similarity between diseases, coercing these entities into a homonet made sense. However, Dr. Baranzini and I became interested in applications where projecting data onto a single axis would discard essential information.

The first application was an interactive browser offering a systems-level view of human complex traits. By displaying the different types of nodes and relationships, the goal was to allow researchers to explore whatever networks components interested them. In addition, we hoped to facilitate observations that span multiple domains to provide a more immersive systems experience. The user-facing product was a Cytoscape application called iCTNet2. I constructed the hetnet underlying the project but did not work on the application. This work is published in [19]:

> Wang L, **Himmelstein DS**, Santaniello A, Parvin M, Baranzini SE (2015) **iCTNet2: integrating heterogeneous biological interactions to understand complex traits**. *F1000Research*. DOI: 10.12688/f1000research.6836.2

iCTNet2 contained six different types of entities and nine different types of relationships. The entities (nodes) consisted of phenotypes, genes, miRNAs, tissues, drugs, and side effects. This project was a learning experience. I became acquainted with the challenges of biomedical data integration. Specifically, mapping entities between vocabularies and eliminating duplication was a major challenge which necessitated using controlled vocabularies and ontologies. At that time in 2012, tools for data manipulation, such as `dplyr` in R and `pandas` in Python, were still in their infancy. Each resource took between several days to several weeks to integrate.

Even after a dataset was processed and ready for inclusion into the hetnet, I struggled with storing and operating on the hetnet. What we needed was a data structure for storing hetnets and an API for manipulating and analyzing them. I evaluated the `networkx` package, but it had poor support for type, especially with respect to nodes. I also tried out several graph databases but found them unwieldy. Therefore I began development of `hetio`, a Python package for hetnets. Unlike most existing network software, which was built for homonets and may have had some hetnet support, `hetio` exclusively supports hetnets. Designing a framework for hetnets made me reflect on what properties hetnets should possess. Deviating from precedent, I decided to support both directed and undirected edges in the same hetnet. One potential use case would be to have an undirected interaction edge for when two genes produce proteins that bind together while having a directed transcription factor edge for when a gene's protein product binds to the promoter or enhancer region of another gene.

Development of `hetio` was initially driven by our study to prioritize disease-associated genes (Chapter 4). In this project, we extracted features from the hetnet that quantified the relationship between a specific gene and disease. I implemented the algorithm for extracting features, named degree-weighted path count ($DWPC$), in `hetio`. This algorithm must traverse the hetnet to find all paths of a specified type between two nodes, which can be computationally intensive. One drawback of our implementation was that the hetnet must be loaded entirely into memory before performing analysis. This limits network size by system RAM and led to several minute wait times to read our hetnet into Python. For our project to repurpose drugs (Chapter 5), I began exploring database alternatives.

While I had tried out the Neo4j graph database in 2012, we revisited the technology which had improved substantially. Specifically, a complete implementation of the Cypher query language for interacting with hetnets was released in 2013. Cypher is like SQL for hetnets, but several decades newer. We migrated to Neo4j for storing and operating on our hetnets [20]. We've been porting more and more functionality to Neo4j, such as network permutation [20]. We created an interactive GraphGist to exhibit our project and use of Neo4j. We submitted our GraphGist to the 2016 Neo4j Challenge, where it won the *Open/Government Data and Politics* category. The migration to Neo4j allowed us to tap into a larger hetnet ecosystem and focus our development more on applications. For this reason, I suspect the adoption of graph databases in bioinformatics, which has so far been limited [21], will see considerable gains in the next several years.

Once data has been integrated into a hetnet, new applications became manageable that would otherwise be too laborious. Our algorithm for relationship prediction is one example. For our drug repuporposing project, we built a predictive classifier that incorporates information from 24 types of relationships and captures connectivity spanning many types of relationships. After a new relationship type is added to a hetnet, no new implementation work is required to incorporate that information into the predictions. Another example of how our hetnets allow analysis at scale is [22]:

> Greene CS, **Himmelstein DS** (2016) **Association-guided analysis of gene networks to discover the genetic basis of complex traits**. *Circulation: Cardiovascular Genetics*. DOI: 10.1161/CIRCGENETICS.115.001181

For this review, I used a prerelease of Hetionet (our drug repurposing hetnet) to characterize a potential bias affecting network biology. Many network approaches convert from SNP-level to gene-level input as part of their analysis pipeline. Then the approaches construct a gene network and analyze the topology. Such approaches go awry when an unequal dispersion of SNPs creates spurious signals in the network. We found that genes with more SNPs tend to also be more connected in common networks (Figure 3.1). The correlation affected genotyping arrays as well as sequencing indicating that effects were not solely due to biased coverage of genotyping arrays. Physical protein interactions — a popular input for GWAS prioritization techniques — showed less correlation than other types. However, GO annotations — a community favorite for gene set enrichment techniques — increased sharply with SNP abundance. We conclude that the potential for erroneous conclusions when gene scores are biased by SNP abundance is high. Ideally, permutation testing should be applied on the SNP level to ensure that SNP to gene conversion biases are not the cause of any positive results. Since access to the raw SNP level data needed for permutation is often impractical or unavailable, care should be taken to use unbiased SNP to gene conversion methods.

While hetnets reduce the obstacles to analyzing heterogeneous information, they do require substantial integration efforts. The majority of the time spent on the studies in Chapters 4 and 5 was dedicated to processing, merging, and integrating different resources. Going forward I hope to spend more time analyzing rather than creating hetnets. As knowledge becomes more standardized, the integration burden should decrease. Another inhibiting factor to hetnet science has been fragmentation by discipline. A 2014 analysis identified 78 studies using multilayer networks (hetnets with an optional time dimension). However, these studies relied on 26 different terms to describe their data structures, 9 of which had multiple definitions [23]. We began an effort to standardize terminology across disciplines [20]. On January 26, 2016, we officially adopted the term hetnet. Hetnet is short for heterogeneous network or heterogeneous information network. We hope that a common nomenclature will help bring the study of hetnet together and assist cross-field interoperability.

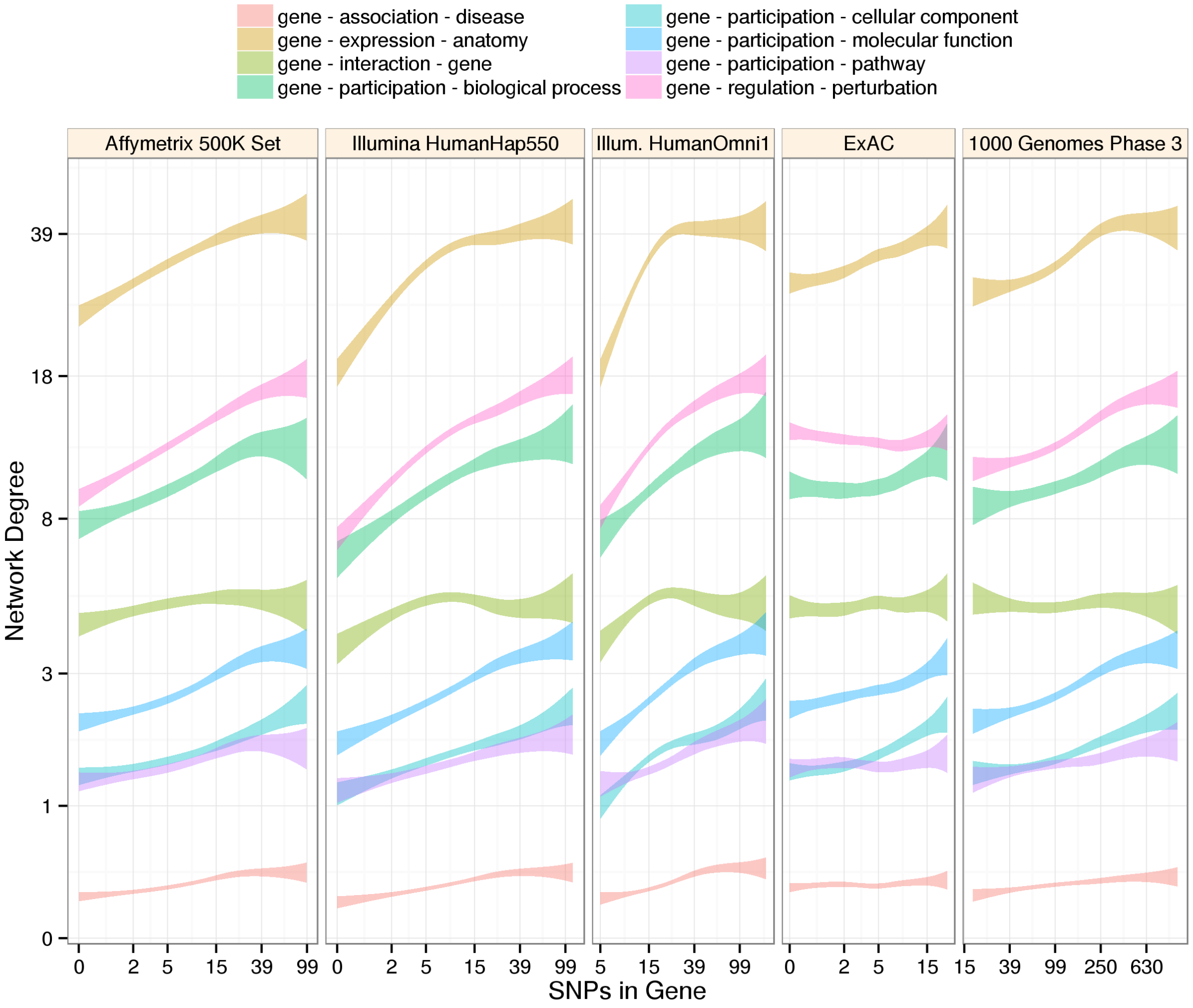

Figure 3.1. Genes with more SNPs tend to have higher network degree. The number of SNPs per gene was calculated for 3 genotyping arrays (Affymetrix 500K Set, Illumina HumanHap550, Illumina HumanOmni1), exome sequencing (ExAC), and whole genome sequencing (1000 Genomes Phase 3). The network degree (number of edges) for each gene was calculated on Rephetio, a network containing multiple types of nodes and edges. Models drawn as 95% confidence bands show the relationship between SNP abundance and network connectivity for 8 types of edges. For most edge types and platforms, genes with more measured SNPs tend to be more connected.

Chapter 4

Prioritizing disease-associated genes

This chapter contains our project to prioritize gene-disease associations. This research was the main focus of years 2–4 of my PhD. The study was covered by Kristin Sainani in her article Unlocking the Genetics of Complex Diseases: GWAS and Beyond for the *Biomedical Computation Review*. The content of this chapter is reprinted from [20]:

> **Himmelstein DS**, Baranzini SE (2015) **Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes**. *PLOS Computational Biology*. DOI: 10.1371/journal.pcbi.1004259

4.1 Abstract

The first decade of Genome Wide Association Studies (GWAS) has uncovered a wealth of disease-associated variants. Two important derivations will be the translation of this information into a multiscale understanding of pathogenic variants, and leveraging existing data to increase the power of existing and future studies through prioritization. We explore edge prediction on heterogeneous networks—graphs with multiple node and edge types—for accomplishing both tasks. First we constructed a network with 18 node types—genes, diseases, tissues, pathophysiologies, and 14 MSigDB (molecular signatures database) collections—and 19 edge types from high-throughput publicly-available resources. From this network composed of 40,343 nodes and 1,608,168 edges, we extracted features that describe the topology between specific genes and diseases. Next, we trained a model from GWAS associations and predicted the probability of association between each protein-coding gene and each of 29 well-studied complex diseases. The model, which achieved 132-fold enrichment in precision at 10% recall, outperformed any individual domain, highlighting the benefit of integrative approaches. We identified pleiotropy, transcriptional signatures of perturbations, pathways, and protein interactions as influential mechanisms explaining pathogenesis. Our method successfully predicted the results (with AUROC = 0.79) from a withheld multiple sclerosis (MS) GWAS despite starting with only 13 previously associated genes. Finally, we combined our network predictions with statistical evidence of association to propose four novel MS genes, three of which (*JAK2*, *REL*, *RUNX3*) validated on the masked GWAS. Furthermore, our predictions provide biological support highlighting *REL* as the causal gene within its gene-rich locus. Users can browse all predictions online (http://het.io). Heterogeneous network edge prediction effectively prioritized genetic associations and provides a powerful new approach for data integration across multiple domains.

4.2 Author Summary

For complex human diseases, identifying the genes harboring susceptibility variants has taken on medical importance. Disease-associated genes provide clues for elucidating disease etiology, predicting disease risk, and highlighting therapeutic targets. Here, we develop a method to predict whether a given gene and disease are associated. To capture the multitude of biological entities underlying pathogenesis, we constructed a heterogeneous network, containing multiple node and edge types. We built on a technique developed for social network analysis, which embraces disparate sources of data to make predictions from heterogeneous networks. Using the compendium of associations from genome-wide studies, we learned the influential mechanisms underlying pathogenesis. Our findings provide a novel perspective about the existence of pervasive pleiotropy across complex diseases. Furthermore, we suggest transcriptional signatures

of perturbations are an underutilized resource amongst prioritization approaches. For multiple sclerosis, we demonstrated our ability to prioritize future studies and discover novel susceptibility genes. Researchers can use these predictions to increase the statistical power of their studies, to suggest the causal genes from a set of candidates, or to generate evidence-based experimental hypothesis.

4.3 Introduction

In the last decade, genome-wide association studies (GWAS) have been established as the main strategy to map genetic susceptibility in dozens of complex diseases and phenotypes. Despite the success of this approach in mapping variation in thousands of loci to hundreds of complex phenotypes, researchers are now confronted with the challenge of maximizing the scientific contribution of existing GWAS datasets, whose undertakings represented a substantial investment of human and monetary resources from the community at large.

A central assumption in GWAS is that every region in the genome (and hence every gene) is *a-priori* equally likely to be associated with the phenotype in question. As a result, small effect sizes and multiple comparisons limit the pace of discovery. However, rational prioritization approaches may afford an increase in study power while avoiding the constraints and expense related to expanded sampling. One such a way forward is the current trend on analyzing the combined contribution of susceptibility variants in the context of biological pathways, rather than single SNPs [24]. For example, Yaspan et al described an approach that aggregates variants of interest from a GWAS into biological pathways using genomic randomization to control for multiple testing and minimize type I error [25]. The popular software PLINK also includes an option to evaluate groups of associations at the gene level, thus enabling pathway analysis by computing enriched gene sets [26]. A less explored but potentially revealing strategy is the integration of diverse sources of data to build more accurate and comprehensive models of disease susceptibility.

Several strategies have been attempted to identify the mechanisms underlying pathogenesis and use these insights to prioritize genes for genetic association analyses. Gene-set enrichment analyses identify prevalent biological functions amongst genes contained in disease-associated loci [27,28]. Gene network approaches search for neighborhoods of genes where disease-associated loci aggregate [29, 30]. Jia et al. reported dmGWAS, a strategy to integrate association signals from GWAS into the human protein interaction network [31]. A similar approach was developed by our group and tested in two large studies comprising more than 15,000 cases [32]. Literature mining techniques aim to chronicle the relatedness of genes to identify a subset of highly-related associated genes. For example, Raychaudhuri et al. reported the Gene Relationships Among Implicated Loci (GRAIL) algorithm, an approach to assess relationships among genomic disease regions by text mining of PubMed abstracts [33].

Prioritization strategies generally rely on user-provided loci as the sole input and do not incorporate broader disease-specific knowledge. Typically, the proportion of genome-wide significant discoveries in a given GWAS is low, thus leaving little high-confidence signal for seed-based approaches to build from. To overcome this limitation, here we aimed at characterizing the ability of various information domains to identify pathogenic variants across the entire compendium of complex disease associations. Using this multiscale approach, we developed a framework to prioritize both existing and future GWAS analyses and highlight candidate genes for further analysis.

To approach this problem, we resorted to a method that integrated diverse information domains naturally. Heterogeneous (or multipartite) networks are a class of networks which contain multiple types of entities (nodes) and relationships (edges or links), and provide a data structure capable of expressing diversity in an intuitive and scalable fashion. Most existing techniques available for network analysis have been developed for homogeneous networks [34–36] and are not directly extensible to heterogenous networks. Accordingly, the early network analyses in disease biology concentrated on homogeneous networks [37]. However in the last half-decade, the complexity of biological systems has spurred interest in heterogeneous approaches.

While still a developing field, network-based biological data integration has been pursued using a variety of techniques. Approaches such as GeneMANIA, weight and then project individual data sources onto a single dimension, enabling homogeneous network algorithms to be used to characterize the resulting graphs [38–40]. Other techniques operate on multi-relational (single node type, multiple edge type) networks, for example by taking into account relationships among local clusters and considering the full topology of weighted gene association networks [30,39,41]. Bipartite networks contain two node types and therefore work well for predicting relationships between entities of two different types (such as disease-gene associations or drug-disease indications) following a 'guilt-by-association' paradigm [42–44]. Other approaches incorporate greater-dimension heterogeneous networks as input but conflate types and, while improving predictions compared to simpler approaches, cannot effectively identify influential network components [45, 46]. Heterogeneous networks of arbitrary complexity have also been applied for edge prediction without a formalized feature extraction methodology, which requires manual descriptor determination for each new network design [47]. Recently, new types of edge prediction methods were reported that naturally accommodate any size heterogeneous network. These include data fusion by matrix-factorization [48–51] and metapath-based techniques [52,53]. This type of intermediate data fusion can treat all data sources directly (i.e. without transforming data into "disease space") and has been successfully used to infer disease similarities [50] and predict gene function in slime mold and baker's yeast [49]. A metapath-based approach was recently developed by researchers studying social sciences to predict future coauthorship [52] and provides an intuitive framework and interpretable models and results. An advantage of metapath-based approaches is that they preserve the network structure and provide the flexibility to explore a diverse set of descriptors. In this work, we extended this methodology to predict the probability that an association between a gene and disease exists.

4.4 Results

4.4.1 Constructing a heterogeneous network to integrate diverse information domains

Using publicly-available databases and standardized vocabularies, we constructed a heterogeneous network with 40,343 nodes and 1,608,168 edges (Fig. 4.1). Databases were selected based on quality, reusability, throughput, and their aggregate ability to provide a diversified, multiscale portrayal of biology. The network was designed to encode entities and relationships relevant to pathogenesis. The network contained 18 node types (metanodes) and 19 edge types (metaedges), displayed in Fig. 4.2A. Entities represented by metanodes consisted of diseases, genes, tissues, pathophysiologies, and gene sets for 14 MSigDB collections [54, 55] including pathways [56, 57], perturbation signatures, motifs [58, 59], and Gene Ontology (GO) domains [60] (Table 4.1). Relationships represented by metaedges consisted of gene-disease association, disease pathophysiology, disease localization, tissue-specific gene expression, protein interaction, and gene-set membership for each MSigDB collection (Table 4.2).

Gene-disease associations were extracted from the GWAS Catalog [67] by overlapping associations into disease-specific loci. Loci were classified as low or high-confidence based on p-value and sample size of the corresponding GWAS. When possible, for each loci, the most-commonly reported gene across studies was designated as primary and subsequently considered responsible for the association. Additional genes reported for the loci were considered secondary. Only high-confidence primary associations were included in the network yielding 938 associations between 99 diseases and 711 genes.

4.4.2 Features quantify the network topology between a gene and disease

To describe the network topology connecting a specific gene and disease, we computed 24 features, each describing a different aspect of connectivity. Each feature corresponds to a type of path (metapath) originating in a given source gene and terminating in a given target dis-

Metanode	**Count**	**Source**	**References**
Disease	99	Disease Ontology	[61]
Gene	19,116	HGNC (coding)	[62]
Tissue	77	BRENDA (BTO)	[63]
Pathophysiology	8	manual	–
Positional	326	MSigDB (C1)	[62]
Perturbation	3,402	MSigDB (C2)	[54, 55]
BioCarta	217	MSigDB (C2)	–
KEGG	186	MSigDB (C2)	[56]
Reactome	674	MSigDB (C2)	[57]
miRNA Target	221	MSigDB (C3)	[58]
TF Target	615	MSigDB (C3)	[58, 59]
Cancer Hood	427	MSigDB (C4)	[64]
Cancer Module	431	MSigDB (C4)	[65]
GO Process	825	MSigDB (C5)	[60]
GO Component	233	MSigDB (C5)	[60]
GO Function	396	MSigDB (C5)	[60]
Oncogenic	189	MSigDB (C6)	[66]
Immunologic	1,910	MSigDB (C7)	[66]

Table 4.1. Metanodes. The kind, number of corresponding nodes, and data source for each type of node.

ease. The biological interpretation of a feature derives from its metapath, and features simply quantify the prevalence of a specific metapath between any gene-disease pair. To quantify metapath prevalence, we adapted an existing method originally developed for social network analysis (*PathPredict*) [52], and developed a new metric called degree-weighted path count (*DWPC*, Fig. 4.2D), which we employed in all but two features. The *DWPC* downweights paths through high-degree nodes when computing metapath prevalence. The strength of downweighting depends on a single parameter (w), which we optimized to $w = 0.4$ and that outperformed the top metric resulting from *PathPredict* [52]. We calculated *DWPC* features for the 22 metapaths of length 3 or less that originated with a gene and terminated with disease. Two non-*DWPC* features were included to assess the pleiotropy of the source gene and the polygenicity of the target disease. Referred to as 'path count' features, they respectively equal the number of diseases associated with the source gene and the number of genes associated with the target disease. For all features, paths with duplicate nodes were excluded, and, if present, the association edge

Metaedge	**Count**	**Source**	**References**
Disease - association - Gene	938	GWAS Catalog	[67]
Disease - membership - Pathophysiology	90	manual	–
Disease - localization - Tissue	1,086	CoPub 5.0	[68]
Gene - expression - Tissue	251,366	GNF BodyMap	[69]
Gene - interaction - Gene	97,938	iRefIndex	[70]
Gene - membership - Positional	18,343	MSigDB (C1)	[62]
Gene - membership - Perturbation	366,211	MSigDB (C2)	[54, 55]
Gene - membership - BioCarta	4,456	MSigDB (C2)	–
Gene - membership - KEGG	12,656	MSigDB (C2)	[56]
Gene - membership - Reactome	35,597	MSigDB (C2)	[57]
Gene - membership - miRNA Target	33,455	MSigDB (C3)	[58]
Gene - membership - TF Target	161,258	MSigDB (C3)	[58, 59]
Gene - membership - Cancer Hood	41,913	MSigDB (C4)	[64]
Gene - membership - Cancer Module	48,220	MSigDB (C4)	[65]
Gene - membership - GO Process	75,155	MSigDB (C5)	[60]
Gene - membership - GO Component	34,880	MSigDB (C5)	[60]
Gene - membership - GO Function	23,578	MSigDB (C5)	[60]
Gene - membership - Oncogenic	30,166	MSigDB (C6)	[66]
Gene - membership - Immunologic	370,862	MSigDB (C7)	[66]

Table 4.2. Metaedges. The kind, number of corresponding edges, and data source for each type of edge.

between the source gene and target disease was masked.

4.4.3 Machine learning approach to predict the probability of association of gene-disease pairs

Further analysis focused on the 29 diseases with at least ten associated genes (Table 4.3). The 698 high-confidence primary associations of these 29 diseases were considered positives—gene-disease pairs with positive experimental relationships. The remaining 551,823 (i.e. unassociated) gene-disease pairs were considered negatives. Low-confidence or secondary associations were excluded from either set. We partitioned gene-disease pairs into training (75%) and testing (25%) sets and created a training network with the testing associations removed.

