Integrating disease genetics and drug bioassays to discover drug impacts on the human phenome

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

Abstract

Motivation: Unintended effects of medications on diverse diseases are often identified many years after these drugs enter common use. This may be because drugs can have effects on multiple molecular targets, influencing unexpected biological processes. Discovering how biological effects of drugs relate to disease biology can provide insight into the basis for these latent drug effects, and help predict new effects. Rich data now comprehensively profile both the biological processes impacted by common drugs, and the human phenotypes known to be affected by these drugs. At the same time, systematic phenome-wide genetic studies associate each common phenotype with its genetic drivers. Here, we develop a method to integrate this data to learn how drug molecular effects can explain drug effects on the phenome.
Results: We develop a supervised approach to quantify how a drug's effect on phenotype can be explained by learned connections between the drug's molecular effects and the genetic drivers of phenotypes. Our predictions of drug phenotype relationships outperform a baseline model. But more importantly, by projecting each drug to the space of its influence on phenotypes, we present evidence that our learned interaction matrix captures information about drug biology. We use the results to propose biological mechanisms by which drugs that share a target influence disease biology.
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/Integrating_disease_genetics_and_drug_bioassays_to_discover_drug_impacts_on_the_human_phenome/157731

1. 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 chemical structure[6], the LINCS Connectivity Map of drug-induced gene expression, and the EPA ToxCast/Tox21 assays of drug biological effects[7,8]. 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 have an expression profile that contrasts with

the disease expression expression profile[9–11]. In a variation on this approach, So, et al., contrasted drug gene expression with disease gene expression for neuropsychiatric diseases using results of genome-wide association studies (GWAS) of these diseases[12]. Specifically, they used the S-PrediXcan method[13,14], 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[15]. The same premise has motivated recent work using matrix completion to find drug effects[16–20].

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 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 or disease biology. Then, we create a model that aims to learn an interaction matrix between these molecular profiles that can explain drug effect on disease. Learning an interpretable model 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.

To estimate this interaction matrix, we take a supervised learning approach, training the matrix based on known drug-disease relationships. We formulate the first application of the affinity regression method[21,22] to uncovering drug biology. Affinity regression was developed and applied to explain 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. We apply the affinity regression approach to model drug effects on phenotypes, recorded in SIDER[3], as a function of their molecular profiles. To summarize the molecular effects of drugs, ToxCast represents a promising resource: each drug is assayed for a curated set of "endpoints" representing a range of biological processes with a possible role in human disease[23]. In order to develop a rich model relating drug to their effect on disease, we need a systematic characterization of the molecular profile associated with each disease. We propose the first phenome-wide use of PhenomeXcan results associating genes with phenotypes, in order to learn the biological basis of the effect of 429 drugs on diverse human phenotypes.

2. Methods

2.1 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[24]. For each GWAS, the PhenomeXcan resource compiles gene-based associations for dozens of human tissues using S-PrediXcan, and combines these results using the S-MultiXcan method[25]. We convert each S-MultiXcan p-value associating a gene with a phenotype to a z-score using the inverse normal cumulative distribution. We Then, following the method in PhenomeXcan, we obtain the S-PrediXcan sign of the association with each tissue, and determine the consensus

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.

To match the UK Biobank phenotypes to the SIDER phenotypes, we used the Unified Medical Language System (UMLS) to match phenotype names to UMLS concept unique identifiers (CUIs)[26]. SIDER includes both CUIs indicating phenotypes for each drug, allowing us to match the UK Biobank phenotypes to drug indication and side effect profiles.

2.2 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 example, one assay tests the androgen receptor agonist potential of a compound, while another tests the androgen receptor antagonist potential. For each such endpoint, a series of modeling, normalization and post-processing steps have already been performed. We obtain the level 5 data, which estimates the fraction of models that call a compound as a "hit" for a particular endpoint.

To perform dimensionality reduction of this data, we use the SoftImpute package[27]. 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 $U_D S_D$.

2.3 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 calculated 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=4e-41), 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 EPA 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 similar 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, Figure 1A).

