Systematic integration of biomedical knowledge prioritizes drugs for repurposing

Daniel S. Himmelstein (daniel.himmelstein@gmail.com), Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L. Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, Sergio E. Baranzini (Sergio.Baranzini@ucsf.edu)

Abstract

The ability to computationally predict whether a compound treats a disease would improve the economy and success rate of drug approval. This study describes the Rephetio Project to systematically model drug efficacy based on 755 existing treatments. First, we constructed Hetionet (neo4j.het.io), an integrative network encoding knowledge from millions of biomedical studies. Hetionet v1.0 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, pharmacologic classes, side effects, and symptoms. Next, we identified network patterns that distinguish treatments from non-treatments. Then we predicted the probability of treatment for 209,168 compound–disease pairs (het.io/repurpose). Our predictions validated in two external datasets, suggesting they will help prioritize drug repurposing candidates. This study was entirely open and received realtime feedback from 36 community members.

Introduction

The cost of developing a new therapeutic drug has been estimated at 1.4 billion dollars [1], the process typically takes 15 years from lead compound to market [2], and the likelihood of success is stunningly low [3]. Strikingly, the costs have been doubling every 9 years since 1970, a sort of inverse Moore's law, which is far from an optimal strategy from both a business and public health perspective [4]. Drug repurposing — identifying novel uses for existing therapeutics — can drastically reduce the duration, failure rates, and costs of approval [5]. These benefits stem from the rich preexisting information on approved drugs, including extensive toxicology profiling performed during development, preclinical models, clinical trials, and postmarketing surveillance.

Drug repurposing is poised to become more efficient as mining of electronic health records (EHRs) to retrospectively assess the effect of drugs gains feasibility [6, 7, 8, 9].

However, systematic approaches to repurpose drugs based on mining EHRs alone will likely lack power due to multiple testing. Similar to the approach followed to increase the power of genome-wide association studies (GWAS) [10, 11], integration of biological knowledge to prioritize drug repurposing will help overcome limited EHR 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 [12]. Network pharmacology assumes diseases consist of a multitude of molecular alterations resulting in a robust disease state. Network pharmacology seeks to uncover multiple points of intervention into a specific pathophysiological state that together rehabilitate an otherwise resilient disease process [13, 14]. 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 [15]. 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 [16, 17, 18]. Furthermore, integration of multiple channels of evidence, particularly diverse types of data, can overcome the limitations and weak performance inherent to data of a single domain [19]. 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 [20, 21].

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 hetnet (Hetionet v1.0) by integrating knowledge and experimental findings from decades of biomedical research spanning millions of publications. We adapted an algorithm originally developed for social network analysis and applied it to Hetionet v1.0 to identify patterns of efficacy and predict new uses for drugs. The algorithm performs edge prediction through a machine learning framework that accommodates the breadth and depth of information contained in Hetionet v1.0 [22, 23]. 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 (i.e. a "gold standard"). Next, we 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 [24]. 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 [25]. 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 [26].

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. By doing this, we were able to capture a multitude of effects that have been suggested as influential for drug repurposing including drug-drug similarity [24, 27], disease-disease similarity [24, 28], transcriptional signatures [25, 29, 17, 30, 18], protein interactions [26], genetic association [31, 32], drug side effects [33, 34], disease symptoms [35], and molecular pathways [36]. Our ability to create such an integrative model of drug efficacy relies on the hetnet data structure to unite diverse information. On Hetionet v1.0, 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.

We refer to this study as Project Rephetio (pronounced as rep-*het*-*ee*-oh). Both Rephetio and Hetionet are portmanteaus combining the words repurpose, heterogeneous, and network with the URL het.io.

Results

Hetionet v1.0

We obtained and integrated data from 29 publicly available resources to create Hetionet v1.0 (Figure 1). The hetnet contains 47,031 nodes of 11 types (Table 1) and 2,250,197 relationships of 24 types (Table 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.

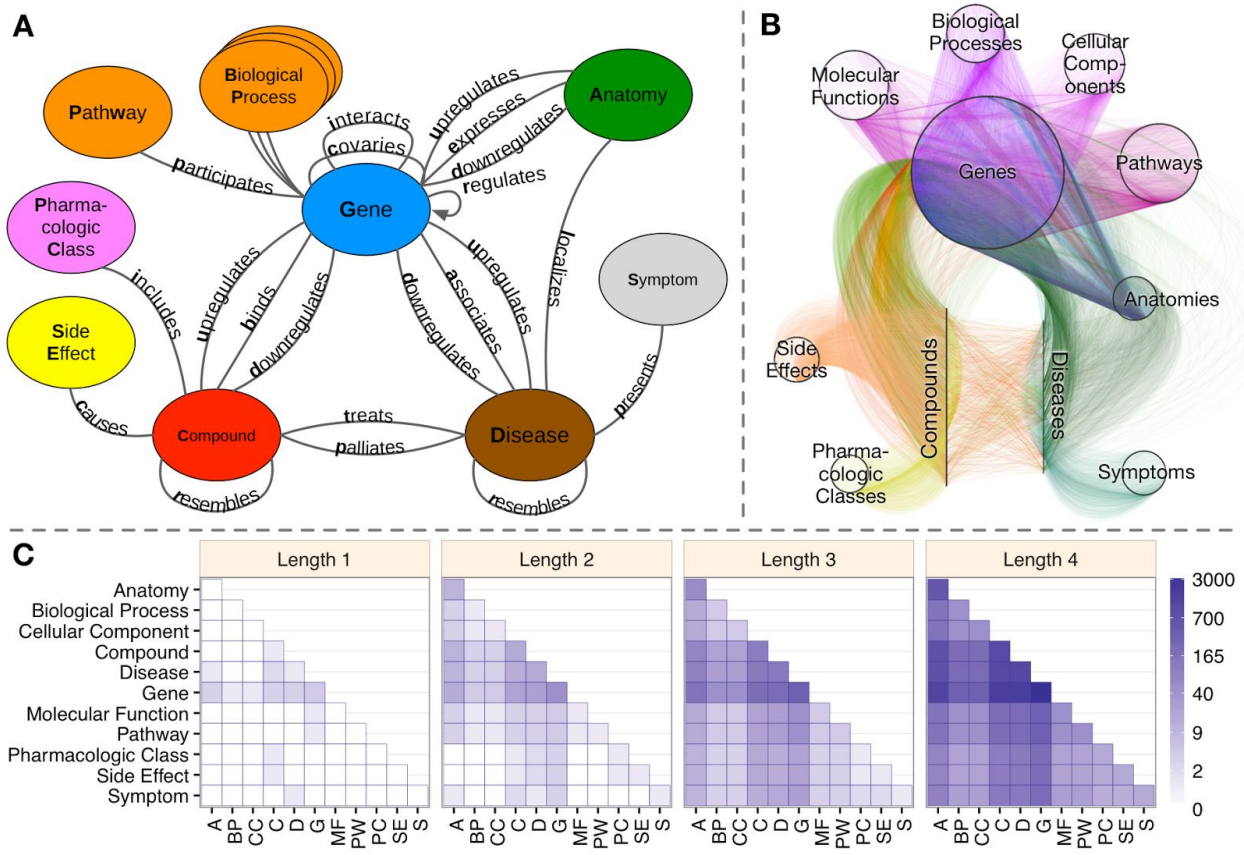

Figure 1. Hetionet v1.0

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-left 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.

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 BindingDB. 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.

Table 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.

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 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. Note that all metaedges besides *Gene regulates Gene* are undirected.

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	Gr>G	265,672	4,634	7,048
Pharmacologic Class–includes–Compound	PCiC	1,029	345	724

While Hetionet v1.0 is ideally suited for drug repurposing, the network has broader biological applicability. For example, we have prototyped queries for a) identifying drugs that target a specific pathway, b) identifying biological processes involved in a specific disease, c) identifying the drug targets responsible for causing a specific side effect, and

d) identifying anatomies with transcriptional relevance for a specific disease [37]. Each of these queries was simple to write and took less than a second to run on our publicly available Hetionet Browser. While it is possible that existing services provide much of the aforementioned functionality, they offer less versatility. Hetionet differentiates itself in its ability to flexibly query across multiple domains of information. As a proof of concept, we enhanced the biological process query (b), which identified processes that were enriched for disease-associated genes, using multiple sclerosis (MS) as an example disease. The enhanced query identified genes that interact with MS GWAS-genes. However, interacting genes were discarded unless they were upregulated in an MS-related anatomy (i.e. anatomical structure, e.g. organ or tissue). Then relevant biological processes were identified. Thus, the single query spanned 4 node and 5 relationship types. Furthermore, the portion of the query to identify paths meeting the above specification required only four lines of Cypher code.

The integrative potential of Hetionet v1.0 is reflected by its connectivity. 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 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 accessible via a Neo4j Browser at https://neo4j.het.io. This public Neo4j instance provides users an installation-free method to query and visualize the network. The Browser contains a tutorial guide as well as guides with the details of each Project Rephetio prediction. Hetionet v1.0 is also available for download 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.

Systematic mechanisms of efficacy

One aim of Project Rephetio was to systematically evaluate how drugs exert their therapeutic potential. To address this question, we compiled 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 traverse from compound to disease and have length of 2–4 (Figure 2A). To control for the different degrees of nodes, we used the degree-weighted path count (*DWPC*) — which downweights paths going through highly-connected nodes [22] — 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.

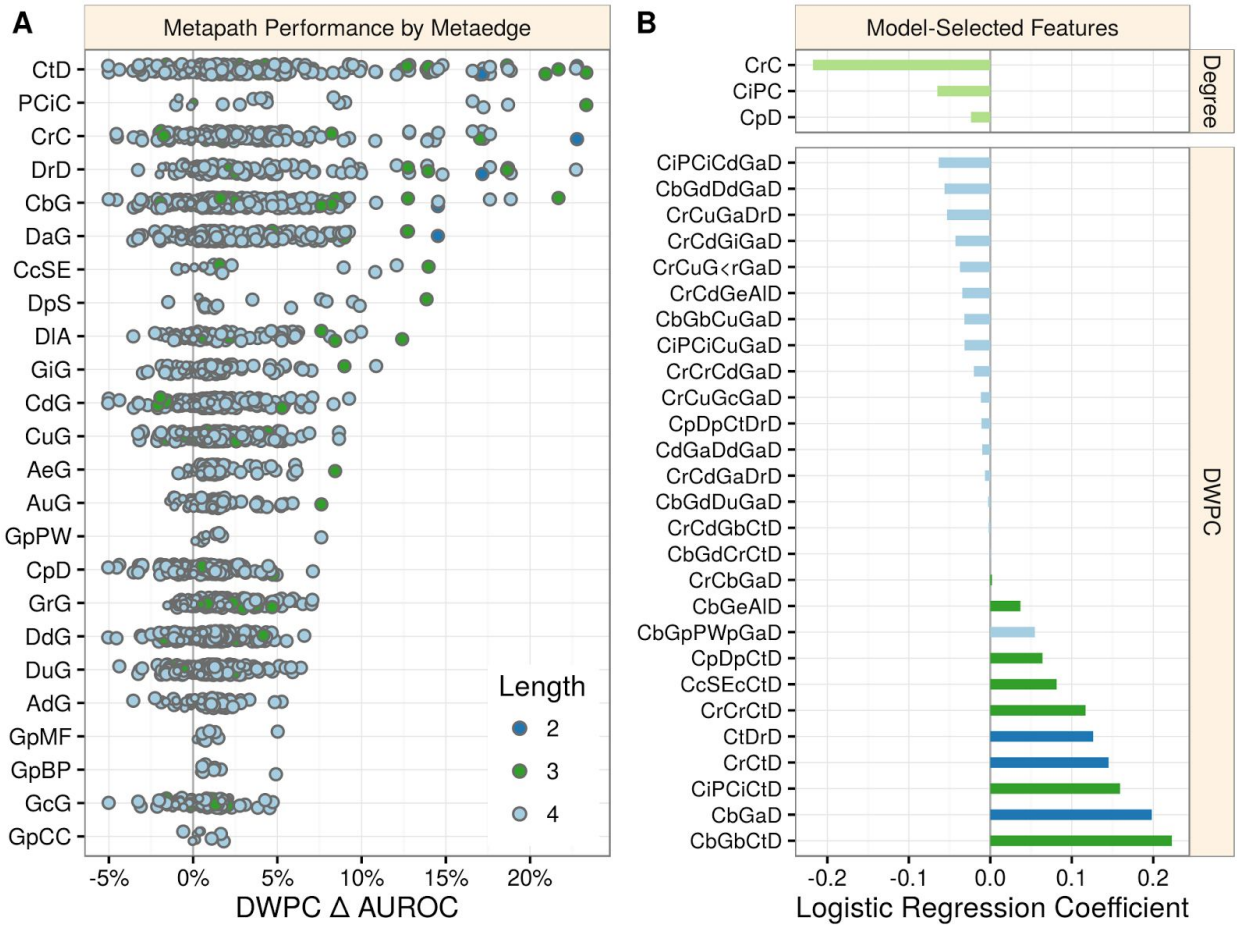

Figure 2. Performance by type and model coefficients

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 [38].

