Bioinformatics, 36(16), 2020, 4490–4497
doi: 10.1093/bioinformatics/btaa495
Advance Access Publication Date: 12 May 2020
Original Paper

Data and text mining

Revealing new therapeutic opportunities through drug target prediction: a class imbalance-tolerant machine learning approach

Siqi Liang [1,2] **and Haiyuan Yu** [1,2,*]

[1]Department of Computational Biology, Cornell University, Ithaca, NY 14853, USA, and [2]Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, NY 14853, USA

*To whom correspondence should be addressed.
Associate Editor: Jonathan Wren

Received on October 10, 2019; revised on February 18, 2020; editorial decision on May 5, 2020; accepted on May 6, 2020

Abstract

Motivation: *In silico* drug target prediction provides valuable information for drug repurposing, understanding of side effects as well as expansion of the druggable genome. In particular, discovery of actionable drug targets is critical to developing targeted therapies for diseases.

Results: Here, we develop a robust method for drug target prediction by leveraging a class imbalance-tolerant machine learning framework with a novel training scheme. We incorporate novel features, including drug–gene phenotype similarity and gene expression profile similarity that capture information orthogonal to other features. We show that our classifier achieves robust performance and is able to predict gene targets for new drugs as well as drugs that potentially target unexplored genes. By providing newly predicted drug–target associations, we uncover novel opportunities of drug repurposing that may benefit cancer treatment through action on either known drug targets or currently undrugged genes.

Contact: haiyuan.yu@cornell.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Target identification is a crucial step during drug development. As the cost of bringing a single new drug to market skyrockets to over 2.7 billion dollars on average (DiMasi *et al.*, 2016), alternative approaches, such as drug repurposing, have been pursued with increasing efforts. For example, the drug aspirin, commonly used for treating fever and acute pain, has been found in recent years to show anti-cancer activities through attenuation of *EGFR* expression (Li *et al.*, 2015), inhibition of COX-2 (Tsujii *et al.*, 1998) and suppression of NF-κB activation by TNF (Kutuk and Basaga, 2004). As a result, the efficacy of aspirin in treating multiple types of cancers, including breast cancer, prostate cancer and colorectal cancer, is being actively evaluated in clinical trials. By repurposing approved drugs for new indications through novel target discovery, the cost of drug development can be substantially reduced, especially in the preclinical and earlier clinical phases where the toxicity and dosage of the drug is assessed (Pushpakom *et al.*, 2019). In addition to benefiting drug repurposing efforts, identifying unknown targets of drugs can facilitate understanding of their side effects, which are often caused by drugs binding to unintended targets. The serotonin receptor agonist cisapride, as an example, is a gastroprokinetic agent used for treating gastric reflux, but it can cause serious cardiac events including arrhythmia and even lead to death. The mechanism behind the cardiac effects of cisapride was discovered in 1997 to be its high-affinity blocking of the human cardiac potassium channel (Rampe *et al.*, 1997). And this resulted in its withdrawal from the US market 3 years later. Furthermore, out of over 4400 genes estimated to be druggable in the human genome (Finan *et al.*, 2017), only less than half of them are currently targeted by approved drugs. Therefore, identification of novel gene targets can help with expanding the druggable genome, opening up new avenues for drug development.

Experimental methods for determining drug–target associations provide direct evidence and information on the mode of action of drugs. However, their high cost and long timeframe have prohibited them from large-scale application. As an alternative, computational approaches, including docking-based methods and machine learning-based methods, have been developed to predict new drug–target associations (Chen *et al.*, 2016). In particular, machine learning-based methods that exploit the chemogenomic space have yielded considerable success in drug target prediction without requiring 3D protein structures of the targets (Ezzat *et al.*, 2018; Jacob and Vert, 2008; Yamanishi *et al.*, 2008; Yu *et al.*, 2012; Zhao and Li, 2010). Various features, including chemical similarity

(Keiser *et al.*, 2007; 2009) and side effect similarity (Campillos *et al.*, 2008; Lounkine *et al.*, 2012), have proved valuable in identifying new associations between drugs and targets. Nevertheless, two fallacies are commonly overlooked: conventional train–test splitting and cross-validation schemes are flawed for pair-input prediction tasks (Park and Marcotte, 2012); extreme class imbalance in drug target datasets is not satisfactorily addressed by commonly used methods such as sampling from the majority class (Ezzat *et al.*, 2016). Moreover, most methods lack the ability to predict drug–target interactions for genes that are not yet known to be druggable.