To learn the importance of each feature and model the probability of association of a given gene-disease pair, we used regularized logistic regression which is designed to prevent overfit-

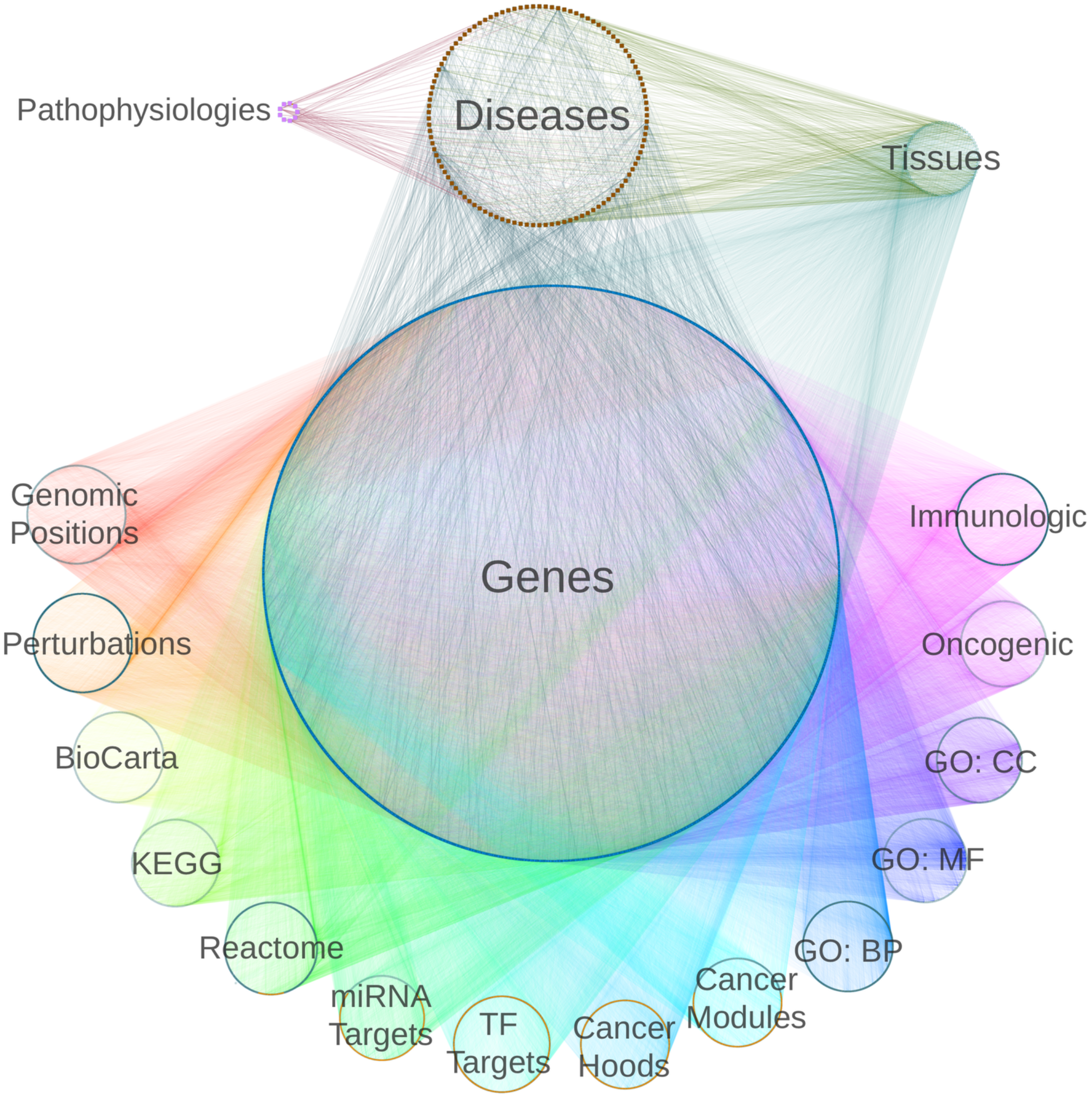

Figure 4.1. Heterogeneous network integrates diverse information domains. We constructed a heterogeneous network with 18 metanodes (denoted with labels) and 19 metaedges (denoted by color). For each metanode, nodes are laid out circularly. Incorporating type information adds structure to a network which would otherwise appear as an undecipherable agglomeration of 40,343 nodes and 1,608,168 edges.

ting and accurately estimate regression coefficients when models include many features. Elastic net regression is a regression method that balances two regularization techniques: ridge (which performs coefficient shrinkage) and lasso (which performs coefficient shrinkage and variable se-

Disease	Pathophysiology	HC-P	HC-S	LC-P	LC-S
Crohn's disease	immunologic	67	179	4	2
multiple sclerosis	immunologic	50	43	38	29
type 2 diabetes mellitus	immunologic	49	49	20	15
breast carcinoma	neoplastic	43	65	2	6
ulcerative colitis	immunologic	40	96	2	3
prostate carcinoma	neoplastic	34	202	3	4
type 1 diabetes mellitus	immunologic	33	56	9	6
rheumatoid arthritis	immunologic	30	27	20	11
coronary artery disease	metabolic	29	43	15	9
obesity	metabolic	28	22	34	18
celiac disease	immunologic	24	32	9	8
systemic lupus erythematosus	immunologic	22	35	14	8
refractive error	degenerative	21	11	2	1
primary biliary cirrhosis	immunologic	20	16	2	0
vitiligo	immunologic	20	27	4	0
age related macular degeneration	degenerative	18	30	11	18
metabolic syndrome X	metabolic	17	11	1	0
asthma	immunologic	17	23	13	4
psoriasis	immunologic	16	14	5	5
schizophrenia	psychiatric	15	27	20	13
chronic lymphocytic leukemia	neoplastic	14	16	3	4
migraine	unspecific	13	15	38	58
Alzheimer's disease	degenerative	12	11	27	18
Graves' disease	immunologic	12	15	1	1
Parkinson's disease	degenerative	12	21	8	13
atopic dermatitis	immunologic	11	15	5	1
bipolar disorder	psychiatric	11	34	26	74
lung carcinoma	neoplastic	10	14	6	6
ankylosing spondylitis	immunologic	10	5	6	6

Table 4.3. Diseases. Associations were predicted for 29 diseases with at least 10 positives. For these diseases, the number of high-confidence primary (HC-P), high-confidence secondary (HC-S), low-confidence primary (LC-P), and low-confidence secondary associations (LC-S) that were extracted from the GWAS Catalog is indicated.

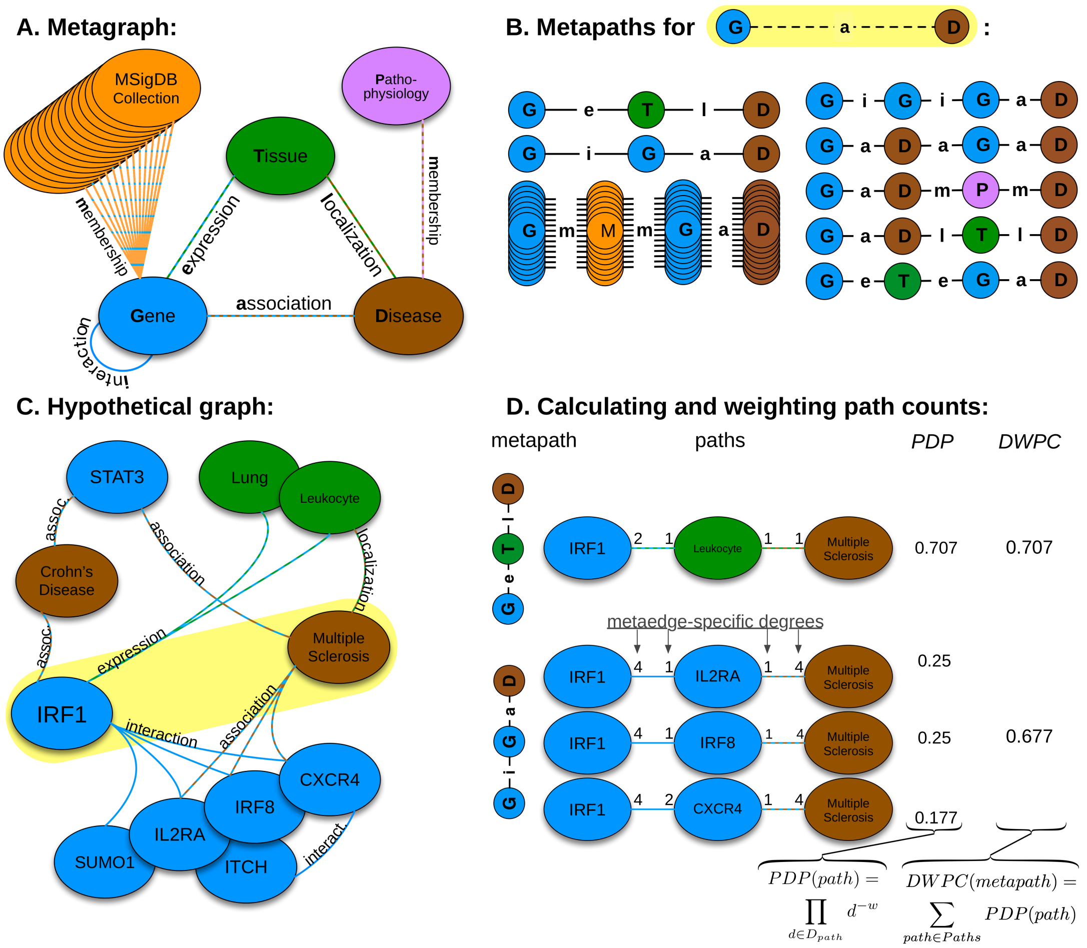

Figure 4.2. Heterogeneous network edge prediction methodology. A) We constructed the network according to a schema, called a metagraph, which is composed of metanodes (node types) and metaedges (edge types). B) The network topology connecting a gene and disease node is measured along metapaths (types of paths). Starting on Gene and ending on Disease, all metapaths length three or less are computed by traversing the metagraph. C) A hypothetical graph subset showing select nodes and edges surrounding *IRF1* and multiple sclerosis. To characterize this relationship, features are computed that measure the prevalence of a specific metapath between *IRF1* and multiple sclerosis. D) Two features (for the *GeTlD* and *GiGaD* metapaths) are calculated to describe the relationship between *IRF1* and multiple sclerosis. The metric underlying the features is degree-weighted path count (*DWPC*). First, for the specified metapath, all paths are extracted from the network. Next, each path receives a path-degree product (*PDP*) measuring its specificity (calculated from node-degrees along the path, Dpath). This step requires a damping exponent (here $w = 0.5$), which adjusts how severely high-degree paths are downweighted. Finally, the path-degree products are summed to produce the *DWPC*.

lection) [71]. On the training set, we optimized the elastic net mixing parameter, a single parameter behind the *DWPC* metric, and two edge-inclusion thresholds. While cross-validated performance was similar across elastic net mixing parameters, ridge demonstrated the greatest consistency, and thus we proceeded with logistic ridge regression as the primary model for predictions.

4.4.4 Method prioritizes associations withheld for testing

We extracted network-based features for gene-disease pairs from the training network and modeled the training set. We next evaluated performance on the 25% of gene-disease pairs (175 positives, 137,956 negatives) withheld for testing. Our predictions achieved an area under the ROC curve (AUROC) of 0.83 (Fig. 4.3A) demonstrating an excellent performance in retrieving hidden associations. Importantly, we did not observe any significant degradation of performance from training to testing (Fig. 4.3A), indicating that our disciplined regularization approach avoided overfitting and that predictions for associations included in the network were not biased by their presence in the network. Furthermore, we observed that at 10% recall (the classification threshold where 10% of true positives were predicted as positives), our predictions achieved 16.7% precision (the proportion of predicted positives that were correct). Since the prevalence of positives in our dataset was 0.13%, the observed precision represents a 132-fold enrichment over the expected probability under a uniform distribution of priors (as in GWAS).

4.4.5 Predicting associations on the complete network

As a next step in our analysis, we recomputed features on the complete network, which now included the previously withheld testing associations. On all positives and negatives, we fit a ridge model (the primary model for predictions) and a lasso model (for comparison). Standardized coefficients (Fig. 4.4) indicate the effect attributed to each feature by the models. The lasso highlighted features that captured pleiotropy (4 features), pathways (2), transcriptional signatures of perturbations (1) and protein interactions (1). Despite the parsimony of the lasso,

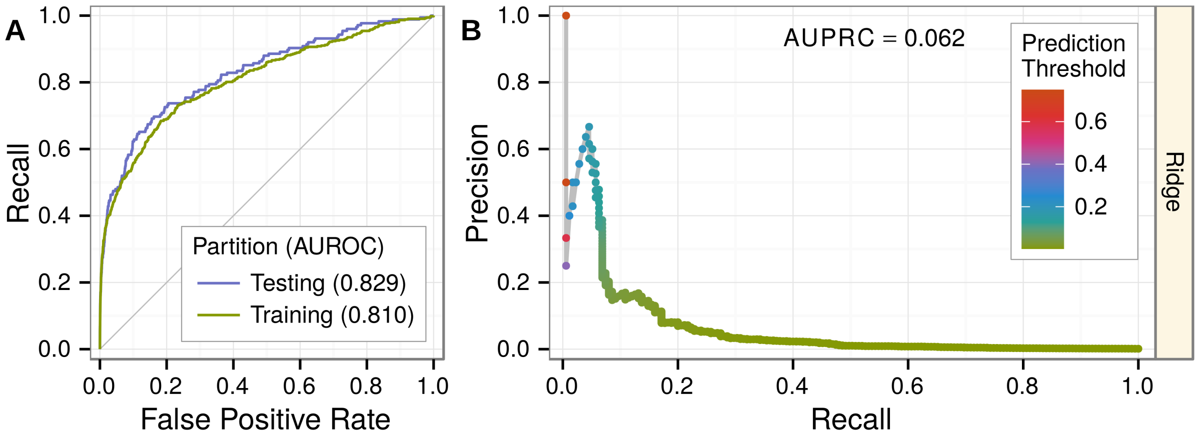

Figure 4.3. Predicting associations withheld for testing. Performance was evaluated on 25% of gene-disease pairs withheld for testing. A) Testing and training ROC curves. At top prediction thresholds, associated gene-disease pairs are recalled at a much higher rate than unassociated pairs are incorrectly classified as positives. The testing area under the curve (AUROC) is slightly greater than the training AUROC, demonstrating the method's lack of overfitting. Performance greatly exceeds random denoted by gray line. B) The precision-recall curve showing performance in the context of the low prevalence of associated gene-disease pairs (0.13%). Nevertheless, at top prediction thresholds, a high percentage of pairs classified as positives are truly associated. Prediction thresholds, shown as points and colored by value, align with the observed precision at that threshold.

performance was similar between models with training AUROCs of 0.83 (ridge) and 0.82 (lasso). However, since multiple features from a correlated group may be causal, the lasso model risks oversimplifying. Ridge regression disperses an effect across a correlated group of features, providing users greater flexibility when interpreting predictions. From the ridge model, we predicted the probability that each protein-coding gene was associated with each analyzed disease and built a webapp to display the predictions (http://het.io/disease-genes/browse).

4.4.6 Degree-preserving network permutations highlight the importance of edge-specificity for top predictions and ten features

Using Markov chain randomized edge-swaps, we created 5 permuted networks. Since metaedge-specific node degree is preserved, features extracted from the permuted network retain unspecific

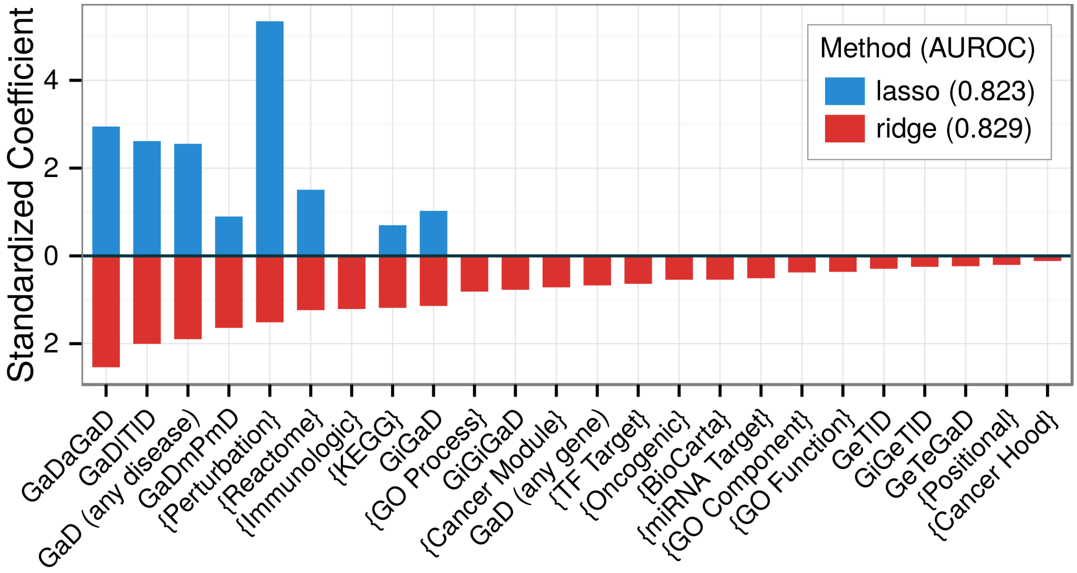

Figure 4.4. Feature selection identifies a parsimonious yet predictive model. Ridge and lasso models were fit from the complete network. The resulting standardized coefficients (y-axis) assess the effect size of each feature (x-axis). Brackets indicate features from MSigDB-traversing metapaths ($Gm\{\}mGaD$). The ridge model disperses effects amongst features whereas the lasso concentrates effects. The lasso identifies an 8-feature model with minimal performance loss compared to the ridge model. Besides *KEGG*, gene-set based features were largely captured by *Perturbations*. The lasso retains several measures of pleiotropy as well as the one-step interactome feature ($GiGaD$).

effects. These effects include general measures a disease's polygenicity and a gene's pleiotropy, multifunctionality, and tissue-specificity. On the first permuted network, we partitioned associations into training and testing sets. Testing associations were masked from the network, features were computed, and a ridge model was fit on the training gene-disease pairs.

Compared to the unpermuted-network model, testing performance was noticeably inferior: the AUROC declined from 0.83 (Fig. 4.3A) to 0.79 and the AUPRC (area under the precision-recall curve) declined from 0.06 (Fig. 4.3B) to 0.02. We interpret the modest decline in AUROC but marked reduction in AUPRC as a direct consequence of the permutation's particularly detrimental effect on top predictions. In other words, edge-specificity was crucial for top predictions, while general effects gleaned from node degree performed reasonably well when ranking the entire spectrum of protein-coding genes for association. A commonly-overlooked finding is that the discriminatory ability of gene networks largely relies on node-degree rather than the edge-specificity [72]. However, we found that for top predictions—which are the only predictions considered by many applications—edge-specificity was critical.

Interestingly, predictions from the permuted-network model displayed a reduced dynamic range with none exceeding 4%, while predictions from the unpermuted-network model exceeded 75%. Therefore, even though they achieve reasonable AUROC, the permuted-network predictions would have little utility as prior probabilities in a bayesian analysis where dynamic range is crucial. Furthermore, the signal present in permuted-network features was greatly diminished: few features survived the lasso's selection resulting in an average lasso AUROC of 0.70 versus 0.80 for ridge. Permuting the network significantly reduced the predictiveness of features based on pleiotropy (2 features), protein interactions (2), transcriptional signatures of perturbations (1), tissue-specificity (1), pathways (3), and immunologic signatures (1). Six of the eight features selected by the lasso and eight of the top ten ridge features (ranked by standardized coefficients) were negatively affected by the permutation. Since our modeling technique preferentially selected/weighted features affected by permutation, we can infer that network components where edge-specificity matters underlie a large portion of predictions.

4.4.7 Feature importance identifies the mechanisms underlying associations

We assessed the informativeness of each feature by calculating feature-specific AUROCs. Feature-specific AUROCs universally exceeded 0.5, indicating that network connectivity, regardless of type, positively discriminates associations. However, performance varied widely by feature and within feature from disease to disease (Fig. 4.5). Top performing domains consisted of transcriptional signatures of perturbations (AUROC = 0.74), immunologic signatures (0.70), and pleiotropy (0.68, 0.67, 0.64, 0.63). Notably, the models greatly outperformed any individual feature, highlighting the importance of an integrative approach.

Features whose metapaths originate with an association (GaD) metaedge measure pleiotropy.

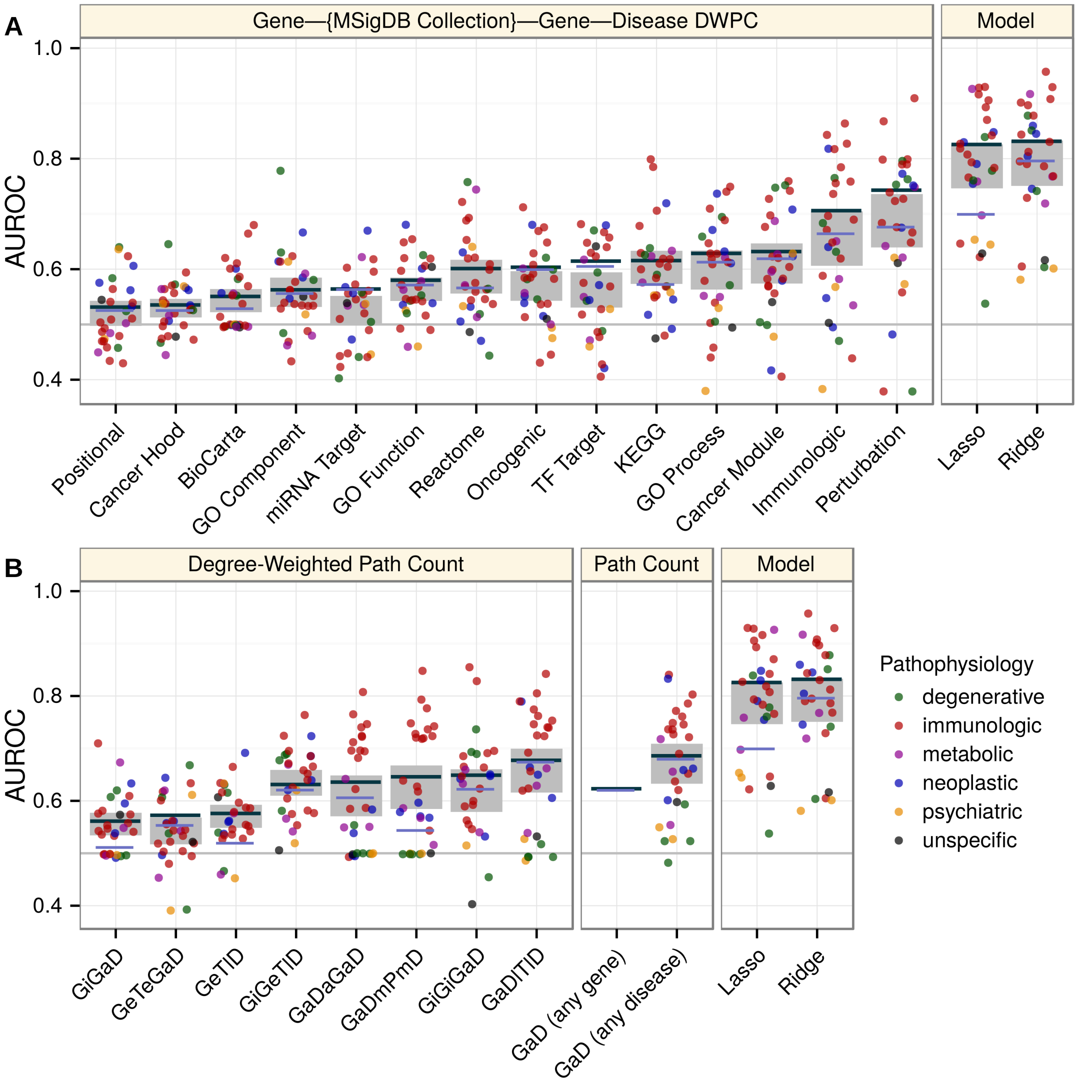

Figure 4.5. Decomposing performance shows the superiority of the integrative model and compares individual features. Disease, feature, and model-specific performance on the complete network. The AUROC (y-axis) was calculated for each classifier (x-axis). In addition to the ridge and lasso models (rightmost panels), each feature was considered as a classifier. Line segments show the classifier's global performance (average performance across permuted networks shown in violet as opposed to dark gray). Points indicate disease-specific performance and are colored by the disease's pathophysiology. Grey rectangles show the 95% confidence interval for mean disease-specific performance. A) Features from metapaths that traverse an MSigDB collection. B) Features from non-MSigDB-traversing metapaths. Metapaths are abbreviated using first letters of metanodes (uppercase) and metaedges (lowercase).