Overall, 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 provided at least some therapeutic 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 3 lists the predictiveness of several metapaths of interest. Refer to the Discussion for our interpretation of these findings.

Table 3. The predictiveness of select metapaths

A small selection of the 1,206 metapaths is provided. Len. refers to number of metaedges composing the metapath. Δ AUROC and −log10(*p*) assess the performance of a metapath's DWPC in discriminating treatments from non-treatments (in the all-features stage as described in Methods). *p* assesses whether permutation affected AUROC. For reference, $p = 0.05$ corresponds to $-\log_{10}(p) = 1.30$. Note that several metapaths shown here provided little evidence that Δ AUROC ≠ 0 underscoring their poor ability to predict whether a compound treated a disease. Coef. reports a metapath's logistic regression coefficient as seen in Figure 2B. Metapaths removed in feature selection have missing coefficients whereas metapaths given zero-weight by the elastic net have coef. = 0.0.

Abbrev.	Len.	Δ AUROC	−log10(*p*)	Coef.	Metapath
CbGaD	2	14.5%	6.2	0.20	Compound–binds–Gene–associates–Disease
CdGuD	2	1.7%	4.5		Compound–downregulates–Gene–upregulates–Disease
CrCtD	2	22.8%	6.9	0.15	Compound–resembles–Compound–treats–Disease
CtDrD	2	17.2%	5.8	0.13	Compound–treats–Disease–resembles–Disease
CuGdD	2	1.1%	2.6		Compound–upregulates–Gene–downregulates–Disease
CbGbCtD	3	21.7%	6.5	0.22	Compound–binds–Gene–binds–Compound–treats–Disease
CbGeAlD	3	8.4%	5.2	0.04	Compound–binds–Gene–expresses–Anatomy–localizes–Disease
CbGiGaD	3	9.0%	4.4	0.00	Compound–binds–Gene–interacts–Gene–associates–Disease

CcSEcCtD	3	14.0%	6.8	0.08	Compound–causes–Side Effect–causes–Compound–treats–Disease
CdGdCtD	3	3.8%	4.6	0.00	Compound–downregulates–Gene–downregulates–Compound–treats–Disease
CdGuCtD	3	-2.1%	2.4		Compound–downregulates–Gene–upregulates–Compound–treats–Disease
CiPCiCtD	3	23.3%	7.5	0.16	Compound–includes–Pharmacologic Class–includes–Compound–treats–Disease
CpDpCtD	3	4.3%	3.9	0.06	Compound–palliates–Disease–palliates–Compound–treats–Disease
CrCrCtD	3	17.0%	5.0	0.12	Compound–resembles–Compound–resembles–Compound–treats–Disease
CtDdGdD	3	4.2%	3.9		Compound–treats–Disease–downregulates–Gene–downregulates–Disease
CtDdGuD	3	0.5%	1.0		Compound–treats–Disease–downregulates–Gene–upregulates–Disease
CtDlAlD	3	12.4%	6.0		Compound–treats–Disease–localizes–Anatomy–localizes–Disease
CtDpSpD	3	13.9%	6.1		Compound–treats–Disease–presents–Symptom–presents–Disease
CtDuGdD	3	0.7%	1.3		Compound–treats–Disease–upregulates–Gene–downregulates–Disease
CtDuGuD	3	1.1%	1.4		Compound–treats–Disease–upregulates–Gene–upregulates–Disease
CuGdCtD	3	-1.6%	2.9		Compound–upregulates–Gene–downregulates–Compound–treats–Disease
CuGuCtD	3	4.4%	3.5	0.00	Compound–upregulates–Gene–upregulates–Compound–treats–Disease
CbGiGiGaD	4	7.0%	5.1	0.00	Compound–binds–Gene–interacts–Gene–interacts–Gene–associates–Disease
CbGpBPpGaD	4	4.9%	3.8	0.00	Compound–binds–Gene–participates–Biological Process–participates–Gene–associates–Disease
CbGpPWpGaD	4	7.6%	7.9	0.05	Compound–binds–Gene–participates–Pathway–participates–Gene–associates–Disease

Predictions of drug efficacy

We implemented a machine learning approach to translate the network connectivity between a compound and a disease into a probability of treatment. The approach relies on the 755 known 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 minimize overfitting, yielding a model with 31 features (Figure 2B). The DWPC features with negative coefficients appear to be included as node-degree-capturing covariates, i.e. they reflect the general connectivity of the compound and disease rather than specific paths between them. However, the 11 DWPC features with non-negligible positive coefficients represent the most salient types of connectivity for systematically modeling drug efficacy. See the metapaths with positive coefficients in Table 3 for unabbreviated names. As an example, the *CcSEcCtD* feature assesses whether the compound causes the same side effects as compounds that treat the disease. Alternatively, the *CbGeAlD* feature assesses whether the compound binds to genes that are expressed in the anatomies affected by the disease.

We applied this 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 [39], available at http://het.io/repurpose/. The 755 known disease-modifying indications were highly ranked (AUROC = 97.4%, Figure 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%). Together, these findings indicate that Project Rephetio has the ability to recognize efficacious compound–disease pairs.

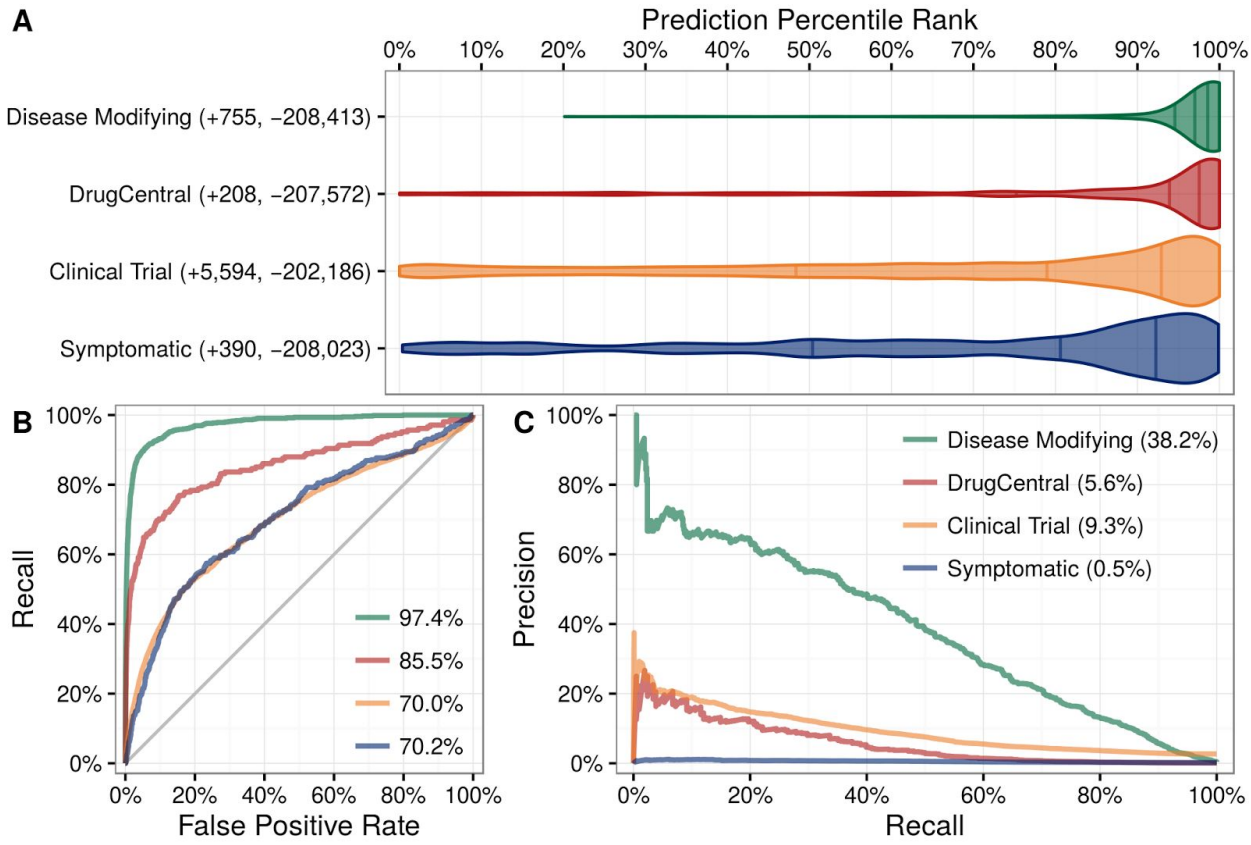

Figure 3. Predictions performance on four indication sets

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% represents a 2-fold enrichment over the null probability. Of the 3,980 predictions with a probability exceeding 1%, 586 corresponded to known disease-modifying indications, leaving 3,394 repurposing candidates. For a given compound or disease, we provide the percentile rank of each prediction. Therefore, users can assess whether a given prediction is a top prediction for the compound or disease. In addition, our table-based prediction browser links to a custom guide for each prediction, which displays in the Neo4j Hetionet Browser. Each guide includes a query to display the top paths supporting the prediction and lists clinical trials investigating the indication.

Nicotine dependence case study

There are currently two FDA-approved medications for smoking cessation (varenicline and bupropion) that are not nicotine replacement therapies. PharmacotherapyDB v1.0 lists varenicline as a disease-modifying indication and nicotine itself as a symptomatic indication for nicotine dependence, but is missing bupropion. Bupropion was first approved for depression in 1985. Owing to the serendipitous observation that it decreased smoking in depressed patients taking this drug, Bupropion was approved for smoking cessation in 1997 [40]. Therefore we looked whether Project Rephetio could have predicted this repurposing. Bupropion was the 9th best prediction for nicotine dependence (99.5th percentile) with a probability 2.50-fold greater than the null. Figure 4 shows the top paths supporting the repurposing of bupropion.