2.4 Implementation of affinity regression for binary outcomes

Next, we adapt affinity regression to our setting. In this method, the bilinear regression problem $DWP^t = logit(p(Y))$ is transformed to a standard regression by taking the Kroneker product: $(P \otimes D) \times stack(W) = logit(p(Y))$(p(Y). In this way, the matrix *W* can be learned using a standard regularized logistic regression, where 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. 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, similar to what is outlined in Pelosoff, et. al.[22], 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)

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.

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 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.

2.5 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 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 empirically based p-values for mutiple tests (10,027 genes for each drug). For this, we use

Benjamini-Yekutili[28,29] 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.

2.6 Drug target and therapeutic class analysis

We obtain drug targets information from DrugBank[4] and Therapeutic Targets Database (TTD)[30]. In analysis of the drug phenome effect matrix, we assess the correlation between pairs of drugs in terms of their phenome effect vectors. First, we ask whether the set of drugs that share a target have correlation that is higher between each other than between those drugs and drugs not sharing the target (Figure 2B). Because the phenome effect vectors are the result of projecting the drug ToxCast endpoint data onto the space of phenome effects, we must control for the expected similarity in endpoint data between drugs that share the same target. We do this by evaluating whether we can distinguish drugs that share targets more effectively than a null model. Our null model projects each drug using $\widetilde{W}_{DP}$, an interaction matrix fitted on scrambled data. We compare pairwise distances between drugs in the null drug phenome effect matrix $U_D S_D \widetilde{W}_{DP}$ versus in the true matrix (Figure 2B).

To assess whether the projections have increased similarity between drugs that share a target, we use the same null model (Figure 2C). That is, we test whether these pairs of drugs have closer phenome effect vector in the true phenome effect matrix than in one projected using $\widetilde{W}_{DP}$.

The disease genome matrix assesses the significance of the effect of each drug on each disease gene. We ask for each drug target in DrugBank, and for each disease gene associated with one or more of the drugs, if drugs that share that target are enriched for drugs associated with that disease gene. We quantify this using the hypergeometric test, and test 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 permuted the assignment of drugs to targets and repeated the procedure. For each true drug target and permuted version of that target, we obtain the p-value for the most significant association. Figure 3A shows this significance level is much higher for true drug-target associations than for the null drug target associations.

3. Results

3.1 Data curation and initial assessment

We prepare data from three primary sources, detailed in the Methods. We obtain drug side effects and drug indications as binary (present or absent) from SIDER. From EPA ToxCast we compile a range of 1391 endpoints for 429 drugs with available indication and side effect data. It is important to note that drugs not are 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=4e-41, see Method for details).

We expect that many of these endpoints are correlated with each other–for example, PPARα and PPARγ endpoints may be stimulated by some of the same drugs. Some of the 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 desired to reduce the dimensionality in order to create a lower-dimensional representation of the drug molecular profile that was not missing any data. To this end, we used SoftImpute[27], 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 $U_D S_D$ of drugs (Spearman correlation=0.21 comparing similarity of pairs of drugs from the matrix *D* versus $U_D S_D$).

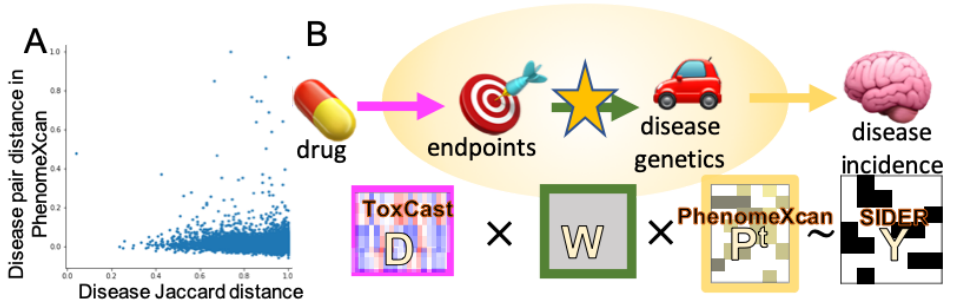

Fig 1 A. Each point is one pair of diseases. The x-axis shows the Jaccard distance (1-Jaccard similarity index) and the y-axis shows the Spearman correlation of the disease pairs in terms of PhenomeXcan genomic profile. Phenotype pairs with higher PhenomeXcan similarity have lower Jaccard distance (Spearman correlation=-0.13). **B.** Design of 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-disese relationships (SIDER, *Y*).