The four pleiotropic features were among the top performing features that did not rely on set-based gene categorization (Fig. 4.5). Of the four features, *GaD (any disease)* had the highest AUROC despite its lack of disease-specificity, reflecting both the sparsity of disease-specific features and the existence of genetic overlap between seemingly disparate diseases. *GaDmPmD* and *GaDaGaD* performed best for immunologic diseases and were affected by permutation, indicating that genetic overlap was greatest between immunologic diseases. On the other hand, the performance of *GaDlTlD* did not decrease after permutation indicating disease colocalization was not a primary driver of genetic overlap.

We also observed that the lasso regression model discarded the majority of features with a minimal performance deficit, suggesting redundancy among features. Indeed, pairwise feature correlations showed moderate collinearity among features. Collinearity was especially pervasive with respect to the *Perturbations* feature, explaining its threefold increase in standardized coefficient in the lasso versus ridge model. The disappearance of all but one other MSigDB-based feature in the lasso model indicated that *Perturbations*—the feature traversing chemical and genetic transcriptional signatures of perturbations—exhausted meaningful gene-set characterization. In other words, the faulty molecular processes behind pathogenesis align with and are encapsulated by the processes perturbed by chemical and genetic modifications. The *Immunologic signatures* feature—traversing gene-sets characterizing "cell types, states, and perturbations within the immune system"—was highly predictive and correlated with *Perturbations*. As expected this feature performed best for diseases with an immune pathophysiology. The one well-performing neoplastic disease (Fig. 4.5) was chronic lymphocytic leukemia, a hematologic cancer with a strong immune component [73]. Additionally, the performance of both the *Perturbation* and *Immunologic* features was affected by permutation indicating information beyond the extent of a gene's multifunctionality was encoded.

Existing network-based gene-prioritization methods, frequently rely solely on protein-protein interactions. Our results supported incorporating protein interactions as the two interactome-based features were discriminatory (AUROCs = 0.65, 0.56) and affected by permutation. However, when compared to the integrative models or other top-performing features, performance of features that relied solely on the interactome was severely limited. Pathways, another founding resource for many approaches, proved important with *KEGG* selected by the lasso and all three pathway resources (AUROCs = 0.61 for *KEGG*, 0.60 for *Reactome*, 0.55 for *BioCarta*) affected by permutation. The *GeTlD* feature—measuring to what extent a gene is expressed in tissues affected by the disease in question—peaked in performance around AUROC = 0.58, was affected by permutation, and required no preexisting knowledge of associated genes. In other words, while approaches based on tissue-specificity may have limited predictive ability on their own, they are broadly applicable (i.e. less susceptible to knowledge bias) and provide orthogonal information that could enhance the overall performance of a model.

Gene set robustness. For each type of gene set, we evaluated the effect of increased sparsity on performance by randomly subsampling gene set nodes or edges and measuring the resulting AUROC of the affected feature. Robustness refers to a gene set collection's ability to withstand a high extent of masking with little performance deficit. Several of the top gene sets had this property, especially GO processes (where supersets are common), which may indicate nodal redundancy. Contrastingly, the MSigDB gene set with the fewest nodes, KEGG, experienced a more immediate and linear decline in performance. Since KEGG avoids duplication and is stringently and manually curated, this finding is expected. To investigate whether the high predictivity of certain gene set collections was due only to size, we compared performance when subsampling nodes to the KEGG level. The two top performing collections, perturbations and immunologic signatures, which also happen to be large, continued to perform better than the majority of complete collections. While performance benefited from increasing densities, a resource's sparsity often reflects an intrinsic property of the underlying information type. Therefore, when identifying influential mechanisms of pathogenesis, we prefered unadjusted comparisons using the complete network.

4.4.8 Case study: prioritizing multiple sclerosis associations

The WTCCC2 multiple sclerosis (MS) GWAS tested 465,434 SNPs for 9,772 cases and 17,376 controls and identified over 50 independently associated loci [74]. Since the GWAS Catalog excludes targeted arrays (such as ImmunoChip), this study remains the largest MS GWAS in the Catalog. To evaluate our method's ability to prioritize associations identified in a future study, we masked the WTCCC2 MS study from the GWAS Catalog and created a pre-WTCCC2 network. The number of high-confidence primary MS associations was thus reduced from 50 to 13, with the 37 novel genes identified by WTCCC2 available to evaluate performance. On the pre-WTCCC2 network, we extracted features, fit a ridge model, and predicted each gene's probability of association with MS. Amongst all 18,993 potentially novel genes, the 37 WTCCC2 genes were ranked highly (AUROC = 0.79, Fig. 4.6).

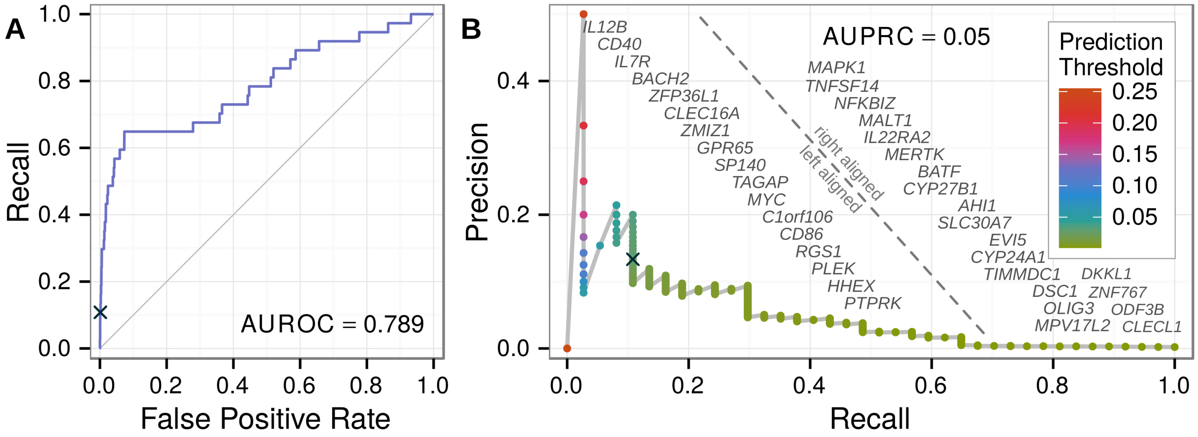

Figure 4.6. Prioritizing multiple sclerosis associations identified by a masked GWAS. From a network with the WTCCC2 MS associations omitted, we predicted probabilities of association for all potentially novel genes. The 37 novel genes identified by the WTCCC2 GWAS were considered positives, and the resulting performance was plotted. The ROC (A) and precision-recall (B) curves show performance, with AUCs in line with the testing performance across all diseases (Fig.˜4.3). A prediction threshold (black cross) that resulted in high performance was selected as the discovery threshold for further analysis. As the classification threshold decreases along the precision-recall curve, the advent of each true positive is denoted by its gene symbol.

4.4.9 Prioritizing statistical candidates with network-based predictions identifies novel multiple sclerosis genes

Finally, we designed a framework for discovering and validating novel MS genes that incorporates our network-based predictions. Meta2.5 is a meta-analysis of all MS GWAS prior to the WTCCC2 study [75]. We calculated genewise p-values for Meta2.5 using VEGAS [76] and observed a large enrichment in nominally significant ($p < 0.05$) genes, suggesting multiple potential associations. We combined this set of experimental candidates with the top predictions from the pre-WTCCC2 network to discover genes with both strong statistical and biological evidence of association. To ensure novelty, we excluded genes from GWAS-established MS loci and the extended MHC region. We chose a threshold for network-based predictions that performed well in prioritizing the genes identified by WTCCC2 (Fig. 4.6).

This strategy discovered four genes, three of which—*JAK2*, *REL*, *RUNX3*—achieved Bonferroni validation on VEGAS-converted WTCCC2 p-values (Table 4.4). The probability of the observed validation rate occurring under random prioritization is 0.01, demonstrating that incorporating our network-based predictions as a prior increased study power. *JAK2* displays overexpression in MS-affected Th17 cells [77] and was implicated in an interactome-based prioritization of GWAS [32]. *RUNX3*, a transcription factor influencing T lymphocyte development, has been associated with celiac disease [78] and ankylosing spondylitis [79] and was hypermethylated in systemic lupus erythematosus patients [80]. The region containing *REL* was uncovered in a recent MS ImmunoChip-based study with 14,498 cases [81, p. S40]. For the gene-dense region containing *REL*, the ImmunoChip study reported a long non-coding RNA, *LINC01185*, overlapping the lead-SNP, rs842639. However, since greater than 80% of the genome shows evidence of transcription [82], the probability of incidental overlap with long non-coding RNA is high. *REL*, however, is an essential transcription factor for lymphocyte development [83] and plays a critical role in autoimmune inflammation [84]. Hence, gene prioritization through integrative analyses offers not only to streamline loci discovery but also subsequent causal gene identification.

Gene	**Meta2.5**	**HNEP**	**WTCCC2**
JAK2	0.047	0.102	**0.0015**
REL	0.001	0.040	**0.0003**
SH2B3	0.012	0.034	0.0130
RUNX3	0.016	0.025	**0.0073**

Table 4.4. Multiple sclerosis gene discovery. Four genes showed nominal statistical evidence of association (Meta2.5 column) and exceeded the network prediction threshold (HNEP column). Three genes achieved Bonferroni validation (bold) in an independent GWAS (WTCCC2 column).

4.5 Discussion

In this work, we developed a framework to predict the probability that each protein-coding gene is associated with each of 29 complex diseases. Our predictions draw on a diverse set of pathogenically-relevant relationships encoded in a heterogeneous network. The predictions successfully prioritized associations hidden from the network. Using MS as a representative example, we were able to combine our predictions with statistical evidence of association to increase study power and identify three novel susceptibility genes in this disease. The disease-specific performance (measured by the AUROC) for MS was exceeded by twelve other diseases suggesting that our predictions have broad applicability for prioritizing genetic association analyses. Prioritization can range from a genome-wide scale to a single loci where this approach can highlight the causal gene from several candidates within the same association block. For researchers focused on a specific disease, these predictions can be used to propose genes for experimental investigation. Inversely, researchers focused on a specific gene can use this resource to find suggestions for relevant complex disease phenotypes.

Most previous explorations of the factors underlying pathogenicity have focused on a single domain such as tissue-specificity [85], protein interactions [15], pathways [24], or disease similarity [86]. The method presented here integrates disparate data sources, learns their importance, and unifies them under a common framework enabling comparison. Therefore, we can conclude that perturbation gene sets—the core of our top-performing feature—are an underutilized resource for disease-associated gene prioritization. Not only did perturbations encompass other set-based gene categorizations, but they greatly outperformed features based on protein interactions, pathways, and tissue-specificity, which form the basis of several prominent prioritization techniques. In addition to characterizing the overall importance of each feature, our online prediction browser visually decomposes an individual prediction into its components.

We observed a prominent influence of pleiotropy, consistent with previous studies that identified pervasive overlap of susceptibility loci across complex diseases [87], especially those of autoimmune nature [88]. Since many existing prioritization techniques are agnostic to the compendia of GWAS associations, they fail to adequately leverage pleiotropy. Unlike approaches initiated from a user-provided gene list, our study only provides predictions for 29 diseases. By not relying on user-provided input, our predictions can serve as independent priors for future analyses. By predicting probabilities, we provide an extensible and interpretable assessment of association that circumvents the limitations inherent to frequentist analyses [89]. Many approaches return no assessment for the majority of genes which fall outside of their set of predicted positives. Here, we overcome this issue and provide a comprehensive and genome-wide output by returning a probability of association for each protein-coding gene.

High-throughput biological data is frequently noisy and incomplete [90]. Combining orthogonal resources can help overcome these issues. Accordingly, we found that our integrative model outperformed any individual domain. While this method has shown encouraging performance, some limitations are worth noticing. For example, many biological networks preferentially cover well-studied vicinities [91]. Knowledge biases that span multiple, presumably-orthogonal resources could diminish the benefits of integration. Here, several of the literature-derived domains were removed by the lasso, suggesting redundancy. In addition, biases in network completeness can lead to high-quality predictions for well-studied vicinities and low-quality predictions for poorly-studied vicinities. The permutation analysis provided evidence of this disparity: edge-specificity was critical for top predictions yet only moderately beneficial for the remainder. Subsequently, we caution users to avoid overinterpreting predictions for poorly-characterized genes. To help place predictions in context, the online browser provides a gene's mean prediction across all diseases and a disease's mean prediction across all genes. However, we recognize that false negatives will continue to persist in our predictions, and users should be mindful of this limitation when interpreting results. As more systematic and unbiased resources become available [90], high-quality predictions will emerge for more network vicinities.

We reason that the desirable qualities of our predictions are the consequence of the heterogenous network edge prediction methodology. The approach is versatile (most biological phenomena are decomposable into entities connected by relationships), scalable (no theoretical limit to metagraph complexity or graph size), and efficient (low marginal cost to including an additional network component). We have extended the previous metapath-based framework set forth by *PathPredict* [52], by: 1) incorporating regularization allowing coefficient estimation for more features without overfitting; 2) designing a framework for predicting a metaedge that is included in the network; 3) developing an improved metric for assessing path specificity; and 4) implementing a degree-preserving permutation. Metapath-based heterogeneous network edge prediction provides a powerful new platform for bioinformatic discovery.

4.6 Methods

4.6.1 Ethics Statement

This study was approved by the UCSF institutional review board on human subjects under protocol #10-00104.

4.6.2 Heterogeneous networks

We created a general framework and open source software package for representing heterogeneous networks. Like traditional graphs, heterogeneous networks consist of nodes connected by edges, except that an additional meta layer defines type. Node type signifies the kind of entity encoded, whereas edge type signifies the kind of relationship encoded. Edge types are comprised of a source node type, target node type, kind (to differentiate between multiple edge types connecting the same node types), and direction (allowing for both directed and undirected edge types). The user defines these types and annotates each node and edge, upon creation, with its corresponding type. The meta layer itself can be represented as a graph consisting of node types connected by edge types. When referring to this graph of types, we use the prefix 'meta'. Metagraphs—called schemas in previous work [52, 53]—consist of metanodes connected by metaedges. In a heterogeneous network, each path, a series of edges with common intermediary nodes, corresponds to a metapath representing the type of path. A path's metapath is the series of metaedges corresponding to that path's edges. The possible metapaths within a heterogeneous network can be enumerated by traversing the metagraph. We implemented this framework as an object-oriented data structure in python and named the resulting package *hetio*. Users are free to browse, use, or contribute to the software, through the online repository (https://github.com/dhimmel/hetio).

4.6.3 Network construction

Resource selection. The included resources, and hence the metaedges and metanodes composing our network, were selected empirically based on a balance among the following properties: 1) **quality** – *relevance to human pathogenesis; high accuracy and an optimal trade-off between false positives and false negatives.* In some cases, quality concerns prevented the inclusion of a desired metaedge. For example, we omitted ontology-based disease similarly due to an inaccurate Disease Ontology hierarchy [61], and we omitted disease comorbidity due to high measurement error for uncommon diseases [92]. For included metaedges, we attempted to select the highest quality resource in that domain. 2) **reusability** – *easily retrievable and parsable; mapped to controlled vocabularies; well documented; amenable to reproducible (scripted) analysis; free of prohibitive reuse stipulations.* 3) **throughput** – *broad domain-specific coverage generated using systematic platforms that minimize bias.* While genetic interactions have previously proven informative [50], their sparse characterization in humans was deemed unfavorable for our approach. 4) **diversified, multiscale portrayal of biology** – *capturing, in aggregate, many aspects of* *pathophysiology across multiple levels of biological complexity.* Levels of the hierarchical architecture of biological complexity include the genome, transcriptome, proteome, interactome, metabolome, cell and tissue organization, and phenome. Balancing these considerations, we integrated as many resources as possible within our computational runtime constraints.

Nodes. Protein-coding genes were extracted from the HGNC database [62]. Resources were mapped to HGNC terms via gene symbol (ambiguous symbols were resolved in the order: approved, previous, synonyms) or Entrez identifiers. Disease nodes were taken from the Disease Ontology (DO) [61]. Due to the limited number of diseases with GWAS, relevant disease references were manually mapped to the DO. Tissues were taken from the BRENDA Tissue Ontology (BTO) [63]. Only tissues with profiled expression were included enabling manual mapping. Nodes for the 14 MSigDB metanodes were directly imported from the Molecular Signature Database version 4.0 [54, 55]. All MSigDB collections were included except those that were supersets of other collections. For example, 'C3: motif gene sets' was the union of two disjoint collections ('C3: microRNA targets' and 'C3: transcription factor targets') and was therefore excluded. Diseases were classified manually into 10 categories according to pathophysiology. The 'idiopathic' and 'unspecific' categories were not included as pathophysiology nodes, since they do not signify meaningful similarities between member diseases.

Associations. Disease-gene associations were extracted from the GWAS Catalog [67], a compilation of GWAS associations where $p < 10^{-5}$. First, associations were segregated by disease. GWAS Catalog phenotypes were converted to Experimental Factor Ontology (EFO) terms using mappings produced by the European Bioinformatics Institute. Associations mapping to multiple EFO terms were excluded to eliminate cross-phenotype studies. We manually mapped EFO to DO terms (now included in the DO as cross-references) and annotated each DO term with its associations.

Associations were classified as either high or low-confidence, where exceeding two thresholds granted high-confidence status. First, $p \leq 5 \times 10^{-8}$ corresponding to $p \leq 0.05$ after Bonferroni adjustment for one million comparisons (an approximate upper bound for the number of independent SNPs evaluated by most GWAS). Second, a minimum sample size (counting both cases and controls) of 1,000 was required, since studies below this size are underpowered [93]—i.e. any discovered associations are more likely than not to be false—for the majority of true effect size distributions commonly assumed to underlie complex disease etiology [89].

Lead-SNP were assigned windows—regions wherein the causal SNPs are assumed to lie—retrieved from the DAPPLE server [29]. Windows were calculated for each lead-SNP by finding the furthest upstream and downstream SNPs where $r^2 > 0.5$ and extending outwards to the next recombination hotspot. Associations were ordered by confidence, sorting on following criteria: high/low confidence, p-value (low to high), and recency. In order of confidence, associations were overlapped by their windows into disease-specific loci. By organizing associations into loci, associations from multiple studies tagging the same underlying signal were condensed. A locus was classified as high-confidence if any of its composite associations were high-confidence and low-confidence otherwise.

For each disease-specific loci, we attempted to identify a primary gene. The primary gene was resolved in the following order: 1) the mode author-reported gene; 2) the containing gene for an intragenic lead-SNP; 3) the mode author-reported gene for an intragenic lead-SNP (in the case of overlapping genes); 4) the mode author-reported gene of the most proximal up and downstream genes. Steps 2–4 were repeated on each association composing the loci, in order of confidence, until a single gene resolved as primary. Loci where ambiguity was unresolvable or where no genes were returned did not receive a primary gene. All non-primary genes—genes that were author-reported, overlapping the lead-SNP, or immediately up or downstream from the lead-SNP—were considered secondary.

Accordingly, four categories of processed associations were created: high-confidence primary, high-confidence secondary, low-confidence primary, and low-confidence secondary. We assume that our primary gene annotation for each loci represents the single causal gene responsible for the association. To investigate the validity of this assumption, we evaluated the performance of our predictions separately using each category of association as positives. For both confidence levels, primary associations outperformed secondary associations suggesting our method succeeded at categorizing causal genes as primary. However, for high-confidence secondary associations, the AUROC equaled 0.74, which could result from multiple causal genes per loci or categorizing sole causal genes as secondary. The performance decline from high to low confidence associations was severe, pointing to a preponderance of falsely identified loci in the GWAS Catalog when $p > 5 \times 10^{-8}$ or sample size drops below 1000.

Protein interactions. Physical protein-protein interactions were extracted from iRefIndex 12.0, a compilation of 15 primary interaction databases [70]. The iRefIndex was processed with ppiTrim to convert proteins to genes, remove protein complexes, and condense duplicated entries [94].

Tissue-specific gene expression. Tissue-specific gene expression levels were extracted from the GNF Gene Expression Atlas [69]. Starting with the GCRMA-normalized and multisample-averaged expression values, 44,775 probes were converted to 16,466 HGNC genes and 84 tissues were manually mapped and converted to 77 BTO terms. For both conversions, the geometric mean was used to average expression values. The log base 10 of expression value was used as the threshold criteria for GeT edge inclusion.

Disease localization. Disease localization was calculated for the 77 tissues with expression profiles. Literature co-occurrence was used to assess whether a tissue is affected by a disease. We used CoPub 5.0 to extract R-scaled scores between tissues and diseases measuring whether two terms occurred together in Medline abstracts more than would be expected by chance [68]. DO terms for diseases with GWAS and BTO tissues with expression profiles were manually mapped to the ‘biological identifier’ terminology used by CoPub. The R-scaled score was used as the threshold criteria for *TlD* edge inclusion.

Feature computation metrics. The simplest metapath-based metric is path count (PC): the number of paths, of a specified metapath, between a source and target node. However, PC does not adjust for the extent of graph connectivity along the path. Paths traversing high-degree nodes will account for a large portion of the PC, despite high-degree nodes frequently representing a biologically broad or vague entity with little informativeness. The previous work evaluated several metrics that include a PC denominator to adjust for connectivity and reported that normalized path count (NPC) performed best [52]. The denominator for NPC equals the number of paths from the source to any target plus the number of paths from any target to the source.

$$NPC_m(s,t) = \frac{PC_m(s,t)}{\sum\limits_{t_i \in T_m} PC_m(s,t_i) + \sum\limits_{s_i \in Sm} PC_m(s_i,t)},$$

where m is the metapath, s is the source node, t is the target node, S_m is the set of nodes corresponding to the source metanode of m, and T_m is the set of nodes corresponding to the target metanode of m. We adopt the any source/target concept to compute the two GaD features. However, dividing the PC by a denominator is flawed because each path composing the PC deserves a distinct degree adjustment. If two paths—one traversing only high-degree nodes and one traversing only low-degree nodes—compose the PC, the network surrounding the high-degree path will monopolize the NPC denominator and overwhelm the contribution of the low-degree path despite its specificity. Therefore, we developed the degree-weighted path count ($DWPC$) which individually downweights each path between a source and target node. Each path receives a path-degree product (PDP) calculated by: 1) extracting all metaedge-specific degrees along the path (D_{path}), where each edge composing the path contributes two degrees; 2) raising each degree to the $-w$ power, where $w \geq 0$ and is called the damping exponent; 3) multiplying all exponentiated degrees to yield the PDP.

$$PDP(path) = \prod_{d \in D_{path}} d^{-w}$$

The *DWPC* equals the sum of *PDPs*.

$$DWPC_m(s,t) = \sum_{path \in Paths_m(s,t)} PDP(path)$$

See Fig. 4.2C–D for a visual description of the *DWPC*.

4.6.4 Machine learning approach

PathPredict relied on basic logistic regression to predict coauthorship status from features corresponding to nine distinct metapaths [52]. However, faced with fewer positives to train our model and a large number of features, we adopted a regularized approach, which aims to contain the overfitting tendencies inherent to regression. Regularization penalizes complexity, a trademark of overfitting. We chose the elastic net technique of regularization [71], which is efficiently implemented for logistic regression by the R *glmnet* package [95].

Regularized logistic regression requires a parameter, λ, setting the strength of regularization. We optimized λ separately for each model fit. Using 10-fold cross-validation and the "one-standard-error" rule to choose the optimal λ from deviance, we adopted a conservative approach designed to prevent overfitting [95].

On the training set of gene-disease pairs, we optimized the elastic net mixing parameter (α), the *DWPC* damping exponent (w), and two edge inclusion thresholds. First, we optimized α and w on the 20 features whose metapaths did not include threshold-dependent metaedges. For each combination of α and w, we calculated average testing AUROC using 20-fold cross-validation repeated for 10 randomized partitionings. After setting α and w, we jointly optimized the two edge-inclusion thresholds using the AUROC for the *GeTlD* feature, whose metapath is composed from the two edges requiring thresholds.

We adopt standardized coefficients as a measure of feature effect size. Standardized coefficients refer to the coefficients from logistic regression when all features have been transformed to z-scores. Standardization provides a common scale to assess feature effect, both within and across models [96].