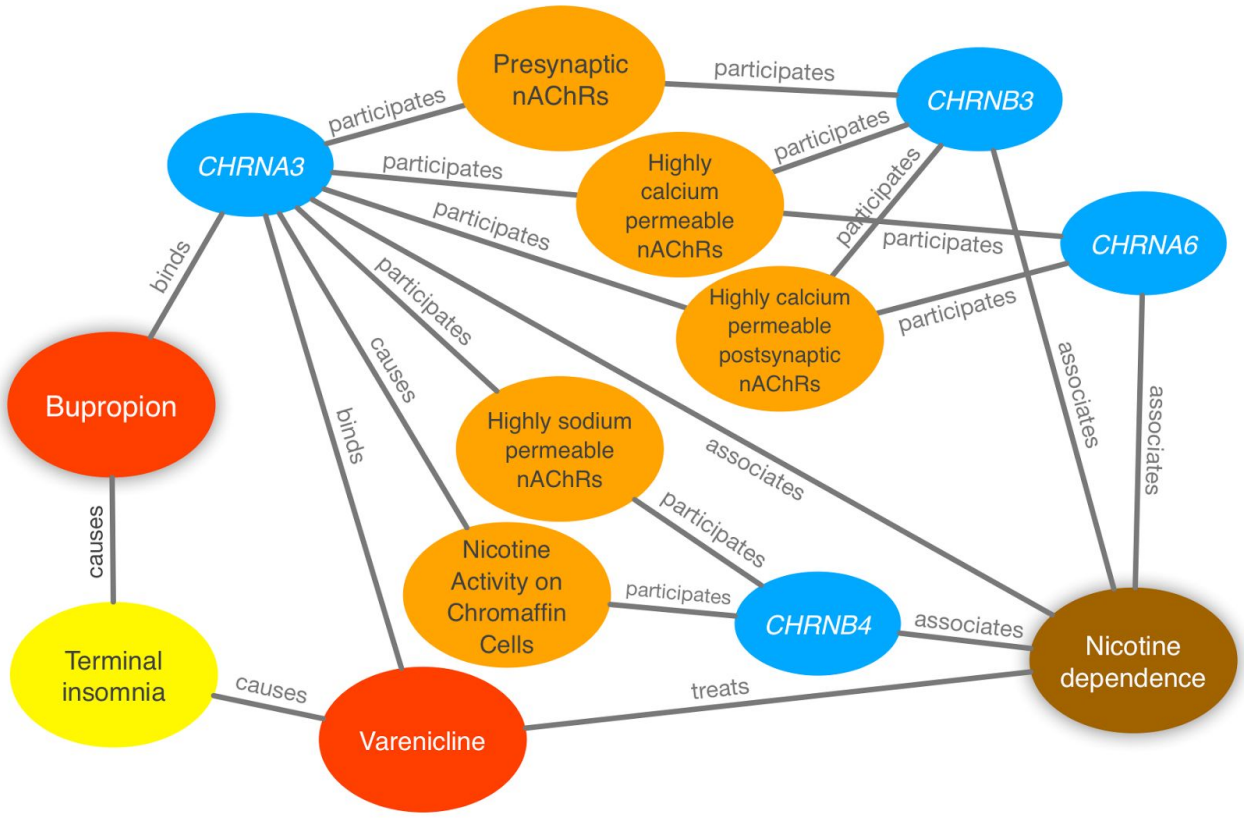

Figure 4. Evidence supporting the repurposing of bupropion for smoking cessation

This figure shows the 10 most supportive paths (out of 365 total) for treating nicotine dependence with bupropion, as available in this prediction's Neo4j Browser guide. Our method detected that bupropion targets the *CHRNA3* gene, which is also targeted by the known-treatment varenicline [41]. Furthermore, *CHRNA3* is associated with nicotine dependence [42] and participates in several pathways that contain other nicotinic-acetylcholine-receptor (nAChRs) genes associated with nicotine

dependence. Finally, bupropion causes terminal insomnia [43] as does varenicline [44], which could indicate an underlying common mechanism of action.

Atop the nicotine dependence predictions were nicotine (10.97-fold over null), cytisine (10.58-fold), and galantamine (9.50-fold). Cytisine is widely used in Eastern Europe for smoking cessation due to its availability at a fraction of the cost of other pharmaceutical options [45]. In the last half decade, large scale clinical trials have confirmed cytisine's efficacy [46, 47]. Galantamine, an approved Alzheimer's treatment, is currently in Phase 2 trial for smoking cessation and is showing promising results [48]. In summary, nicotine dependence illustrates Project Rephetio's ability to predict efficacious treatments and prioritize historic and contemporary repurposing opportunities.

Epilepsy case study

Several factors make epilepsy an interesting disease for evaluating repurposing predictions [49]. Antiepileptic drugs work by increasing the seizure threshold — the amount of electric stimulation that is required to induce seizure. The effect of a drug on the seizure threshold can be cheaply and reliably tested in rodent models. As a result, the viability of most approved drugs in treating epilepsy is known.

We focused our evaluation on the top 100 scoring compounds — referred to as the epilepsy predictions in this section — after discarding a single combination drug. We classified each compound as anti-ictogenic (seizure suppressing), unknown (no established effect on the seizure threshold), or ictogenic (seizure generating) according to medical literature [49]. Of the epilepsy predictions, 77 were anti-ictogenic, 8 were unknown, and 15 were ictogenic (Figure 5). Notably, the predictions contained 23 of the 25 disease-modifying antiepileptics in PharamcotherapyDB v1.0.

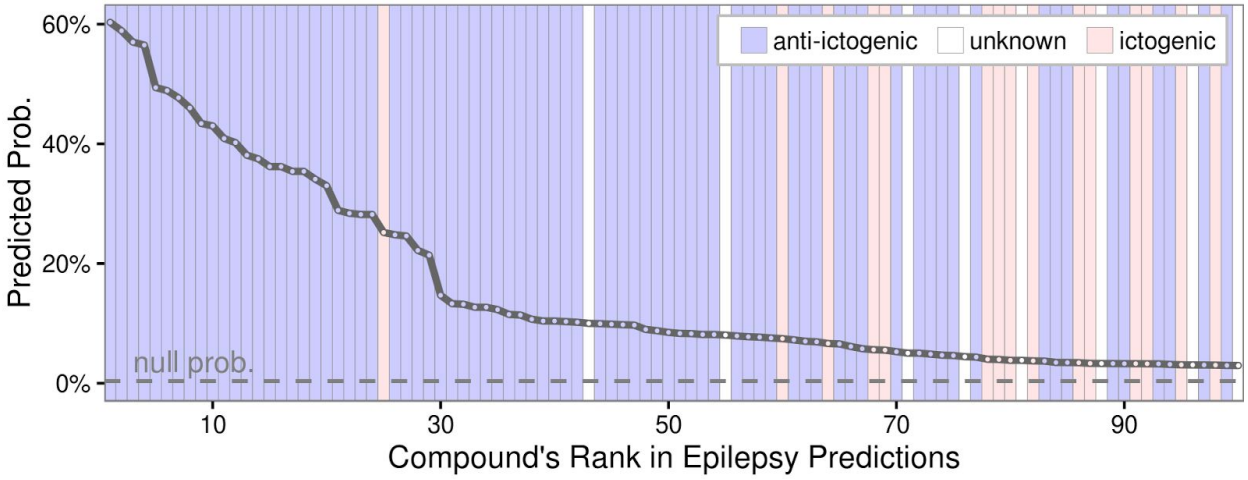

Figure 5. Top 100 epilepsy predictions colored by their effect on seizures

The predicted probability of treatment versus prediction rank is plotted for the top 100 epilepsy predictions. Note that all compounds shown received probabilities far exceeding the null probability of treatment. Furthermore, the highest predictions are almost exclusively anti-ictogenic. Further down the prediction list the prevalence of drugs with an ictogenic (contraindications) or unknown (novel repurposing candidates) effect on epilepsy increases.

Many of the 77 anti-ictogenic compounds were not first-line antiepileptic drugs. Instead, they were used as ancillary drugs in the treatment of status epilepticus. For example, we predicted four halogenated ethers, two of which (isoflurane and desflurane) are used clinically to treat life-threatening seizures that persist despite treatment [50]. As inhaled anesthetics, these compounds are not appropriate as daily epilepsy medications, but are feasible for refractory status epilepticus where patients are intubated.

Given this high precision (77%), the 8 compounds of unknown effect are promising repurposing candidates. For example, acamprosate — whose top prediction was epilepsy — is a taurine analog that promotes alcohol abstinence. Support for this repurposing arose from acamprosate's positive modulation of the $GABA_A$ receptor and inhibition of the glutamate receptor. If effective against epilepsy, acamprosate could serve a dual benefit for recovering alcoholics who experience seizures from alcohol withdrawal.

Also notable are the 15 ictogenic compounds in our top 100 predictions. As an example, we predicted five tricyclic antidepressants primarily based on their binding to the $GABA_A$ receptor. However, these compounds are $GABA_A$ antagonists, rather than agonists, likely resulting in their ictogenic properties.

As isoflurane, desflurane, and acamprosate demonstrate, Project Rephetio is capable of predicting repurposing candidates that fulfil a therapeutic niche. In addition, a portion of Rephetio's predictions are likely contraindications. However, in the case of epilepsy, where the effect of most approved drugs is known, our approach was still able to overwhelmingly prioritize ictogenic compounds.

Discussion

We created Hetionet v1.0 by integrating 29 resources into a single data structure — the hetnet. Consisting of 11 types of nodes and 24 types of relationships, Hetionet v1.0 brings more types of information together than previous leading-studies in biological data integration [51]. Moreover, we strove to create a reusable, extensible, and property-rich network. While all of the resources we include are publicly available, their

integration was a time-intensive undertaking. Hetionet allows researchers to begin answering integrative questions without having to first spend months processing data.

Our public Neo4j instance allows users to immediately interact with Hetionet. Through the Cypher language, users can perform highly specialized graph queries with only a few lines of code. Queries can be executed in the web browser or programmatically from a language with a Neo4j driver. For users that are unfamiliar with Cypher, we include several example queries in a Browser guide. In contrast to traditional REST APIs, our public Neo4j instance provides users with maximal flexibility to construct custom queries by exposing the underlying database.

As data has grown more plentiful and diverse, so has the applicability of hetnets. Unfortunately, network science has been naturally fragmented by discipline resulting in relatively slow progress in integrating heterogeneous data. 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 [52, 53]. Nonetheless, core infrastructure and algorithms for hetnets are emerging. One goal of our project has been to unite hetnet research across disciplines. We approached this goal by making Project Rephetio entirely available online and inviting community feedback throughout the process [54].

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 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, namely the metaedges *Disease–associates–Gene*, *Compound–causes–Side Effect*, *Disease–presents–Symptom*, *Disease–localizes–Anatomy*, and *Gene–interacts–Gene*.

The lower-middle tier contains the transcriptomics 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. However, our findings suggest that these resources are only moderately 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. Unlike the top tier resources, this knowledge has historically been 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.

The performance of specific metapaths (Table 3) provides further insight. For example, significant emphasis has been put on the use of transcriptional data for drug repurposing [30]. One common approach has been to identify compounds with opposing transcriptional signatures to a disease [55, 18]. However, several systematic studies report underwhelming performance of this approach [25, 24, 26] — a finding supported by the low performance of the *CuGdD* and *CdGuD* metapaths in Project Rephetio. Nonetheless, other transcription-based methods showed some promise. Compounds with similar transcriptional signatures were prone to treating the same disease (*CuGuCtD* and *CdGdCtD* metapaths), while compounds with opposing transcriptional signatures were slightly averse to treating the same disease (*CuGdCtD* and *CdGuCtD* metapaths). In contrast, diseases with similar transcriptional profiles were not prone to treatment by the same compound (*CtDdGuD* and *CtDuGdD*).

By comparably assessing the informativeness of different metaedges and metapaths, Project Rephetio aims to guide future research towards promising data types and analyses. Encouragingly, most data types were at least weakly informative and hence suitable for further study. Ideally, different data types would provide orthogonal information. However, our model for whether a compound treats a disease focused on 11 metapaths — a small portion of the hundreds of metapaths available. While parsimony aids interpretation, our model did not draw on the weakly-predictive high-throughput data types — which are intriguing for their novelty, scalability, and cost-effectiveness — as much as we had hypothesized.

Instead our model selected types of information traditionally considered in pharmacology. However unlike a pharmacologist whose area of expertise may be limited to a few drug classes, our model was able to predict probabilities of treatment for all 209,168 compound–disease pairs. Furthermore, our model systematically learned the importance of each type of network connectivity. For any compound–disease pair, we now can immediately provide the top network paths supporting its therapeutic efficacy. A traditional pharmacologist may be able to produce a similar explanation, but likely not until spending substantial time researching the compound's pharmacology, the disease's pathophysiology, and the molecular relationships in between. Accordingly, we hope

certain predictions will spur further research, such as trials to investigate the off-label use of acamprosate for epilepsy.

As demonstrated by the 15 ictogenic compounds in our top 100 epilepsy predictions, Project Rephetio's predictions can include contraindications in addition to indications. Since many of Hetionet v1.0's relationship types are general (e.g. the *Compound–binds–Gene* relationship type conflates antagonist with agonist effects), we expect some high scoring predictions to exacerbate rather than treat the disease. However, the predictions made by Hetionet v1.0 represent such substantial relative enrichment over the null that uncovering the correct directionality is a logical next step and worth undertaking. Going forward, advances in automated mining of the scientific literature could enable extraction of precise relationship types at omics scale [56].

Future research should focus on gleaning orthogonal information from data types that are so expansive that computational methods are the only option. Our *CuGuCtD* feature — measuring whether a compound upregulates the same genes as compounds which treat the disease — is a good example. This metapath was informative by itself (Δ AUROC = 4.4%) but was not selected by the model, despite its orthogonal origin (gene expression) to selected metapaths. Using a more extensive catalog of treatments as the gold standard would be one possible approach to increase the power of feature selection.

