Mapping drug biology to disease genetics to discover drug impacts on the human phenome

Short title: Learning the drug-disease network explaining drug phenotypic effects

Mamoon Habib, UMass Lowell Department of Computer Science
Panagiotis Nikolaos Lalagkas, UMass Lowell Department of Biological Science
Rachel D. Melamed, UMass Lowell Department of Biological Science
Rachel_Melamed@uml.edu

Author contributions: MH implemented the affinity regression method; PNL interpreted drug associations and edited the manuscript; RDM designed the project, implemented analyses, and wrote the manuscript.

Abstract

Unintended effects of medications on diverse diseases are widespread, resulting in not only harmful drug side effects, but also beneficial drug repurposing. This implies that drugs can unexpectedly influence disease networks. Then, discovering how biological effects of drugs relate to disease biology can both provide insight into the biological basis for latent drug effects, and can help predict new effects. Rich data now comprehensively profile both drug impacts on biological processes, and known drug associations with human phenotypes. At the same time, systematic phenome-wide genetic studies have linked each common phenotype with putative disease driver genes. Here, we develop Draphnet, a supervised linear model that integrates in vitro data on 429 drugs and gene associations of nearly 200 common phenotypes to learn a network connecting these molecular signals to explain drug effects on disease. The approach uses the -omics level similarity among drugs, and among phenotypes, to extrapolate impacts of drug on disease. Our predicted drug-phenotype relationships outperform a baseline predictive model. But more importantly, by projecting each drug to the space of its influence on disease driver genes, we propose the biological mechanism of unexpected effects of drugs on disease phenotypes. We show that drugs sharing downstream predicted biological effects share known biology (i.e., gene targets), supporting the potential of our method to provide insights into the biology of unexpected drug effects on disease. Using Draphnet to map a drug's known molecular effects to their downstream effect on the disease genome, we put forward disease genes impacted by drug targets, and we suggest new grouping of drugs based on shared effects on the disease genome. Our approach has multiple applications, including predicting drug uses and learning about drug biology, with potential implications for personalized medicine.

Author summary

Medications can impact a number of cellular processes, resulting in both their intended treatment of a health condition, and also unintended harmful or beneficial effects on other diseases. We aim to understand and predict these drug effects by learning the network connecting the biological processes altered by drugs to the genes driving disease. Our model, called Draphnet, can predict drug side effects and indications, but beyond prediction we show that it is also able to learn a drug's expected effect on the disease genome. Using Draphnet to summarize the biological impact of each drug, we put forward the disease genes impacted by drugs or drug targets. For instance, both anti-inflammatories and some PPARα-agonists share downstream effect on the cholesterol ester transfer protein (*CETP*), a gene previously experimentally supported as an effector of fenofibrate. Our approach provides a biological basis for drug repurposing, potentially accelerating clinical advances.

Introduction

Thousands of drugs are FDA-approved, and some unexpected health benefits and risks have been uncovered only after these drugs come into common use. Notably, some drugs have hidden influence on diseases of major public health importance(1,2). This suggests opportunities for drug repurposing, or for disease prevention. An increasing number of data sources describe the effects of drugs: SIDER(3), DrugBank(4) , and the Drug Repurposing Hub(5) compile known drug effects on human disease. Systematic information on drug molecular properties are also cataloged, including drug gene targets, and the EPA ToxCast/Tox21 assays of drug biological effects(6,7). Therefore, methods that can exploit existing data to understand and predict drug effects could have a significant impact on public health.

A number of methods mine data to predict drug effects. A popular method, the connectivity score, proposes that effective drugs for a disease will induce an expression profile that contrasts with the disease expression profile(8–10). In a variation on this approach, So, et al., contrasted drug-induced gene expression with disease gene expression for neuropsychiatric diseases using results of genome-wide association studies (GWAS) of these diseases(11). Specifically, they used the S-PrediXcan method(12,13), which estimates the association of disease risk with regulation of expression of each gene. Other computational methods for discovering drug effects propose that drugs with more similar molecular effects will have more similar phenotypic effects(14). The same premise has motivated recent work using matrix completion to find drug effects(15–19).

The approaches described above have focused only on predicting drug effects, rather than learning how drug molecular effects relate to disease biology. Here, we aim to learn how drug effects on disease can be explained by the relationship between the biological effects of the drug and the genetic alterations driving disease. Therefore, we represent each drug and disease by a profile of the molecular changes associated with drug (from ToxCast) or disease phenotype (S-PrediXcan). We propose that linear models connecting a drug's molecular effects to a disease's genetic drivers can explain the drug's effect on phenotype. To estimate this interaction matrix, we take a supervised learning approach, training the matrix based on known drug impacts on disease. Importantly, by simultaneously training this model to predict relationships between tens of thousands of drug-disease pairs in a multitask fashion, we aim to learn an interpretable network connecting drugs to phenotypes. We call this method Draphnet, or Drug and Phenotype Network.

Learning this network has a number of advantages. First, interpretability provides a rationale for predictions, increasing confidence in these predictions. Second, this model can provide testable hypotheses for future analysis of drug-disease biology. Third, these findings can provide new insight into the biological basis of known drug phenotypic effects, which are often poorly understood. This can allow a new classification of drugs based on their downstream effects on disease biology. We describe 1) our development and implementation of this model, 2) evidence supporting its ability to predict drug-phenotype relationships and recapitulate known drug biology, and 3) its application for gaining new insight into drug biology.

Results

Data curation and initial assessment

We explore the premise that the effects of drugs on key biological processes propagate to their effects on disease genes, explaining the effect of drug on phenotype (*Fig 1A).* To implement and test this hypothesis, we aim to learn a model linking the biological processes altered by each drug (summarized in the matrix D) to the gene drivers of phenotype risk (in the matrix P), where the model is trained to predict the (incomplete) matrix Y of drug-phenotype association from SIDER (*Fig 1B*). We obtain the drug-phenotype association matrices Y (drug side effects and drug indications) as binary (present or unrecorded) from SIDER.