4.6.5 Degree-preserving permutation

Starting from the complete network, a permuted network was created by swapping edges separately for each metaedge. Edge swaps were performed by switching the target nodes for two randomly selecting edges. For each metaedge, the number of attempted swaps was ten times the corresponding edge count. We adopted a Markov Chain strategy where additional rounds of permutation were initiated from the most-recently permuted network. A training network was generated from the first permuted network by masking 25% of the associations for testing. When contrasting this performance with the unpermuted-network model, we employed the Condensed-ROC curve to magnify the importance of top predictions [97]. Using the exponential transformation with a magnification factor of 460—the value which maps a FPR of 0.01 to 0.99—we concentrated on the top 1% of predictions. A one-sided unpaired DeLong test [98] was used to assess whether feature-specific AUROCs from the complete network exceeded those from the first permuted network.

4.6.6 Gene Set Subsampling

We performed a subsampling analysis for 15 gene sets—the 14 MSigDB gene sets and tissues—to assess the effect of sparsity on feature-specific performance. Two without-replacement subsampling schemes were investigated: node masking and edge masking. For a specific gene set and scheme, we masked a percentage of the gene set and calculated the corresponding feature's AUROC. We evaluated a range of percentages and performed ten subsampling repetitions for each percentage.

4.6.7 Multiple sclerosis gene discovery

We excluded 588 genes from the discovery phase of the multiple sclerosis analysis. First we excluded genes in the extended MHC region (spanning from *SCGN* to *SYNGAP1* on chromosome 6 [99]) due to the complex pattern of linkage characterizing this region containing several highly-penetrant MS-risk alleles [74]. Second, we excluded putative MS genes: high-confidence primary genes from the GWAS Catalog and reported genes for the WTCCC2-replicated loci. We omitted genes in linkage disequilibrium with the putative genes by excluding: 1) consecutive sequences of nominally significant genes (using the WTCCC2-VEGAS p-values) that included a putative gene; and 2) high-confidence secondary genes from the GWAS catalog. Post exclusion, 1211 genes were nominally significant in Meta2.5, four of which exceeded the network-based discovery threshold. Using a hypergeometric test for overrepresentation, we calculated the probability of randomly selecting 4 of the 1211 genes and Bonferroni validating at least 3 of the 4 on WTCCC2.

Chapter 5

Identifying drug repurposing candidates

This chapter contains our project to predict drug repurposing. I proposed this project for my qualifying exam. However, I decided to pursue disease-associated gene prediction first. For the final project of my PhD, I returned to the project. This chapter is reprinted from the preprint:

> **Himmelstein DS**, Lizee A, Khankhanian P, Brueggeman L, Chen SL, Hadley D, Hessler CS, Green AJ, Baranzini SE (2016) **Rephetio: Repurposing drugs on a hetnet**. *Thinklab*.

5.1 Abstract

This study describes Project Rephetio – a systematic investigation of drug efficacy. We constructed Hetionet v1.0, an integrative network for drug repurposing. Hetionet consists of 47,031 nodes of 11 types and 2,250,197 relationships of 24 types. Data was integrated from 29 public resources to connect compounds, diseases, genes, anatomies, pathways, biological processes, molecular functions, cellular components, perturbations, pharmacologic classes, drug side effects, and disease symptoms. In the process, we created *PharmacotherapyDB* – a physician-curated

catalog of medical indications, which differentiates between disease-modifying and symptomatic therapy. We used the 755 disease-modifying treatments to ground our analysis and enable a systematic inspection of pharmacology. First, we identified network patterns that were predictive of treatment. Then we predicted the probability of treatment for 209,168 compound–disease pairs. Our predictions performed well in two external validations, suggesting that our predictions will help prioritize drug repurposing candidates. Project Rephetio was open notebook and included contributions from 35 community members who provided feedback in realtime.

5.2 Introduction

The cost of developing a new therapeutic drug has been estimated at 1.4 billion dollars [100], the process typically takes 15 years from lead compound to market [101], and the likelihood of success is stunningly low [102]. Strikingly, the costs have been doubling every 9 years since 1970 [103]. Drug repurposing – identifying novel uses for existing therapeutics – can drastically reduce the duration, failure rates, and costs of approval [104]. These benefits stem from the rich preexisting information on approved drugs, including extensive toxicology profiling performed during development, clinical trials, and postmarketing surveillance.

Drug repurposing is poised to become more efficient as mining of electronic medical records (EMRs) to retrospectively assess the effect of drugs gains feasibility [19, 105–107]. However, systematic approaches to repurpose drugs based on mining EMRs alone will likely lack power due to multiple testing. Similar to the approach followed to increase the power of genome-wide association studies (GWAS) [89, 93], integration of biological knowledge to prioritize drug repurposing will help overcome limited EMR sample size and data quality.

In addition to repurposing, several other paradigm shifts in drug development have been proposed to improve efficiency. Since small molecules tend to bind to many targets, polypharmacology aims to find synergy in the multiple effects of a drug [108]. Network pharmacology assumes diseases consist of a multitude of molecular corruptions resulting in a robust disease state. Network pharmacology seeks to uncover multiple points of intervention into a specific pathophysiology that together rehabilitate an otherwise recalcitrant disease state [109, 110]. Although target-centric drug discovery has dominated the field for decades, phenotypic screens have more recently resulted in a comparatively higher number of first-in-class small molecules [111]. Recent technological advances have enabled a new paradigm in which mid- to high-throughput assessment of intermediate phenotypes, such as the molecular response to drugs, is replacing the classic target discovery approach [112–114]. Modern computational approaches offer a convenient platform to tie these developments together as the reduced cost and increased velocity of *in silico* experimentation massively lowers the barriers to entry and price of failure [115, 116].

Hetnets (short for heterogeneous networks) are networks with multiple types of nodes and relationships. They offer an intuitive, versatile and powerful structure for data integration. In this study, we developed a heterogeneous network (Hetionet v1.0) to prioritize drug indications and facilitate their repurposing. Specifically, we integrated knowledge and experimental findings from decades of biomedical study spanning millions of publications. We adapted an algorithm originally developed for social network analysis and applied it to the network to identify patterns of efficacy and predict new uses for drugs. The algorithm performs hetnet edge prediction through a machine learning framework that accommodates the multitude of types in a hetnet [52, 117]. Our approach represents an *in silico* implementation of network pharmacology that natively incorporates polypharmacology and high-throughput phenotypic screening.

One fundamental characteristic of our method is that it learns and evaluates itself on existing medical indications. Here we'll introduce previous approaches that also performed comprehensive evaluation on existing treatments. A 2011 study, named PREDICT, compiled 1,933 treatments between 593 drugs and 313 diseases [118]. Starting from the premise that similar drugs treat similar diseases, PREDICT trained a classifier that incorporates 5 types of drug-drug and 2 types of disease-disease similarity. A 2014 study compiled 890 treatments between 152 drugs and 145 diseases with transcriptional signatures [119]. The authors found that compounds triggering an opposing transcriptional response to the disease were more likely to be treatments, although this effect was weak and limited to cancers. A 2016 study compiled 402 treatments between 238 drugs and 78 diseases and used a single proximity score – the average shortest path distance between a drug's targets and disease's associated proteins on the interactome – as a classifier [120].

We build on these successes by creating a framework for incorporating the effects of any biological relationship into the prediction of whether a drug treats a disease. Thus, we're able to capture a multitude of effects that have been suggested as influential for drug repurposing including drug-drug similarity [46, 118], disease-disease similarity [118, 121], transcriptional signatures [113,114,119,122,123], protein interactions [120], genetic association [124,125], drug side effects [126, 127], disease symptoms [128], and molecular pathways [129]. Our ability to create such an integrative model of drug efficacy relies on the hetnet data structure to unite diverse information. On the hetnet, our algorithm learns which types of compound–disease paths discriminate treatments from non-treatments in order to predict the probability that a compound treats a disease.

5.3 Results

5.3.1 Hetionet v1.0

We obtained and integrated data from 29 publicly-available resources to create Hetionet v1.0 (Figure 5.1). The hetnet contains 47,031 nodes of 11 types (Table 5.1) and 2,250,197 relationships of 24 types (Table 5.2). The nodes consist of 1,552 small molecule compounds and 137 complex diseases, as well as genes, anatomies, pathways, biological processes, molecular functions, cellular components, perturbations, pharmacologic classes, drug side effects, and disease symptoms. The edges represent relationships between these nodes and encompass the collective knowledge produced by millions of studies over the last half century.

For example, *Compound–binds–Gene* edges represent when a compound binds to a protein encoded by a gene. This information has been extracted from the literature by human curators and compiled into databases such as DrugBank, ChEMBL, DrugCentral, and Bind-

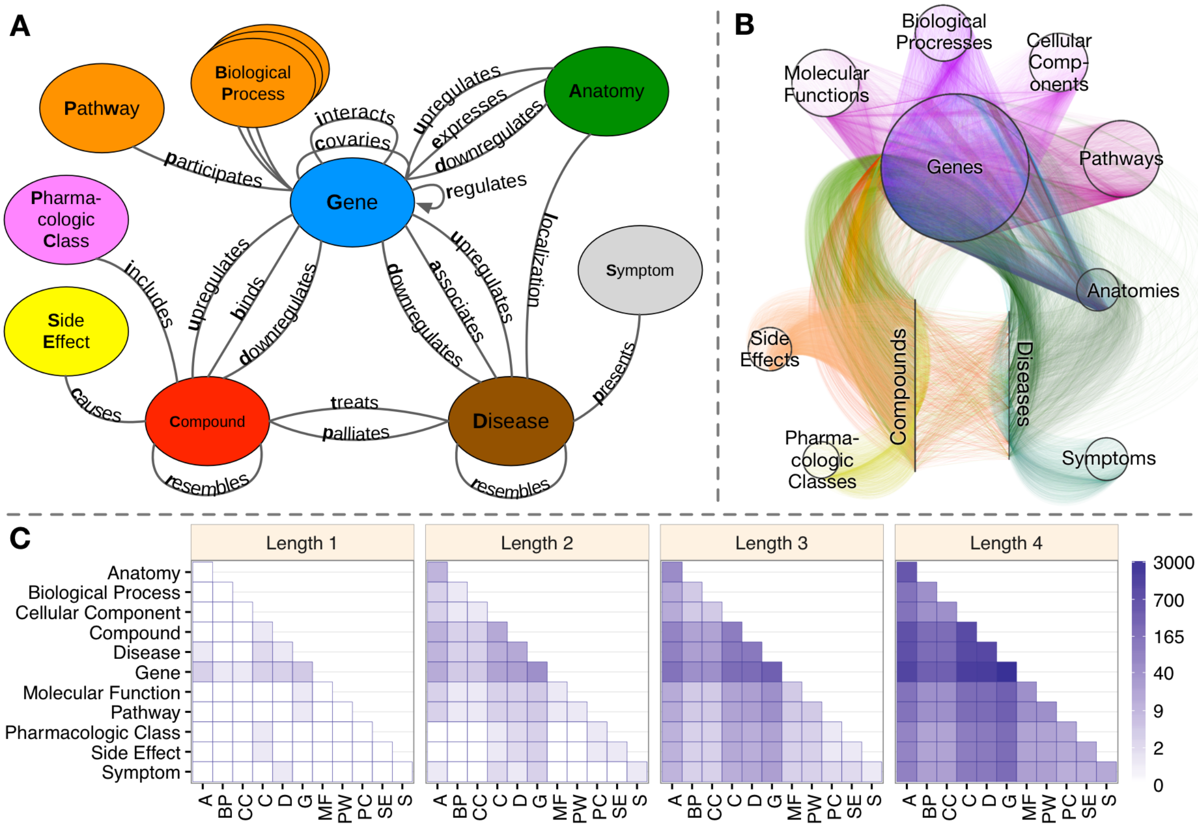

Figure 5.1. Hetionet v1.0 A) The metagraph, a schema of the network types. B) The hetnet visualized. Nodes are drawn as dots and laid out orbitally, thus forming circles. Edges are colored by type. C) Metapath counts by path length. The number of different types of paths of a given length that connect two node types is shown. For example, the top-right tile in the Length 1 panel denotes that Anatomy nodes are not connected to themselves (i.e. no edges connect nodes of this type between themselves). However, the bottom-left tile of the Length 4 panel denotes that 88 types of length-four paths connect Symptom to Anatomy nodes.

ingDB. We combined these databases to create 11,571 binding edges between 1,389 compounds and 1,689 genes. These edges were compiled from 10,646 distinct publications, which Hetionet binding edges reference as an attribute. Binding edges represent a comprehensive catalog constructed from low throughput experimentation. However, we also integrated findings from high throughput technologies – many of which have only recently become available. For example, we generated consensus transcriptional signatures for compounds in LINCS L1000 and diseases in STARGEO.

Metanode	Abbr	Nodes	Disconnected	Metaedges
Anatomy	A	402	2	4
Biological Process	BP	11,381	0	1
Cellular Component	CC	1,391	0	1
Compound	C	1,552	14	8
Disease	D	137	1	8
Gene	G	20,945	1,800	16
Molecular Function	MF	2,884	0	1
Pathway	PW	1,822	0	1
Pharmacologic Class	PC	345	0	1
Side Effect	SE	5,734	33	1
Symptom	S	438	23	1

Table 5.1. Metanodes. Hetionet v1.0 includes 11 node types (metanodes). For each metanode, this table shows the abbreviation, number of nodes, number of nodes without any edges, and the number of metaedges connecting the metanode.

While Hetionet v1.0 is ideally suited for drug repurposing, the network has broader biological applicability. Among the 11 metanodes, there are 66 possible source–target pairs. However, only 11 of them have at least one direct connection. In contrast, for paths of length 2, 50 pairs have connectivity (paths types that start on the source node type and end on the target node type, see Figure 5.1C). At length 3, all 66 pairs are connected. At length 4, the source–target pair with the fewest types of connectivity (Side Effect to Symptom) has 13 metapaths, while the pair with the most connectivity types (Gene to Gene) has 3,542 pairs. This high level of connectivity across a diversity of biomedical entities forms the foundation for automated translation of knowledge into biomedical insight.

Hetionet v1.0 is available online in JSON, Neo4j, and TSV formats. The JSON and Neo4j database formats include node and edge properties – such as URLs, source and license information, and confidence scores – and are thus recommended. In addition, a read-only Neo4j Browser is available at http://neo4j.het.io, providing users an installation-free method to query and visualize the network.

Metaedge	**Abbr**	**Edges**	**Sources**	**Targets**
Anatomy–downregulates–Gene	AdG	102,240	36	15,097
Anatomy–expresses–Gene	AeG	526,407	241	18,094
Anatomy–upregulates–Gene	AuG	97,848	36	15,929
Compound–binds–Gene	CbG	11,571	1,389	1,689
Compound–causes–Side Effect	CcSE	138,944	1,071	5,701
Compound–downregulates–Gene	CdG	21,102	734	2,880
Compound–palliates–Disease	CpD	390	221	50
Compound–resembles–Compound	CrC	6,486	1,042	1,054
Compound–treats–Disease	CtD	755	387	77
Compound–upregulates–Gene	CuG	18,756	703	3,247
Disease–associates–Gene	DaG	12,623	134	5,392
Disease–downregulates–Gene	DdG	7,623	44	5,745
Disease–localizes–Anatomy	DlA	3,602	133	398
Disease–presents–Symptom	DpS	3,357	133	415
Disease–resembles–Disease	DrD	543	112	106
Disease–upregulates–Gene	DuG	7,731	44	5,630
Gene–covaries–Gene	GcG	61,690	9,043	9,532
Gene–interacts–Gene	GiG	147,164	9,526	14,084
Gene–participates–Biological Process	GpBP	559,504	14,772	11,381
Gene–participates–Cellular Component	GpCC	73,566	10,580	1,391
Gene–participates–Molecular Function	GpMF	97,222	13,063	2,884
Gene–participates–Pathway	GpPW	84,372	8,979	1,822
Gene–regulates–Gene	GrG	265,672	4,634	7,048
Pharmacologic Class–includes–Compound	PCiC	1,029	345	724

Table 5.2. Metaedges. Hetionet v1.0 contains 24 edge types (metaedges). For each metaedge, the table reports the abbreviation, the number of edges, the number of source nodes connected by the edges, and the number of target nodes connected by the edges.

5.3.2 Systematic mechanisms of efficacy

One aim of Project Rephetio was to systematically evaluate why drugs work. To address this question, we created a gold standard of 755 disease-modifying indications, which form the *Compound–treats–Disease* edges in Hetionet v1.0. Next, we identified types of paths (metapaths) that occurred more frequently between treatments than non-treatments (any compound–disease pair that is not a treatment). The advantage of this approach is that metapaths naturally correspond to mechanisms of pharmacological efficacy. For example, the Compound–binds–Gene–associates–Disease (*CbGaD*) metapath identifies when a drug binds to a protein corresponding to a gene involved in the disease.

We evaluated all 1,206 metapaths that go from compound to disease and have length of 2–4 (Figure 5.2A). To control the different degrees of nodes, we used the degree-weighted path count ($DWPC$) – which downweights paths through high degree nodes [117] – to assess path prevalence. In addition, we compared the performance of each metapath to a baseline computed from permuted networks. Hetnet permutation preserves node degree while eliminating edge specificity, allowing us to isolate the portion of unpermuted metapath performance resulting from actual network paths. We refer to the permutation-adjusted performance measure as Δ AUROC.

709 of the 1,206 metapaths exhibited a statistically significant Δ AUROC at a false discovery rate cutoff of 5%. These 709 metapaths included all 24 metaedges, suggesting that each type of relationship we integrated had some pharmacological utility. However, not all metaedges were equally present in significant metapaths: 259 significant metapaths included a *Compound–binds–Gene* metaedge, whereas only 4 included a *Gene–participates–Cellular Component* metaedge. Table 5.3 provides the predictiveness of several interesting metapaths. Refer to the Discussion for our interpretation of these findings.

5.3.3 Predictions of drug efficacy

We implemented a machine learning approach to translate the network connectivity between a compound and disease into a probability of treatment. The approach relies on the 755 treatments as positives and 29,044 non-treatments as negatives to train a logistic regression model. The features consisted of a prior probability of treatment, node degrees for 14 metaedges, and DWPCs for 123 metapaths that were well suited for modeling. A cross-validated elastic net was used to prevent overfitting, yielding a model with 31 features (Figure 5.2B). The DWPC features with negative coefficients appear to be included as node-degree-capturing covariates. However, the 11 DWPC features with non-negligible positive coefficients embody the most salient types of connectivity for systematically modeling pharmacology.

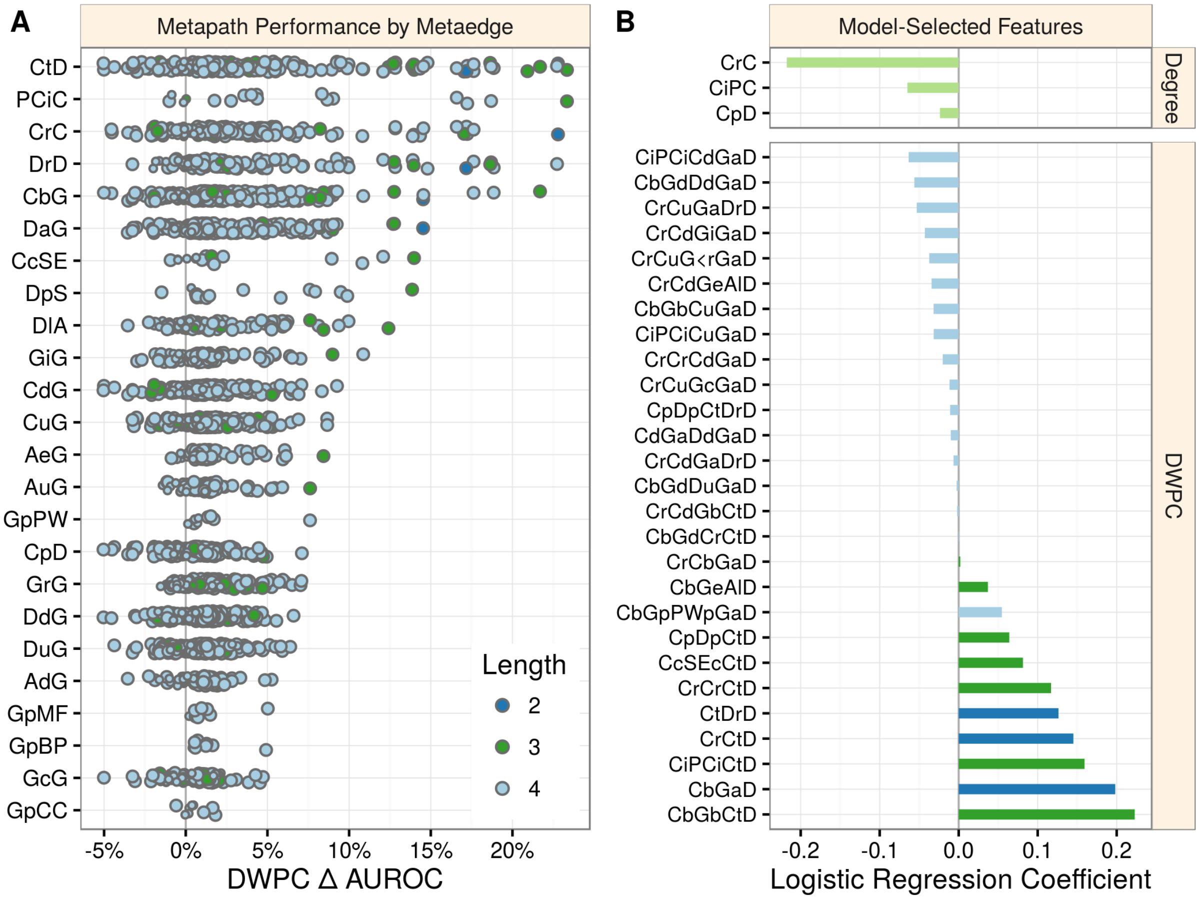

Figure 5.2. Performance by type and model coefficients. A) The performance of the DWPCs for 1,206 metapaths, organized by their composing metaedges. The larger dots represent metapaths that were significantly affected by permutation (false discovery rate $< 5\%$). Metaedges are ordered by their best performing metapath. Since a metapath's performance is limited by its least informative metaedge, the best performing metapath for a metaedge provides a lower bound on the pharmacologic utility of a given domain of information. B) Barplot of the model coefficients. Features were standardized prior to model fitting to make the coefficients comparable [130].

These 11 features correspond to the following types of connectivity: whether the compound binds to the same genes as compounds which treat the disease ($CbGbCtD$); whether the compound binds to genes that are associated with the disease ($CbGaD$); whether the compound belongs to the same pharmacologic classes as compounds that treat the disease ($CiPCiCtD$); whether the compound chemically resembles compounds that treat the disease ($CrCtD$); whether

Metapath	Δ AUROC	-log10(p)
CbGaD	14.5%	6.2
CbGiGaD	9.0%	4.4
CbGiGiGaD	7.0%	5.1
CbGpPWpGaD	7.6%	7.9
CbGpBPpGaD	4.9%	3.8
CcSEcCtD	14.0%	6.8
CtDpSpD	13.9%	6.1
CbGeAlD	8.4%	5.2
CtDlAlD	12.4%	6.0
CuGdD	1.1%	2.6
CdGuD	1.7%	4.5
CuGuCtD	4.4%	3.5
CdGdCtD	3.8%	4.6
CuGdCtD	-1.6%	2.9
CdGuCtD	-2.1%	2.4
CtDuGuD	1.1%	1.4
CtDdGdD	4.2%	3.9
CtDdGuD	0.5%	1.0
CtDuGdD	0.7%	1.3

Table 5.3. The predictiveness of select metapaths. The performance of several interesting metapaths is shown.

the compound treats diseases which resemble the disease (*CtDrD*); whether the compound resembles compounds that resemble compounds that treat the disease (*CrCrCtD*); whether the compound causes the same side effects as compounds that treat the disease (*CcSEcCtD*); whether the compound palliates the same diseases as compounds that treat the disease (*CpDpCtD*); whether the compound binds to genes that participate in the same pathways as genes associated with the disease (*CbGpPWpGaD*); whether the compound binds to genes that are expressed in the anatomies affected by the disease (*CbGeAlD*).

We applied the model to predict the probability of treatment between each of 1,538 connected compounds and each of 136 connected diseases, resulting in predictions for 209,168 compound–disease pairs [131]. The 755 known disease-modifying indications were highly ranked (AUROC = 97.4%, Figure 5.3). The predictions also successfully prioritized two external validation sets: novel indications from DrugCentral (AUROC = 85.5%) and novel indications in clinical trial

(AUROC = 70.0%).