Integrating more types of information into Hetionet should also be a future priority. The "network effect" phenomenon suggests that the addition of each new piece of information will enhance the value of Hetionet's existing information. We envision a future where all biological knowledge is encoded into a single hetnet. Hetionet v1.0 was an early attempt, and we hope the strong performance of Project Rephetio in repurposing drugs foreshadows the many applications that will thrive from encoding biology in hetnets.

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 [57, 58].

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 [22], 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.

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 [59].

We selected 137 terms from the Disease Ontology [60, 61] (which we refer to as DO Slim [62, 63]) 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` [64]. 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* [65, 66].

Approved small molecule compounds with documented chemical structures were extracted from DrugBank version 4.2 [67, 68, 69]. 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 [70, 71, 72]. SIDER codes side effects using UMLS identifiers [73], which we also adopted. Pharmacologic Classes were extracted from the DrugCentral data repository [74].

Protein-coding human genes were extracted from Entrez Gene [75, 76, 77]. Anatomical structures, which we refer to as anatomies, were extracted from Uberon [78]. 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 [79, 80].

Pathways were extracted by combining human pathways from WikiPathways [81, 82], Reactome [83], and the Pathway Interaction Database [84]. The latter two resources were retrieved from Pathway Commons [85], which compiles pathways from several providers. Duplicate pathways and pathways without multiple participating genes were removed [86, 87]. Biological processes, cellular components, and molecular functions were extracted from the Gene Ontology [88]. Only terms with 2–1000 annotated genes were included.

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 [89]. 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 [68], since these are poorly standardized [90]. UniChem's [91] Connectivity Search [92] was used to map compounds, which maps by atomic connectivity (based on First InChIKey Hash Blocks [93]) and ignores small molecular differences.

Edges

Anatomy–downregulates–Gene and *Anatomy–upregulates–Gene* edges [94, 95, 96] were extracted from Bgee [97], which computes differentially expressed genes by anatomy in post-juvenile adult humans. *Anatomy–expresses–Gene* edges were extracted from Bgee and TISSUES [98, 99, 100].

Compound–binds–Gene edges were aggregated from BindingDB [101, 102], DrugBank [103, 67], and DrugCentral. Only binding relationships to single proteins with affinities of at least 1 μM (as determined by Kd, Ki, or IC50) were selected from the October 2015 release of BindingDB [104, 105]. Target, carrier, transporter, and enzyme interactions with single proteins (i.e. excluding protein groups) were extracted from DrugBank 4.2 [106, 69]. In addition, all mapping DrugCentral target relationships were included [74].

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 [70, 71, 72], which uses natural language processing to identify side effects in drug labels. *Compound–resembles–Compound* relationships [107, 69, 108] represent chemical similarity and correspond to a Dice coefficient ≥ 0.5 [109] between extended connectivity fingerprints [110, 111].*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 [112], DISEASES [113, 114], DisGeNET [115, 116], and DOAF [117, 118]. The GWAS Catalog compiles disease–SNP associations from published GWAS [119]. We aggregated overlapping loci associated with each disease and identified the mode reported gene for each high confidence locus [120, 121]. DISEASES integrates evidence of association from text mining, curated catalogs, and experimental data [122]. 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 [123]. Associations with scores ≥ 0.06 were included. DOAF mines Entrez Gene GeneRIFs (textual annotations of gene function) for disease mentions [124]. Associations with 3 or more supporting GeneRIFs were included. *Disease–downregulates–Gene* and *Disease–upregulates–Gene* relationships [125, 126] were computed using STARGEO as described in Intermediate resources.

Disease–localizes–Anatomy, *Disease–presents–Symptom*, and *Disease–resembles–Disease*edges were calculated from MEDLINE co-occurrence [65, 127]. 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`) [128]; and 363,928 articles with both a disease topic (`majr`) and a symptom topic (`noexp`). We used a Fisher's exact test [129] to identify pairs of terms that occurred together more than would be expected by chance in their respective corpus. We included co-occuring terms with $p < 0.005$ in Hetionet v1.0.

Gene–covaries–Gene edges represent evolutionary rate covariation ≥ 0.75 [130, 131, 132]. *Gene–interacts–Gene* edges [133, 134] represent when two genes produce physically-interacting proteins. We compiled these interactions from the Human Interactome Database [135, 136, 137, 138], the Incomplete Interactome [139], and our

previous study [22]. *Gene–participates–Biological Process*, *Gene–participates–Cellular Component*, and *Gene–participates–Molecular Function* edges are from Gene Ontology annotations [140]. As described in Intermediate resources, annotations were propagated [141, 142].

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 [143, 144]. 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 [125, 126]. 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 of the 49 diseases 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 [145, 146].

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 [147, 148]. 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 [149]. 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 chance 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 [150].

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 [151, 152, 153]. 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 [154]: MEDI-HPS which mined indications from RxNorm, SIDER 2, MedlinePlus, and Wikipedia [155]; LabeledIn which extracted indications from drug labels via human curation [156, 157, 158]; EHRLink which identified medication–problem pairs that clinicians linked together in electronic health records [159, 160]; and indications from PREDICT, which were compiled from UMLS relationships, drugs.com, and drug labels [24]. 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 [161]. 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 κ = 49.9%). We then recruited a third practicing physician, who reviewed all 1,388 calls and created a detailed explanation of his methodology [161]. 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. We ultimately considered 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 (http://git.dhimmel.com/gene-ontology/) to provide straightforward access to Gene Ontology annotations [142, 141]. 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.

Data copyright and licensing

We committed to openly releasing our data and analyses from the origin of the project [162]. Our goals were to contribute to the advancement of science [163, 164], maximize our impact [165], and enable reproducibility [166, 167, 168]. These objectives required publicly distributing and openly licensing Hetionet and Project Rephetio data and analyses [169, 170].

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 [171], 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 [172]. 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 [173]. Therefore, we sought input from legal experts and chronicled our progress [174, 175, 176, 177].

Ultimately, we did not find an ideal solution. We had to choose between absolute compliance and Hetionet: strictly adhering to copyright and licensing arrangements would have decimated the 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 [178], since its license was highly problematic [175].

Permuted Hetnets

From Hetionet, we derived five permuted hetnets [179]. The permutations preserve node degree but eliminate edge specificity by employing an algorithm called XSwap to randomly swap edges [180]. Permuted networks are useful for computing the baseline performance of meaningless edges while preserving node degree [181].

Neo4j

Graph database adoption in bioinformatics has thus far been limited [182]. We used the Neo4j graph database for storing and operating on Hetionet and noticed major benefits from tapping into this large open source ecosystem [183]. Persistent storage with immediate access and the Cypher query language — a sort of SQL for hetnets — were two of the biggest benefits. To facilitate our migration to Neo4j, we updated `hetio` — our existing Python package for hetnets [184] — to export networks into Neo4j and DWPC queries to Cypher. In addition, we created an interactive GraphGist for Project Rephetio, which introduces our approach and showcases its Cypher queries. Finally, we created a public Neo4j instance [185], which leverages several modern technologies such Neo4j Browser guides, cloud hosting with HTTPS, and Docker deployment [186, 187].

Machine learning approach

We made several refinements to metapath-based hetnet edge prediction compared to previous studies [22, 23]. First, we transformed DWPCs by mean scaling and then taking the inverse hyperbolic sine [188] to make them more amenable to modeling [189]. Second, we bifurcated the workflow into an all-features stage and an all-observations stage [190]. 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 (4 × 755) 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 [191], 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 [192]. 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 [193]. 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).

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 [192, 194], 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 indications [74]. 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 [195]. This set contains 5,594 indications. Since these indications were not manually curated and clinical trials often show a lack of efficacy, we expected lower performance on this set.

- 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.

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 [196]. 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, 36 non-team members commented across 80 discussions, which generated 488 comments and 161 notes (Figure 6). The Thinklab content for this project totaled 111,425 words or 698,830 characters [197]. Using an estimated 7,000 words per academic publication as a benchmark, Project Rephetio generated written content comparable in volume to 15.9 publications prior to its completion. We noticed several other benefits from using Thinklab including forging a community of contributors [198]; receiving feedback during the early stages when feedback is the most actionable [199]; disseminating our research without delay [200, 201]; opening avenues for external input [202]; facilitating problem-oriented teaching [203, 204]; and improving our documentation by maintaining a publication-grade digital lab notebook [205].

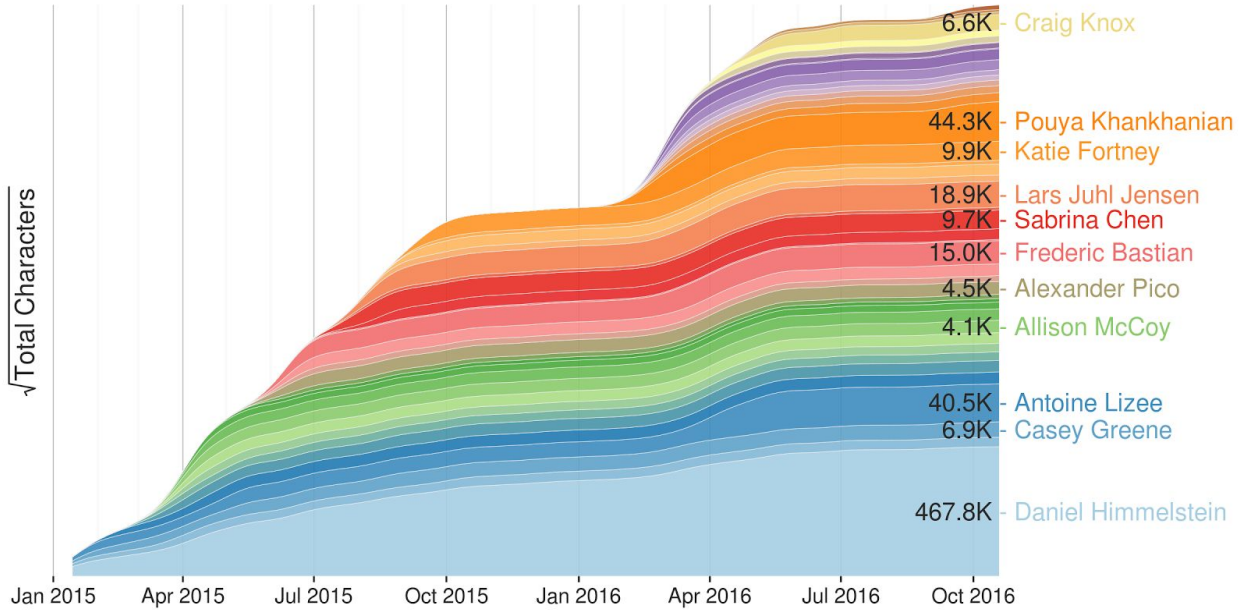

Figure 6. The growth the Project Rephetio corpus on Thinklab over time

This figure shows Project Rephetio contributions by user over time. Each band represented the cumulative contribution of a Thinklab user to discussions in the Rephetio project [197]. Users are ordered by date of first contribution. Users who contributed over 4,000 characters are named. The square root transformation of characters written per user accentuates the activity of new contributors, thereby emphasizing collaboration and diverse input.

Acknowledgements

We are immensely grateful to our Thinklab contributors who joined us in our experiment of radically open science. The following non-team members provided contributions that received 5 or more Thinklab points: Lars Juhl Jensen, Frederic Bastian, Alexander Pico, Casey Greene, Craig Knox, Benjamin Good, Chris Mungall, Katie Fortney, Venkat Malladi, MacKenzie Smith, Caty Chung, Mike Gilson, Tudor Oprea, Allison McCoy, Alexey Strokach, Ritu Khare, Marina Sirota, Greg Way, Raghavendran Partha, Jesse Spaulding, Alessandro Didonna, Alex Pankov, Janet Piñero, Oleg Ursu, Tong Shu Li. We would also like to thank Neo Technology, whose staff provided excellent technical support.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1144247 to DSH. SEB is supported by the Heidrich Family and Friends Foundation.