To address these challenges, in this study, we design a novel training scheme that prevents possible overfitting caused by overlapping drugs or targets in the training and test sets and at the same time solves the class imbalance problem with an ensemble method. Additionally, we exploit two new types of features, namely the phenotype similarity between a drug and a gene, and the expression profile similarity between two genes across different tissues. We show that they confer considerable predictive power and provide orthogonal information that is not captured by other features. Incorporating these features, we build a classifier and demonstrate that it achieves robust performance. Further, our classifier is able to make predictions for drugs without known targets and for genes that are not yet known to be druggable. By predicting new potential drug–target associations, we reveal unexplored opportunities of drug discovery and repurposing for cancer treatment.

2 Materials and methods

2.1 Data collection

We collected a comprehensive dataset of known drug–gene associations by extracting relevant information for all drugs with human gene targets from the Probes and Drugs database (version 10.2018) (Skuta *et al.*, 2017). Side effects of drugs were obtained from SIDER 4.1 (Kuhn *et al.*, 2016) and OFFSIDES (Tatonetti *et al.*, 2012), both of which used Unified Medical Language System (UMLS) concept IDs as identifiers of side effects. However, as similar side effect terms could cause biases in calculating side effect similarity, we mapped all UMLS concept IDs to MedDRA concept IDs using the 2017AB release of UMLS (Bodenreider, 2004). This allowed us to map UMLS concept IDs to a specific level, Preferred Term (PT), of the MedDRA hierarchy obtained from MedDRA (version 21.0) (Brown *et al.*, 1999). Gene–disease associations were obtained from DisGeNET (version 5.0) (Pinero *et al.*, 2017). Similar to side effects, disease phenotype terms were also mapped to the PT level of the MedDRA hierarchy, allowing direct comparison with side effects. We only considered drugs with available side effect information and at least one human gene target whose association with disease phenotypes is known. This resulted in a final set of 11 556 drug–gene associations involving 1262 drugs and 1062 human genes. Since there is not a gold-standard dataset of non-targets of drugs, non-associated drug–gene pairs were obtained by taking all drug–gene combinations not known to be associated using these sets of drugs and genes.

2.2 Feature extraction and selection

Similarity-based features have been widely used for drug target prediction (Ding *et al.*, 2014). Behind them is a simple motivating hypothesis: similar drugs tend to have the same gene targets, and vice versa, similar genes tend to be targeted by the same drugs. Among various drug–drug similarity metrics, chemical similarity and side effect similarity have been most extensively employed (Campillos *et al.*, 2008; Keiser *et al.*, 2007, 2009; Lounkine *et al.*, 2012). Chemical similarity was calculated by taking the Tanimoto similarity of the fingerprints of the drugs, which are bit vectors of fixed sizes where each bit characterizes the drug by indicating presence or absence of a defined structural fragment:

$$\text{Tanimoto}(V_a,\ V_b) = \frac{V_a \cdot V_b}{|V_a|_1 + |V_b|_1 - V_a \cdot V_b} \tag{1}$$