To represent the molecular profile of each drug in the matrix *D,* we compile EPA ToxCast assays recording a range of 1391 endpoints for hundreds of common medications. For example, one endpoint assay tests the androgen receptor agonist potential of a compound, while another tests the androgen receptor antagonist potential. It is important to note that drugs are not assayed for all endpoints–on average, each drug is assayed for around 100 endpoints. Despite this, we found that drugs with more similar ToxCast endpoint profiles were more likely to be associated with the same phenotypes (p=1e-22, see Method for details), even when accounting for the number of endpoints assayed. We expect that many of these endpoints are correlated with each other: some endpoints belong to the same pathway, and others represent the same readout at two time points. In order to use this sparse data source for our model, we reduce the dimensionality to create a lower-dimensional representation of the drug molecular profile without missing data. To this end, we use SoftImpute(20), a method for dimensionality reduction and matrix completion (see Method). This allows us to project each drug molecular profile to a lower-dimensional representation $U_D S_D$. Although we doubtless lose some information about each drug's biological effects, we find a strong correlation between the pairwise similarity of drugs before dimensionality reduction and as projected on the $U_D S_D$ (Spearman correlation=0.21

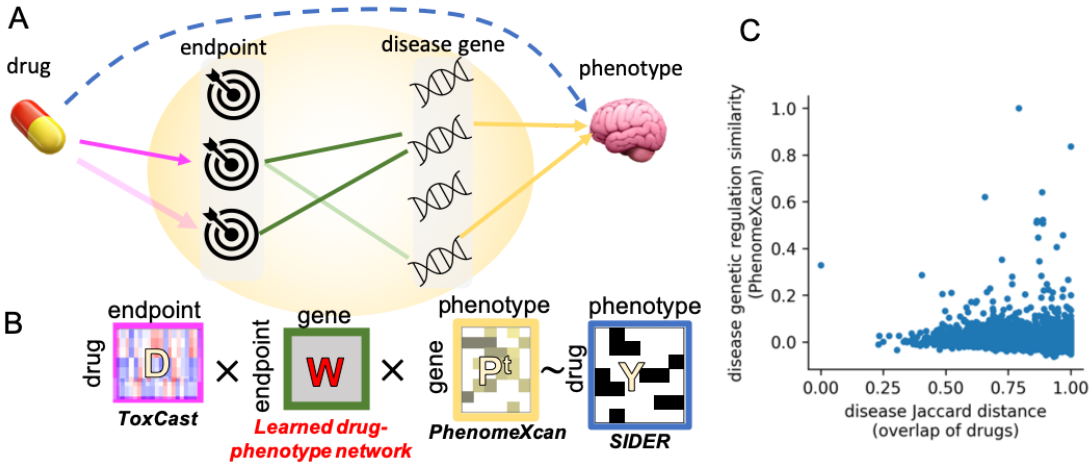

Fig 1 Design and support for the method. **A.** Proposed model where a particular drug's effects on endpoints (pink) can be propagated to impact on disease genes (green), genes in turn are associated with a phenotype (yellow), explaining drug effect on phenotype (dashed blue line). The same endpoint to disease gene network (green) is learned across thousands of pairs of drug and phenotype. **B.** Translating the proposed model into an affinity regression integrating 1) similarity of a drug to all other drugs (ToxCast, *D*), 2) similarity of a disease to all other diseases (PhenomeXcan, P^t), and 3) known drug-disease relationships (SIDER, *Y*). **C.** Each point is one pair of diseases. The x-axis shows the Jaccard distance (1-Jaccard similarity index) and the y-axis shows the Spearman correlation of the disease pairs in terms of PhenomeXcan genomic profile. Phenotype pairs with higher PhenomeXcan similarity have lower Jaccard distance (Spearman correlation=-0.11).

comparing similarity of pairs of drugs from the matrix *D* versus $U_D S_D$).

To represent disease biology in the matrix *P*, we use gene-based results derived from UKBiobank. GWAS of this data has associated loci with risk of thousands of phenotypes (including all common diseases as well as many health traits such as smoking)(21). The PhenomeXcan project integrates these GWAS results with expression quantitative trait locus results linking the same risk loci to expression of each gene(22). Therefore, the matrix *P* contains estimated association of regulation of each gene with presence of disease. Keeping only the genes that vary most highly across phenotypes, we obtain 10,027 genes for 197 phenotypes that can be matched to SIDER (either as indications or side effects). Analogous to the evaluation we performed with the drugs, we ask whether diseases with more similar genome wide gene regulation occur as side effects for overlapping sets of drugs. We found a strong relationship (p=4e-81, *Fig 1C*).

Therefore, we conclude that drugs with more similar molecular profiles are associated with more similar side effects and indications. As well, diseases with more similar PhenomeXcan molecular profiles are also impacted by more similar sets of drugs. To exploit this signal, we adapt the affinity regression method(23,24). Affinity regression was developed to model gene regulation. In that context as well, the goal was to learn an interaction matrix describing how molecular relationships manifest in the resulting (molecular) phenotypic data. Here, we anticipate, affinity regression can leverage the predictive potential of the similarity among drugs and among diseases for predicting drug-disease effects. While affinity regression has previously been applied to predict continuous (normally distributed) data, here we model a binary outcome (drug-disease relation) in the model: $DWP^t = logit(p(Y))$ (*Fig 1B*, see Methods). In effect, we are learning the weighted network *W* connecting each drug molecular profile to each disease genetic driver. Although the matrix *W* has many parameters, we impose sparsity constraints, and we additionally reduce the model size by factorizing both *D* and P^t to lower dimensional representations (Methods). Therefore, we instead learn the smaller matrix W_{DP}. We train Draphnet separately to predict either drug side effects or indications. Matrices are summarized in Table 1.

Table 1: Summary of matrices

Quantity/ Notation	Dimensions	Description/name
D	429 drugs x 1391 endpoints	Toxcast endpoints matrix
P	197 phenotypes x 10027 genes	PhenomeXcan phenotype genetics matrix (filtered to match Y)
Y	212 indication phenotypes x 1113 drug *or* 197 side effect phenotypes x 1402 drugs	SIDER indications or side effects matrix (filtered to match P and D)
U_DS_D	429 drugs x R_D	Reduced dimension representation of drug endpoint matrix from SoftImpute
$U_PS_PV_P^t$	Size of P	SVD of P
W_{DP}	R_D x R_P (95 x 131)	Compressed drug-phenotype interaction matrix
$U_DS_DW_{DP}$	429 drugs x R_P	Drug-phenome matrix
$U_DS_DW_{DP}S_P^{-1}U_P^t$	429 drugs x 10027 genes	Drug-disease genome projection (note, actual matrix is not the product, described in methods)

Assessment of the model's predictive performance

As an initial assessment of Draphnet, we test its ability to predict drug side effects. We ask whether the predictive performance can be explained by the input data alone, or if the model outperforms its input data. Predicting the drug side effects for held out drugs, we find that for a majority of drugs, our predictive model outperforms a nearest neighbor model as baseline (lower Jaccard distance between the predictions and the actual side effect profile, as compared to the nearest neighbor method, p=3e-8, rank sum test) (*Fig 2A*). This shows that Draphnet's phenotype predictions can generalize to new drugs.

Some interesting drug-phenotype pairs not present in SIDER are ranked highly. For example, among drugs not known to treat eczema, fludrocortisone is most strongly predicted to treat eczema. This drug is an oral corticosteroid, while eczema is typically treated by topical steroids. The highest ranked non-indicated drug for glaucoma is methyclothiazide, a diuretic. As glaucoma's main cause is fluid retention in the eye, this indication is plausible. In another promising example, dasatinib's primary indication (leukemias) was not available in our training matrix *P*, but despite this, the top predicted indications were cancers.