References

1. **Innovation in the pharmaceutical industry: New estimates of R&D costs**

   Joseph A. DiMasi, Henry G. Grabowski, Ronald W. Hansen (2016) *Journal of Health Economics*. doi:10.1016/j.jhealeco.2016.01.012

2. **A guide to drug discovery: Trends in development and approval times for new therapeutics in the United States**

   Janice M. Reichert (2003) *Nature Reviews Drug Discovery*. doi:10.1038/nrd1178

3. **Clinical development success rates for investigational drugs**

   Michael Hay, David W Thomas, John L Craighead, Celia Economides, Jesse Rosenthal (2014) *Nat Biotechnol*. doi:10.1038/nbt.2786

4. **Diagnosing the decline in pharmaceutical R&D efficiency**

   Jack W. Scannell, Alex Blanckley, Helen Boldon, Brian Warrington (2012) *Nature Reviews Drug Discovery*. doi:10.1038/nrd3681

5. **Drug repositioning: identifying and developing new uses for existing drugs**

   Ted T. Ashburn, Karl B. Thor (2004) *Nat Rev Drug Discov*. doi:10.1038/nrd1468

6. **A method for systematic discovery of adverse drug events from clinical notes**

   G. Wang, K. Jung, R. Winnenburg, N. H. Shah (2015) *Journal of the American Medical Informatics Association*. doi:10.1093/jamia/ocv102

7. **Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality**

   H. Xu, M. C. Aldrich, Q. Chen, H. Liu, N. B. Peterson, Q. Dai, M. Levy, A. Shah, X. Han, X. Ruan, M. Jiang, Y. Li, J. S. Julien, J. Warner, C. Friedman, D. M. Roden, J. C. Denny (2014) *Journal of the American Medical Informatics Association*. doi:10.1136/amiajnl-2014-002649

8. **Mining Retrospective Data for Virtual Prospective Drug Repurposing: L-DOPA and Age-related Macular Degeneration**

   Murray H. Brilliant, Kamyar Vaziri, Thomas B. Connor, Stephen G. Schwartz, Joseph J. Carroll, Catherine A. McCarty, Steven J. Schrodi, Scott J. Hebbring, Krishna S. Kishor, Harry W. Flynn, Andrew A. Moshfeghi, Darius M. Moshfeghi, M. Elizabeth Fini, Brian S. McKay (2016) *The American Journal of Medicine*. doi:10.1016/j.amjmed.2015.10.015

9. **Data-Driven Prediction of Drug Effects and Interactions**

N. P. Tatonetti, P. P. Ye, R. Daneshjou, R. B. Altman (2012) *Science Translational Medicine*. doi:10.1126/scitranslmed.3003377

10. **Bayesian statistical methods for genetic association studies**

    Matthew Stephens, David J. Balding (2009) *Nat Rev Genet*. doi:10.1038/nrg2615

11. **The complex genetics of multiple sclerosis: pitfalls and prospects**

    S. Sawcer (2008) *Brain*. doi:10.1093/brain/awn081

12. **Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia**

    Bryan L. Roth, Douglas J. Sheffler, Wesley K. Kroeze (2004) *Nat Rev Drug Discov*. doi:10.1038/nrd1346

13. **Network pharmacology: the next paradigm in drug discovery**

    Andrew L Hopkins (2008) *Nature Chemical Biology*. doi:10.1038/nchembio.118

14. **Network pharmacology**

    Andrew L Hopkins (2007) *Nat Biotechnol*. doi:10.1038/nbt1007-1110

15. **How were new medicines discovered?**

    David C. Swinney, Jason Anthony (2011) *Nat Rev Drug Discov*. doi:10.1038/nrd3480

16. **Drug discovery in the age of systems biology: the rise of computational approaches for data integration**

    Murat Iskar, Georg Zeller, Xing-Ming Zhao, Vera van Noort, Peer Bork (2012) *Current Opinion in Biotechnology*. doi:10.1016/j.copbio.2011.11.010

17. **The Connectivity Map: a new tool for biomedical research**

    Justin Lamb (2007) *Nature Reviews Cancer*. doi:10.1038/nrc2044

18. **Applications of Connectivity Map in drug discovery and development**

    Xiaoyan A. Qu, Deepak K. Rajpal (2012) *Drug Discovery Today*. doi:10.1016/j.drudis.2012.07.017

19. **In silicomethods for drug repurposing and pharmacology**

Rachel A. Hodos, Brian A. Kidd, Khader Shameer, Ben P. Readhead, Joel T. Dudley (2016) *WIREs Syst Biol Med*. doi:10.1002/wsbm.1337

20. **Computational Drug Repositioning: From Data to Therapeutics**

M R Hurle, L Yang, Q Xie, D K Rajpal, P Sanseau, P Agarwal (2013) *Clin Pharmacol Ther*. doi:10.1038/clpt.2013.1

21. **In silico drug repositioning – what we need to know**

Zhichao Liu, Hong Fang, Kelly Reagan, Xiaowei Xu, Donna L. Mendrick, William Slikker, Weida Tong (2013) *Drug Discovery Today*. doi:10.1016/j.drudis.2012.08.005

22. **Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes**

Daniel S. Himmelstein, Sergio E. Baranzini (2015) *PLOS Computational Biology*. doi:10.1371/journal.pcbi.1004259

23. **Co-author Relationship Prediction in Heterogeneous Bibliographic Networks**

Yizhou Sun, Rick Barber, Manish Gupta, Charu C. Aggarwal, Jiawei Han (2011) *2011 International Conference on Advances in Social Networks Analysis and Mining*. doi:10.1109/ASONAM.2011.112

24. **PREDICT: a method for inferring novel drug indications with application to personalized medicine**

A. Gottlieb, G. Y. Stein, E. Ruppin, R. Sharan (2011) *Molecular Systems Biology*. doi:10.1038/msb.2011.26

25. **Systematic evaluation of connectivity map for disease indications**

Jie Cheng, Lun Yang, Vinod Kumar, Pankaj Agarwal (2014) *Genome Medicine*. doi:10.1186/s13073-014-0095-1

26. **Network-based in silico drug efficacy screening**

Emre Guney, Jörg Menche, Marc Vidal, Albert-László Barábasi (2016) *Nature Communications*. doi:10.1038/ncomms10331

27. **A new method for computational drug repositioning using drug pairwise similarity**

Jiao Li, Zhiyong Lu (2012) *2012 IEEE International Conference on Bioinformatics and Biomedicine*. doi:10.1109/BIBM.2012.6392722

28. **Systematic Evaluation of Drug–Disease Relationships to Identify Leads for Novel Drug Uses**

A P Chiang, A J Butte (2009) *Clin Pharmacol Ther*. doi:10.1038/clpt.2009.103

29. **The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease**

J. Lamb (2006) *Science*. doi:10.1126/science.1132939

30. **Transcriptional data: a new gateway to drug repositioning?**

Francesco Iorio, Timothy Rittman, Hong Ge, Michael Menden, Julio Saez-Rodriguez (2013) *Drug Discovery Today*. doi:10.1016/j.drudis.2012.07.014

31. **The support of human genetic evidence for approved drug indications**

Matthew R Nelson, Hannah Tipney, Jeffery L Painter, Judong Shen, Paola Nicoletti, Yufeng Shen, Aris Floratos, Pak Chung Sham, Mulin Jun Li, Junwen Wang, Lon R Cardon, John C Whittaker, Philippe Sanseau (2015) *Nature Genetics*. doi:10.1038/ng.3314

32. **Use of genome-wide association studies for drug repositioning**

Philippe Sanseau, Pankaj Agarwal, Michael R Barnes, Tomi Pastinen, J Brent Richards, Lon R Cardon, Vincent Mooser (2012) *Nat Biotechnol*. doi:10.1038/nbt.2151

33. **Drug Target Identification Using Side-Effect Similarity**

M. Campillos, M. Kuhn, A.-C. Gavin, L. J. Jensen, P. Bork (2008) *Science*. doi:10.1126/science.1158140

34. **Computational drug repositioning based on side-effects mined from social media**

Timothy Nugent, Vassilis Plachouras, Jochen L. Leidner (2016) *PeerJ Computer Science*. doi:10.7717/peerj-cs.46

35. **Human symptoms–disease network**

XueZhong Zhou, Jörg Menche, Albert-László Barabási, Amitabh Sharma (2014) *Nat Comms*. doi:10.1038/ncomms5212

36. **Pathway-based Bayesian inference of drug–disease interactions**

Naruemon Pratanwanich, Pietro Lió (2014) *Mol. BioSyst.*. doi:10.1039/C4MB00014E

37. **Exploring the power of Hetionet: a Cypher query depot**

Daniel Himmelstein (2016) *Thinklab*. doi:10.15363/thinklab.d220

38. **Computing standardized logistic regression coefficients**

Daniel Himmelstein, Antoine Lizee (2016) *Thinklab*. doi:10.15363/thinklab.d205

39. **Predictions of whether a compound treats a disease**

Daniel Himmelstein, Chrissy Hessler, Pouya Khankhanian (2016) *Thinklab*. doi:10.15363/thinklab.d203

40. **Development of Novel Pharmacotherapeutics for Tobacco Dependence: Progress and Future Directions**

D. Harmey, P. R. Griffin, P. J. Kenny (2012) *Nicotine & Tobacco Research*. doi:10.1093/ntr/nts201

41. **Varenicline Is a Partial Agonist at 4beta2 and a Full Agonist at 7 Neuronal Nicotinic Receptors**

K. B. Mihalak (2006) *Molecular Pharmacology*. doi:10.1124/mol.106.025130

42. **A variant associated with nicotine dependence, lung cancer and peripheral arterial disease**

Thorgeir E. Thorgeirsson, Frank Geller, Patrick Sulem, Thorunn Rafnar, Anna Wiste, Kristinn P. Magnusson, Andrei Manolescu, Gudmar Thorleifsson, Hreinn Stefansson, Andres Ingason, Simon N. Stacey, Jon T. Bergthorsson, Steinunn Thorlacius, Julius Gudmundsson, Thorlakur Jonsson, Margret Jakobsdottir, Jona Saemundsdottir, Olof Olafsdottir, Larus J. Gudmundsson, Gyda Bjornsdottir, Kristleifur Kristjansson, Halla Skuladottir, Helgi J. Isaksson, Tomas Gudbjartsson, Gregory T. Jones, Thomas Mueller, Anders Gottsäter, Andrea Flex, Katja K. H. Aben, Femmie de Vegt, Peter F. A. Mulders, Dolores Isla, Maria J. Vidal, Laura Asin, Berta Saez, Laura Murillo, Thorsteinn Blondal, Halldor Kolbeinsson, Jon G. Stefansson, Ingunn Hansdottir, Valgerdur Runarsdottir, Roberto Pola, Bengt Lindblad, Andre M. van Rij, Benjamin Dieplinger, Meinhard Haltmayer, Jose I. Mayordomo, Lambertus A. Kiemeney, Stefan E. Matthiasson, Hogni Oskarsson, Thorarinn Tyrfingsson, Daniel F. Gudbjartsson, Jeffrey R. Gulcher, Steinn Jonsson, Unnur Thorsteinsdottir, Augustine Kong, Kari Stefansson (2008) *Nature*. doi:10.1038/nature06846

43. **Evaluation of the safety of bupropion (Zyban) for smoking cessation from experience gained in general practice use in England in 2000**

Andrew Boshier, Lynda V. Wilton, Saad A. W. Shakir (2003) *European Journal of Clinical Pharmacology*. doi:10.1007/s00228-003-0693-0

44. **Efficacy and Safety of Varenicline for Smoking Cessation**

J. Taylor Hays, Jon O. Ebbert, Amit Sood (2008) *The American Journal of Medicine*. doi:10.1016/j.amjmed.2008.01.017

45. **Nicotine receptor partial agonists for smoking cessation**

Kate Cahill, Nicola Lindson-Hawley, Kyla H Thomas, Thomas R Fanshawe, Tim Lancaster (2016) *Cochrane Database of Systematic Reviews*. doi:10.1002/14651858.CD006103.pub7

46. **Placebo-Controlled Trial of Cytisine for Smoking Cessation**

Robert West, Witold Zatonski, Magdalena Cedzynska, Dorota Lewandowska, Joanna Pazik, Paul Aveyard, John Stapleton (2011) *New England Journal of Medicine*. doi:10.1056/NEJMoa1102035

47. **Cytisine versus Nicotine for Smoking Cessation**

Natalie Walker, Colin Howe, Marewa Glover, Hayden McRobbie, Joanne Barnes, Vili Nosa, Varsha Parag, Bruce Bassett, Christopher Bullen (2014) *New England Journal of Medicine*. doi:10.1056/NEJMoa1407764