While 2D similarity utilizes Morgan fingerprints, which represent planar chemical substructures (Rogers and Hahn, 2010), a method for encoding the 3D structure of molecules has been developed and has been shown to enhance the performance of conventional 2D fingerprinting methods in binding prediction (Axen *et al.*, 2017). We calculated Morgan fingerprints of compounds with the RDKit Python package, and generated 3D fingerprints with the E3FP Python package. Side effect similarity was calculated by taking the Jaccard index of the sets of side effects of the drugs mapped to the PT level of the MedDRA hierarchy:

$$\text{Jaccard}(S_a,\ S_b) = \frac{|S_a \cap S_b|}{|S_a \cup S_b|} \tag{2}$$

For each type of drug–drug similarity (2D chemical, 3D chemical, side effect), we calculated two groups of features: (i) similarity between the drug in question and drugs that are known to target the gene in question (Fig. 1a); and (ii) similarity between the drug in question and drugs that are known to target protein interactors of the gene in question (Fig. 1b). Since similarity for multiple drug pairs were calculated for each drug–gene pair, aggregation functions were applied to obtain feature values for drug–gene instances. For the former group, four different aggregation functions were applied to each type of feature: min, mean, median and max, while for the latter group, mean was replaced by first applying the mean to each set of drugs that were known to target a single protein interactor before applying a second mean function to obtain a single value.

In addition to aforementioned similarity metrics, which have already been incorporated in previous drug–target prediction methods, here we introduce two novel types of features: drug–gene phenotype similarity and expression profile similarity between two genes. Drugs that act directly on a protein and alter its activity may lead to similar phenotypic changes as mutations on the corresponding gene. On this account, we designed a drug–gene phenotype similarity metric by taking the Jaccard index of the side effects of the drugs and disease phenotypes of the gene (Fig. 2a). We also considered protein interactors of the gene in question and calculated their phenotypic resemblance to our drug (Fig. 2b). Similar to drug–drug similarity metrics, we obtained a group of four features by aggregation with min, mean, median and max. In addition to drug–gene

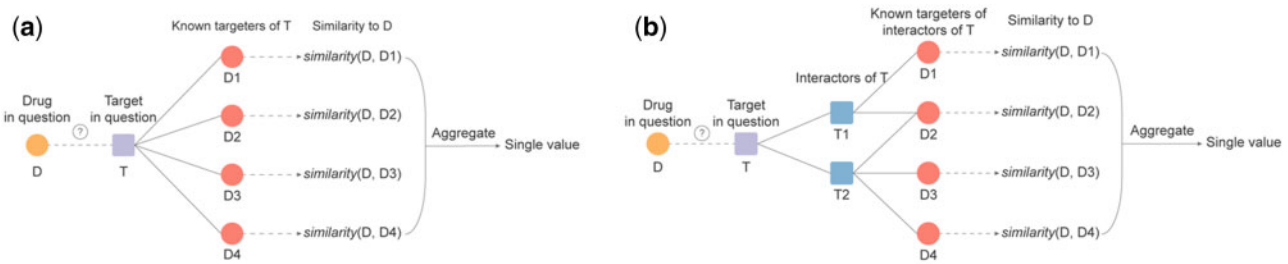

Fig. 1. Calculation of drug–drug similarity features. (**a**) Schematics of calculating drug–drug similarity features for each drug–gene pair. Each group of features consists of four features corresponding to four different aggregation functions. (**b**) Schematics of calculating drug–drug similarity features considering known targeters of protein–protein interaction partners of the target in question