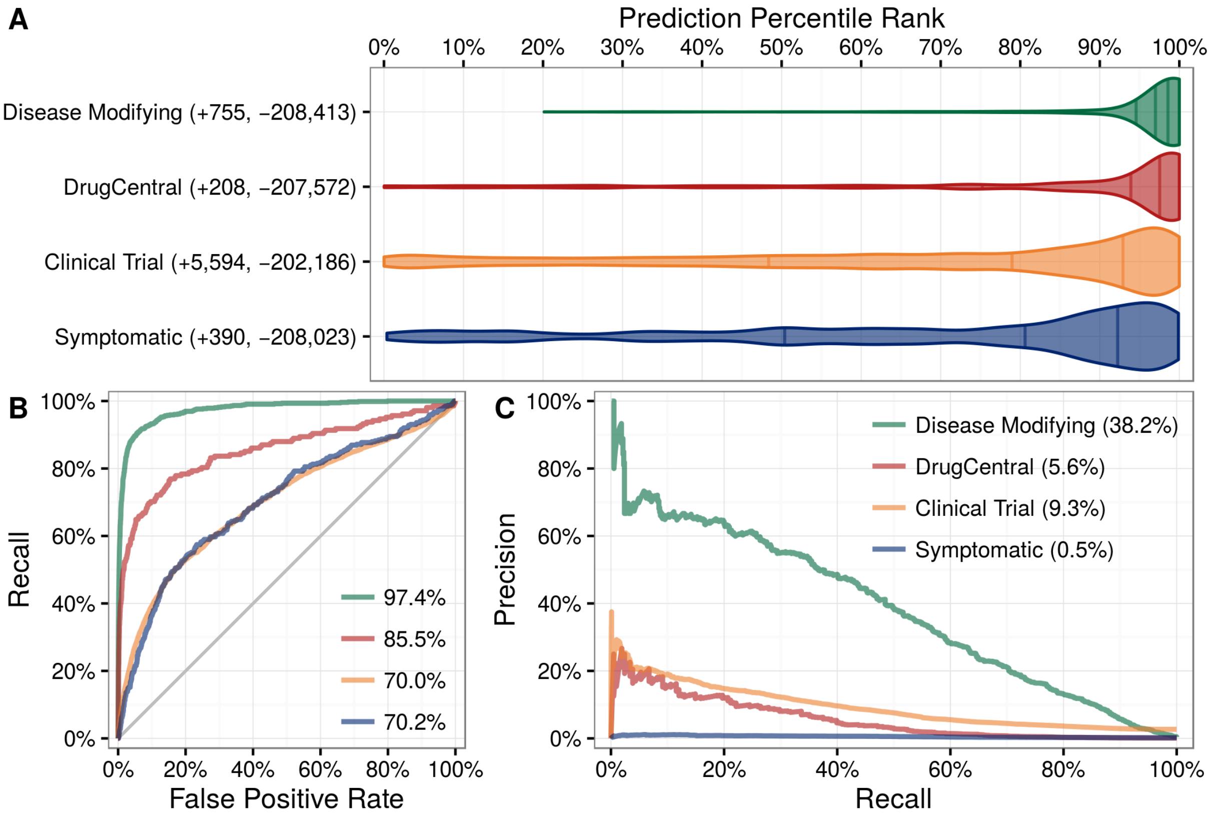

Figure 5.3. Predictions performance on four indication sets. We assess how well our predictions prioritize four sets of indications. A) The y-axis labels denote the number of indications (+) and non-indications (-) composing each set. Violin plots with quartile lines show the distribution of indications when compound–disease pairs are ordered by their prediction. In all four cases, the actual indications were ranked highly by our predictions. B) ROC Curves with AUROCs in the legend. C) Precision–Recall Curves with AUPRCs in the legend.

Predictions were scaled to the overall prevalence of treatments (0.36%). Hence a compound–disease pair that received a prediction of 1% shows a 3-fold enrichment over random. Of the 3,980 predictions over 1%, 586 were disease-modifying indications leaving 3,394 repurposing candidates. One such example is the application of clofarabine to treat multiple sclerosis (Figure 5.4). Clofarabine is chemically similar to cladribine [132], which showed promising phase 3 results and was approved in Australia and Russia before being withdrawn due to safety concerns

[133, 134]. Our method picked up on the similarities between cladribine and clofarabine, both in terms of structure and targets. In addition, clofarabine is a nucleic acid synthesis inhibitor like azathioprine – an effective multiple sclerosis treatment [135].

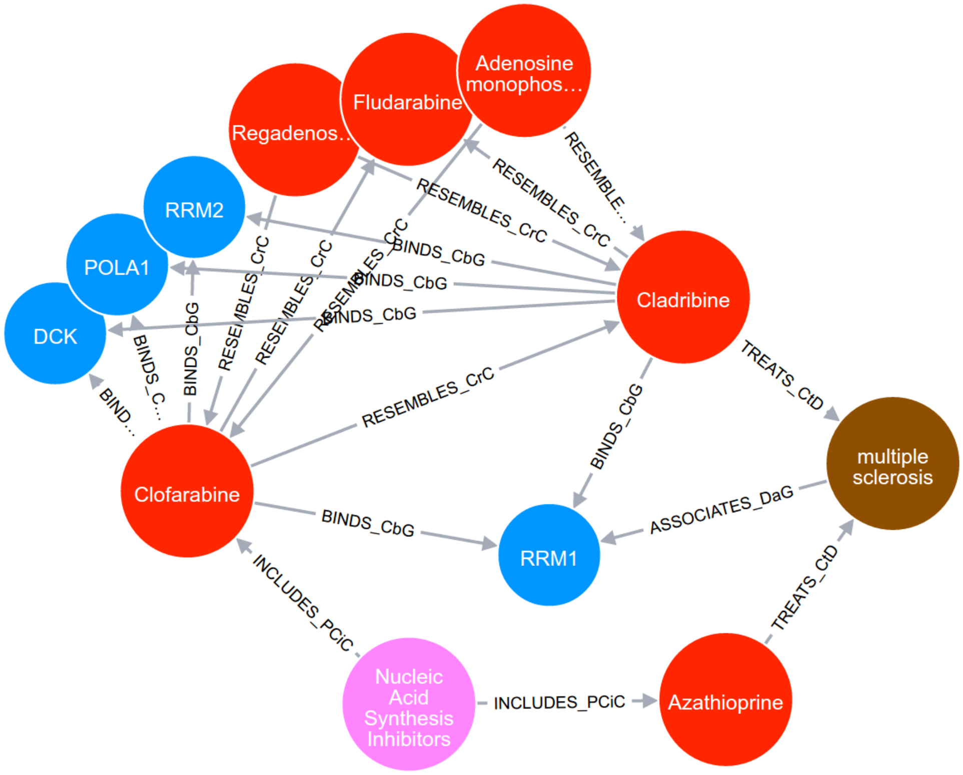

Figure 5.4. Visualizing the prediction that Clofarabine treats multiple sclerosis. Clofarabine was the top prediction for multiple sclerosis that wasn't already a known treatment. It received a probability of 8.80%, representing a 24-fold enrichment in probability over random. The ten paths that provide the greatest support for the efficacy of clofarabine in treating multiple sclerosis are shown.

5.4 Discussion

We created Hetionet v1.0, which integrates 29 resources into a single hetnet. Consisting of 11 types of nodes and 24 types of relationships, Hetionet brings more types of information together than previous leading-studies in biological data integration [10]. Moreover, we strove to create a reusable, extensible, and property-rich network.

As data has grown more plentiful and diverse, so have the applications for hetnets. Unfortunately, network science has been fragmented by discipline and slow to adapt. A 2014 analysis identified 78 studies using multilayer networks – a superset of hetnets with the potential for a time dimension. However, the studies relied on 26 different terms, 9 of which had multiple definitions [23]. Nonetheless, core infrastructure and algorithms for hetnets are emerging. One goal of our project has been to unite hetnet research across disciplines [20]. We approached this goal by making Project Rephetio entirely available online and inviting community feedback throughout the process.

Integrating every resource into a single interconnected data structure allowed us to assess systematic mechanisms of drug efficacy. Using the max performing metapath to assess the pharmacological utility of a metaedge (Figure 5.2A), we can divide our relationships into tiers of informativeness. The top tier consists of the types of information traditionally considered by pharmacology: *Compound–treats–Disease*, *Pharmacologic Class–includes–Compound*, *Compound–resembles–Compound*, *Disease–resembles–Disease*, and *Compound–binds–Gene*. The upper-middle tier consists of types of information that have been the focus of substantial medical study, but have only recently started to play a bigger role in drug development: *Disease–associates–Gene*, *Compound–causes–Side Effect*, *Disease–presents–Symptom*, *Disease–localizes–Anatomy*, and *Gene–interacts–Gene* metaedges.

The lower-middle tier contains the transciptomics metaedges such as *Compound–downregulates–Gene*, *Anatomy–expresses–Gene*, *Gene–regulates–Gene*, and *Disease–downregulates–Gene*. Much excitement surrounds these resources due to their high throughput and genome-wide scope, which offers a route to drug discovery that is less biased by existing knowledge. Our findings suggest that these resources are weakly informative of drug efficacy. Other lower-middle tier metaedges were the product of time-intensive biological experimentation, such as *Gene–participates–Pathway*, *Gene–participates–Molecular Function*, and *Gene–participates–Biological Process*. However, unlike the top tier resources, this knowledge was historically pursued for basic science rather than primarily medical applications. The weak yet appreciable performance of the *Gene–covaries–Gene* suggests the synergy between the fields of evolutionary genomics and disease biology. The lower tier included the *Gene–participates–Cellular Component* metaedge, which may reflect that the relevance of cellular location to pharmacology is highly case dependent and not amenable to systematic profiling.

5.5 Methods

Hetionet was built entirely from publicly-available resources with the goal of integrating a broad diversity of information types of medical relevance, ranging in scale from molecular to organismal. Practical considerations such as data availability, licensing, reusability, documentation, throughput, and standardization informed our choice of resources. We abided by a simple litmus test for determining how to encode information in a hetnet: nodes represent nouns, relationships represent verbs [136,137].

Our method for relationship prediction creates a strong incentive to avoid redundancy, which increases the computational burden without improving performance. In a previous study to predict disease–gene associations using a hetnet of pathophysiology [117], we found that different types of gene sets contributed highly redundant information. Therefore, in Hetionet v1.0 we reduced the number of gene set node types from 14 to 3 by omitting several gene set collections and aggregating all pathway nodes.

5.5.1 Nodes

Nodes encode entities. We extracted nodes from standard terminologies, which provide curated vocabularies to enable data integration and prevent concept duplication. The ease of mapping external vocabularies, adoption, and comprehensiveness were primary selection criteria. Hetionet v1.0 includes nodes from 5 ontologies – which provide hierarchy of entities for a specific domain – selected for their conformity to current best practices [138].

We selected 137 terms from the Disease Ontology [61,139] (which we refer to as DO Slim [140, 141]) as our disease set. Our goal was to identify complex diseases that are distinct and specific enough to be clinically relevant yet general enough to be well annotated. To this end, we included diseases that have been studied by GWAS and cancer types from `TopNodes_DOcancerslim` [142]. We ensured that no DO Slim disease was a subtype of another DO Slim disease. Symptoms were extracted from MeSH by taking the 438 descendants of *Signs and Symptoms* [143, 144].

Approved small molecule compounds with documented chemical structures were extracted from DrugBank version 4.2 [145–147]. Unapproved compounds were excluded because our focus was repurposing. In addition, unapproved compounds tend to be less studied than approved compounds making them less attractive for our approach where robust network connectivity is critical. Finally, restricting to small molecules with known documented structures enabled us to map between compound vocabularies (see Mappings).

Side effects were extracted from SIDER version 4.1 [148–150]. SIDER codes side effects using UMLS identifiers [151], which we also adopted. Pharmacologic Classes were extracted from the DrugCentral data repository (`olegursu/drugtarget`) [152].

Protein-coding human genes were extracted from Entrez Gene [153–155]. Anatomical structures, which we refer to as anatomies, were extracted from Uberon [156]. We selected a subset of 402 Uberon terms by excluding terms known not to exist in humans and terms that were overly broad or arcane [157, 158].

Pathways were extracted by combining human pathways from WikiPathways [159, 160], Reactome [161], and the Pathway Interaction Database [162]. The latter two resources were retrieved from Pathway Commons [163], which compiles pathways from several providers. Duplicate pathways and pathways without multiple participating genes were removed [164, 165]. Biological processes, cellular components, and molecular functions were extracted from the Gene Ontology [60]. Only terms with 2–1000 annotated genes were included.

5.5.2 Mappings

Before adding relationships, all identifiers needed to be converted into the vocabularies matching that of our nodes. Oftentimes, our node vocabularies included external mappings. For example, the Disease Ontology includes mappings to MeSH, UMLS, and the ICD, several of which we submitted during the course of this study [166]. In a few cases, the only option was to map using gene symbols, a disfavored method given that it can lead to ambiguities.

When mapping external disease concepts onto DO Slim, we used transitive closure. For example, the UMLS concept for primary progressive multiple sclerosis (`C0751964`) was mapped to the DO Slim term for multiple sclerosis (`DOID:2377`).

Chemical vocabularies presented the greatest mapping challenge [146], since these are poorly standardized [167]. UniChem's [168] Connectivity Search [169] was used to map compounds, which maps by atomic connectivity (based on First InChIKey Hash Blocks [170]) and ignores small molecular differences.

5.5.3 Edges

Anatomy–downregulates–Gene and *Anatomy–upregulates–Gene* edges [171–173] were extracted from Bgee [174], which computes differentially expressed genes by anatomy in post-juvenile adult humans. *Anatomy–expresses–Gene* edges were extracted from Bgee and TISSUES [175–177].

Compound–binds–Gene edges were aggregated from BindingDB [178, 179], DrugBank [145, 180], and DrugCentral.

Only binding relationships to single proteins with affinities of at least 1 M (as determined by Kd, Ki, or IC50) were selected from the October 2015 release of BindingDB [181, 182]. Target, carrier, transporter, and enzyme interactions with single proteins (i.e. excluding protein groups) were extracted from DrugBank 4.2 [147, 183]. In addition, all mapping DrugCentral target relationships were included [152].

Compound–treats–Disease (disease-modifying indications) and *Compound–palliates–Disease* (symptomatic indications) edges are from PharmacotherapyDB as described in Intermediate resources. *Compound–causes–Side Effect* edges were obtained from SIDER 4.1 [148–150], which uses natural language processing to identify side effects in drug labels. *Compound–resembles–Compound* relationships [147,184,185] represent chemical similarity and correspond to a Dice coefficient ≥ 0.5 [186] between extended connectivity fingerprints [187,188]. *Compound–downregulates–Gene* and *Compound–upregulates–Gene* relationships were computed from LINCS L1000 as described in Intermediate resources.

Disease–associates–Gene edges were extracted from the GWAS Catalog [189], DISEASES [190,191], DisGeNET [192,193], and DOAF [194,195]. The GWAS Catalog compiles disease–SNP associations from published GWAS [67]. We aggregated overlapping loci associated with each disease and identified the mode reported gene for each high confidence locus [196,197]. DISEASES integrates evidence of association from text mining, curated catalogs, and experimental data [198]. Associations from DISEASES with integrated scores ≥ 2 were included after removing the contribution of DistiLD. DisGeNET integrates evidence from over 10 sources and reports a single score for each association [199]. Associations with scores ≥ 0.06 were included. DOAF mines Entrez Gene GeneRIFs (textual annotations of gene function) for disease mentions [200]. Associations with 3 or more supporting GeneRIFs were included. *Disease–downregulates–Gene* and *Disease–upregulates–Gene* relationships [201,202] were computed using STARGEO as described in Intermediate resources.

Disease–localizes–Anatomy, *Disease–presents–Symptom*, and *Disease–resembles–Disease* edges were calculated from MEDLINE cooccurrence [143,203]. MEDLINE is a subset of 21 million PubMed articles for which designated human curators have assigned topics. When retrieving articles for a given topic (MeSH term), we activated two non-default search options as specified below: `majr` for selecting only articles where the topic is major and `noexp` for suppressing explosion (returning articles linked to MeSH subterms). We identified 4,161,769 articles with two or more disease topics; 696,252 articles with both a disease topic (`majr`) and an anatomy topic (`noexp`) [204]; and 363,928 articles with both a disease topic (`majr`) and a symptom topic (`noexp`). We used a Fisher's exact test [205] to identify pairs of terms that occurred together more than would be expected by chance in their respective corpus. We included cooccuring terms with $p < 0.005$ in Hetionet v1.0.

Gene–covaries–Gene edges represent evolutionary rate covariation ≥ 0.75 [206–208]. *Gene–interacts–Gene* edges [209, 210] represent when two genes produce physically-interacting proteins. We compiled these interactions from the Human Interactome Database [90, 211–213], the Incomplete Interactome [214], and our previous study [117]. *Gene–participates–Biological Process*, *Gene–participates–Cellular Component*, and *Gene–participates–Molecular Function* edges are from Gene Ontology annotations [215]. As described in Intermediate resources, annotations were propagated [216, 217].

5.5.4 Intermediate resources

In the process of creating Hetionet, we produced several datasets with broad applicability that extended beyond Project Rephetio. These resources are referred to as intermediate resources and described below.

Transcriptional signatures of disease using STARGEO STARGEO is a nascent platform for annotating and meta-analyzing differential gene expression experiments. The STAR acronym stands for Search-Tag-Analyze Resources, while GEO refers to the Gene Expression Omnibus [218, 219]. STARGEO is a layer on top of GEO that crowdsources sample annotation and automates meta-analysis.

Using STARGEO, we computed differentially expressed genes between healthy and diseased samples for 49 diseases [201, 202]. First, we and others created case/control tags for 66 diseases. After combing through GEO series and tagging samples, 49 diseases had sufficient data for case-control meta-analysis: multiple series with at least 3 cases and 3 controls. For each disease, we performed a random effects meta-analysis on each gene to combine log2 fold-change across series. These analyses incorporated 27,019 unique samples from 460 series on 107 platforms.

Differentially expressed genes (false discovery rate ¡ 0.05) were identified for each disease.

The median number of upregulated genes per disease was 351 and the median number of downregulated genes was 340. Endogenous depression was the only disease without any significantly dysregulated genes.

Transcriptional signatures of perturbation from LINCS L1000 LINCS L1000 profiled the transcriptional response to small molecule and genetic interference perturbations. To increase throughput, expression was only measured for 978 genes, which were selected for their ability to impute expression of the remaining genes. A single perturbation was often assayed under a variety of conditions including cell types, dosages, timepoints, and concentrations. Each condition generates a single signature of dysregulation z-scores. We further processed these signatures to fit into our approach [220, 221].

First we computed consensus signatures – which meta-analyze multiple signatures to condense them into one – for DrugBank small molecules, Entrez genes, and all L1000 perturbations [222, 223]. First, we discarded non-gold (non-replicating or indistinct) signatures. Then we meta-analyzed z-scores using Stouffer's method. Each signature was weighted by its average Spearman's correlation to other signatures, with a 0.05 minimum, to de-emphasize discordant signatures. Our signatures include the 978 measured genes and the 6,489 imputed genes from the "best inferred gene subset". To identify significantly dysregulated genes, we selected genes using a Bonferroni cutoff of $p = 0.05$ and limited the number of imputed genes to 1,000.

The consensus signatures for genetic perturbations allowed us to assess various characteristics of the L1000 dataset. First, we looked at whether genetic interference dysregulated its target gene in the expected direction [224]. Looking at measured z-scores for target genes, we found that the knockdown perturbations were highly reliable, while the overexpression perturbations were only moderately reliable with 36% of overexpression perturbations downregulating their target. However, imputed z-scores for target genes barely exceeded random at responding in the expected direction to interference. Hence, we concluded that the imputation quality of LINCS L1000 is poor. However, when restricting to significantly dyseregulated targets, 22 out of 29 imputed genes responded in the expected direction. This provides some evidence that the directional fidelity of imputation is higher for significantly dysregulated genes. Finally, we found that the transcriptional signatures of knocking down and overexpressing the same gene were positively correlated 65% of the time, suggesting the presence of a general stress response [225].

Based on these findings, we performed additional filtering of signifcantly dysregulated genes when building Hetionet v1.0. *Compound–down/up-regulates–Gene* relationships were restricted to the 125 most significant per compound-direction-status combination (status refers to measured versus imputed). For genetic interference perturbations, we restricted to the 50 most significant genes per gene-direction-status combination and merged the remaining edges into a single *Gene–regulates–Gene* relationship type containing both knockdown and overexpression perturbations.

PharmacotherapyDB: physician curated indications We created PharmacotherapyDB, an open catalog of drug therapies for disease [226–228]. Version 1.0 contains 755 disease-modifying therapies and 390 symptomatic therapies between 97 diseases and 601 compounds.

This resource was motivated by the need for a gold standard of medical indications to train and evaluate our approach. Initially, we identified four existing indication catalogs [229]: MEDI-HPS which mined indications from RxNorm, SIDER 2, MedlinePlus, and Wikipedia [230]; LabeledIn which extracted indications from drug labels via human curation [231–233]; EHRLink which identified medication–problem pairs that clinicians linked together in electronic health records [234, 235]; and indications from PREDICT, which were compiled from UMLS relationships, drugs.com, and drug labels [118]. After mapping to DO Slim and DrugBank Slim, the four resources contained 1,388 distinct indications.

However, we noticed that many indications were palliative and hence problematic as a gold standard of pharmacotherapy for our *in silico* approach. Therefore, we recruited two practicing physicians to curate the 1,388 preliminary indications [236]. After a pilot on 50 indications, we defined three classifications: *disease modifying* meaning a drug that therapeutically changes the underlying or downstream biology of the disease; *symptomatic* meaning a drug that treats a significant symptom of the disease; and *non-indication* meaning a drug that neither therapeutically changes the underlying or downstream biology nor treats a significant symptom of the disease. Both curators independently classified all 1,388 indications.

The two curators disagreed on 444 calls (Cohen's = 49.9%). We then recruited a third practicing physician, who reviewed all 1,388 calls and created a detailed explanation of his methodology. We proceeded with the third curator's calls as the consensus curation. The first two curators did have reservations with classifying steroids as disease modifying for autoimmune diseases. However, we were convinced that these indications met our definition of disease modifying, which is based on a pathophysiological rather than clinical standard. Accordingly, therapies we consider disease modifying may not be used to alter long-term disease course in the modern clinic due to a poor risk–benefit ratio.

User-friendly Gene Ontology annotations We created a browser to provide straightforward access to Gene Ontology annotations [216, 217]. Our service provides annotations between Gene Ontology terms and Entrez Genes. The user chooses propagated/direct annotation and all/experimental evidence. Annotations are currently available for 37 species and downloadable as user-friendly TSV files.

5.5.5 Data copyright and licensing

We committed to openly releasing our data and analyses from the origin of the project [237]. Our goals were to contribute to the advancement of science [238,239], maximize our impact [240], and enable reproducibility [241–243]. All three of these objectives require publicly distributing Hetionet. In addition, all three benefit if our hetnet and analyses are openly licensed [244, 245].

Since we integrated only public resources, which were overwhelmingly funded by academic grants, we had assumed that our project and open sharing of our network would not be an issue. However, upon releasing a preliminary version of our hetnet [246], a community reviewer informed us of legal barriers to integrating public data. In essence, both copyright (rights of exclusivity automatically granted to original works) and terms of use (rules that users must agree to in order to use a resource) place legally-binding restrictions on data reuse.

Of the 29 resources we integrated, only 12 had licenses that met the Open Definition with respect to knowledge. 9 did not have a license, which equates to all rights reserved and by default forbids reuse. Several resources had incompatible licenses caused primarily by non-commercial and share-alike stipulations. One resource included terms which explicitly forbid redistribution. In addition, it was often unclear who owned the data [247]. Therefore, we sought input from legal experts and chronicled our progress [248–251].