48. **Repeated administration of an acetylcholinesterase inhibitor attenuates nicotine taking in rats and smoking behavior in human smokers**

R L Ashare, B A Kimmey, L E Rupprecht, M E Bowers, M R Hayes, H D Schmidt (2016) *Translational Psychiatry*. doi:10.1038/tp.2015.209

49. **Prediction in epilepsy**

Pouya Khankhanian, Daniel Himmelstein (2016) *Thinklab*. doi:10.15363/thinklab.d224

50. **Treatment of Refractory Status Epilepticus With Inhalational Anesthetic Agents Isoflurane and Desflurane**

Seyed M. Mirsattari, Michael D. Sharpe, G. Bryan Young (2004) *Archives of Neurology*. doi:10.1001/archneur.61.8.1254

51. **Methods for biological data integration: perspectives and challenges**

Vladimir Gligorijević, Nataša Pržulj (2015) *J. R. Soc. Interface*. doi:10.1098/rsif.2015.0571

52. **Multilayer networks**

M. Kivela, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, M. A. Porter (2014) *Journal of Complex Networks*. doi:10.1093/comnet/cnu016

53. **Renaming 'heterogeneous networks' to a more concise and catchy term**

Daniel Himmelstein, Casey Greene, Sergio Baranzini (2015) *Thinklab*. doi:10.15363/thinklab.d104

54. **Rephetio: Repurposing drugs on a hetnet [project]**

Daniel Himmelstein, Antoine Lizee, Pouya Khankhanian, Leo Brueggeman, Sabrina Chen, Dexter Hadley, Chrissy Hessler, Ari Green, Sergio Baranzini (2015) *Thinklab*. doi:10.15363/thinklab.4

55. **Discovery and Preclinical Validation of Drug Indications Using Compendia of Public Gene Expression Data**

M. Sirota, J. T. Dudley, J. Kim, A. P. Chiang, A. A. Morgan, A. Sweet-Cordero, J. Sage, A. J. Butte (2011) *Science Translational Medicine*. doi:10.1126/scitranslmed.3001318

56. **Data programming with DDLite**

Henry R. Ehrenberg, Jaeho Shin, Alexander J. Ratner, Jason A. Fries, Christopher Ré (2016) *Proceedings of the Workshop on Human-In-the-Loop Data Analytics - HILDA '16*. doi:10.1145/2939502.2939515

57. **English, Chinese and ER diagrams**

Peter Pin-Shan Chen (1997) *Data & Knowledge Engineering*. doi:10.1016/s0169-023x(97)00017-7

58. **Data nomenclature: naming and abbreviating our network types**

Daniel Himmelstein, Lars Juhl Jensen, Pouya Khankhanian (2016) *Thinklab*. doi:10.15363/thinklab.d162

59. **Ten Simple Rules for Selecting a Bio-ontology**

James Malone, Robert Stevens, Simon Jupp, Tom Hancocks, Helen Parkinson, Cath Brooksbank (2016) *PLOS Computational Biology*. doi:10.1371/journal.pcbi.1004743

60. **Disease Ontology: a backbone for disease semantic integration**

L. M. Schriml, C. Arze, S. Nadendla, Y.-W. W. Chang, M. Mazaitis, V. Felix, G. Feng, W. A. Kibbe (2011) *Nucleic Acids Research*. doi:10.1093/nar/gkr972

61. **Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data**

W. A. Kibbe, C. Arze, V. Felix, E. Mitraka, E. Bolton, G. Fu, C. J. Mungall, J. X. Binder, J. Malone, D. Vasant, H. Parkinson, L. M. Schriml (2014) *Nucleic Acids Research*. doi:10.1093/nar/gku1011

62. **Unifying disease vocabularies**

Daniel Himmelstein, Tong Shu Li (2015) *Thinklab*. doi:10.15363/thinklab.d44

63. **User-friendly extensions to the Disease Ontology v1.0**

Daniel S. Himmelstein (2016) *Zenodo*. doi:10.5281/zenodo.45584

64. **Generating a focused view of disease ontology cancer terms for pan-cancer data integration and analysis**

T.-J. Wu, L. M. Schriml, Q.-R. Chen, M. Colbert, D. J. Crichton, R. Finney, Y. Hu, W. A. Kibbe, H. Kincaid, D. Meerzaman, E. Mitraka, Y. Pan, K. M. Smith, S. Srivastava, S. Ward, C. Yan, R. Mazumder (2015) *Database*. doi:10.1093/database/bav032

65. **Mining knowledge from MEDLINE articles and their indexed MeSH terms**

Daniel Himmelstein, Alex Pankov (2015) *Thinklab*. doi:10.15363/thinklab.d67

66. **User-friendly extensions to MeSH v1.0**

Daniel S. Himmelstein (2016) *Zenodo*. doi:10.5281/zenodo.45586

67. **DrugBank 4.0: shedding new light on drug metabolism**

V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z. T. Dame, B. Han, Y. Zhou, D. S. Wishart (2013) *Nucleic Acids Research*. doi:10.1093/nar/gkt1068

68. **Unifying drug vocabularies**

Daniel Himmelstein (2015) *Thinklab*. doi:10.15363/thinklab.d40

69. **User-friendly extensions of the DrugBank database v1.0**

Daniel S. Himmelstein (2016) *Zenodo*. doi:10.5281/zenodo.45579

70. **The SIDER database of drugs and side effects**

Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, Peer Bork (2015) *Nucleic Acids Res*. doi:10.1093/nar/gkv1075

71. **Extracting side effects from SIDER 4**

Daniel Himmelstein (2015) *Thinklab*. doi:10.15363/thinklab.d97

72. **Extracting tidy and user-friendly TSVs from SIDER 4.1**

Daniel S. Himmelstein (2016) *Zenodo*. doi:10.5281/zenodo.45521

73. **The Unified Medical Language System (UMLS): integrating biomedical terminology**

O. Bodenreider (2004) *Nucleic Acids Research*. doi:10.1093/nar/gkh061

74. **Incorporating DrugCentral data in our network**

Daniel Himmelstein, Oleg Ursu (2016) *Thinklab*. doi:10.15363/thinklab.d186

75. **Entrez Gene: gene-centered information at NCBI**

D. Maglott, J. Ostell, K. D. Pruitt, T. Tatusova (2010) *Nucleic Acids Research*. doi:10.1093/nar/gkq1237

76. **Using Entrez Gene as our gene vocabulary**

Daniel Himmelstein, Casey Greene, Alexander Pico (2015) *Thinklab*. doi:10.15363/thinklab.d34

77. **Processed Entrez Gene datasets for humans v1.0**

Daniel S. Himmelstein (2016) *Zenodo*. doi:10.5281/zenodo.45524

78. **Uberon, an integrative multi-species anatomy ontology**

Christopher J Mungall, Carlo Torniai, Georgios V Gkoutos, Suzanna E Lewis, Melissa A Haendel (2012) *Genome Biol*. doi:10.1186/gb-2012-13-1-r5

79. **Tissue Node**

Venkat Malladi, Daniel Himmelstein, Chris Mungall (2015) *Thinklab*. doi:10.15363/thinklab.d41

80. **User-friendly anatomical structures data from the Uberon Ontology v1.0**

Daniel S. Himmelstein (2016) *Zenodo*. doi:10.5281/zenodo.45527

81. **WikiPathways: capturing the full diversity of pathway knowledge**

Martina Kutmon, Anders Riutta, Nuno Nunes, Kristina Hanspers, Egon L. Willighagen, Anwesha Bohler, Jonathan Mélius, Andra Waagmeester, Sravanthi R. Sinha, Ryan Miller, Susan L. Coort, Elisa Cirillo, Bart Smeets, Chris T. Evelo, Alexander R. Pico (2015) *Nucleic Acids Res*. doi:10.1093/nar/gkv1024

82. **WikiPathways: Pathway Editing for the People**

Alexander R. Pico, Thomas Kelder, Martijn P. van Iersel, Kristina Hanspers, Bruce R. Conklin, Chris Evelo (2008) *Plos Biol*. doi:10.1371/journal.pbio.0060184

83. **The Reactome pathway Knowledgebase**

Antonio Fabregat, Konstantinos Sidiropoulos, Phani Garapati, Marc Gillespie, Kerstin Hausmann, Robin Haw, Bijay Jassal, Steven Jupe, Florian Korninger, Sheldon McKay, Lisa

Matthews, Bruce May, Marija Milacic, Karen Rothfels, Veronica Shamovsky, Marissa Webber, Joel Weiser, Mark Williams, Guanming Wu, Lincoln Stein, Henning Hermjakob, Peter D'Eustachio (2015) *Nucleic Acids Res*. doi:10.1093/nar/gkv1351

84. **PID: the Pathway Interaction Database**

C. F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, K. H. Buetow (2009) *Nucleic Acids Research*. doi:10.1093/nar/gkn653

85. **Pathway Commons, a web resource for biological pathway data**

E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, O. Babur, N. Anwar, N. Schultz, G. D. Bader, C. Sander (2010) *Nucleic Acids Research*. doi:10.1093/nar/gkq1039

86. **Adding pathway resources to your network**

Alexander Pico, Daniel Himmelstein (2015) *Thinklab*. doi:10.15363/thinklab.d72

87. **dhimmel/pathways v2.0: Compiling human pathway gene sets**

Daniel S. Himmelstein, Alexander R. Pico (2016) *Zenodo*. doi:10.5281/zenodo.48810

88. **Gene Ontology: tool for the unification of biology**

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, Gavin Sherlock (2000) *Nat Genet*. doi:10.1038/75556

89. **Disease Ontology feature requests**

Daniel Himmelstein (2015) *Thinklab*. doi:10.15363/thinklab.d68

90. **Chemical databases: curation or integration by user-defined equivalence?**

Anne Hersey, Jon Chambers, Louisa Bellis, A. Patrícia Bento, Anna Gaulton, John P. Overington (2015) *Drug Discovery Today: Technologies*. doi:10.1016/j.ddtec.2015.01.005

91. **UniChem: a unified chemical structure cross-referencing and identifier tracking system**

Jon Chambers, Mark Davies, Anna Gaulton, Anne Hersey, Sameer Velankar, Robert Petryszak, Janna Hastings, Louisa Bellis, Shaun McGlinchey, John P Overington (2013) *Journal of Cheminformatics*. doi:10.1186/1758-2946-5-3

92. **UniChem: extension of InChI-based compound mapping to salt, connectivity and stereochemistry layers**

Jon Chambers, Mark Davies, Anna Gaulton, George Papadatos, Anne Hersey, John P Overington (2014) *Journal of Cheminformatics*. doi:10.1186/s13321-014-0043-5

93. **InChI - the worldwide chemical structure identifier standard**

Stephen Heller, Alan McNaught, Stephen Stein, Dmitrii Tchekhovskoi, Igor Pletnev (2013) *Journal of Cheminformatics*. doi:10.1186/1758-2946-5-7

94. **dhimmel/bgee v1.0: Anatomy-specific gene expression in humans from Bgee**

Daniel Himmelstein, Frederic Bastian, Sergio Baranzini (2016) *Zenodo*. doi:10.5281/zenodo.47157

95. **Processing Bgee for tissue-specific gene presence and over/under-expression**

Daniel Himmelstein, Frederic Bastian (2015) *Thinklab*. doi:10.15363/thinklab.d124

96. **Tissue-specific gene expression resources**

Daniel Himmelstein, Frederic Bastian (2015) *Thinklab*. doi:10.15363/thinklab.d81

97. **Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species**

Frederic Bastian, Gilles Parmentier, Julien Roux, Sebastien Moretti, Vincent Laudet, Marc Robinson-Rechavi (2008) *Data Integration in the Life Sciences*. doi:10.1007/978-3-540-69828-9_12

98. **Comprehensive comparison of large-scale tissue expression datasets**

Alberto Santos, Kalliopi Tsafou, Christian Stolte, Sune Pletscher-Frankild, Seán I. O'Donoghue, Lars Juhl Jensen (2015) *PeerJ*. doi:10.7717/peerj.1054

99. **Gene–Tissue relationships from the TISSUES database**

Daniel Himmelstein, Lars Juhl Jensen (2015) *Zenodo*. doi:10.5281/zenodo.27244

100. **The TISSUES resource for the tissue-specificity of genes**

Daniel Himmelstein, Lars Juhl Jensen (2015) *Thinklab*. doi:10.15363/thinklab.d91