Using the model to map drugs to their effect on the phenome

Although predicting drug effect on phenotype is of interest, the primary advantage of our approach is its potential to provide insight into the biology of drug effects. To this end, we use our learned interaction matrix to map drug endpoints to their effects on the phenome. The U_DS_D matrix summarizes the variation in drug endpoints induced by each drug. By multiplying this matrix with the learned matrix W_{DP} we obtain $U_DS_DW_{DP}$, which maps each drug to a compressed summary of its effect across all phenotypes. Therefore, we call this

matrix the *drug phenome effect matrix*. Next, we investigate whether the drug phenome effect matrix reflects known characteristics of drugs.

First, we confirm that drugs that share one or more known gene targets have more similar ToxCast endpoint effects (p=1e-28, rank sum test comparing distribution of Spearman correlation of pairs of drugs sharing targets to those that do not). To show that the drug phenome effect matrix reflects learned information beyond that captured in the ToxCast matrix *D* (and, therefore, in $U_D S_D$), we create a null model for the phenome effect matrix. Our null model is obtained by projecting $U_D S_D$ onto a permuted summary of its effects on the phenome $\widetilde{W}_{DP}$, by multiplying $U_D S_D \widetilde{W}_{DP}$. Comparing the drug phenome effect matrix to this null model allows us to identify the effect of learning the true interaction matrix. It is important to note that this permutation does not nullify the information captured in the endpoint data $U_D S_D$, so in the null projection $U_D S_D \widetilde{W}_{DP}$ drugs that share targets are also more similar than ones that do not. However, for most drug targets, the true projection improves our ability to distinguish drugs sharing targets from those that do not, as compared to the null matrix (Fig 2B).

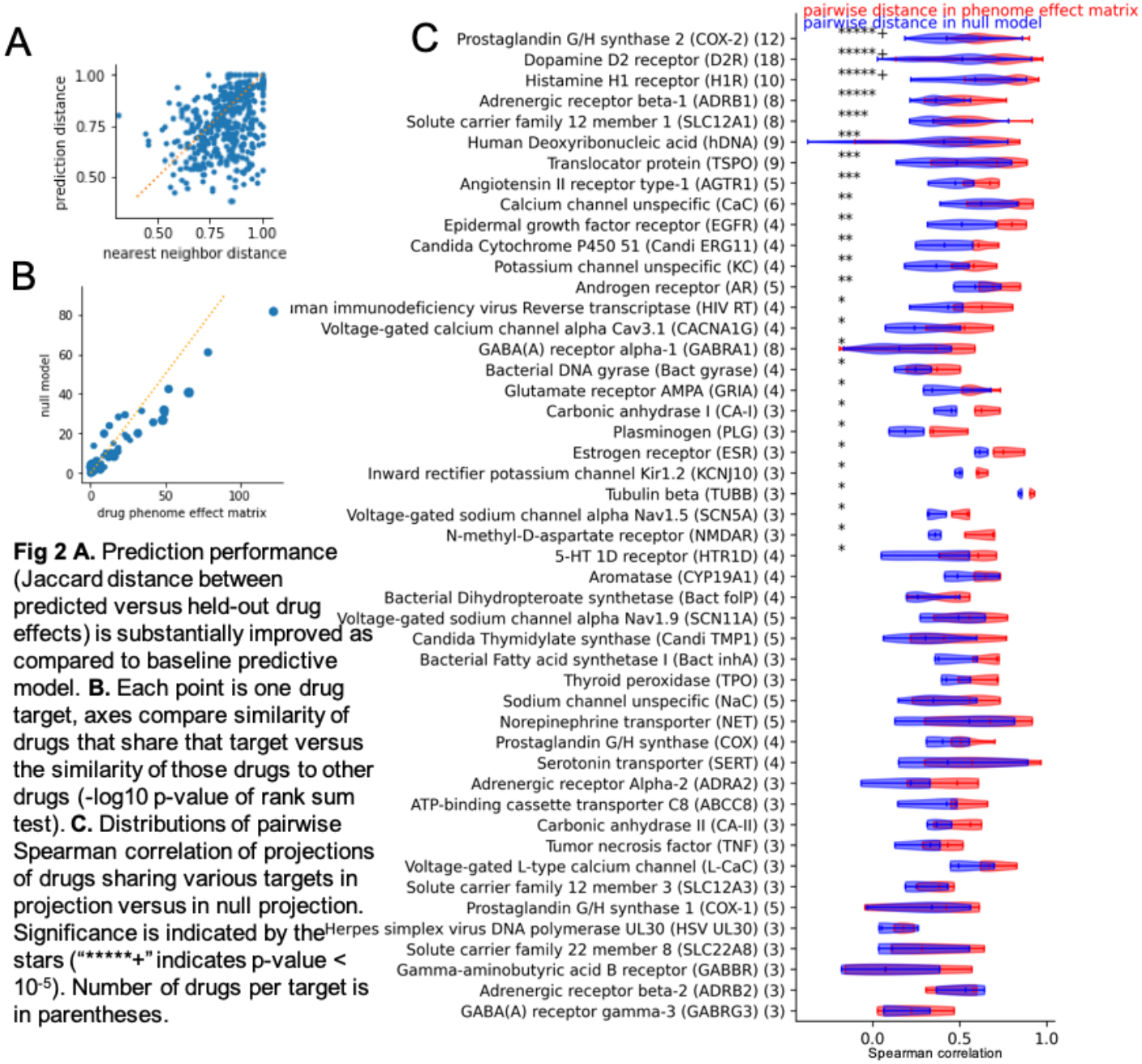

Fig 2 A. Prediction performance (Jaccard distance between predicted versus held-out drug effects) is substantially improved as compared to baseline predictive model. **B.** Each point is one drug target, axes compare similarity of drugs that share that target versus the similarity of those drugs to other drugs (-log10 p-value of rank sum test). **C.** Distributions of pairwise Spearman correlation of projections of drugs sharing various targets in projection versus in null projection. Significance is indicated by the stars ("*****+" indicates p-value < 10^{-5}). Number of drugs per target is in parentheses.

While some drug target classes do not follow this pattern, this may be due to polypharmacology, which describes the complex nature of the biological effects of drugs. Most of these targets are rather broad. For example DrugBank annotates 14 drugs as targeting *CHRM1*, and this list include anticholinergics, neuroleptics, migraine treatments, and ophthalmological preparations. These 14 drugs have a median of 19 other targets. This underlines the need for systematic approaches to better understand the biological effects of drugs.

Therefore, we investigate the similarity of drugs of the phenome effect vectors within individual classes of targets (annotated by Therapeutic Targets Database). Comparing pairs of drugs that share targets, their phenome effect similarity is systematically higher in the true drug phenome effect matrix as compared to the same pairs of drugs in the null projection (*Fig 2C*). This implies that the learned interaction matrix allows us to learn a representation of drugs that is consistent across drugs sharing known mechanisms of effect, and that this is not just due to the similarity of the drugs in the input data.