To represent disease biology, we obtain PhenomeXcan estimates of the 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. Similar 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, Figure 1A).,

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. These results support the application of affinity regression with this data to link molecular properties of drugs and diseases. 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). To this end, we fit a logistic regression model $DWP^t = logit(p(Y))$. Here, *D* is the drug endpoint matrix with 429 drugs and 1391 endpoints, where each row represents the molecular profile of one drug. *P* represents the disease genetic regulation matrix, with 197 phenotypes and 10042 genes. Finally, *Y* is the matrix of drug-phenotype effects, with a binary entry indicating presence or absence of a recorded impact of the drug on the phenotype.

We use logistic regression to fit the matrix *W* that connects drug molecular profiles to disease genetics, predicting *Y* (Fig 1B, see Methods). In effect, we are learning the weighted network connecting each drug molecular effect to each disease genetic driver. Although the matrix *W* has many parameters, the number can be reduced by factorizing both *D* and P^t to lower dimensional

representations. Therefore, we instead learn the smaller matrix W_{DP}. We train this model separately to predict either side effects or drug 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
$U_D S_D$	429 drugs x R_D	Reduced dimension representation of drug endpoint matrix from SoftImpute
$U_P S_P V_P^t$		SVD of P
W_{DP}	R_D x R_P	Compressed drug-phenotype interaction matrix
$U_D S_D W_{DP}$	429 drugs x R_P	Drug-phenome matrix
$U_D S_D W_{DP} S_P^{-1}$	429 drugs x 10027 genes	Drug-disease genome projection (note, actual matrix is described in methods)

3.2 Assessment of the model's predictive performance

As an initial assessment of our model, we ask whether the performance could be explained by the input data alone, or if the model was able to outperform its input data. Predicting the drug side effects for held out drugs, we find that for the majority of drugs, our predictive model outperformed a nearest neighbor model as baseline (lower Jaccard distance between the predictions and the actual side effect profile) (Figure 2A). This shows that our phenotype predictions can generalize to held out drugs. Some interesting drug-phenotype combinations 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.

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

The advantage of our approach is not just in its predictive ability, but in its potential to provide insight into the biology of drug effect on phenotype. To this end, we use our learned interaction matrix to map drug endpoints to their effects related to disease biology. The $U_D S_D$ matrix summarizes the variation in drug endpoints induced by each drug. By multiplying this matrix with the learned matrix W_{DP} we obtain $U_D S_D W_{DP}$, 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 effecct matrix reflects known characteristics of drugs. Using the Spearman correlation on the endpoint vectors for each pair of drugs, we can obtain an estimate of the similarity of drug pairs in the ToxCast data. Then, we obtain molecular targets for each drug from DrugBank and Therapeutic Targets Database[4,31]. We would expect increased similarity of endpoint vectors for drugs that share a target, as they would be expected to have similar biological effects. In fact, we do recover this expected pattern (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 we learn information beyond that captured in the ToxCast matrix, we create a null model for the phenome effect matrix. Our null model is obtained by fitting $\tilde{W}_{DP}$ from permuted versions of the input data, and calculating $U_D S_D \tilde{W}_{DP}$. 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 we still find significantly higher similarity of drugs that share targets as compared to those that do not share targets in the null projection. However, as compared to the randomized projections, the learned drug phenome effect matrix has consistently increased distinction between drugs that share targets and those that do not share targets is increased (Figure 2B).

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

Focusing on drugs that do share targets, we find that the similarity of pairs of drugs that share targets is systematically higher in the true drug phenome effect matrix as compared to the same pairs of drugs in the null versions (Figure 2C). This implies that the learned interaction matrix allows us to create a representation of drugs that is consistent across drugs sharing known mechanisms of effect.