Ultimately, we did not find an ideal solution. We had to choose between absolute compliance and our hetnet: strictly adhering to copyright and licensing arrangements would have decimated our network. Hence we choose a path forward which balanced legal, normative, ethical, and scientific considerations. If a resource was in the public domain, for example works of the US Government, we licensed any derivatives as CC0 1.0. For resources licensed to allow use, redistribution, and modification, we transmitted their licenses as properties on the specific nodes and relationships in our hetnet. For all other resources – for example, resources without licenses or with licenses that forbid redistribution – we sent permission requests to their creators. The median time till first response to our permission requests was 16 days, with only 2 resources affirmatively granting us permission. We did not receive any responses asking us to remove a resource. However, we did voluntarily remove MSigDB [54], since its license was highly problematic [249].

5.5.6 Permuted Hetnets

From Hetionet, we derived five permuted hetnets [252]. The permutations preserve node degree but eliminate edge specificity by employing an algorithm called XSwap to randomly swap edges [253]. Permuted networks are useful for computing the baseline performance of meaningless edges while preserving node degree [254].

5.5.7 Neo4j

While in a previous project, we developed `hetio` – a Python package for hetnets [255] – for this work, we migrated to the Neo4j graph database for storing and operating on hetnets [256]. However, `hetio` was still used to create the network and prepare Neo4j queries. Graph database adoption in bioinformatics has thus far been limited [21]. Nonetheless, we noticed major benefits by tapping into a larger open source ecosystem. Persistent storage with immediate access and the Cypher query language – a sort of SQL for hetnets – were two of the biggest draws. We created an interactive GraphGist on our project, which introduces our approach and showcases our Cypher queries.

5.5.8 Machine learning approach

We made several refinements to metapath-based hetnet edge prediction compared to previous studies [52, 117]. First, we transformed DWPCs to make them more amenable to modeling [257] by mean scaling and then taking the inverse hyperbolic sine [258]. Second, we bifurcated the workflow into an all-features stage and an all-observations stage [259]. The all-features stage assesses feature performance and does not require computing features for all negatives. Here we selected a random subset of 3,020 negatives. Little error was introduced by this optimization, since the predominant limitation to performance assessment was the small number of positives (755) rather than negatives. Based on the all-features performance assessment [260], we selected 142 DWPCs to compute on all observations (all 209,168 compound–disease pairs). The feature selection was designed to remove uninformative features (according to permutation) and guard against edge-dropout contamination [261]. Third, we included 14 degree features, which assess the degree of a specific metaedge for either the source compound or target disease.

Prior probability of treatment The 755 treatments in Hetionet v1.0 are not evenly distributed between all compounds and diseases. For example, methotrexate treats 19 diseases and hypertension is treated by 68 compounds. We estimated a prior probability of treatment

– based only on the treatment degree of the source compound and target disease – on 744,975 permutations of the bipartite treatment network [262]. Methotrexate received a 79.6% prior probability of treating hypertension, whereas a compound and disease that both had only one treatment received a prior of 0.12%.

Across the 209,168 compound–disease pairs, the prior predicted the known treatments with AUROC = 97.9%. The strength of this association threatened to dominate our predictions. However, not modeling the prior can lead to omitted-variable bias and confounded proxy variables. To address the issue, we included the logit-transformed prior, without any regularization, as a term in the model. This restricted model fitting to the 29,799 observations with a nonzero prior – corresponding to the 387 compounds and 77 diseases with at least one treatment. To enable predictions for all 209,168 observations, we set the prior for each compound–disease pair to the overall prevalence of positives (0.36%).

This method succeeded at accommodating the treatment degrees. The prior probabilities performed poorly on the validation sets with AUROC = 54.1% on DrugCentral indications and AUROC = 62.5% on clinical trials. This performance dropoff compared to training shows the danger of encoding treatment degree into predictions. The benefits of our solution are highlighted by the superior validation performance of our predictions compared to the prior (Figure 3).

5.5.9 Indication sets

We evaluated our predictions on four sets of indications as shown in Figure 3.

- **Disease Modifying** – the 755 disease modifying treatments in PharmacotherapyDB v1.0. These indications are included in the hetnet as *treats* edges and used to train the logistic regression model. Due to edge dropout contamination and self-testing [261, 263], overfitting could potentially inflate performance on this set. Therefore, for the three remaining indication sets, we removed any observations that were positives in this set.
- **DrugCentral** – We discovered the DrugCentral database after completing our physician curation for PharmacotherapyDB. This database contained 210 additional indica-

tions [152]. While we didn't curate these indications, we observed a high proportion of disease modifying therapy.

- **Clinical Trial** – We compiled indications that have been investigated by clinical trial from ClinicalTrials.gov [264]. This set contains 5,594 indications.
- **Symptomatic** – 390 symptomatic indications from PharacotherapyDB. These edges are included in the hetnet as *palliates* edges.

Only the Clinical Trial and DrugCentral indication sets were used for external validation, since the Disease Modifying and Symptomatic indications were included in the hetnet.

5.5.10 Realtime open science & Thinklab

We conducted our study using *Thinklab* – a platform for realtime open collaborative science – on which this study was the first project. We began the study by publicly proposing the idea and inviting discussion [265]. We continued by chronicling our progress via discussions. We used Thinklab as the frontend to coordinate and report our analyses and GitHub as the backend to host our code, data, and notebooks. On top of our Thinklab team consisting of core contributors, we welcomed community contribution and review. In areas where our expertise was lacking or advice would be helpful, we sought input from domain experts and encouraged them to respond on Thinklab where their comments would be CC BY licensed and their contribution rated and rewarded.

In total, 35 non-team members commented across 77 discussions, which generated 452 comments and 152 notes. The Thinklab content for this project totaled 101,501 words or 635,151 characters [266]. Using an estimated 7,000 words per academic publication as a benchmark, Project Rephetio generated written content comparable in volume to 14.5 publications prior to its completion. We noticed several other benefits from using Thinklab including forging a community of contributors [267]; receiving feedback during the early stages when feedback is the most actionable [268]; disseminating our research without delay [269, 270]; opening avenues for external input [271]; facilitating problem-oriented teaching [272, 273]; and improving our documentation by maintaining a publication-grade digital lab notebook [274].

Chapter 6

Open science

The entirety of my dissertation was exclusively based on publicly-available data. Hetnets thrive by integrating data of many types, which comes from many sources. The breadth and diversity of the data our hetnets integrate means no one proprietary collection contains all the data. Hence, public data was the *only* option to achieve our desired scale. Additionally, public data has the benefit of being immediately available. Were we dependent on 30 different collaborating research groups rather than 30 public databases to compile our data, our projects would not have been possible. As discussed in Chapter 5, there are still legal barriers to data reuse that consumed substantial time. However, I'm hoping our experience will spur progress in this area. Sharing my research as soon as possible and as openly as possible has been a guiding principle of my PhD. This chapter discusses other open science projects and efforts to improve science, I undertook during my PhD.

6.1 Elevation, Oxygen, and Lung Cancer

During the second year of my PhD in early 2013, my roommate and soon-to-be colleague, Kamen Simeonov, mentioned his observation that lung cancer rates were lower at high altitude. He theorized that breathing oxygen causes cancer. We began designing an epidemiological study to interrogate his hypothesis. As two early career scientists not affiliated with any lung cancer research programs, we were dependent solely on public data. We compiled data from 11 publicly-available databases. The study was published in [275]:

> Simeonov KP, **Himmelstein DS** (2015) **Lung cancer incidence decreases with elevation: evidence for oxygen as an inhaled carcinogen**. *PeerJ*. DOI: 10.7717/peerj.705

Figure 6.1 shows the inverse association between lung cancer and elevation we observed. The abstract of this study follows:

> The level of atmospheric oxygen, a driver of free radical damage and tumorigenesis, decreases sharply with rising elevation. To understand whether ambient oxygen plays a role in human carcinogenesis, we characterized age-adjusted cancer incidence (compiled by the National Cancer Institute from 2005 to 2009) across counties of the elevation-varying Western United States and compared trends displayed by respiratory cancer (lung) and non-respiratory cancers (breast, colorectal, and prostate). To adjust for important demographic and cancer-risk factors, 8–12 covariates were considered for each cancer. We produced regression models that captured known risks. Models demonstrated that elevation is strongly, negatively associated with lung cancer incidence ($p < 10^{-16}$), but not with the incidence of non-respiratory cancers. For every 1,000 m rise in elevation, lung cancer incidence decreased by 7.23 99% CI [5.18–9.29] cases per 100,000 individuals, equivalent to 12.7% of the mean incidence, 56.8. As a predictor of lung cancer incidence, elevation was second only to smoking prevalence in terms of significance and effect size. Furthermore, no evidence of ecological fallacy or of confounding arising from evaluated factors was detected: the lung cancer association was robust to varying regression models, county stratification, and population subgrouping; additionally seven environmental correlates of elevation, such as exposure to sunlight and fine particulate matter, could not capture the association. Overall, our findings suggest the presence of an

> inhaled carcinogen inherently and inversely tied to elevation, offering epidemiological support for oxygen-driven tumorigenesis. Finally, highlighting the need to consider elevation in studies of lung cancer, we demonstrated that previously reported inverse lung cancer associations with radon and UVB became insignificant after accounting for elevation.

While our study was not the first to suggest oxygen-driven tumorigenesis in lung cancer etiology [276, 277], our study turned out to be provocative. *Cancer Research UK* ridiculed the study, but we addressed their criticism in a blog post. George Johnson, in his Raw Data column for the *New York Times*, would later summarize the controversy writing:

> Skeptics were quick to strike back, though not very effectively. A would-be debunking on the Cancer Research UK website was quickly followed by a debunking of the debunking.

In addition to the Times, our study was covered by over 100 news outlets. It was also named a Top Cancer Biology Paper by *PeerJ* and won the Abramson Cancer Center 2015 Basic Research Prize at the University of Pennsylvania.

We made our entire analysis available online. I designed the codebase so the study could be replicated with a single command. One highlight was post publication when we received a GitHub issue reporting an error running our code. Debugging the code led me to discover that a few values in our publication were incorrect. The *PeerJ* Question/Comment feature allowed me to issue a corrigendum in realtime, which traced the error back to the specific line of faulty code. The experience illustrates how the self-correction of science will accelerate from openness.

6.2 Publishing Delays

While waiting for the contents of Chapter 4 to be published in *PLOS Computational Biology*, I grew restless and frustrated by the glacial pace of scientific publishing. Our paper had been

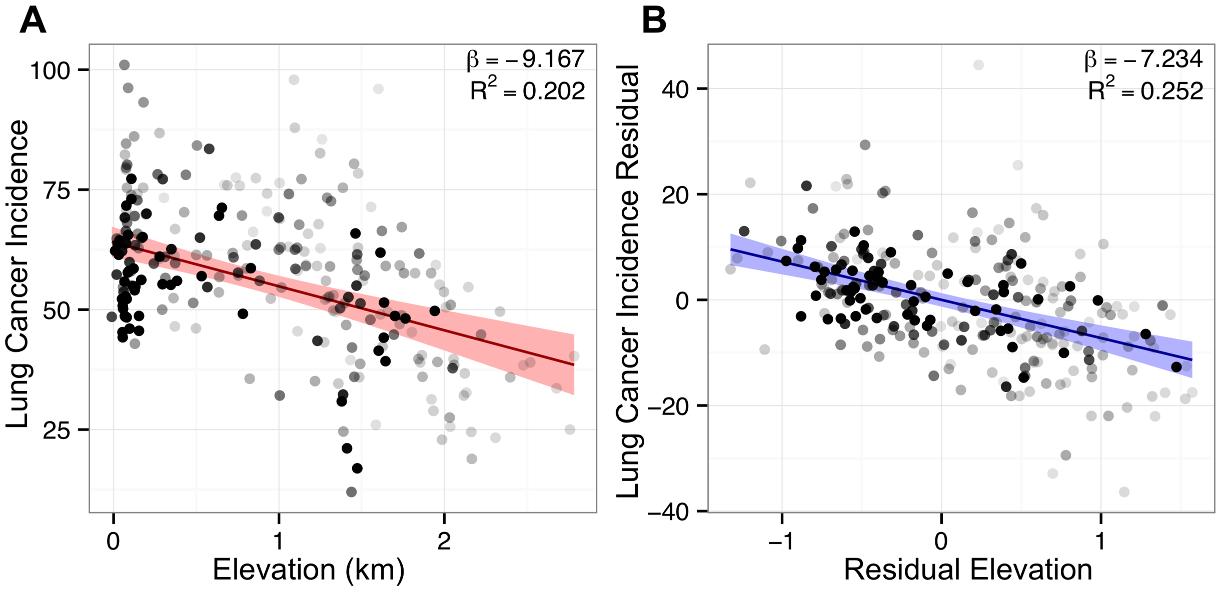

Figure 6.1. Adjustment for covariates sharpens lung cancer's association with elevation. Points represent counties shaded by their regression weight based on population. Bivariate (red) and partial (blue) regression lines are displayed with 99% confidence bands. (A) Bivariate plot of county lung cancer incidence (age-adjusted per 100,000) and elevation (km). (B) Partial regression plot for elevation based on the optimal best subset lung model. Association sharpens after adjustment for covariates, illustrated by the tighter confidence band and higher R2 in the partial plot.

accepted for 68 days but was still not published. I began researching publishing delays for the *PLOS* family of journals. I initially started by scraping the *PLOS* website for article timestamps. However, I soon began downloading the history dates in PubMed for all articles since 2014. I visualized the delays for several open access titles in my field. I posted the findings on my blog (Figure 6.2) and added a table of median delays for the 3,475 journals that submitted PubMed history data [278]. My tweet introducing the analysis received hundreds of retweets and *Nature News* soon covered the story [279].

Later Kendall Powell, writing a feature for *Nature News*, contacted me with additional questions. Her investigation had uncovered a widespread belief that delays were worsening with time. But she wanted data, and the existing data was field specific or anecdotal [280]. So I set out to uncover the history of publishing delays. Using PubMed, I extracted delays for over

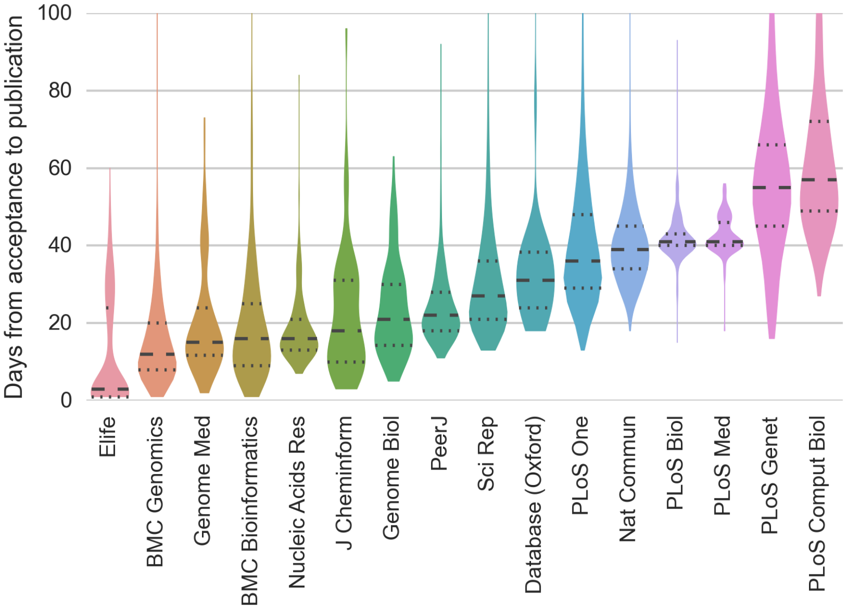

Figure 6.2. Publication delay distributions for 16 journals in my field. Quartiles are drawn as horizontal lines. The two journals where we submitted Chapter 4 for publication had the highest median publication delays.

3 million articles since 1965. I posted my findings in a blog post released in tandem with the feature [269, 281]. The feature drew considerably from my analysis and is available at:

> Powell K (2016) **Does it take too long to publish research?**. *Nature*. DOI: 10.1038/530148a

In short, I found that the median time from submission to acceptance has hovered around 100 days since 1981 (Figure 6.3). However, the median time from acceptance to online publication has decreased over 50 days in the early 2000s to under 25 days in 2015. One caveat with my analysis of acceptance delays is that journals may be resetting the clock (reporting subsequent manuscript receival datas rather than the data of first submission). However, users can select a journal of their choosing and see its specific delay history. I am hoping this increased transparency will help eliminate deceptive timestamping. My goal with this research has been to help researchers avoid excessive delays while replacing anecdote with evidence in the contemporary discussion of scientific publishing.

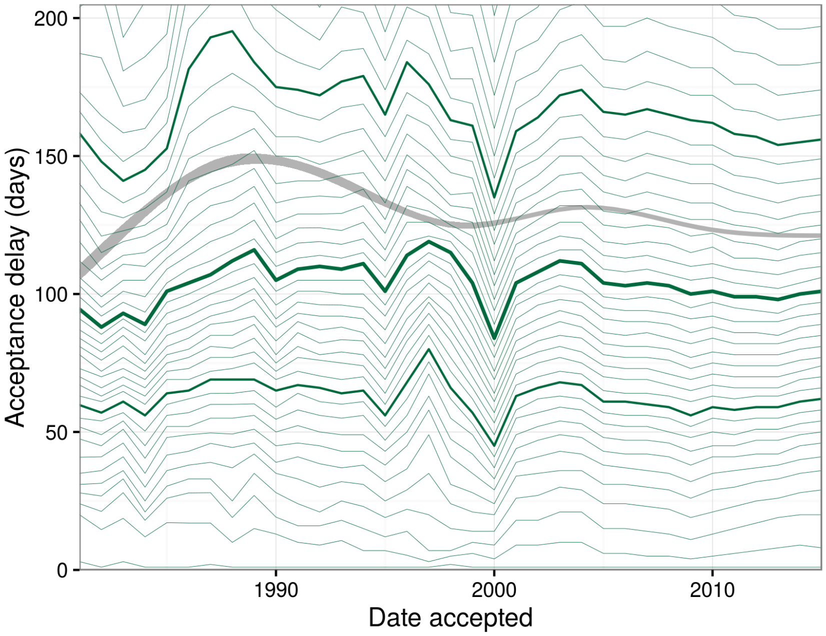

Figure 6.3. 35 years of acceptance delays. The time between submission and acceptance encapsulates editorial decision, peer review, and revision. This figure visualize 35 years of acceptance delays. Each year, the green lines indicate delay percentiles, spaced every 2.5 points with quartiles bolded. The gray band displays a curve fitted for all articles over time.

Chapter 7

Conclusion

My dissertation imagines a future where vast amounts of information are encoded using hetnets. A hetnet is to artificial intelligence what an encyclopedias is to human intelligence. Hetnets provide a medium for integrating diverse information. Structured vocabularies, ontologies, and terminology mappings have enabled large scale biomedical hetnets, such as Hetionet.

Now that the data structure is gaining momentum, there is a large opportunity to create algorithms for extracting insights from hetnets. Our experience indicates machine learning on networks is a complex task. Particularly, it's difficult to assign causality to any observed signals. Great care must be taken to avoid potential confounding factors. We found that permutation is an invaluable tool when analyzing hetnets. However, designing the appropriate permutation is not always trivial and is usually computationally intensive. I hope to continue developing a conceptual framework for handling hetnets.

With respect to hetnet edge prediction, there are many unexplored possibilities. One idea is complex metapaths that mandate several conditions. For example, finding compounds that affect the same genes as a disease, but only considering genes that are expressed in disease-relevant anatomies. Or another example, allowing query-time transitive closure to address hierarchical concepts. Another enhancement would be to support edge weights. Currently, our feature extraction paradigm assumes binary (absent or present) edges. However, performance could improve if we could account for edge confidence scores or probabilities. Finally, we would like to begin comparing our method to potential alternatives [10].

There is also opportunity to grow the hetnet to include additional domains. For example, we could include SNPs, allowing us to integrate an enormous amount of high-throughput genomic data. Another idea would be to differentiate proteins and genes to achieve a more nuanced encoding of the central dogma. For nodes already in Hetionet, there are several promising additional edge types such as disease comorbidity and drug interactions. When adding new edges, we should focus on domains that will contribute orthogonal information. As the hetnet community grows and software solutions such as Neo4j advance, we expect to be able to operate on hetnets with greater efficiency and accommodate larger analyses.

Bibliography

[1] Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, et al. (2015) **Big data: Astronomical or genomical?** *PLOS Biology*. DOI: 10.1371/journal.pbio.1002195

[2] Schatz M (2015) **The next 20 years of genome research**. *Cold Spring Harbor Laboratory Press*. DOI: 10.1101/020289

[3] Cook CE, Bergman MT, Finn RD, Cochrane G, Birney E, et al. (2015) **The european bioinformatics institute in 2016: Data growth and integration**. *Nucleic Acids Res*. DOI: 10.1093/nar/gkv1352

[4] Jensen PB, Jensen LJ, Brunak S (2012) **Mining electronic health records: towards better research applications and clinical care**. *Nat Rev Genet*. DOI: 10.1038/nrg3208

[5] Gomez-Cabrero D, Abugessaisa I, Maier D, Teschendorff A, Merkenschlager M, et al. (2014) **Data integration in the era of omics: current and future challenges**. *BMC Systems Biology*. DOI: 10.1186/1752-0509-8-s2-i1

[6] Marx V (2013) **Biology: The big challenges of big data**. *Nature*. DOI: 10.1038/498255a

[7] Prins P, de Ligt J, Tarasov A, Jansen RC, Cuppen E, et al. (2015) **Toward effective software solutions for big biology**. *Nat Biotechnol*. DOI: 10.1038/nbt.3240

[8] Baker M (2013) **Big biology: The ’omes puzzle**. *Nature*. DOI: 10.1038/494416a

[9] Kalf RR, Mihaescu R, Kundu S, de Knijff P, Green RC, et al. (2013) **Variations in predicted risks in personal genome testing for common complex diseases**. *Genetics in Medicine*. DOI: 10.1038/gim.2013.80

[10] Gligorijević V, Pržulj N (2015) **Methods for biological data integration: perspectives and challenges**. *J R Soc Interface*. DOI: 10.1098/rsif.2015.0571

[11] Lapatas V, Stefanidakis M, Jimenez RC, Via A, Schneider MV (2015) **Data integration in biological research: an overview**. *Journal of Biological Research-Thessaloniki*. DOI: 10.1186/s40709-015-0032-5

[12] Jacomy M, Venturini T, Heymann S, Bastian M (2014) **ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software**. *PLoS ONE*. DOI: 10.1371/journal.pone.0098679

[13] Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) **Fast unfolding of communities in large networks**. *J Stat Mech*. DOI: 10.1088/1742-5468/2008/10/p10008

[14] Khankhanian P, Cozen W, Himmelstein DS, Madireddy L, Din L, et al. (2016) **Meta-analysis of genome-wide association studies reveals genetic overlap between hodgkin lymphoma and multiple sclerosis**. *International Journal of Epidemiology*. DOI: 10.1093/ije/dyv364

[15] Goh KI, Cusick ME, Valle D, Childs B, Vidal M, et al. (2007) **The human disease network**. *Proceedings of the National Academy of Sciences*. DOI: 10.1073/pnas.0701361104

[16] Farh KKH, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, et al. (2014) **Genetic and epigenetic fine mapping of causal autoimmune disease variants**. *Nature*. DOI: 10.1038/nature13835

[17] Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, et al. (2015) **An atlas of genetic correlations across human diseases and traits**. *Nature Genetics*. DOI: 10.1038/ng.3406

[18] Himmelstein D, Khankhanian P, Baranzini S (2015). **A human disease network from gwas loci**. *Zenodo*. DOI: 10.5281/zenodo.23025

[19] Wang G, Jung K, Winnenburg R, Shah NH (2015) **A method for systematic discovery of adverse drug events from clinical notes**. *Journal of the American Medical Informatics Association*. DOI: 10.1093/jamia/ocv102

[20] Himmelstein D, Greene C, Baranzini S (2015). **Renaming 'heterogeneous networks' to a more concise and catchy term**. *Thinklab*. DOI: 10.15363/thinklab.d104

[21] Have CT, Jensen LJ (2013) **Are graph databases ready for bioinformatics?** *Bioinformatics*. DOI: 10.1093/bioinformatics/btt549

[22] Greene CS, Himmelstein DS (2016) **Genetic association–guided analysis of gene networks for the study of complex traits**. *Circulation: Cardiovascular Genetics*. DOI: 10.1161/circgenetics.115.001181