Mapping drugs to their effect on the disease genome

To investigate biological insight that can be gained from these mappings, we further project the drug phenome effect matrix to the space of disease genetic regulation, using the inverses of our matrix decomposition (Methods). This projection maps each drug onto the space of its estimated impact on genetic regulation driving disease. Again, we wish to distinguish drug effect on the phenome learned in our model, as opposed to patterns that simply exist in the input data. Therefore, we again compare each drug-disease gene score against the corresponding drug-gene values observed in empirical null models (see Method). As a result, we create a matrix estimating for each drug the importance of its effect on each disease gene, which we call the *drug-disease genome matrix* (Supplementary Table 1). Because we compare the strength of each drug-gene connection to projections using the same input data (*D* and *P*) but with null models $\widetilde{W}_{DP}$, these connections cannot be due only to the prior data on drug molecular effects: the drug-gene relations must be due to the learned interaction matrix that estimates how molecular effects propagate to impact disease (as outlined in *Fig 1*). In principle, we could estimate the chance a drug affects a particular disease by taking the dot product of the drug's disease genome vector with the disease's PhenomeXcan profile:

$$\sum_g drugGeneEffect_g \times geneDiseaseEffect_g.$$

We obtain a median of 7 disease genes associated with each drug. We then assess for each drug target, whether drugs that share that target also share disease genes (see Methods). Across 132 DrugBank targets shared by at least three drugs, we find 28 targets that have one or more significantly associated disease genes at adjusted $p < 0.01$ (Methods, Supplementary Table 2). Figure 3A shows that this level of association of drug disease genetics and drug targets is not likely to happen by chance. Therefore, we conclude that drugs sharing targets are more likely to impact the same disease genes, supporting the biological relevance of the drug-disease genome matrix.

We visualize the variation in drug-gene associations across drugs in these target groups in Figure 3B, where each drug is labeled by its ATC therapeutic subgroup. This visualization shows that drugs in the same therapeutic category have more similar gene associations: calcium channel blockers cluster together in one area, and antiinflammatories and analgesics are in another cluster. We investigate some of the drug-disease gene relationships in Figure 3C. For example *PPM1M* is associated with a number of neuroleptic drugs that target *HTR2C,* involved in serotonin signaling. It is plausible that *PPM1M* could be a key driver of the effect of these drugs: it is the top PhenomeXcan gene for bipolar disorder; a recent study found loci in this gene to be associated with schizophrenia(25); and another study linked the locus to rare mental illness(26). Another interesting finding was the association of disease driver *CETP,* or cholesterol ester transfer protein, with fenofibrate and other drugs targeting lipid metabolism. This gene is associated with high cholesterol in the PhenomeXcan results (though not one of the top associated genes). Supporting a true effect of drugs on this

driver, this gene has been associated with the effects of fenofibrate and PPARα agonism in experimental work(27,28).

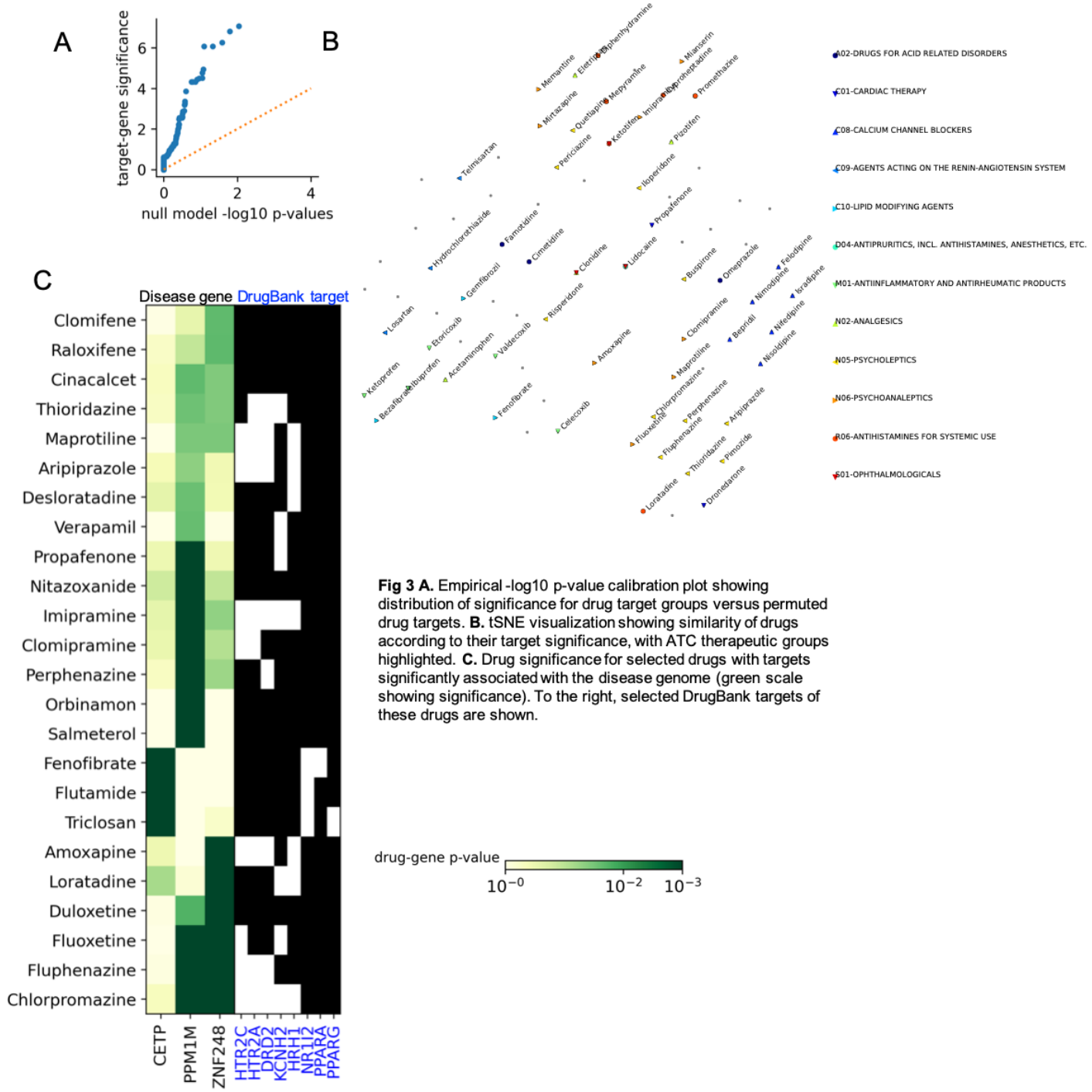

Fig 3 A. Empirical -log10 p-value calibration plot showing distribution of significance for drug target groups versus permuted drug targets. **B.** tSNE visualization showing similarity of drugs according to their target significance, with ATC therapeutic groups highlighted. **C.** Drug significance for selected drugs with targets significantly associated with the disease genome (green scale showing significance). To the right, selected DrugBank targets of these drugs are shown.