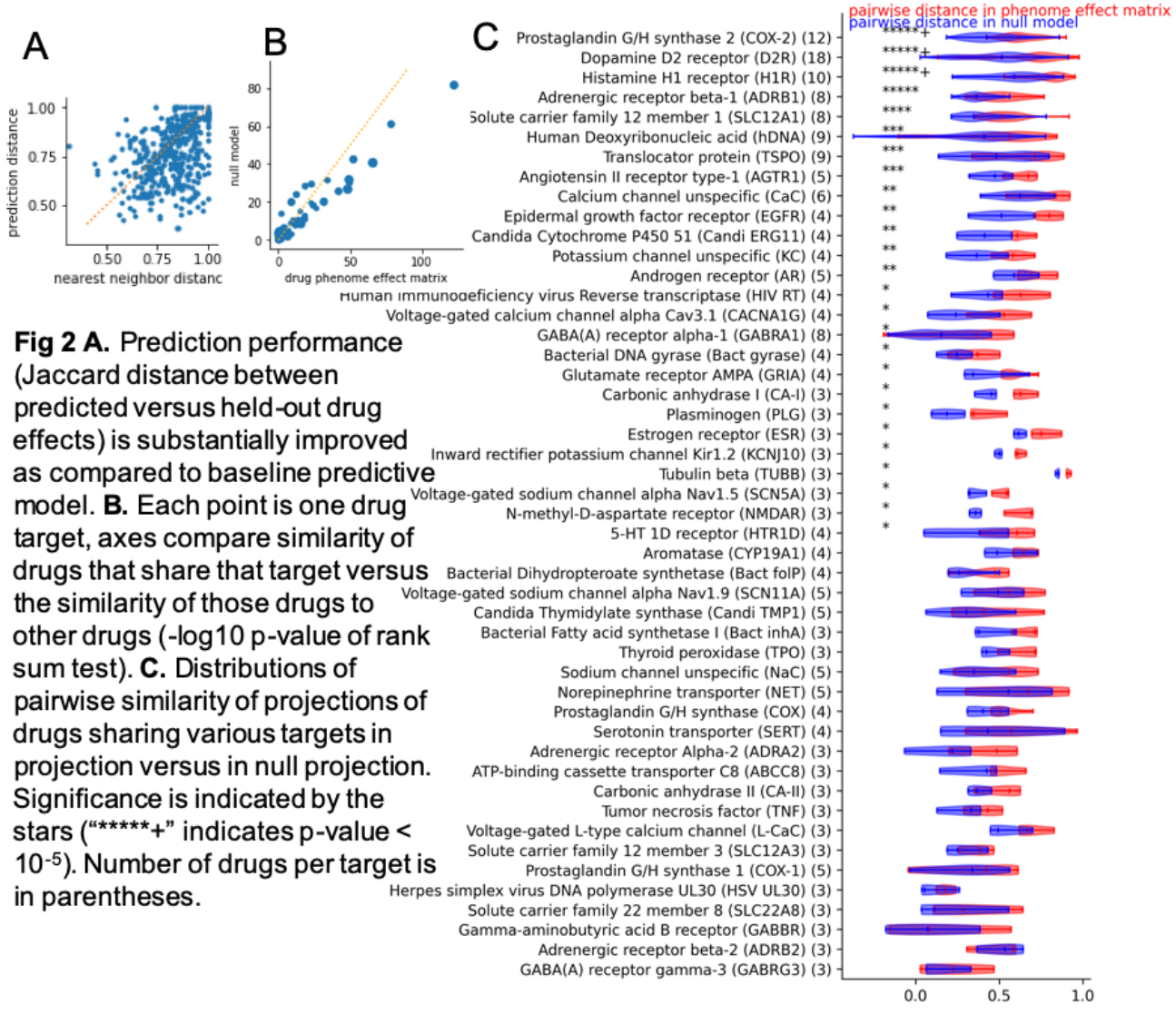

Fig 2 A. Prediction performance (Jaccard distance between predicted versus held-out drug effects) is substantially improved as compared to baseline predictive model. **B.** Each point is one drug target, axes compare similarity of drugs that share that target versus the similarity of those drugs to other drugs (-log10 p-value of rank sum test). **C.** Distributions of pairwise similarity 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.