101. **BindingDB: A Web-Accessible Molecular Recognition Database**

Xi Chen, Ming Liu, Michael Gilson (2001) *Combinatorial Chemistry & High Throughput Screening*. doi:10.2174/1386207013330670

102. **BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology**

Michael K. Gilson, Tiqing Liu, Michael Baitaluk, George Nicola, Linda Hwang, Jenny Chong (2015) *Nucleic Acids Res*. doi:10.1093/nar/gkv1072

103. **DrugBank: a comprehensive resource for in silico drug discovery and exploration**

D. S. Wishart (2006) *Nucleic Acids Research*. doi:10.1093/nar/gkj067

104. **Integrating drug target information from BindingDB**

Daniel Himmelstein, Mike Gilson (2015) *Thinklab*. doi:10.15363/thinklab.d53

105. **Processing the October 2015 BindingDB**

Daniel Himmelstein, Michael Gilson, Sergio Baranzini (2015) *Zenodo*. doi:10.5281/zenodo.33987

106. **Protein (target, carrier, transporter, and enzyme) interactions in DrugBank**

Daniel Himmelstein, Sabrina Chen (2015) *Thinklab*. doi:10.15363/thinklab.d65

107. **Calculating molecular similarities between DrugBank compounds**

Daniel Himmelstein, Sabrina Chen (2015) *Thinklab*. doi:10.15363/thinklab.d70

108. **Pairwise molecular similarities between DrugBank compounds**

Daniel Himmelstein, Leo Brueggeman, Sergio Baranzini (2015) *Figshare*. doi:10.6084/m9.figshare.1418386

109. **Measures of the Amount of Ecologic Association Between Species**

Lee R. Dice (1945) *Ecology*. doi:10.2307/1932409

110. **Extended-Connectivity Fingerprints**

David Rogers, Mathew Hahn (2010) *Journal of Chemical Information and Modeling*. doi:10.1021/ci100050t

111. **The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service.**

H. L. Morgan (1965) *J. Chem. Doc.*. doi:10.1021/c160017a018

112. **dhimmel/gwas-catalog v1.0: Extracting geneâdisease associations from the GWAS Catalog**

Daniel S. Himmelstein, Sergio E. Baranzini (2016) *Zenodo*. doi:10.5281/zenodo.48428

113. **Processing the DISEASES resource for disease–gene relationships**

Daniel Himmelstein, Lars Juhl Jensen (2015) *Thinklab*. doi:10.15363/thinklab.d106

114. **dhimmel/diseases v1.0: Processing the DISEASES database of geneâdisease associations**

Daniel S. Himmelstein, Lars Juhl Jensen (2016) *Zenodo*. doi:10.5281/zenodo.48425

115. **Processing DisGeNET for disease-gene relationships**

Daniel Himmelstein, janet piñero (2015) *Thinklab*. doi:10.15363/thinklab.d105

116. **dhimmel/disgenet v1.0: Processing the DisGeNET database of geneâdisease associations**

Daniel S. Himmelstein, Janet PiÃąero (2016) *Zenodo*. doi:10.5281/zenodo.48426

117. **Functional disease annotations for genes using DOAF**

Daniel Himmelstein (2015) *Thinklab*. doi:10.15363/thinklab.d94

118. **dhimmel/doaf v1.0: Processing the DOAF database of geneâdisease associations**

Daniel S. Himmelstein (2016) *Zenodo*. doi:10.5281/zenodo.48427

119. **The NHGRI GWAS Catalog, a curated resource of SNP-trait associations**

D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, H. Parkinson (2013) *Nucleic Acids Research*. doi:10.1093/nar/gkt1229

120. **Extracting disease-gene associations from the GWAS Catalog**

Daniel Himmelstein (2015) *Thinklab*. doi:10.15363/thinklab.d80

121. **Calculating genomic windows for GWAS lead SNPs**

Daniel Himmelstein, Marina Sirota, Greg Way (2015) *Thinklab*. doi:10.15363/thinklab.d71

122. **DISEASES: Text mining and data integration of disease–gene associations**

Sune Pletscher-Frankild, Albert Pallejà, Kalliopi Tsafou, Janos X. Binder, Lars Juhl Jensen (2015) *Methods*. doi:10.1016/j.ymeth.2014.11.020

123. **DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes**

J. Pinero, N. Queralt-Rosinach, A. Bravo, J. Deu-Pons, A. Bauer-Mehren, M. Baron, F. Sanz, L. I. Furlong (2015) *Database*. doi:10.1093/database/bav028

124. **A Framework for Annotating Human Genome in Disease Context**

Wei Xu, Huisong Wang, Wenqing Cheng, Dong Fu, Tian Xia, Warren A. Kibbe, Simon M. Lin (2012) *PLoS ONE*. doi:10.1371/journal.pone.0049686

125. **STARGEO: expression signatures for disease using crowdsourced GEO annotation**

Daniel Himmelstein, Frederic Bastian, Dexter Hadley, Casey Greene (2015) *Thinklab*. doi:10.15363/thinklab.d96

126. **dhimmel/stargeo v1.0: differentially expressed genes for 48 diseases from STARGEO**

Daniel Himmelstein, Dexter Hadley, Alexander Schepanovski (2016) *Zenodo*. doi:10.5281/zenodo.46866

127. **dhimmel/medline v1.0: Disease, symptom, and anatomy cooccurence in MEDLINE**

Daniel S. Himmelstein (2016) *Zenodo*. doi:10.5281/zenodo.48445

128. **Disease similarity from MEDLINE topic cooccurrence**

Daniel Himmelstein (2015) *Thinklab*. doi:10.15363/thinklab.d93

129. **On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P**

R. A. Fisher (1922) *Journal of the Royal Statistical Society*. doi:10.2307/2340521

130. **Evolutionary Signatures amongst Disease Genes Permit Novel Methods for Gene Prioritization and Construction of Informative Gene-Based Networks**

Nolan Priedigkeit, Nicholas Wolfe, Nathan L. Clark (2015) *PLOS Genetics*. doi:10.1371/journal.pgen.1004967

131. **Selecting informative ERC (evolutionary rate covariation) values between genes**

Daniel Himmelstein, Raghavendran Partha (2015) *Thinklab*. doi:10.15363/thinklab.d57

132. **dhimmel/erc v1.0: Processing human evolutionary rate covaration data**

Daniel S. Himmelstein (2016) *Zenodo*. doi:10.5281/zenodo.48444

133. **Creating a catalog of protein interactions**

Daniel Himmelstein, Dexter Hadley, Alexey Strokach (2015) *Thinklab*. doi:10.15363/thinklab.d85

134. **dhimmel/ppi v1.0: Compiling a human protein interaction catalog**

Daniel S. Himmelstein, Sergio E. Baranzini (2016) *Zenodo*. doi:10.5281/zenodo.48443

135. **Towards a proteome-scale map of the human protein–protein interaction network**

Jean-François Rual, Kavitha Venkatesan, Tong Hao, Tomoko Hirozane-Kishikawa, Amélie Dricot, Ning Li, Gabriel F. Berriz, Francis D. Gibbons, Matija Dreze, Nono Ayivi-Guedehoussou, Niels Klitgord, Christophe Simon, Mike Boxem, Stuart Milstein, Jennifer Rosenberg, Debra S. Goldberg, Lan V. Zhang, Sharyl L. Wong, Giovanni Franklin, Siming Li, Joanna S. Albala, Janghoo Lim, Carlene Fraughton, Estelle Llamosas, Sebiha Cevik, Camille Bex, Philippe Lamesch, Robert S. Sikorski, Jean Vandenhaute, Huda Y. Zoghbi, Alex Smolyar, Stephanie Bosak, Reynaldo Sequerra, Lynn Doucette-Stamm, Michael E. Cusick, David E. Hill, Frederick P. Roth, Marc Vidal (2005) *Nature*. doi:10.1038/nature04209

136. **An empirical framework for binary interactome mapping**

Kavitha Venkatesan, Jean-François Rual, Alexei Vazquez, Ulrich Stelzl, Irma Lemmens, Tomoko Hirozane-Kishikawa, Tong Hao, Martina Zenkner, Xiaofeng Xin, Kwang-Il Goh, Muhammed A Yildirim, Nicolas Simonis, Kathrin Heinzmann, Fana Gebreab, Julie M Sahalie, Sebiha Cevik, Christophe Simon, Anne-Sophie de Smet, Elizabeth Dann, Alex Smolyar, Arunachalam Vinayagam, Haiyuan Yu, David Szeto, Heather Borick, Amélie Dricot, Niels Klitgord, Ryan R Murray, Chenwei Lin, Maciej Lalowski, Jan Timm, Kirstin Rau, Charles Boone, Pascal Braun, Michael E Cusick, Frederick P Roth, David E Hill, Jan Tavernier, Erich E Wanker, Albert-László Barabási, Marc Vidal (2008) *Nat Meth*. doi:10.1038/nmeth.1280

137. **Next-generation sequencing to generate interactome datasets**

Haiyuan Yu, Leah Tardivo, Stanley Tam, Evan Weiner, Fana Gebreab, Changyu Fan, Nenad Svrzikapa, Tomoko Hirozane-Kishikawa, Edward Rietman, Xinping Yang, Julie Sahalie, Kourosh Salehi-Ashtiani, Tong Hao, Michael E Cusick, David E Hill, Frederick P Roth, Pascal Braun, Marc Vidal (2011) *Nat Meth*. doi:10.1038/nmeth.1597

138. **A Proteome-Scale Map of the Human Interactome Network**

Thomas Rolland, Murat Taşan, Benoit Charloteaux, Samuel J. Pevzner, Quan Zhong, Nidhi Sahni, Song Yi, Irma Lemmens, Celia Fontanillo, Roberto Mosca, Atanas Kamburov, Susan D. Ghiassian, Xinping Yang, Lila Ghamsari, Dawit Balcha, Bridget E. Begg, Pascal Braun, Marc Brehme, Martin P. Broly, Anne-Ruxandra Carvunis, Dan Convery-Zupan, Roser Corominas, Jasmin Coulombe-Huntington, Elizabeth Dann, Matija Dreze, Amélie Dricot, Changyu Fan, Eric Franzosa, Fana Gebreab, Bryan J. Gutierrez, Madeleine F. Hardy, Mike Jin, Shuli Kang, Ruth Kiros, Guan Ning Lin, Katja Luck, Andrew MacWilliams, Jörg Menche, Ryan R. Murray, Alexandre Palagi, Matthew M. Poulin, Xavier Rambout, John Rasla, Patrick Reichert, Viviana

Romero, Elien Ruyssinck, Julie M. Sahalie, Annemarie Scholz, Akash A. Shah, Amitabh Sharma, Yun Shen, Kerstin Spirohn, Stanley Tam, Alexander O. Tejeda, Shelly A. Trigg, Jean-Claude Twizere, Kerwin Vega, Jennifer Walsh, Michael E. Cusick, Yu Xia, Albert-László Barabási, Lilia M. Iakoucheva, Patrick Aloy, Javier De Las Rivas, Jan Tavernier, Michael A. Calderwood, David E. Hill, Tong Hao, Frederick P. Roth, Marc Vidal (2014) *Cell*. doi:10.1016/j.cell.2014.10.050

139. **Uncovering disease-disease relationships through the incomplete interactome**

J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, A.-L. Barabasi (2015) *Science*. doi:10.1126/science.1257601

140. **The GOA database: Gene Ontology annotation updates for 2015**

R. P. Huntley, T. Sawford, P. Mutowo-Meullenet, A. Shypitsyna, C. Bonilla, M. J. Martin, C. O'Donovan (2014) *Nucleic Acids Research*. doi:10.1093/nar/gku1113

141. **Compiling Gene Ontology annotations into an easy-to-use format**

Daniel Himmelstein, Casey Greene, Venkat Malladi, Frederic Bastian (2015) *Thinklab*. doi:10.15363/thinklab.d39

142. **gene-ontology: Initial zenodo release**

Daniel Himmelstein, Casey Greene, Venkat Malladi, Frederic Bastian, Sergio Baranzini (2015) *Zenodo*. doi:10.5281/zenodo.21711