Learning a new categorization of drugs

Discovering biological effects of drugs is an open area of research. The results above point to some unexpected groupings of drugs, which we further investigate. We focus on the disease genes linked to one or more targets, and we create a novel categorization of drugs based on overlap of these disease gene associations (Supplementary Table 3). This exploratory categorization is intended to be a simple application of our drug disease genome matrix. First, we connect pairs of drugs with a significant overlap in their disease

genes to create a drug-drug network (Methods). Then, we identify fully connected cliques of drugs, containing at least 3 drugs that are all connected with each other.

As expected, this categorization overlaps with known drug targets and drug effects. For example, anti-inflammatories such as ketoprofen, ibuprofen, and indomethacin cluster together, but they also connect to PPARα-agonists (fenofibrate, bezafibrate). PPARα-agonists are known to modulate inflammation(29). Another connected drug is pyridoxine (vitamin B-6), a nutrient with diverse effects including anti-inflammation(30). In a cluster containing calcium channel blockers (bepridil, verapimil), we surprisingly find connections to pimozide, a psychiatric drug known to cause arrhythmia as a side effect. Another intriguing drug cluster connects mecamylamine, a largely discontinued antihypertensive, to pindolol, an antihypertensive beta blocker, as well as gabapentin, a central nervous system drug for seizures and nerve pain. Mecamylamine was discontinued as an antihypertensive in part because of its unintended diverse central nervous system effects, but more recently has been investigated for seizure and behavioral disorders(31). We present these drug connections as

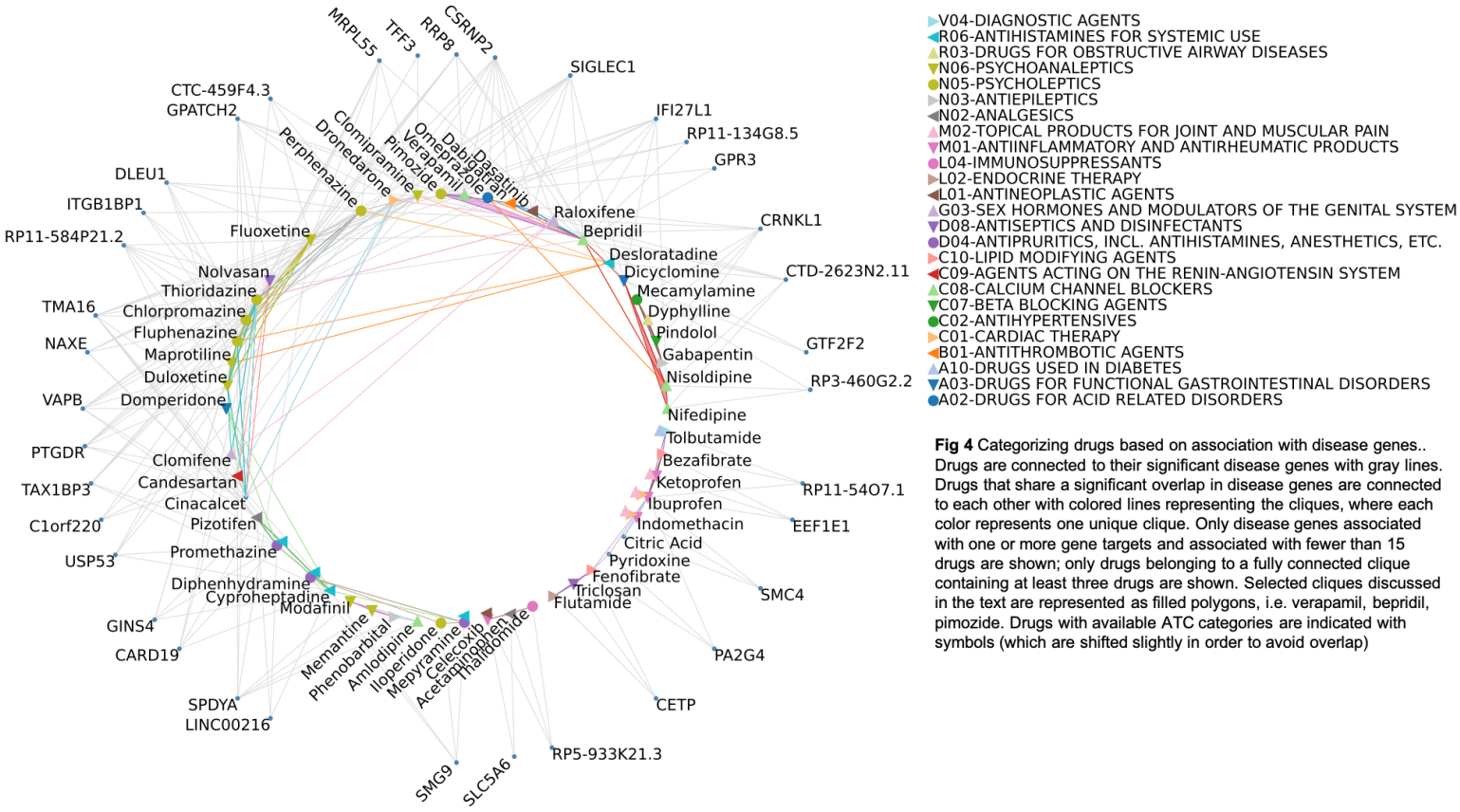

Fig 4 Categorizing drugs based on association with disease genes.. Drugs are connected to their significant disease genes with gray lines. Drugs that share a significant overlap in disease genes are connected to each other with colored lines representing the cliques, where each color represents one unique clique. Only disease genes associated with one or more gene targets and associated with fewer than 15 drugs are shown; only drugs belonging to a fully connected clique containing at least three drugs are shown. Selected cliques discussed in the text are represented as filled polygons, i.e. verapamil, bepridil, pimozide. Drugs with available ATC categories are indicated with symbols (which are shifted slightly in order to avoid overlap)

a resource for further exploration.

Discussion

Our approach learns how drug molecular effects impact disease genes and result in drug effects on phenotype. We have demonstrated that Draphnet results both reflect known drug biology, and they have the potential to provide new insights into the biological basis of unexpected drug effects on phenotypes.

While neural networks and other supervised approaches could outperform our predictions on the same data, we focus not on prediction but on biological interpretability. It is worth noting that the drug-effect matrix

used to train the model is, of necessity, always incomplete: we expect our putative negative training examples include some drug-phenotype relationships that have not yet been discovered. Then, accuracy may not be the best metric for evaluating the performance of the model(32).

We have also shown the potential of two untapped data sources for drug side effect and indication discovery: ToxCast and PhenomeXcan. While PhenomeXcan has been used to suggest possible drugs for a few diseases(11,33), no previous method has integrated this information across a range of diseases to build a drug-phenotype model. To our knowledge, ToxCast has not been used in a systematic analysis to discover new drug effects. Future work could extend the method to use the LINCS Connectivity Map data, which comprehensively profiles the effects of compounds on gene expression in multiple cancer cell lines.