[23] Kivela M, Arenas A, Barthelemy M, Gleeson JP, Moreno Y, et al. (2014) **Multilayer networks**. *Journal of Complex Networks*. DOI: 10.1093/comnet/cnu016

[24] Wang K, Li M, Hakonarson H (2010) **Analysing biological pathways in genome-wide association studies**. *Nat Rev Genet*. DOI: 10.1038/nrg2884

[25] Yaspan BL, Bush WS, Torstenson ES, Ma D, Pericak-Vance MA, et al. (2011) **Genetic analysis of biological pathway data through genomic randomization**. *Hum Genet*. DOI: 10.1007/s00439-011-0956-2

[26] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) **PLINK: A tool set for whole-genome association and population-based linkage analyses**. *The American Journal of Human Genetics*. DOI: 10.1086/519795

[27] Segrè AV, Groop L, Mootha VK, Daly MJ, Altshuler D (2010) **Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits**. *PLoS Genetics*. DOI: 10.1371/journal.pgen.1001058

[28] Holmans P, Green EK, Pahwa JS, Ferreira MA, Purcell SM, et al. (2009) **Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder**. *The American Journal of Human Genetics*. DOI: 10.1016/j.ajhg.2009.05.011

[29] Rossin EJ, Lage K, Raychaudhuri S, Xavier RJ, Tatar D, et al. (2011) **Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology**. *PLoS Genetics*. DOI: 10.1371/journal.pgen.1001273

[30] Taşan M, Musso G, Hao T, Vidal M, MacRae CA, et al. (2014) **Selecting causal genes from genome-wide association studies via functionally coherent subnetworks**. *Nature Methods*. DOI: 10.1038/nmeth.3215

[31] Jia P, Zheng S, Long J, Zheng W, Zhao Z (2010) **dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks**. *Bioinformatics*. DOI: 10.1093/bioinformatics/btq615

[32] Baranzini SE, Khankhanian P, Patsopoulos NA, Li M, Stankovich J, et al. (2013) **Network-based multiple sclerosis pathway analysis with GWAS data from 15,000 cases and 30,000 controls**. *The American Journal of Human Genetics*. DOI: 10.1016/j.ajhg.2013.04.019

[33] Raychaudhuri S, Plenge RM, Rossin EJ, Ng ACY, Purcell SM, et al. (2009) **Identifying relationships among genomic disease regions: Predicting genes at pathogenic SNP associations and rare deletions**. *PLoS Genetics*. DOI: 10.1371/journal.pgen.1000534

[34] Jungnickel D (2013) Graphs, Networks and Algorithms. Springer Berlin Heidelberg. DOI: 10.1007/978-3-642-32278-5

[35] L L, Zhou T (2011) **Link prediction in complex networks: A survey**. *Physica A: Statistical Mechanics and its Applications*. DOI: 10.1016/j.physa.2010.11.027

[36] Tong H, Faloutsos C, yu Pan J (2006) DOI: 10.1109/icdm.2006.70

[37] Cho DY, Kim YA, Przytycka TM (2012) **Chapter 5: Network biology approach to complex diseases**. *PLoS Comput Biol*. DOI: 10.1371/journal.pcbi.1002820

[38] Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, et al. (2010) **The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function**. *Nucleic Acids Research*. DOI: 10.1093/nar/gkq537

[39] Gonçalves JP, Francisco AP, Moreau Y, Madeira SC (2012) **Interactogeneous: Disease gene prioritization using heterogeneous networks and full topology scores**. *PLoS ONE*. DOI: 10.1371/journal.pone.0049634

[40] Valentini G, Paccanaro A, Caniza H, Romero AE, Re M (2014) **An extensive analysis of disease-gene associations using network integration and fast kernel-based gene prioritization methods**. *Artificial Intelligence in Medicine*. DOI: 10.1016/j.artmed.2014.03.003

[41] Davis DA, Chawla NV (2011) **Exploring and exploiting disease interactions from multi-relational gene and phenotype networks**. *PLoS ONE*. DOI: 10.1371/journal.pone.0022670

[42] Guo X, Gao L, Wei C, Yang X, Zhao Y, et al. (2011) **A computational method based on the integration of heterogeneous networks for predicting disease-gene associations**. *PLoS ONE*. DOI: 10.1371/journal.pone.0024171

[43] Davis D, Lichtenwalter R, Chawla NV (2012) **Supervised methods for multi-relational link prediction**. *Soc Netw Anal Min*. DOI: 10.1007/s13278-012-0068-6

[44] Wang W, Yang S, Li J (2012) DOI: 10.1142/9789814447973_0006

[45] Li Y, Patra JC (2010) **Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network**. *Bioinformatics*. DOI: 10.1093/bioinformatics/btq108

[46] Li J, Lu Z (2012) DOI: 10.1109/bibm.2012.6392722

[47] Radivojac P, Peng K, Clark WT, Peters BJ, Mohan A, et al. (2008) **An integrated approach to inferring gene-disease associations in humans**. *Proteins*. DOI: 10.1002/prot.21989

[48] Zitnik M, Zupan B (2015) **Data fusion by matrix factorization**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. DOI: 10.1109/tpami.2014.2343973

[49] Žitnik M, Zupan B (2013) DOI: 10.1142/9789814583220_0038

[50] Žitnik M, Janjić V, Larminie C, Zupan B, Pržulj N (2013) **Discovering disease-disease associations by fusing systems-level molecular data**. *Sci Rep*. DOI: 10.1038/srep03202

[51] Gligorijevi V, Janji V, ulj NP (2014) **Integration of molecular network data reconstructs gene ontology**. *Bioinformatics*. DOI: 10.1093/bioinformatics/btu470

[52] Sun Y, Barber R, Gupta M, Aggarwal CC, Han J (2011) DOI: 10.1109/asonam.2011.112

[53] Sun Y, Han J (2012) **Mining heterogeneous information networks: Principles and methodologies**. *Synthesis Lectures on Data Mining and Knowledge Discovery*. DOI: 10.2200/s00433ed1v01y201207dmk005

[54] Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, et al. (2011) **Molecular signatures database (MSigDB) 3.0**. *Bioinformatics*. DOI: 10.1093/bioinformatics/btr260

[55] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles**. *Proceedings of the National Academy of Sciences*. DOI: 10.1073/pnas.0506580102

[56] Kanehisa M (2000) **KEGG: Kyoto encyclopedia of genes and genomes**. *Nucleic Acids Research*. DOI: 10.1093/nar/28.1.27

[57] Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, et al. (2009) **Reactome knowledgebase of human biological pathways and processes**. *Nucleic Acids Research*. DOI: 10.1093/nar/gkn863

[58] Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, et al. (2005) **Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals**. *Nature*. DOI: 10.1038/nature03441

[59] Matys V (2006) **TRANSFAC(r) and its module TRANSCompel(r): transcriptional gene regulation in eukaryotes**. *Nucleic Acids Research*. DOI: 10.1093/nar/gkj143

[60] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) **Gene ontology: tool for the unification of biology**. *Nature Genetics*. DOI: 10.1038/75556

[61] Schriml LM, Arze C, Nadendla S, Chang YWW, Mazaitis M, et al. (2011) **Disease ontology: a backbone for disease semantic integration**. *Nucleic Acids Research*. DOI: 10.1093/nar/gkr972

[62] Gray KA, Daugherty LC, Gordon SM, Seal RL, Wright MW, et al. (2012) **Genenames.org: the HGNC resources in 2013**. *Nucleic Acids Research*. DOI: 10.1093/nar/gks1066

[63] Gremse M, Chang A, Schomburg I, Grote A, Scheer M, et al. (2010) **The BRENDA tissue ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources**. *Nucleic Acids Research*. DOI: 10.1093/nar/gkq968

[64] Brentani H, Caballero OL, Camargo AA, da Silva AM, da Silva WA, et al. (2003) **The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags**. *Proceedings of the National Academy of Sciences*. DOI: 10.1073/pnas.1233632100

[65] Segal E, Friedman N, Koller D, Regev A (2004) **A module map showing conditional activity of expression modules in cancer**. *Nature Genetics*. DOI: 10.1038/ng1434

[66] Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, et al. (2009) **NCBI GEO: archive for high-throughput functional genomic data**. *Nucleic Acids Research*. DOI: 10.1093/nar/gkn764

[67] Welter D, MacArthur J, Morales J, Burdett T, Hall P, et al. (2013) **The NHGRI GWAS catalog, a curated resource of SNP-trait associations**. *Nucleic Acids Research*. DOI: 10.1093/nar/gkt1229

[68] Fleuren WWM, Verhoeven S, Frijters R, Heupers B, Polman J, et al. (2011) **CoPub update: CoPub 5.0 a text mining system to answer biological questions**. *Nucleic Acids Research*. DOI: 10.1093/nar/gkr310

[69] Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, et al. (2004) **A gene atlas of the mouse and human protein-encoding transcriptomes**. *Proceedings of the National Academy of Sciences*. DOI: 10.1073/pnas.0400782101

[70] Razick S, Magklaras G, Donaldson IM (2008) **iRefIndex: A consolidated protein interaction database with provenance**. *BMC Bioinformatics*. DOI: 10.1186/1471-2105-9-405

[71] Zou H, Hastie T (2005) **Regularization and variable selection via the elastic net**. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. DOI: 10.1111/j.1467-9868.2005.00503.x

[72] Gillis J, Pavlidis P (2011) **The impact of multifunctional genes on "guilt by association" analysis**. *PLoS ONE*. DOI: 10.1371/journal.pone.0017258

[73] Chiorazzi N, Rai KR, Ferrarini M (2005) **Chronic lymphocytic leukemia**. *New England Journal of Medicine*. DOI: 10.1056/nejmra041720

[74] Sawcer S, Hellenthal G, Pirinen M, Spencer CCA, Patsopoulos NA, et al. (2011) **Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis**. *Nature*. DOI: 10.1038/nature10251

[75] Patsopoulos NA, de Bakker PIW (2011) **Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci**. *Annals of Neurology*. DOI: 10.1002/ana.22609

[76] Liu JZ, Mcrae AF, Nyholt DR, Medland SE, Wray NR, et al. (2010) **A versatile gene-based test for genome-wide association studies**. *The American Journal of Human Genetics*. DOI: 10.1016/j.ajhg.2010.06.009

[77] Conti L, Palma RD, Rolla S, Boselli D, Rodolico G, et al. (2012) **Th17 cells in multiple sclerosis express higher levels of JAK2, which increases their surface expression of IFN- r2**. *The Journal of Immunology*. DOI: 10.4049/jimmunol.1004013

[78] Dubois PCA, Trynka G, Franke L, Hunt KA, Romanos J, et al. (2010) **Multiple common variants for celiac disease influencing immune gene expression**. *Nature Genetics*. DOI: 10.1038/ng.543

[79] Evans DM, Spencer CCA, Pointon JJ, Su Z, Harvey D, et al. (2011) **Interaction between ERAP1 and HLA-b27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-b27 in disease susceptibility**. *Nature Genetics*. DOI: 10.1038/ng.873

[80] Jeffries M, Dozmorov M, Tang Y, Merrill JT, Wren JD, et al. (2011) **Genome-wide DNA methylation patterns in CD4+ t cells from patients with systemic lupus erythematosus**. *Epigenetics*. DOI: 10.4161/epi.6.5.15374

[81] Beecham AH, Patsopoulos NA, Xifara DK, Davis MF, Kemppinen A, et al. (2013) **Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis**. *Nature Genetics*. DOI: 10.1038/ng.2770

[82] Hangauer MJ, Vaughn IW, McManus MT (2013) **Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs**. *PLoS Genetics*. DOI: 10.1371/journal.pgen.1003569

[83] Gilmore TD, Kalaitzidis D, Liang MC, Starczynowski DT (2004) **The c-rel transcription factor and b-cell proliferation: a deal with the devil**. *Oncogene*. DOI: 10.1038/sj.onc.1207410

[84] Hilliard BA, Mason N, Xu L, Sun J, Lamhamedi-Cherradi SE, et al. (2002) **Critical roles of c-rel in autoimmune inflammation and helper t cell differentiation**. *Journal of Clinical Investigation*. DOI: 10.1172/jci0215254

[85] Lage K, Hansen NT, Karlberg EO, Eklund AC, Roque FS, et al. (2008) **A large-scale analysis of tissue-specific pathology and gene expression of human dis-**

ease genes and complexes. *Proceedings of the National Academy of Sciences*. DOI: 10.1073/pnas.0810772105

[86] van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JAM (2006) **A text-mining analysis of the human phenome**. *Eur J Hum Genet*. DOI: 10.1038/sj.ejhg.5201585

[87] Sivakumaran S, Agakov F, Theodoratou E, Prendergast JG, Zgaga L, et al. (2011) **Abundant pleiotropy in human complex diseases and traits**. *The American Journal of Human Genetics*. DOI: 10.1016/j.ajhg.2011.10.004

[88] Cotsapas C, Voight BF, Rossin E, Lage K, Neale BM, et al. (2011) **Pervasive sharing of genetic effects in autoimmune disease**. *PLoS Genetics*. DOI: 10.1371/journal.pgen.1002254

[89] Stephens M, Balding DJ (2009) **Bayesian statistical methods for genetic association studies**. *Nat Rev Genet*. DOI: 10.1038/nrg2615

[90] Venkatesan K, Rual JF, Vazquez A, Stelzl U, Lemmens I, et al. (2008) **An empirical framework for binary interactome mapping**. *Nature Methods*. DOI: 10.1038/nmeth.1280

[91] Gillis J, Ballouz S, Pavlidis P (2014) **Bias tradeoffs in the creation and analysis of protein–protein interaction networks**. *Journal of Proteomics*. DOI: 10.1016/j.jprot.2014.01.020

[92] Hidalgo CA, Blumm N, Barabási AL, Christakis NA (2009) **A dynamic network approach for the study of human phenotypes**. *PLoS Comput Biol*. DOI: 10.1371/journal.pcbi.1000353

[93] Sawcer S (2008) **The complex genetics of multiple sclerosis: pitfalls and prospects**. *Brain*. DOI: 10.1093/brain/awn081

[94] Stojmirovic A, Yu YK (2011) **ppiTrim: constructing non-redundant and up-to-date interactomes**. *Database*. DOI: 10.1093/database/bar036

[95] Friedman J, Hastie T, Tibshirani R (2010) **Regularization paths for generalized linear models via coordinate descent**. *Journal of Statistical Software*. DOI: 10.18637/jss.v033.i01

[96] Schielzeth H (2010) **Simple means to improve the interpretability of regression coefficients**. *Methods in Ecology and Evolution*. DOI: 10.1111/j.2041-210x.2010.00012.x

[97] Swamidass SJ, Azencott CA, Daily K, Baldi P (2010) **A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval**. *Bioinformatics*. DOI: 10.1093/bioinformatics/btq140

[98] DeLong ER, DeLong DM, Clarke-Pearson DL (1988) **Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach**. *Biometrics*. DOI: 10.2307/2531595

[99] Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, et al. (2004) **Gene map of the extended human MHC**. *Nat Rev Genet*. DOI: 10.1038/nrg1489

[100] DiMasi JA, Grabowski HG, Hansen RW (2016) **Innovation in the pharmaceutical industry: New estimates of r&d costs**. *Journal of Health Economics*. DOI: 10.1016/j.jhealeco.2016.01.012

[101] Reichert JM (2003) **A guide to drug discovery: Trends in development and approval times for new therapeutics in the united states**. *Nature Reviews Drug Discovery*. DOI: 10.1038/nrd1178

[102] Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J (2014) **Clinical development success rates for investigational drugs**. *Nat Biotechnol*. DOI: 10.1038/nbt.2786

[103] Scannell JW, Blanckley A, Boldon H, Warrington B (2012) **Diagnosing the decline in pharmaceutical r&d efficiency**. *Nature Reviews Drug Discovery.* DOI: 10.1038/nrd3681

[104] Ashburn TT, Thor KB (2004) **Drug repositioning: identifying and developing new uses for existing drugs**. *Nature Reviews Drug Discovery.* DOI: 10.1038/nrd1468

[105] Xu H, Aldrich MC, Chen Q, Liu H, Peterson NB, et al. (2014) **Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality**. *Journal of the American Medical Informatics Association.* DOI: 10.1136/amiajnl-2014-002649

[106] Brilliant MH, Vaziri K, Connor TB, Schwartz SG, Carroll JJ, et al. (2016) **Mining retrospective data for virtual prospective drug repurposing: L-DOPA and age-related macular degeneration**. *The American Journal of Medicine.* DOI: 10.1016/j.amjmed.2015.10.015

[107] Tatonetti NP, Ye PP, Daneshjou R, Altman RB (2012) **Data-driven prediction of drug effects and interactions**. *Science Translational Medicine.* DOI: 10.1126/scitranslmed.3003377

[108] Roth BL, Sheffler DJ, Kroeze WK (2004) **Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia**. *Nature Reviews Drug Discovery.* DOI: 10.1038/nrd1346

[109] Hopkins AL (2008) **Network pharmacology: the next paradigm in drug discovery**. *Nature Chemical Biology.* DOI: 10.1038/nchembio.118

[110] Hopkins AL (2007) **Network pharmacology**. *Nat Biotechnol.* DOI: 10.1038/nbt1007-1110

[111] Swinney DC, Anthony J (2011) **How were new medicines discovered?** *Nature Reviews Drug Discovery.* DOI: 10.1038/nrd3480

[112] Iskar M, Zeller G, Zhao XM, van Noort V, Bork P (2012) **Drug discovery in the age of systems biology: the rise of computational approaches for data integration**. *Current Opinion in Biotechnology*. DOI: 10.1016/j.copbio.2011.11.010

[113] Lamb J (2007) **The connectivity map: a new tool for biomedical research**. *Nature Reviews Cancer*. DOI: 10.1038/nrc2044

[114] Qu XA, Rajpal DK (2012) **Applications of connectivity map in drug discovery and development**. *Drug Discovery Today*. DOI: 10.1016/j.drudis.2012.07.017

[115] Hurle MR, Yang L, Xie Q, Rajpal DK, Sanseau P, et al. (2013) **Computational drug repositioning: From data to therapeutics**. *Clin Pharmacol Ther*. DOI: 10.1038/clpt.2013.1

[116] Liu Z, Fang H, Reagan K, Xu X, Mendrick DL, et al. (2013) **In silico drug repositioning – what we need to know**. *Drug Discovery Today*. DOI: 10.1016/j.drudis.2012.08.005

[117] Himmelstein DS, Baranzini SE (2015) **Heterogeneous network edge prediction: A data integration approach to prioritize disease-associated genes**. *PLOS Computational Biology*. DOI: 10.1371/journal.pcbi.1004259

[118] Gottlieb A, Stein GY, Ruppin E, Sharan R (2014) **PREDICT: a method for inferring novel drug indications with application to personalized medicine**. *Molecular Systems Biology*. DOI: 10.1038/msb.2011.26

[119] Cheng J, Yang L, Kumar V, Agarwal P (2014) **Systematic evaluation of connectivity map for disease indications**. *Genome Medicine*. DOI: 10.1186/s13073-014-0095-1

[120] Guney E, Menche J, Vidal M, Barábasi AL (2016) **Network-based in silico drug efficacy screening**. *Nature Communications*. DOI: 10.1038/ncomms10331

[121] Chiang AP, Butte AJ (2009) **Systematic evaluation of drug–disease relationships to identify leads for novel drug uses**. *Clin Pharmacol Ther*. DOI: 10.1038/clpt.2009.103

[122] Lamb J (2006) **The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease**. *Science*. DOI: 10.1126/science.1132939

[123] Iorio F, Rittman T, Ge H, Menden M, Saez-Rodriguez J (2013) **Transcriptional data: a new gateway to drug repositioning?** *Drug Discovery Today*. DOI: 10.1016/j.drudis.2012.07.014

[124] Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, et al. (2015) **The support of human genetic evidence for approved drug indications**. *Nature Genetics*. DOI: 10.1038/ng.3314

[125] Sanseau P, Agarwal P, Barnes MR, Pastinen T, Richards JB, et al. (2012) **Use of genome-wide association studies for drug repositioning**. *Nat Biotechnol*. DOI: 10.1038/nbt.2151

[126] Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P (2008) **Drug target identification using side-effect similarity**. *Science*. DOI: 10.1126/science.1158140

[127] Nugent T, Plachouras V, Leidner JL (2016) **Computational drug repositioning based on side-effects mined from social media**. *PeerJ Computer Science*. DOI: 10.7717/peerj-cs.46

[128] Zhou X, Menche J, Barabási AL, Sharma A (2014) **Human symptoms–disease network**. *Nature Communications*. DOI: 10.1038/ncomms5212

[129] Pratanwanich N, Lió P (2014) **Pathway-based bayesian inference of drug–disease interactions**. *Mol BioSyst*. DOI: 10.1039/c4mb00014e

[130] Himmelstein D, Lizee A (2016). **Computing standardized logistic regression coefficients**. *Thinklab*. DOI: 10.15363/thinklab.d205

[131] Himmelstein D (2016). **Predictions of whether a compound treats a disease**. *Thinklab*. DOI: 10.15363/thinklab.d203

[132] Bonate PL, Arthaud L, Cantrell WR, Stephenson K, Secrist JA, et al. (2006) **Discovery and development of clofarabine: a nucleoside analogue for treating cancer**. *Nature Reviews Drug Discovery*. DOI: 10.1038/nrd2055

[133] Giovannoni G, Comi G, Cook S, Rammohan K, Rieckmann P, et al. (2010) **A placebo-controlled trial of oral cladribine for relapsing multiple sclerosis**. *New England Journal of Medicine*. DOI: 10.1056/nejmoa0902533

[134] Pfeuffer S, Ruck T, Kleinschnitz C, Wiendl H, Meuth SG (2016) **Failed, interrupted and inconclusive trials on relapsing multiple sclerosis treatment: update 2010–2015**. *Expert Review of Neurotherapeutics*. DOI: 10.1080/14737175.2016.1176531

[135] Massacesi L, Tramacere I, Amoroso S, Battaglia MA, Benedetti MD, et al. (2014) **Azathioprine versus beta interferons for relapsing-remitting multiple sclerosis: A multicentre randomized non-inferiority trial**. *PLoS ONE*. DOI: 10.1371/journal.pone.0113371

[136] Chen PPS (1997) **English, chinese and ER diagrams**. *Data & Knowledge Engineering*. DOI: 10.1016/s0169-023x(97)00017-7

[137] Himmelstein D, Jensen LJ, Khankhanian P (2016). **Data nomenclature: naming and abbreviating our network types**. *Thinklab*. DOI: 10.15363/thinklab.d162

[138] Malone J, Stevens R, Jupp S, Hancocks T, Parkinson H, et al. (2016) **Ten simple rules for selecting a bio-ontology**. *PLOS Computational Biology*. DOI: 10.1371/journal.pcbi.1004743

[139] Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, et al. (2014) **Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data**. *Nucleic Acids Research*. DOI: 10.1093/nar/gku1011

[140] Himmelstein D (2015). **Unifying disease vocabularies**. *Thinklab*. DOI: 10.15363/thinklab.d44

[141] Himmelstein DS (2016). **User-friendly extensions to the disease ontology v1.0**. *Zenodo*. DOI: 10.5281/zenodo.45584

[142] Wu TJ, Schriml LM, Chen QR, Colbert M, Crichton DJ, et al. (2015) **Generating a focused view of disease ontology cancer terms for pan-cancer data integration and analysis**. *Database*. DOI: 10.1093/database/bav032

[143] Himmelstein D, Pankov A (2015). **Mining knowledge from MEDLINE articles and their indexed MeSH terms**. *Thinklab*. DOI: 10.15363/thinklab.d67

[144] Himmelstein DS (2016). **User-friendly extensions to mesh v1.0**. *Zenodo*. DOI: 10.5281/zenodo.45586

[145] Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, et al. (2013) **DrugBank 4.0: shedding new light on drug metabolism**. *Nucleic Acids Research*. DOI: 10.1093/nar/gkt1068

[146] Himmelstein D (2015). **Unifying drug vocabularies**. *Thinklab*. DOI: 10.15363/thinklab.d40

[147] Himmelstein DS (2016). **User-friendly extensions of the drugbank database v1.0**. *Zenodo*. DOI: 10.5281/zenodo.45579

[148] Kuhn M, Letunic I, Jensen LJ, Bork P (2015) **The SIDER database of drugs and side effects**. *Nucleic Acids Res*. DOI: 10.1093/nar/gkv1075