3.4 Mapping drugs to their effect on the disease genome

To investigate the biological insight that can be gained from these mappings, we project each drug onto the space of its estimated impact on genetic regulation driving disease. Briefly, we use the inverse of our matrix decomposition to project the compact representation of each drug back to the space of disease genetic regulation. Then for each entry in the drug-genetics matrix, we compare the projected value against the projections obtained from null models (see Method). As a result, we create a matrix estimating for each drug the importance of its effect on each disease gene. Therefore,

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.** (Left) Drug significance for selected drugs with targets significantly associated with the disease genome. To the right, selected DrugBank targets of these drugs are shown. Orange lines highlight particular disease genes discussed in the text.

we call this estimated matrix the *drug-disease genome matrix.* This matrix is the result of connecting a drug's molecular endpoint profile (from *D*) to that drug's phenome effect, and then projecting the components of the phenome back to the gene level. Because we compare the strength of each drug-gene connection to a null model, these connections cannot be due only to the prior data on drug molecular effects, but must be due to the learned interaction matrix that estimates how molecular effects propagate to impact disease. 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, if drugs that share that target also share disease genes (see Methods). Across 132 DrugBank targets shared by at least three drugs, we find 28 that have one or more significantly associated disease genes at adjusted $p < 0.01$. Figure 3A shows that this level of association of drug disease genetics and drug targets is not likely to happen by chance. 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 therapeutic categories 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[32]; and another study linked its locus to rare mental illness[33]. Another interesting finding was the association of disease driver *CETP,* or cholesterol ester transfer protein, with fenofribrate 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[34,35]

4. Discussion

Our approach learns how drug molecular effects impact disease genes and result in drug effects on phenotype. We have demonstrated that our model both reflects known drug biology, and has 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[36].

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[12,37], no previous method has integrated this information across a range of diseases to build a drug-phenotype model. To our knowledge, ToxCast data 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, perhaps in a multi-task setting across multiple 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 are aiming to estimate. To address this issue and assess its affect on our results, we have taken steps including cross-validation and regularization; reducing the feature space to minimize the number of parameters possible in the model; and rigorous assessment of the resulting model.

The results we have already provided can be a starting point for multiple new analyses. It will be of interest to investigate the association of each ToxCast endpoint with disease genetics. As well, projecting the drug phenome effect matrix to biological pathways can further interpret the effects of drugs. It is possible to pursue a new categorization of drugs based on their effects on disease genes. Similarly, our results could be used to analyze how unexpected diseases can be linked by shared pathways related to drug mechanisms. 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.

1. Geifman, N., Brinton, R. D., Kennedy, R. E., Schneider, L. S. & Butte, A. J. Evidence for benefit of statins to modify cognitive decline and risk in Alzheimer's disease. *Alzheimer's Research & Therapy* **9**, 10 (2017).
2. Gronich, N. & Rennert, G. Beyond aspirin—cancer prevention with statins, metformin and bisphosphonates. *Nat Rev Clin Oncol* **10**, 625–642 (2013).
3. Kuhn, M., Letunic, I., Jensen, L. J. & Bork, P. The SIDER database of drugs and side effects. *Nucleic Acids Res* **44**, D1075–D1079 (2016).
4. Law, V. *et al.* DrugBank 4.0: Shedding new light on drug metabolism. *Nucleic Acids Research* **42**, 1091–1097 (2014).
5. Corsello, S. M. *et al.* The Drug Repurposing Hub: a next-generation drug library and information resource. *Nat Med* **23**, 405–408 (2017).
6. Cherkasov, A. *et al.* QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **57**, 4977–5010 (2014).
7. Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171**, 1437-1452.e17 (2017).
8. Richard, A. M. *et al.* ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. *Chem. Res. Toxicol.* **29**, 1225–1251 (2016).
9. Dudley, J. T. *et al.* Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Sci Transl Med* **3**, 96ra76 (2011).
10. Chen, B. *et al.* Reversal of cancer gene expression correlates with drug efficacy and reveals therapeutic targets. *Nature Communications* **8**, 16022 (2017).
11. Sirota, M. *et al.* Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci Transl Med* **3**, 96ra77 (2011).
12. So, H.-C. *et al.* Analysis of genome-wide association data highlights candidates for drug repositioning in psychiatry. *Nature Neuroscience* **20**, 1342–1349 (2017).
13. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics* **47**, 1091–1098 (2015).
14. Barbeira, A. N. *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature Communications* **9**, 1825 (2018).
15. Gottlieb, A., Stein, G. Y., Ruppin, E. & Sharan, R. PREDICT: a method for inferring novel drug