A limitation of our study is that although we aim to maximize the number of drug-phenotype pairs included, the training data size remains low considering the number of parameters we could aim to estimate. To address this issue and assess its effect on our results, we have taken steps including cross-validation and regularization; reducing the feature space to minimize the number of parameters in the model; and rigorous assessment of the resulting model, including with independent drug target data. Another limitation is the ambiguity of assigning GWAS signals to disease genes. For instance, one gene connected to PPARγ agonists, the long noncoding RNA RP11-54O7.17 is adjacent to *PERM1*, a known effector of PPARγ(34). But, *PERM1* has fewer recorded eQTLs, possibly explaining its lack of signal in the PhenomeXcan data. Other comprehensive resources linking genes to disease may also be explored(35).

The results we have already provided can be a starting point for multiple new analyses. Certain ToxCast endpoints could be connected to particular disease genes. As well, projecting the drug phenome effect matrix to biological pathways can further interpret the effects of drugs. Similar to our work on categorizing drugs, our results could be used to analyze how unexpected diseases can be linked by shared pathways related to drug mechanisms. As well, this approach could provide insight into the connections between drug indications and drug side effects. One application could adapt our model to predict personalized drug effects based on the risk allele profile of a given individual, impacting precision medicine. Both analysis of our current results, and future improvements on the method, promise to improve our understanding of the biological basis of unexpected medication effects on human health.

Materials and Methods

Preparation of the disease genetic gene expression profiles and linking to drug phenotype data

Genome-wide association study (GWAS) results for many UK Biobank phenotypes have been made publicly available(36). For each GWAS, the PhenomeXcan resource compiles gene-based associations for dozens of human tissues using S-PrediXcan, and combines these results across tissues using the S-MultiXcan method(22). We convert each S-MultiXcan p-value associating a gene with a phenotype to a z-score using the inverse normal cumulative distribution. This quantity is signed, representing phenotypic association of variants associated with increased or decreased gene expression. Following the method in PhenomeXcan, we obtain the consensus (majority) S-PrediXcan sign for a gene across all tissues. Therefore, our gene-based score for each phenotype is: $|\Phi^{-1}(multiXcanP_{gene,phenotype})| \times sign_{gene,phenotype}$. For simplicity, we refer to these estimates of gene-disease association as PhenomeXcan results. This creates the preliminary matrix *P* with rows as each gene, and columns comprising hundreds of UK Biobank phenotypes.

To match the UK Biobank phenotypes in *P* to the SIDER phenotypes in *Y*, we use the Unified Medical Language System (UMLS) to match phenotype names to UMLS concept unique identifiers (CUIs)(37). SIDER includes CUIs indicating phenotypes associated with each drug, allowing us to match the UK Biobank

phenotypes to drug indication and side effect profiles. Finally, we keep only the top half of genes most variably associated with phenotype (highest standard deviation).

Preprocessing of ToxCast data

We obtain the ToxCast data from https://www.epa.gov/chemical-research/exploring-toxcast-data. Each assay tests the effect of multiple concentrations of a compound against some readout. For each such endpoint, a series of normalization, post-processing, and modeling steps have already been performed. We obtain the public Level 5 data, which estimates the fraction of models that call a compound as a "hit" for a particular endpoint. This matrix *D* contains 1391 endpoints for more than 429 compounds, but this matrix contains many missing entries where a compound was not tested for an endpoint.

To perform dimensionality reduction of this data, we use the SoftImpute package(20). This method finds a singular value decomposition of a matrix $D = U_D S_D V_D^t$ that can impute the missing values in the matrix *D*. The method requires the user to specify the rank of the decomposition, as well as a regularization parameter. To choose these values, we perform a cross-validation-like approach, setting 5% of the non-missing values to be missing, and quantifying the mean squared error of imputation of these values. After picking these hyperparameters, we project each drug onto this lower dimensional space using the product $U_D S_D$.

Assessing similarity between drug-phenotype relationships and molecular profiles

To establish the premise of our approach, we assess whether pairs of drugs with more similar molecular profiles also have more similar phenome-wide associations. For each pair of drugs, we calculate the Jaccard index between the two drugs' binary profiles denoting presence or absence of association with each disease. Then, we calculate the Spearman correlation of the ToxCast endpoint scores for each pair of drugs, when considering only the endpoints in which both drugs were evaluated. Finally, we calculate the association between Jaccard index and endpoint correlation across drug pairs, using the Spearman correlation coefficient (p=1e-22, described in Results), as well as a linear model that accounts for the number of endpoints a pair has in common (p=1.7e-43). These results show that drugs with similar ToxCast endpoint profiles have more similar phenotypic associations.

Similarly, we estimate whether diseases that are impacted by similar drugs have a similar molecular profile. We perform the analogous calculation: for each pair of phenotypes, we calculate how overlapping their sets of drugs are using the Jaccard index, and we compare this quantity to how correlated their PhenomeXcan gene associations are. We found that for all tissues, the correlation between disease genetic similarity and disease drug similarity was high (p=4e-81, *Fig 1C*, Results).

Implementation of affinity regression for binary outcomes

Next, we adapt affinity regression to our setting. In this method, the bilinear regression problem $DWP^t = Y$ is transformed to a standard regression by taking the Kronecker product: $(P \otimes D) \times stack(W) = Y$. In this way, the matrix *W* can be learned using a standard regularized regression. Because in our case *Y* is not continuous but binary we instead used regularized logistic regression to learn the model $DWP^t = logit(p(Y))$. The regularization parameter is tuned by holding out data on 10% of drugs in each fold of the cross validation.

Because of the missing values and high dimension of *D,* as mentioned above, we instead represent each drug using the lower rank matrix $U_D S_D$ learned using SoftImpute. The matrix *P* has no missing values, but it is very high dimensional, with thousands of genes that are typically highly correlated with each other(38). Therefore, we decompose this matrix as well using standard singular value decomposition (SVD): $P^t = U_P S_P V_P^t$. As a result, analogous to what is outlined in Pelosoff, et. al.(24), we instead reformulate the

regression as: $DWP^t = logit(p(Y)) = (U_D S_D V_D^t) W (U_P S_P V_P^t) = (U_D S_D) W_{DP} (V_P^t)$ where $W_{DP} = (V_D^t) W (U_P S_P)$ (equation 1). Therefore, we train the model: $(V_P \otimes (U_D S_D)) \times stack(W_{DP}) = logit(p(Y))$

We experiment with truncating the rank of the SoftImpute and SVD decompositions to find the best performance of the model. Again using a cross validation strategy, we find the ranks r_P and r_D that result in the best prediction accuracy on held out drugs. Our final hyperparameters are r_P = 131, and r_D=95 and the best regularization parameter for the lasso logistic regression was 1.

To compare the predictive performance of our model against the baseline nearest neighbor method, we perform a 20-fold cross validation analysis. For each fold, we obtain the predictions of drug side effects for the held-out 5% of drugs. As well, we obtain predictions for each drug by using the drug's nearest neighbor in the *D* matrix as a predictor of that drug's side effects. We compare these two types of predictions in Figure 2A.