[149] Himmelstein D (2015). **Extracting side effects from SIDER 4**. *Thinklab*. DOI: 10.15363/thinklab.d97

[150] Himmelstein DS (2016). **Extracting tidy and user-friendly tsvs from sider 4.1**. *Zenodo*. DOI: 10.5281/zenodo.45521

[151] Bodenreider O (2004) **The unified medical language system (UMLS): integrating biomedical terminology**. *Nucleic Acids Research*. DOI: 10.1093/nar/gkh061

[152] Himmelstein D (2016). **Incorporating DrugCentral data in our network**. *Thinklab*. DOI: 10.15363/thinklab.d186

[153] Maglott D, Ostell J, Pruitt KD, Tatusova T (2010) **Entrez gene: gene-centered information at NCBI**. *Nucleic Acids Research*. DOI: 10.1093/nar/gkq1237

[154] Himmelstein D, Greene C, Pico A (2015). **Using entrez gene as our gene vocabulary**. *Thinklab*. DOI: 10.15363/thinklab.d34

[155] Himmelstein DS (2016). **Processed entrez gene datasets for humans v1.0**. *Zenodo*. DOI: 10.5281/zenodo.45524

[156] Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA (2012) **Uberon, an integrative multi-species anatomy ontology**. *Genome Biol*. DOI: 10.1186/gb-2012-13-1-r5

[157] Malladi V, Himmelstein D, Mungall C (2015). **Tissue node**. *Thinklab*. DOI: 10.15363/thinklab.d41

[158] Himmelstein DS (2016). **User-friendly anatomical structures data from the uberon ontology v1.0**. *Zenodo*. DOI: 10.5281/zenodo.45527

[159] Kutmon M, Riutta A, Nunes N, Hanspers K, Willighagen EL, et al. (2015) **WikiPathways: capturing the full diversity of pathway knowledge**. *Nucleic Acids Res*. DOI: 10.1093/nar/gkv1024

[160] Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, et al. (2008) **WikiPathways: Pathway editing for the people**. *PLoS Biology*. DOI: 10.1371/journal.pbio.0060184

[161] Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, et al. (2015) **The reactome pathway knowledgebase**. *Nucleic Acids Res*. DOI: 10.1093/nar/gkv1351

[162] Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, et al. (2009) **PID: the pathway interaction database**. *Nucleic Acids Research*. DOI: 10.1093/nar/gkn653

[163] Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, et al. (2010) **Pathway commons, a web resource for biological pathway data**. *Nucleic Acids Research*. DOI: 10.1093/nar/gkq1039

[164] Pico A, Himmelstein D (2015). **Adding pathway resources to your network**. *Thinklab*. DOI: 10.15363/thinklab.d72

[165] Himmelstein DS, Pico AR (2016). **dhimmel/pathways v2.0: Compiling human pathway gene sets**. *Zenodo*. DOI: 10.5281/zenodo.48810

[166] Himmelstein D (2015). **Disease ontology feature requests**. *Thinklab*. DOI: 10.15363/thinklab.d68

[167] Hersey A, Chambers J, Bellis L, Bento AP, Gaulton A, et al. (2015) **Chemical databases: curation or integration by user-defined equivalence?** *Drug Discovery Today: Technologies*. DOI: 10.1016/j.ddtec.2015.01.005

[168] Chambers J, Davies M, Gaulton A, Hersey A, Velankar S, et al. (2013) **UniChem: a unified chemical structure cross-referencing and identifier tracking system**. *Journal of Cheminformatics*. DOI: 10.1186/1758-2946-5-3

[169] Chambers J, Davies M, Gaulton A, Papadatos G, Hersey A, et al. (2014) **UniChem: extension of InChI-based compound mapping to salt, connectivity and stereochemistry layers**. *J Cheminform*. DOI: 10.1186/s13321-014-0043-5

[170] Heller S, McNaught A, Stein S, Tchekhovskoi D, Pletnev I (2013) **InChI - the worldwide chemical structure identifier standard**. *Journal of Cheminformatics*. DOI: 10.1186/1758-2946-5-7

[171] Himmelstein D, Bastian F, Baranzini S (2016). **dhimmel/bgee v1.0: Anatomy-specific gene expression in humans from bgee**. *Zenodo*. DOI: 10.5281/zenodo.47157

[172] Himmelstein D, Bastian F (2015). **Processing bgee for tissue-specific gene presence and over/under-expression**. *Thinklab*. DOI: 10.15363/thinklab.d124

[173] Himmelstein D, Bastian F (2015). **Tissue-specific gene expression resources**. *Thinklab*. DOI: 10.15363/thinklab.d81

[174] Bastian F, Parmentier G, Roux J, Moretti S, Laudet V, et al. (2008) **Bgee: Integrating and comparing heterogeneous transcriptome data among species**. In: Lecture Notes in Computer Science, Springer Science + Business Media. pp. 124–131. DOI: 10.1007/978-3-540-69828-9_12

[175] Santos A, Tsafou K, Stolte C, Pletscher-Frankild S, O'Donoghue SI, et al. (2015) **Comprehensive comparison of large-scale tissue expression datasets**. *PeerJ*. DOI: 10.7717/peerj.1054

[176] Himmelstein D, Jensen LJ (2015). **Genetissue relationships from the tissues database**. *Zenodo*. DOI: 10.5281/zenodo.27244

[177] Himmelstein D, Jensen LJ (2015). **The TISSUES resource for the tissue-specificity of genes**. *Thinklab*. DOI: 10.15363/thinklab.d91

[178] Chen X, Liu M, Gilson M (2001) **BindingDB: A web-accessible molecular recognition database**. *Combinatorial Chemistry & High Throughput Screening*. DOI: 10.2174/1386207013330670

[179] Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, et al. (2015) **BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology**. *Nucleic Acids Res*. DOI: 10.1093/nar/gkv1072

[180] Wishart DS (2006) **DrugBank: a comprehensive resource for in silico drug discovery and exploration**. *Nucleic Acids Research*. DOI: 10.1093/nar/gkj067

[181] Himmelstein D, Gilson M (2015). **Integrating drug target information from BindingDB**. *Thinklab*. DOI: 10.15363/thinklab.d53

[182] Himmelstein D, Gilson M, Baranzini S (2015). **Processing the october 2015 bindingdb**. *Zenodo*. DOI: 10.5281/zenodo.33987

[183] Himmelstein D, Chen S (2015). **Protein (target, carrier, transporter, and enzyme) interactions in DrugBank**. *Thinklab*. DOI: 10.15363/thinklab.d65

[184] Himmelstein D, Chen S (2015). **Calculating molecular similarities between DrugBank compounds**. *Thinklab*. DOI: 10.15363/thinklab.d70

[185] Himmelstein D, Brueggeman L, Baranzini S (2015). **Pairwise molecular similarities between drugbank compounds**. *Figshare*. DOI: 10.6084/m9.figshare.1418386

[186] Dice LR (1945) **Measures of the amount of ecologic association between species**. *Ecology*. DOI: 10.2307/1932409

[187] Rogers D, Hahn M (2010) **Extended-connectivity fingerprints**. *Journal of Chemical Information and Modeling*. DOI: 10.1021/ci100050t

[188] Morgan HL (1965) **The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service.** *J Chem Doc*. DOI: 10.1021/c160017a018

[189] Himmelstein DS, Baranzini SE (2016). **dhimmel/gwas-catalog v1.0: Extracting genedisease associations from the gwas catalog**. *Zenodo*. DOI: 10.5281/zenodo.48428

[190] Himmelstein D, Jensen LJ (2015). **Processing the DISEASES resource for disease–gene relationships**. *Thinklab*. DOI: 10.15363/thinklab.d106

[191] Himmelstein DS, Jensen LJ (2016). **dhimmel/diseases v1.0: Processing the diseases database of genedisease associations**. *Zenodo*. DOI: 10.5281/zenodo.48425

[192] Himmelstein D, janet piñero (2015). **Processing DisGeNET for disease-gene relationships**. *Thinklab*. DOI: 10.15363/thinklab.d105

[193] Himmelstein DS, Piero J (2016). **dhimmel/disgenet v1.0: Processing the disgenet database of genedisease associations**. *Zenodo*. DOI: 10.5281/zenodo.48426

[194] Himmelstein D (2015). **Functional disease annotations for genes using DOAF**. *Thinklab*. DOI: 10.15363/thinklab.d94

[195] Himmelstein DS (2016). **dhimmel/doaf v1.0: Processing the doaf database of genedisease associations**. *Zenodo*. DOI: 10.5281/zenodo.48427

[196] Himmelstein D (2015). **Extracting disease-gene associations from the GWAS catalog**. *Thinklab*. DOI: 10.15363/thinklab.d80

[197] Himmelstein D, Sirota M, Way G (2015). **Calculating genomic windows for GWAS lead SNPs**. *Thinklab*. DOI: 10.15363/thinklab.d71

[198] Pletscher-Frankild S, Pallejà A, Tsafou K, Binder JX, Jensen LJ (2015) **DISEASES: Text mining and data integration of disease–gene associations**. *Methods*. DOI: 10.1016/j.ymeth.2014.11.020

[199] Pinero J, Queralt-Rosinach N, Bravo A, Deu-Pons J, Bauer-Mehren A, et al. (2015) **DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes**. *Database*. DOI: 10.1093/database/bav028

[200] Xu W, Wang H, Cheng W, Fu D, Xia T, et al. (2012) **A framework for annotating human genome in disease context**. *PLoS ONE*. DOI: 10.1371/journal.pone.0049686

[201] Himmelstein D, Bastian F, Hadley D, Greene C (2015). **STARGEO: expression signatures for disease using crowdsourced GEO annotation**. *Thinklab*. DOI: 10.15363/thinklab.d96

[202] Himmelstein D, Hadley D, Schepanovski A (2016). **dhimmel/stargeo v1.0: differentially expressed genes for 48 diseases from stargeo**. *Zenodo*. DOI: 10.5281/zenodo.46866

[203] Himmelstein DS (2016). **dhimmel/medline v1.0: Disease, symptom, and anatomy cooccurence in medline**. *Zenodo*. DOI: 10.5281/zenodo.48445

[204] Himmelstein D (2015). **Disease similarity from MEDLINE topic cooccurrence**. *Thinklab*. DOI: 10.15363/thinklab.d93

[205] Fisher RA (1922) **On the interpretation of chi-squared from contingency tables, and the calculation of p**. *Journal of the Royal Statistical Society*. DOI: 10.2307/2340521

[206] Priedigkeit N, Wolfe N, Clark NL (2015) **Evolutionary signatures amongst disease genes permit novel methods for gene prioritization and construction of informative gene-based networks**. *PLoS Genet*. DOI: 10.1371/journal.pgen.1004967

[207] Himmelstein D, Partha R (2015). **Selecting informative ERC (evolutionary rate covariation) values between genes**. *Thinklab*. DOI: 10.15363/thinklab.d57

[208] Himmelstein DS (2016). **dhimmel/erc v1.0: Processing human evolutionary rate covaration data**. *Zenodo*. DOI: 10.5281/zenodo.48444

[209] Himmelstein D, Hadley D, Strokach A (2015). **Creating a catalog of protein interactions**. *Thinklab*. DOI: 10.15363/thinklab.d85

[210] Himmelstein DS, Baranzini SE (2016). **dhimmel/ppi v1.0: Compiling a human protein interaction catalog**. *Zenodo*. DOI: 10.5281/zenodo.48443

[211] Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, et al. (2005) **Towards a proteome-scale map of the human protein–protein interaction network**. *Nature*. DOI: 10.1038/nature04209

[212] Yu H, Tardivo L, Tam S, Weiner E, Gebreab F, et al. (2011) **Next-generation sequencing to generate interactome datasets**. *Nature Methods*. DOI: 10.1038/nmeth.1597

[213] Rolland T, Taşan M, Charloteaux B, Pevzner SJ, Zhong Q, et al. (2014) **A proteome-scale map of the human interactome network**. *Cell*. DOI: 10.1016/j.cell.2014.10.050

[214] Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, et al. (2015) **Uncovering disease-disease relationships through the incomplete interactome**. *Science*. DOI: 10.1126/science.1257601

[215] Huntley RP, Sawford T, Mutowo-Meullenet P, Shypitsyna A, Bonilla C, et al. (2014) **The GOA database: Gene ontology annotation updates for 2015**. *Nucleic Acids Research*. DOI: 10.1093/nar/gku1113

[216] Himmelstein D, Greene C, Malladi V, Bastian F (2015). **Compiling gene ontology annotations into an easy-to-use format**. *Thinklab*. DOI: 10.15363/thinklab.d39

[217] Himmelstein D, Greene C, Malladi V, Bastian F, Baranzini S (2015). **gene-ontology: Initial zenodo release**. *Zenodo*. DOI: 10.5281/zenodo.21711

[218] Edgar R (2002) **Gene expression omnibus: NCBI gene expression and hybridization array data repository**. *Nucleic Acids Research*. DOI: 10.1093/nar/30.1.207

[219] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. (2012) **NCBI GEO: archive for functional genomics data sets–update**. *Nucleic Acids Research*. DOI: 10.1093/nar/gks1193

[220] Himmelstein D, Brueggeman L, Baranzini S (2016). **dhimmel/lincs v2.0: Refined consensus signatures from lincs l1000**. *Zenodo*. DOI: 10.5281/zenodo.47223

[221] Himmelstein D, Brueggeman L, Baranzini S (2016). **l1000.db: Sqlite database of lincs l1000 metadata**. *Figshare*. DOI: 10.6084/m9.figshare.3085837.v1

[222] Himmelstein D, Chung C (2015). **Computing consensus transcriptional profiles for LINCS l1000 perturbations**. *Thinklab*. DOI: 10.15363/thinklab.d43

[223] Himmelstein D, Brueggeman L, Baranzini S (2016). **Consensus signatures for lincs l1000 perturbations**. *Figshare*. DOI: 10.6084/m9.figshare.3085426.v1

[224] Himmelstein D (2016). **Assessing the imputation quality of gene expression in LINCS l1000**. *Thinklab*. DOI: 10.15363/thinklab.d185

[225] Himmelstein D, Greene C, Jensen LJ (2016). **Positive correlations between knockdown and overexpression profiles from LINCS l1000**. *Thinklab*. DOI: 10.15363/thinklab.d171

[226] Himmelstein D (2016). **Announcing PharmacotherapyDB: the open catalog of drug therapies for disease**. *Thinklab*. DOI: 10.15363/thinklab.d182

[227] Himmelstein D, Khankhanian P, Hessler CS, Green AJ, Baranzini S (2016). **PharmacotherapyDB 1.0: the open catalog of drug therapies for disease**. *Figshare*. DOI: 10.6084/m9.figshare.3103054

[228] Himmelstein DS, Khankhanian P, Hessler CS, Green AJ, Baranzini SE (2016). **dhimmel/indications v1.0. PharmacotherapyDB: the open catalog of drug therapies for disease**. *Zenodo*. DOI: 10.5281/zenodo.47664

[229] Himmelstein D, Good B, Oprea T, McCoy A, Lizee A (2015). **How should we construct a catalog of drug indications?** *Thinklab*. DOI: 10.15363/thinklab.d21

[230] Wei WQ, Cronin RM, Xu H, Lasko TA, Bastarache L, et al. (2013) **Development and evaluation of an ensemble resource linking medications to their indi-**

cations. *Journal of the American Medical Informatics Association.* DOI: 10.1136/amiajnl-2012-001431

[231] Khare R, Li J, Lu Z (2014) **LabeledIn: Cataloging labeled indications for human drugs**. *Journal of Biomedical Informatics.* DOI: 10.1016/j.jbi.2014.08.004

[232] Khare R, Burger JD, Aberdeen JS, Tresner-Kirsch DW, Corrales TJ, et al. (2015) **Scaling drug indication curation through crowdsourcing**. *Database.* DOI: 10.1093/database/bav016

[233] Himmelstein D, Khare R (2015). **Processing LabeledIn to extract indications**. *Thinklab.* DOI: 10.15363/thinklab.d46

[234] McCoy AB, Wright A, Laxmisan A, Ottosen MJ, McCoy JA, et al. (2012) **Development and evaluation of a crowdsourcing methodology for knowledge base construction: identifying relationships between clinical problems and medications**. *Journal of the American Medical Informatics Association.* DOI: 10.1136/amiajnl-2012-000852

[235] Himmelstein D (2015). **Extracting indications from the ehrlink resource**. *Thinklab.* DOI: 10.15363/thinklab.d62

[236] Himmelstein D, Khankhanian P, Hessler C (2015). **Expert curation of our indication catalog for disease-modifying treatments**. *Thinklab.* DOI: 10.15363/thinklab.d95

[237] Spaulding J, Himmelstein D, Greene C, Good B (2015). **Enabling reproducibility and reuse**. *Thinklab.* DOI: 10.15363/thinklab.d23

[238] Hrynaszkiewicz I (2011) **The need and drive for open data in biomedical publishing**. *Serials: The Journal for the Serials Community.* DOI: 10.1629/2431

[239] Molloy JC (2011) **The open knowledge foundation: Open data means better science**. *PLoS Biology.* DOI: 10.1371/journal.pbio.1001195

[240] Piwowar HA, Vision TJ (2013) **Data reuse and the open data citation advantage**. *PeerJ*. DOI: 10.7717/peerj.175

[241] Stodden V, Miguez S (2014) **Best practices for computational science: Software infrastructure and environments for reproducible and extensible research**. *Journal of Open Research Software*. DOI: 10.5334/jors.ay

[242] Baggerly K (2010) **Disclose all data in publications**. *Nature*. DOI: 10.1038/467401b

[243] Vanpaemel W, Vermorgen M, Deriemaecker L, Storms G (2015) **Are we wasting a good crisis? the availability of psychological research data after the storm**. *Collabra*. DOI: 10.1525/collabra.13

[244] Hrynaszkiewicz I, Cockerill MJ (2012) **Open by default: a proposed copyright license and waiver agreement for open access research and data in peer-reviewed journals**. *BMC Research Notes*. DOI: 10.1186/1756-0500-5-494

[245] Hagedorn G, Mietchen D, Morris R, Agosti D, Penev L, et al. (2011) **Creative commons licenses and the non-commercial condition: Implications for the re-use of biodiversity information**. *ZooKeys*. DOI: 10.3897/zookeys.150.2189

[246] Himmelstein D, Jensen LJ (2015). **One network to rule them all**. *Thinklab*. DOI: 10.15363/thinklab.d102

[247] Elliott R (2005) **Who owns scientific data? the impact of intellectual property rights on the scientific publication chain**. *Learned Publishing*. DOI: 10.1087/0953151053584984

[248] Himmelstein D, Jensen LJ, Smith M, Fortney K, Chung C (2015). **Integrating resources with disparate licensing into an open network**. *Thinklab*. DOI: 10.15363/thinklab.d107

[249] Himmelstein D (2015). **MSigDB licensing**. *Thinklab*. DOI: 10.15363/thinklab.d108

[250] Himmelstein D (2015). **Incomplete interactome licensing**. *Thinklab*. DOI: 10.15363/thinklab.d111

[251] Himmelstein D (2015). **LINCS l1000 licensing**. *Thinklab*. DOI: 10.15363/thinklab.d110

[252] Himmelstein D (2016). **Assessing the effectiveness of our hetnet permutations**. *Thinklab*. DOI: 10.15363/thinklab.d178

[253] Hanhijrvi S, Garriga GC, Puolamki K (2009) **Randomization techniques for graphs**. In: Proceedings of the 2009 SIAM International Conference on Data Mining, Society for Industrial & Applied Mathematics (SIAM). pp. 780–791. DOI: 10.1137/1.9781611972795.67

[254] Himmelstein D (2015). **Permuting hetnets and implementing randomized edge swaps in cypher**. *Thinklab*. DOI: 10.15363/thinklab.d136

[255] Himmelstein D (2015). **Hetnets in python: hetio v0.1.0 initial release**. *Zenodo*. DOI: 10.5281/zenodo.31763

[256] Himmelstein D (2015). **Using the neo4j graph database for hetnets**. *Thinklab*. DOI: 10.15363/thinklab.d112

[257] Himmelstein D, Khankhanian P, Lizee A (2016). **Transforming DWPCs for hetnet edge prediction**. *Thinklab*. DOI: 10.15363/thinklab.d193

[258] Burbidge JB, Magee L, Robb AL (1988) **Alternative transformations to handle extreme values of the dependent variable**. *Journal of the American Statistical Association*. DOI: 10.1080/01621459.1988.10478575

[259] Himmelstein D (2016). **Our hetnet edge prediction methodology: the modeling framework for project rephetio**. *Thinklab*. DOI: 10.15363/thinklab.d210

[260] Himmelstein D (2015). **Assessing the informativeness of features**. *Thinklab*. DOI: 10.15363/thinklab.d115

[261] Himmelstein D (2016). **Edge dropout contamination in hetnet edge prediction**. *Thinklab*. DOI: 10.15363/thinklab.d215

[262] Lizee A, Himmelstein D (2016). **Network edge prediction: Estimating the prior**. *Thinklab*. DOI: 10.15363/thinklab.d201

[263] Lizee A, Himmelstein D (2016). **Network edge prediction: how to deal with self-testing**. *Thinklab*. DOI: 10.15363/thinklab.d194

[264] Himmelstein D (2016). **Cataloging drug–disease therapies in the ClinicalTrials.gov database**. *Thinklab*. DOI: 10.15363/thinklab.d212

[265] Himmelstein D, Lizee A, Brueggeman L, Chen S, Khankhanian P, et al. (2015) **Repurposing drugs on a hetnet**. *Thinklab*. DOI: 10.15363/thinklab.a5

[266] Himmelstein D, Lizee A (2016). **Measuring user contribution and content creation**. *Thinklab*. DOI: 10.15363/thinklab.d200

[267] Patil C, Siegel V (2009) **This revolution will be digitized: online tools for radical collaboration**. *Disease Models & Mechanisms*. DOI: 10.1242/dmm.003285

[268] Mietchen D, Mounce R, Penev L (2015) **Publishing the research process**. *Research Ideas and Outcomes*. DOI: 10.3897/rio.1.e7547

[269] Powell K (2016) **Does it take too long to publish research?** *Nature*. DOI: 10.1038/530148a

[270] Vale RD (2015) **Accelerating scientific publication in biology**. *Proceedings of the National Academy of Sciences*. DOI: 10.1073/pnas.1511912112

[271] Allison DB, Brown AW, George BJ, Kaiser KA (2016) **Reproducibility: A tragedy of errors**. *Nature*. DOI: 10.1038/530027a

[272] Himmelstein D, Keough K, Vysotskiy M, Kim J, Norgeot B, et al. (2016). **Workshop to analyze LINCS data for the systems pharmacology course at UCSF**. *Thinklab*. DOI: 10.15363/thinklab.d181

[273] Waldrop MM (2015) **Why we are teaching science wrong, and how to make it right**. *Nature*. DOI: 10.1038/523272a

[274] Giles J (2012) **Going paperless: The digital lab**. *Nature*. DOI: 10.1038/481430a

[275] Simeonov KP, Himmelstein DS (2015) **Lung cancer incidence decreases with elevation: evidence for oxygen as an inhaled carcinogen**. *PeerJ*. DOI: 10.7717/peerj.705

[276] Weinberg CR, Brown KG, Hoel DG (1987) **Altitude, radiation, and mortality from cancer and heart disease**. *Radiation Research*. DOI: 10.2307/3577265

[277] Pelt WRV (2003) **Epidemiological associations among lung cancer, radon exposure and elevation above sea level—a reassessment of cohen's county level radon study**. *Health Physics*. DOI: 10.1097/00004032-200310000-00002

[278] Himmelstein D (2015). **Publication delays at plos and 3,475 other journals**. *Zenodo*. DOI: 10.5281/zenodo.19117

[279] Woolston C (2015) **Long wait for publication plagues many journals**. *Nature*. DOI: 10.1038/523131f

[280] Vosshall LB (2012) **The glacial pace of scientific publishing: Why it hurts everyone and what we can do to fix it**. *The FASEB Journal*. DOI: 10.1096/fj.12-0901ufm

[281] Himmelstein DS, Powell K (2016). **Analysis for "the history of publishing delays" blog post v1.0**. *Zenodo*. DOI: 10.5281/zenodo.45516

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature

June 2, 2016

Date