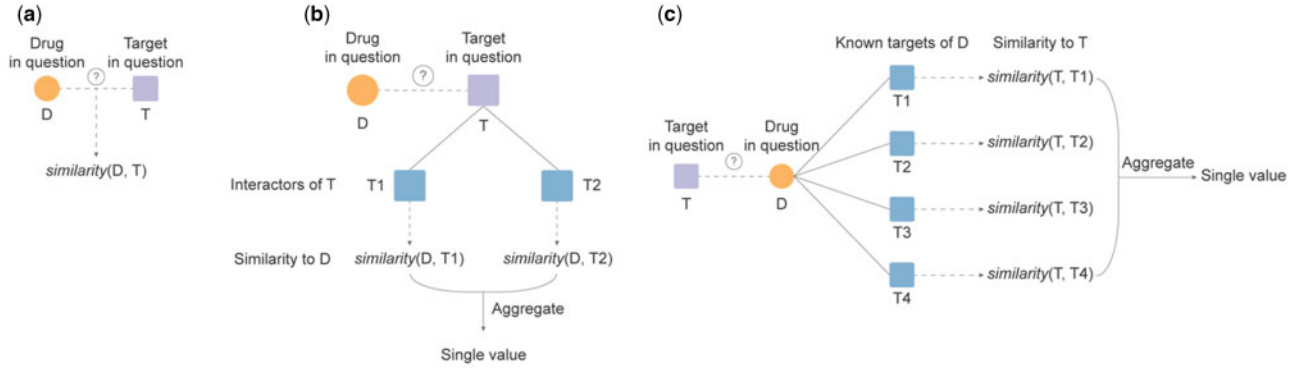

Fig. 2. Calculation of drug–gene similarity features and gene–gene similarity features. (**a**) Schematics of calculating the drug–gene phenotype similarity feature. (**b**) Schematics of calculating drug–gene phenotype similarity features considering protein–protein interaction partners of the target in question. (**c**) Schematics of calculating gene expression profile similarity features. This group of features consists of four features corresponding to four different aggregation functions

similarity, we calculated similarity between two genes by taking their Spearman correlation coefficient in expression levels across different tissues using gene expression data from GTEx (Consortium, 2013). For each drug–gene pair, we considered the similarity between the gene in question and known gene targets of the drug in question. Application of the same aggregation functions resulted in another group of four features (Fig. 2c). Overall, we calculated five types of similarity metrics: 2D chemical similarity between drugs, side effect similarity between drugs, 3D chemical similarity between drugs, drug–gene phenotype similarity and expression profile similarity between genes. In total, we extracted a total of 33 features across 9 groups for each drug–gene pair (Supplementary Table S1).

Since these similarity features utilize the whole drug–gene association network as well as the protein–protein interaction network, care needs to be taken when calculating features for the training/validation/test sets. When calculating the feature matrix for the full training set, only drugs and genes used for training and connections among them were regarded as known (Supplementary Fig. S1a). On the other hand, when calculating the feature matrix for the test set, all associations except those between test drugs and test targets were treated as known (Supplementary Fig. S1b). This principle also applied to training and validation sets during hold-out validation, where the set used for fitting the classifier was analogous to the training set and the validation set was equivalent to the test set.

To obtain an optimal feature combination, we calculated all features for the full training set and applied group maximum concave penalty (MCP) (Breheny and Huang, 2011) with the *grpreg* R package for feature selection using default parameters. All subsequent training was done with this optimal set of 14 features (Supplementary Table S1).

2.3 The training scheme and hyperparameter optimization

In order to build a machine learning model for drug–target prediction, we divided all drug–gene pairs into a training set and a test set. If the split is random in this pair-input prediction setting, the overall test set performance is not representative of the different classes within all test set instances (Park and Marcotte, 2012). More specifically, in the problem of drug–target prediction, test pairs sharing no drugs or targets with the training set would perform much more poorly than those sharing both drugs and targets with the training set. In order to build a classifier that can be applied to the most general scenario where either the drug or the gene target, or even both, may have no known drug–gene association, we applied a splitting scheme where the drugs were first randomly divided into 'train drugs' and 'test drugs' with a 2:1 ratio, and the genes were similarly split into 'train targets' and 'test targets' with the same ratio (Fig. 3a), guaranteeing that there is no overlap between the training set and the test set in terms of either drugs or genes. The training set then consisted of all drug–gene pairs where the drug is a 'train drug' and the target is a 'train target'; the test set consisted of all drug–gene pairs where the drug is a 'test drug' and the target is a 'test target' (Fig. 3a).