Mapping drugs to their phenome and disease genome effects

In order to map each drug onto the space of its effects on diseases, we multiply the transformed drug endpoint data $U_D S_D$ with the learned lower-dimensional matrix W_{DP}. We call this product $U_D S_D W_{DP}$ the *drug phenome matrix* because it maps each drug to the space r_P representing a compressed summary of the effects of drugs on the phenome.

We can further decompress this representation to reconstruct the higher dimensional *drug disease genome matrix.* Using the inverses of the matrices from the SVD of *P,* we calculate the product $(U_D S_D W_{DP}) S_P^{-1} U_P^t$ (equation 2) . Note that U_P^t is an orthogonal matrix (see equation 1). As a result, we project each drug onto the space of phenotype genetics (here, 10,027 genes with variation in regulation associated with UK Biobank phenotypes).

In order to assess the importance of each connection between a drug and a disease gene, we create a null distribution through permutation. Specifically, we permute the values of *Y* and then train the model again. We obtain a null drug disease genome matrix $\widetilde{W}_{DP}$ for each of 10,000 permutations, and create the null drug disease genome matrix using the procedure in Equation 2. Then, for each entry in the drug disease genome matrix, we test whether the true value is lower (or higher) than the distribution of the corresponding drug-gene pair in the permuted data. Finally, we adjust these empirical permutation based p-values for multiple tests (10,027 genes for each drug). For this, we use Benjamini-Yekutili(39,40) method, which is appropriate for non-independent hypotheses. In result, we have a p-value for the importance of the connection of each drug to each disease gene. Note that these p-values represent the significance of the association between a drug and disease gene that is not just due the input data *D*, as the input data remains the same across all permutations.

Drug target and therapeutic class analysis

We obtain molecular targets for each drug from DrugBank and Therapeutic Targets Database(4,41,42). For the input ToxCast data, and for the projected phenome effect matrix, we obtain the Spearman correlation of the feature vectors for pairs of drugs that share, or do not share, recorded targets. We compare these correlation distributions using the rank sum test. As described in Results, drugs that share targets have significantly more correlated endpoint effects than those that do not share targets. To demonstrate that our model learns new information beyond that contained in the input data, we evaluate whether our projected drug phenome matrix is able to distinguish drugs that share targets more effectively than a null model. The null model projects each drug using $\widetilde{W}_{DP}$, a scrambled interaction matrix. We compare rank sum test results for each drug target in the null drug phenome effect matrix $U_D S_D \widetilde{W}_{DP}$ versus in the true matrix (*Fig. 2B*). To assess whether the projections

have increased similarity between drugs that share a target, we use the same null model. Finally, we directly assess correlations of the phenome effect vectors among pairs of drugs sharing a target, comparing correlations in the true phenome effect matrix as compared to the vectors projected using $\widetilde{W}_{DP}$ (*Fig. 2C*)

The disease genome matrix assesses which disease genes each drug is significantly associated with. We ask for each drug target in DrugBank, and for each disease gene associated with one or more of the drugs targeting that gene, if drugs that share that target are enriched for drugs associated with that disease gene. We quantify this overlap using the hypergeometric test, and the results are adjusted for the number of genes tested for each target using the Benjamini-Hochberg procedure. To assess whether drug targets have more significant gene associations than expected by chance, we permute the assignment of drugs to targets and repeat the procedure. For each true drug target and permuted version of that target, we obtain the p-value for the most significant association. *Fig 3A* shows this significance level is much higher for true drug-target associations than for the null drug target associations.

Categorizing drugs by association with disease genome

To make the visualization in *Fig 4*, we filter the disease genes to keep those significantly associated with one or more targets, and associated with fewer than 15 total drugs. First, we connect pairs of drugs that share more disease genes than expected by chance, to create a drug-drug network. Then, we identify fully connected cliques of drugs from this network.

To test whether a given pair of drugs share a significant number of disease genes, we create a null distribution by simulation. Specifically, we sample random disease genes for one drug of the pair, where disease genes are sampled in proportion to how many drugs they are significantly associated with. Then, we assess the size of the overlap in this random set, and repeat this 1000 times to obtain a nominal significance level based on comparing the true overlap to this distribution. Next we use the networkx package to create a drug-drug graph based on these connections, and retrieve all fully connected cliques of drugs from this graph. Finally, we use the igraph package to visualize drugs and genes as a bipartite graph.

Code availability

Code to reproduce the analysis is available at https://github.com/RDMelamed/drug-phenome Predicted phenotypic effects for each drug, and drug disease genomics matrix are available at https://figshare.com/projects/Mapping_drug_biology_to_disease_genetics_to_discover_drug_impacts_on_the_human_phenome/157731

References

1. Geifman N, Brinton RD, Kennedy RE, Schneider LS, Butte AJ. Evidence for benefit of statins to modify cognitive decline and risk in Alzheimer's disease. Alzheimer's Research & Therapy. 2017 Feb 17;9:10.
2. Gronich N, Rennert G. Beyond aspirin—cancer prevention with statins, metformin and bisphosphonates. Nat Rev Clin Oncol. 2013 Nov;10(11):625–42.
3. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res. 2016 Jan 4;44(Database issue):D1075–9.
4. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, et al. DrugBank 4.0: Shedding new light on drug metabolism. Nucleic Acids Research. 2014;42(D1):1091–7.
5. Corsello SM, Bittker JA, Liu Z, Gould J, McCarren P, Hirschman JE, et al. The Drug Repurposing Hub: a next-generation drug library and information resource. Nat Med. 2017 Apr;23(4):405–8.
6. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell. 2017 Nov;171(6):1437-1452.e17.