143. **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository**

R. Edgar (2002) *Nucleic Acids Research*. doi:10.1093/nar/30.1.207

144. **NCBI GEO: archive for functional genomics data sets--update**

T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, A. Soboleva (2012) *Nucleic Acids Research*. doi:10.1093/nar/gks1193

145. **dhimmel/lincs v2.0: Refined consensus signatures from LINCS L1000**

Daniel Himmelstein, Leo Brueggeman, Sergio Baranzini (2016) *Zenodo*. doi:10.5281/zenodo.47223

146. **l1000.db: SQLite database of LINCS L1000 metadata**

Daniel Himmelstein, Leo Brueggeman, Sergio Baranzini (2016) *Figshare*. doi:10.6084/m9.figshare.3085837.v1

147. **Computing consensus transcriptional profiles for LINCS L1000 perturbations**

Daniel Himmelstein, Caty Chung (2015) *Thinklab*. doi:10.15363/thinklab.d43

148. **Consensus signatures for LINCS L1000 perturbations**

Daniel Himmelstein, Leo Brueggeman, Sergio Baranzini (2016) *Figshare*. doi:10.6084/m9.figshare.3085426.v1

149. **Assessing the imputation quality of gene expression in LINCS L1000**

Daniel Himmelstein (2016) *Thinklab*. doi:10.15363/thinklab.d185

150. **Positive correlations between knockdown and overexpression profiles from LINCS L1000**

Daniel Himmelstein, Casey Greene, Lars Juhl Jensen (2016) *Thinklab*. doi:10.15363/thinklab.d171

151. **Announcing PharmacotherapyDB: the Open Catalog of Drug Therapies for Disease**

Daniel Himmelstein (2016) *Thinklab*. doi:10.15363/thinklab.d182

152. **PharmacotherapyDB 1.0: the open catalog of drug therapies for disease**

Daniel Himmelstein, Pouya Khankhanian, Christine S. Hessler, Ari J. Green, Sergio Baranzini (2016) *Figshare*. doi:10.6084/m9.figshare.3103054

153. **dhimmel/indications v1.0. PharmacotherapyDB: the open catalog of drug therapies for disease**

Daniel S. Himmelstein, Pouya Khankhanian, Christine S. Hessler, Ari J. Green, Sergio E. Baranzini (2016) *Zenodo*. doi:10.5281/zenodo.47664

154. **How should we construct a catalog of drug indications?**

Daniel Himmelstein, Benjamin Good, Tudor Oprea, Allison McCoy, Antoine Lizee (2015) *Thinklab*. doi:10.15363/thinklab.d21

155. **Development and evaluation of an ensemble resource linking medications to their indications**

W.-Q. Wei, R. M. Cronin, H. Xu, T. A. Lasko, L. Bastarache, J. C. Denny (2013) *Journal of the American Medical Informatics Association*. doi:10.1136/amiajnl-2012-001431

156. **LabeledIn: Cataloging labeled indications for human drugs**

Ritu Khare, Jiao Li, Zhiyong Lu (2014) *Journal of Biomedical Informatics*. doi:10.1016/j.jbi.2014.08.004

157. **Scaling drug indication curation through crowdsourcing**

R. Khare, J. D. Burger, J. S. Aberdeen, D. W. Tresner-Kirsch, T. J. Corrales, L. Hirchman, Z. Lu (2015) *Database*. doi:10.1093/database/bav016

158. **Processing LabeledIn to extract indications**

Daniel Himmelstein, Ritu Khare (2015) *Thinklab*. doi:10.15363/thinklab.d46

159. **Development and evaluation of a crowdsourcing methodology for knowledge base construction: identifying relationships between clinical problems and medications**

A. B. McCoy, A. Wright, A. Laxmisan, M. J. Ottosen, J. A. McCoy, D. Butten, D. F. Sittig (2012) *Journal of the American Medical Informatics Association*. doi:10.1136/amiajnl-2012-000852

160. **Extracting indications from the ehrlink resource**

Daniel Himmelstein (2015) *Thinklab*. doi:10.15363/thinklab.d62

161. **Expert curation of our indication catalog for disease-modifying treatments**

Daniel Himmelstein, Pouya Khankhanian, Chrissy Hessler (2015) *Thinklab*. doi:10.15363/thinklab.d95

162. **Enabling reproducibility and reuse**

Jesse Spaulding, Daniel Himmelstein, Casey Greene, Benjamin Good (2015) *Thinklab*. doi:10.15363/thinklab.d23

163. **The need and drive for open data in biomedical publishing**

Iain Hrynaszkiewicz (2011) *Serials: The Journal for the Serials Community*. doi:10.1629/2431

164. **The Open Knowledge Foundation: Open Data Means Better Science**

Jennifer C. Molloy (2011) *PLoS Biology*. doi:10.1371/journal.pbio.1001195

165. **Data reuse and the open data citation advantage**

Heather A. Piwowar, Todd J. Vision (2013) *PeerJ*. doi:10.7717/peerj.175

166. **Best Practices for Computational Science: Software Infrastructure and Environments for Reproducible and Extensible Research**

(2014) *Journal of Open Research Software*. doi:10.5334/jors.ay

167. **Disclose all data in publications**

Keith Baggerly (2010) *Nature*. doi:10.1038/467401b

168. **Are We Wasting a Good Crisis? The Availability of Psychological Research Data after the Storm**

Wolf Vanpaemel, Maarten Vermorgen, Leen Deriemaecker, Gert Storms (2015) *Collabra*. doi:10.1525/collabra.13

169. **Open by default: a proposed copyright license and waiver agreement for open access research and data in peer-reviewed journals**

Iain Hrynaszkiewicz, Matthew J Cockerill (2012) *BMC Research Notes*. doi:10.1186/1756-0500-5-494

170. **Creative Commons licenses and the non-commercial condition: Implications for the re-use of biodiversity information**

Gregor Hagedorn, Daniel Mietchen, Robert Morris, Donat Agosti, Lyubomir Penev, Walter Berendsohn, Donald Hobern (2011) *ZooKeys*. doi:10.3897/zookeys.150.2189

171. **One network to rule them all**

Daniel Himmelstein, Lars Juhl Jensen (2015) *Thinklab*. doi:10.15363/thinklab.d102

172. **Legal confusion threatens to slow data science**

Simon Oxenham (2016) *Nature*. doi:10.1038/536016a

173. **Who owns scientific data? The impact of intellectual property rights on the scientific publication chain**

Roger Elliott (2005) *Learned Publishing*. doi:10.1087/0953151053584984

174. **Integrating resources with disparate licensing into an open network**

Daniel Himmelstein, Lars Juhl Jensen, MacKenzie Smith, Katie Fortney, Caty Chung (2015) *Thinklab*. doi:10.15363/thinklab.d107

175. **MSigDB licensing**

Daniel Himmelstein (2015) *Thinklab*. doi:10.15363/thinklab.d108

176. **Incomplete Interactome licensing**

Daniel Himmelstein (2015) *Thinklab*. doi:10.15363/thinklab.d111

177. **LINCS L1000 licensing**

Daniel Himmelstein (2015) *Thinklab*. doi:10.15363/thinklab.d110

178. **Molecular signatures database (MSigDB) 3.0**

A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdottir, P. Tamayo, J. P. Mesirov (2011) *Bioinformatics*. doi:10.1093/bioinformatics/btr260

179. **Assessing the effectiveness of our hetnet permutations**

Daniel Himmelstein (2016) *Thinklab*. doi:10.15363/thinklab.d178

180. **Randomization Techniques for Graphs**

Sami Hanhijärvi, Gemma C. Garriga, Kai Puolamäki (2009) *Proceedings of the 2009 SIAM International Conference on Data Mining*. doi:10.1137/1.9781611972795.67

181. **Permuting hetnets and implementing randomized edge swaps in cypher**

Daniel Himmelstein (2015) *Thinklab*. doi:10.15363/thinklab.d136

182. **Are graph databases ready for bioinformatics?**

C. T. Have, L. J. Jensen (2013) *Bioinformatics*. doi:10.1093/bioinformatics/btt549

183. **Using the neo4j graph database for hetnets**

Daniel Himmelstein (2015) *Thinklab*. doi:10.15363/thinklab.d112

184. **dhimmel/hetio v0.2.0: Neo4j export, Cypher query creation, hetnet stats, and other enhancements**

Daniel Himmelstein (2016) *Zenodo*. doi:10.5281/zenodo.61571

185. **Hosting Hetionet in the cloud: creating a public Neo4j instance**

Daniel Himmelstein (2016) *Thinklab*. doi:10.15363/thinklab.d216

186. **Bioboxes: standardised containers for interchangeable bioinformatics software**

Peter Belmann, Johannes Dröge, Andreas Bremges, Alice C. McHardy, Alexander Sczyrba, Michael D. Barton (2015) *GigaScience*. doi:10.1186/s13742-015-0087-0

187. **Reproducible Computational Workflows with Continuous Analysis**

Brett K Beaulieu-Jones, Casey S Greene (2016) *Cold Spring Harbor Laboratory Press*. doi:10.1101/056473

188. **Alternative Transformations to Handle Extreme Values of the Dependent Variable**

John B. Burbidge, Lonnie Magee, A. Leslie Robb (1988) *Journal of the American Statistical Association*. doi:10.2307/2288929

189. **Transforming DWPCs for hetnet edge prediction**

Daniel Himmelstein, Pouya Khankhanian, Antoine Lizee (2016) *Thinklab*. doi:10.15363/thinklab.d193

190. **Our hetnet edge prediction methodology: the modeling framework for Project Rephetio**

Daniel Himmelstein (2016) *Thinklab*. doi:10.15363/thinklab.d210

191. **Assessing the informativeness of features**

Daniel Himmelstein (2015) *Thinklab*. doi:10.15363/thinklab.d115

192. **Edge dropout contamination in hetnet edge prediction**

Daniel Himmelstein (2016) *Thinklab*. doi:10.15363/thinklab.d215

193. **Network Edge Prediction: Estimating the prior**

Antoine Lizee, Daniel Himmelstein (2016) *Thinklab*. doi:10.15363/thinklab.d201

194. **Network Edge Prediction: how to deal with self-testing**

Antoine Lizee, Daniel Himmelstein (2016) *Thinklab*. doi:10.15363/thinklab.d194

195. **Cataloging drug–disease therapies in the ClinicalTrials.gov database**

Daniel Himmelstein (2016) *Thinklab*. doi:10.15363/thinklab.d212

196. **Rephetio: Repurposing drugs on a hetnet [proposal]**

Daniel Himmelstein, Antoine Lizee, Pouya Khankhanian, Leo Brueggeman, Sabrina Chen, Dexter Hadley, Chrissy Hessler, Ari Green, Sergio Baranzini (2015) *Thinklab*. doi:10.15363/thinklab.a5

197. **Measuring user contribution and content creation**

Daniel Himmelstein, Antoine Lizee (2016) *Thinklab*. doi:10.15363/thinklab.d200

198. **This revolution will be digitized: online tools for radical collaboration**

C. Patil, V. Siegel (2009) *Disease Models & Mechanisms*. doi:10.1242/dmm.003285

199. **Publishing the research process**

Daniel Mietchen, Ross Mounce, Lyubomir Penev (2015) *Research Ideas and Outcomes*. doi:10.3897/rio.1.e7547

200. **The waiting game**

Kendall Powell (2016) *Nature*. doi:10.1038/530148a

201. **Accelerating scientific publication in biology**

Ronald D. Vale (2015) *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1511912112

202. **Reproducibility: A tragedy of errors**

David B. Allison, Andrew W. Brown, Brandon J. George, Kathryn A. Kaiser (2016) *Nature*. doi:10.1038/530027a

203. **Workshop to analyze LINCS data for the Systems Pharmacology course at UCSF**

Daniel Himmelstein, Kathleen Keough, Misha Vysotskiy, Jeffrey Kim, Beau Norgeot, Julia Cluceru, Marjorie Imperial, Emmalyn Chen, Jasleen Sodhi, Elizabeth Levy (2016) *Thinklab*. doi:10.15363/thinklab.d181

204. **Why we are teaching science wrong, and how to make it right**

M. Mitchell Waldrop (2015) *Nature*. doi:10.1038/523272a

205. **Going paperless: The digital lab**

Jim Giles (2012) *Nature*. doi:10.1038/481430a