Since there was no gold-standard dataset of non-associated drug–gene pairs, all drug–gene pairs not known to be associated were considered as non-associated. This resulted in an extreme class imbalance where negative instances were over 100 folds more than positive instances in quantity. Class imbalance is usually dealt with by providing a weight for each class to place a higher penalty for misclassifying the minority class (Chen *et al.*, 2004). However, an extreme imbalance could be detrimental to classifier fitting. Other common approaches include oversampling the minority class and undersampling the majority class. While these methods do not prevent the classifier from improper fitting, the former introduces copies of data of the minority class, leading to increased likelihood of overfitting (Chawla *et al.*, 2002), while the latter does not make use of all instances of the minority class, leaving out valuable information for the classification task (Ganganwar, 2012). To address the extreme class imbalance problem here, we adopted an approach similar to that proposed by Ezzat *et al.* (2016), splitting all drug–gene pairs with negative labels into subsets, each having a size 5 times that of all the positive labels in the training set expect the last subset (which has a size between 5 to 10 times that of all the positive labels in the training set). Each subset of negative labels was combined with all the instances with positive labels in the training set to obtain a training subset (Fig. 3b). In this way, we made use of all data—especially those with negative labels—while keeping a reasonable class ratio within each training subset, which will be used for classifier fitting.

For each training subset, we trained an extreme gradient boosting (XGBoost) classifier (Chen and Guestrin, 2016). XGBoost is a decision tree ensemble model that additively trains decision trees that predict the prediction error of the existing ensemble. We chose XGBoost for its speed and ability to automatically learn branch directions for missing values. In each classifier the `scale_pos_weight` hyperparameter was set to the negative-to-positive class ratio in the corresponding training subset. At the end, we applied an ensemble approach by taking the average prediction score of all the classifiers trained as the final prediction score.

To find the best set of hyperparameters for each classifier, we adopted the tree-structured Parzen estimator (TPE) approach (Bergstra *et al.*, 2011). Instead of cross-validation, we split the full training set into training and validation sets using the same splitting method as the train–test split to ensure that there was no overlap between data used for training and validation in terms of either drugs or genes (Fig. 3c). For each classifier, the part used for training was then intersected with the corresponding training subset before used as input to the classifier, while the entire validation set was used for performance evaluation. For each classifier and each set of hyperparameter, the split was conducted 15 times, and we selected 1 minus the average area under the precision-recall curve (AUPR) of the 15 trials of hold-out validation as the loss function to minimize