7. Richard AM, Judson RS, Houck KA, Grulke CM, Volarath P, Thillainadarajah I, et al. ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. Chem Res Toxicol. 2016 Aug 15;29(8):1225–51.
8. Dudley JT, Sirota M, Shenoy M, Pai RK, Roedder S, Chiang AP, et al. Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. Sci Transl Med. 2011 Aug 17;3(96):96ra76.
9. Chen B, Ma L, Paik H, Sirota M, Wei W, Chua MS, et al. Reversal of cancer gene expression correlates with drug efficacy and reveals therapeutic targets. Nature Communications. 2017 Jul 12;8(1):16022.
10. Sirota M, Dudley JT, Kim J, Chiang AP, Morgan AA, Sweet-Cordero A, et al. Discovery and preclinical validation of drug indications using compendia of public gene expression data. Sci Transl Med. 2011 Aug 17;3(96):96ra77.
11. So HC, Chau CKL, Chiu WT, Ho KS, Lo CP, Yim SHY, et al. Analysis of genome-wide association data highlights candidates for drug repositioning in psychiatry. Nature Neuroscience. 2017 Oct;20(10):1342–9.
12. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, et al. A gene-based association method for mapping traits using reference transcriptome data. Nature Genetics. 2015 Aug 10;47(9):1091–8.
13. Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, Torres JM, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nature Communications. 2018 May 8;9(1):1825.
14. Gottlieb A, Stein GY, Ruppin E, Sharan R. PREDICT: a method for inferring novel drug indications with application to personalized medicine. Mol Syst Biol. 2011 Jan;7(1):496.
15. Bakal G, Kilicoglu H, Kavuluru R. Non-Negative Matrix Factorization for Drug Repositioning: Experiments with the repoDB Dataset. AMIA Annu Symp Proc. 2020 Mar 4;2019:238–47.
16. Zhang W, Xu H, Li X, Gao Q, Wang L. DRIMC: an improved drug repositioning approach using Bayesian inductive matrix completion. Mathelier A, editor. Bioinformatics. 2020 May 1;36(9):2839–47.
17. Wang L, Li X, Zhang L, Gao Q. Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. BMC Cancer. 2017 Dec;17(1):513.
18. Galeano D, Li S, Gerstein M, Paccanaro A. Predicting the frequencies of drug side effects. Nature Communications. 2020 Sep 11;11(1):4575.
19. Lau A, So HC. Turning genome-wide association study findings into opportunities for drug repositioning. Comput Struct Biotechnol J. 2020;18:1639–50.
20. Hastie T, Mazumder R, Lee J, Zadeh R. Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares [Internet]. arXiv; 2014 [cited 2023 Jan 17]. Available from: http://arxiv.org/abs/1410.2596
21. Details and considerations of the UK Biobank GWAS [Internet]. Neale lab. [cited 2022 Apr 18]. Available from: http://www.nealelab.is/blog/2017/9/11/details-and-considerations-of-the-uk-biobank-gwas
22. Pividori M, Rajagopal PS, Barbeira A, Liang Y, Melia O, Bastarache L, et al. PhenomeXcan: Mapping the genome to the phenome through the transcriptome. Sci Adv. 2020 Sep;6(37):eaba2083.
23. Osmanbeyoglu HU, Pelossof R, Bromberg J, Leslie CS. Linking signaling pathways to transcriptional programs in breast cancer. Genome Research. 2014 Sep;24:1869–80.
24. Pelossof R, Singh I, Yang JL, Weirauch MT, Hughes TR, Leslie CS. Affinity regression predicts the recognition code of nucleic acid–binding proteins. Nature Biotechnology. 2015 Nov 16;33(12):1242–9.
25. Goes FS, McGrath J, Avramopoulos D, Wolyniec P, Pirooznia M, Ruczinski I, et al. Genome-wide association study of schizophrenia in Ashkenazi Jews. Am J Med Genet B Neuropsychiatr Genet. 2015 Dec;168(8):649–59.
26. Mallard TT, Linnér RK, Grotzinger AD, Sanchez-Roige S, Seidlitz J, Okbay A, et al. Multivariate GWAS of psychiatric disorders and their cardinal symptoms reveal two dimensions of cross-cutting genetic liabilities. Cell Genom. 2022 Jun 8;2(6):100140.
27. Hansen MK, McVey MJ, White RF, Legos JJ, Brusq JM, Grillot DA, et al. Selective CETP Inhibition and PPARα Agonism Increase HDL Cholesterol and Reduce LDL Cholesterol in Human ApoB100/Human CETP Transgenic Mice. J Cardiovasc Pharmacol Ther. 2010 Jun;15(2):196–202.
28. Armitage J, Holmes MV, Preiss D. Cholesteryl Ester Transfer Protein Inhibition for Preventing Cardiovascular Events. J Am Coll Cardiol. 2019 Feb 5;73(4):477–87.
29. Zandbergen F, Plutzky J. PPARα in atherosclerosis and inflammation. Biochim Biophys Acta [Internet]. 2007 Aug [cited 2023 May 26];1771(8):972–82. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2083576/
30. Ueland PM, McCann A, Midttun Ø, Ulvik A. Inflammation, vitamin B6 and related pathways. Molecular

Aspects of Medicine [Internet]. 2017 Feb 1 [cited 2023 May 26];53:10–27. Available from: https://www.sciencedirect.com/science/article/pii/S0098299716300395
31. Nickell JR, Grinevich VP, Siripurapu KB, Smith AM, Dwoskin LP. Potential Therapeutic Uses of Mecamylamine and its Stereoisomers. Pharmacol Biochem Behav [Internet]. 2013 Jul [cited 2023 May 26];108:28–43. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3690754/
32. Galeano D, Paccanaro A. A Recommender System Approach for Predicting Drug Side Effects. In: 2018 International Joint Conference on Neural Networks (IJCNN) [Internet]. Rio de Janeiro: IEEE; 2018 [cited 2021 Jan 4]. p. 1–8. Available from: https://ieeexplore.ieee.org/document/8489025/
33. Gerring ZF, Gamazon ER, White A, Derks EM. Integrative Network-Based Analysis Reveals Gene Networks and Novel Drug Repositioning Candidates for Alzheimer Disease. Neurol Genet. 2021 Oct;7(5):e622.
34. Cho Y, Hazen BC, Russell AP, Kralli A. Peroxisome Proliferator-activated Receptor γ Coactivator 1 (PGC-1)- and Estrogen-related Receptor (ERR)-induced Regulator in Muscle 1 (PERM1) Is a Tissue-specific Regulator of Oxidative Capacity in Skeletal Muscle Cells *. Journal of Biological Chemistry [Internet]. 2013 Aug 30 [cited 2023 May 26];288(35):25207–18. Available from: https://www.jbc.org/article/S0021-9258(20)49126-4/abstract
35. Watanabe K, Stringer S, Frei O, Umićević Mirkov M, de Leeuw C, Polderman TJC, et al. A global overview of pleiotropy and genetic architecture in complex traits. Nat Genet. 2019 Sep;51(9):1339–48.
36. Rapid GWAS of thousands of phenotypes for 337,000 samples in the UK Biobank [Internet]. Neale lab. [cited 2023 Jan 17]. Available from: http://www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-for-337000-samples-in-the-uk-biobank
37. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D267-270.
38. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008 Dec 29;9(1):559.
39. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. 1995;57(1):289–300.
40. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics. 2001 Aug;29(4):1165–88.
41. Wang Y, Zhang S, Li F, Zhou Y, Zhang Y, Wang Z, et al. Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. Nucleic Acids Research. 2020 Jan 8;48(D1):D1031–41.
42. Wang Y, Zhang S, Li F, Zhou Y, Zhang Y, Wang Z, et al. Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. Nucleic Acids Research. 2020 Jan 8;48(D1):D1031–41.

Supporting information

Supplementary Table S1. Drug-disease genome matrix, as described in the text. Each entry is the adjusted p-value for the association of a drug and a disease gene.
Supplementary Table S2. Target-disease genome matrix, as described in the text. Each entry is the adjusted p-value for the association of a target and a disease gene.
Supplementary Table S3. Drug-drug network, as described in the text. Each row shows a pair of drugs; the p-value of the overlap of their genes; the overlapping genes.