indications with application to personalized medicine. *Mol Syst Biol* **7**, 496 (2011).
16. Bakal, G., Kilicoglu, H. & Kavuluru, R. Non-Negative Matrix Factorization for Drug Repositioning: Experiments with the repoDB Dataset. *AMIA Annu Symp Proc* **2019**, 238–247 (2020).
17. Zhang, W., Xu, H., Li, X., Gao, Q. & Wang, L. DRIMC: an improved drug repositioning approach using Bayesian inductive matrix completion. *Bioinformatics* **36**, 2839–2847 (2020).
18. Wang, L., Li, X., Zhang, L. & Gao, Q. Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. *BMC Cancer* **17**, 513 (2017).
19. Galeano, D., Li, S., Gerstein, M. & Paccanaro, A. Predicting the frequencies of drug side effects. *Nature Communications* **11**, 4575 (2020).
20. Lau, A. & So, H.-C. Turning genome-wide association study findings into opportunities for drug repositioning. *Comput Struct Biotechnol J* **18**, 1639–1650 (2020).
21. Osmanbeyoglu, H. U., Pelossof, R., Bromberg, J. & Leslie, C. S. Linking signaling pathways to transcriptional programs in breast cancer. *Genome Research* **24**, 1869–1880 (2014).
22. Pelossof, R. *et al.* Affinity regression predicts the recognition code of nucleic acid–binding proteins. *Nature Biotechnology* **33**, 1242–1249 (2015).
23. Judson, R. S. *et al.* In vitro screening of environmental chemicals for targeted testing prioritization: the ToxCast project. *Environ Health Perspect* **118**, 485–492 (2010).
24. Rapid GWAS of thousands of phenotypes for 337,000 samples in the UK Biobank. *Neale lab* http://www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-for-337000-samples-in-the-uk-biobank.
25. Pividori, M. *et al.* PhenomeXcan: Mapping the genome to the phenome through the transcriptome. *Sci. Adv.* **6**, eaba2083 (2020).
26. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**, D267-270 (2004).
27. Hastie, T., Mazumder, R., Lee, J. & Zadeh, R. Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares. Preprint at http://arxiv.org/abs/1410.2596 (2014).
28. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* **57**, 289–300 (1995).
29. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* **29**, 1165–1188 (2001).
30. Wang, Y. *et al.* Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. *Nucleic Acids Research* **48**, D1031–D1041 (2020).
31. Wang, Y. *et al.* Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. *Nucleic Acids Research* **48**, D1031–D1041 (2020).
32. Goes, F. S. *et al.* Genome-wide association study of schizophrenia in Ashkenazi Jews. *Am J Med Genet B Neuropsychiatr Genet* **168**, 649–659 (2015).
33. Mallard, T. T. *et al.* Multivariate GWAS of psychiatric disorders and their cardinal symptoms reveal two dimensions of cross-cutting genetic liabilities. *Cell Genom* **2**, 100140 (2022).
34. Hansen, M. K. *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* **15**, 196–202 (2010).
35. Armitage, J., Holmes, M. V. & Preiss, D. Cholesteryl Ester Transfer Protein Inhibition for Preventing Cardiovascular Events. *J Am Coll Cardiol* **73**, 477–487 (2019).
36. Galeano, D. & Paccanaro, A. A Recommender System Approach for Predicting Drug Side Effects. in *2018 International Joint Conference on Neural Networks (IJCNN)* 1–8 (IEEE, 2018). doi:10.1109/IJCNN.2018.8489025.
37. Gerring, Z. F., Gamazon, E. R., White, A. & Derks, E. M. Integrative Network-Based Analysis Reveals Gene Networks and Novel Drug Repositioning Candidates for Alzheimer Disease. *Neurol Genet* **7**, e622 (2021).