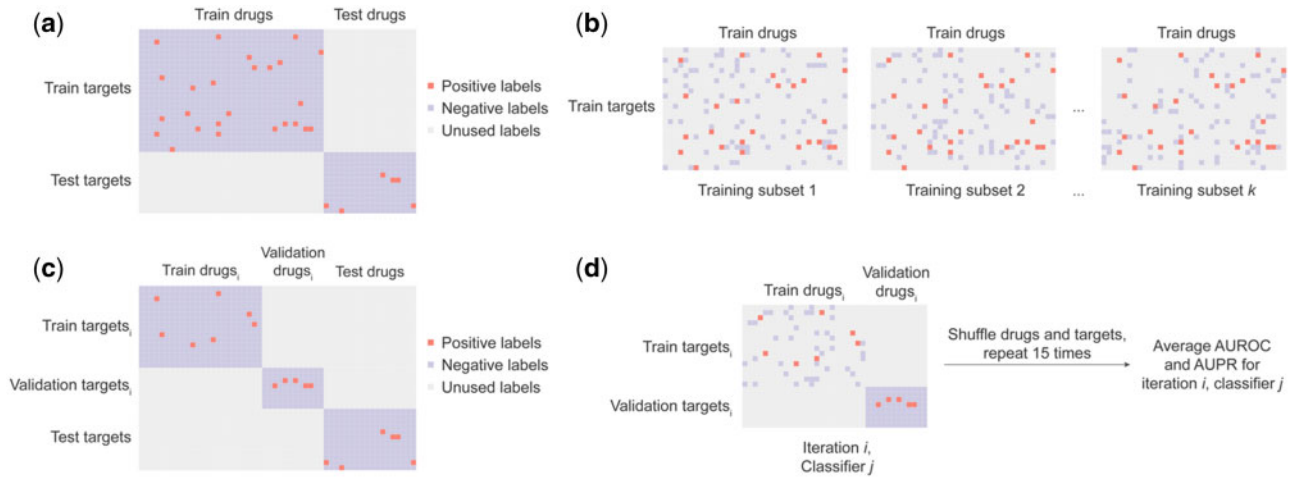

Fig. 3. Data splitting and the training scheme. (**a**) The train–test split. All drugs were split into 'train drugs' and 'test drugs', while all genes were split into 'train targets' and 'test targets'. The training set then consisted of all drug–gene pairs where the drug was a 'train drug' and the target was a 'train target'. Similarly, the test set consisted of all drug–gene pairs where the drug was a 'test drug' and the target was a 'test target'. (**b**) The training set was split into multiple training subsets by splitting all the training examples with a negative label into subsets, each combined with all examples with a positive label to form a training subset. (**c**) The entire training set was split into one portion used for fitting the classifiers and one portion used for validation using a similar splitting method as the train–test split so that there was no overlap between data used for fitting the classifiers and data used for validation in terms of either drugs or genes. (**d**) For each classifier and each set of hyperparameters, the train–validation split was done 15 times. Each time the training data was intersected with the corresponding training subset before used for classifier fitting, while the entire validation set was used for performance evaluation using AUROC and AUPR as metrics. The average AUROC and AUPR over the 15 splits were considered as the performance for that specific set of hyperparameters

for TPE (Fig. 3d). We ran TPE for 1000 iterations to obtain the best set of hyperparameters that minimized the loss function for each classifier in our ensemble. After obtaining the optimal sets of hyperparameters, we retrained each classifier using all data from the corresponding training subset.

2.4 Model evaluation and application

Model performance on the training set was evaluated by taking the average performance across all classifiers in the ensemble using performance metrics calculated from the best sets of hyperparameters during hyperparameter optimization, including average AUROC and AUPR across 15 splits. To evaluate the contribution of each type of feature to model performance, we dropped each type of features and repeated training and hyperparameter optimization procedures. To further evaluate the predictive power of our model, we predicted on the left-out test set which had no overlap with training data in terms of either drugs or genes. In order to evaluate the extent to which our training scheme prevents overfitting, we compared our training and test set performance against that derived from hyperparameter optimization with conventional 5-fold cross-validation on each of the training subsets. Finally, to apply our model to predicting new drug targets, we considered all drug–gene pairs that were not previously known to be associated where the drug had known side effects and the gene had known disease phenotypes. A total of 9958 new drug–gene associations were identified with a precision lower bound cutoff at 10% (Supplementary Table S2).

3 Results

3.1 Predictive power of similarity-based features

To determine whether the newly proposed features are informative for predicting drug targets, we compared feature values of known associated drug–gene pairs with those of other drug–gene pairs. Not surprisingly, when aggregating by the maximum, mean or median, drugs are significantly more chemically similar to known targeters of their gene target than to known targeters of other genes (Fig. 4a), using Morgan fingerprints as representations of molecular structures. On a similar note, measuring drug–drug similarity by taking the set similarity of their side effects gave identical trends (Fig. 4b). Interestingly, when aggregating similarity scores by the minimum, drug–gene pairs that are known to be associated had significantly lower scores than those that are not known to be associated, regardless of the type of similarity metric used (Fig. 4a and b). This can be explained by the fact that genes in associated drug–gene pairs have a significantly higher number of known targeters in a broader chemogenomic space than genes in other drug–gene pairs (Supplementary Fig. S2a). Furthermore, features utilizing 3D molecular fingerprints as the chemical similarity metric uncovered similar trends as 2D chemical similarity and side effect similarity (Fig. 4c). But notably, 3D chemical similarity features are only weakly correlated with 2D chemical similarity and side effect similarity features (Supplementary Fig. S2b). This indicates that 3D chemical similarity brings in information that is not be captured by 2D chemical similarity, highlighting the importance of including both for our prediction task.

Our newly proposed features, drug–gene similarity and gene expression profile similarity, also exhibit classifying power in distinguishing known drug–target pairs from other drug–gene pairs. As expected, drug–gene pairs that are known to be associated have significantly higher phenotype similarity scores than drug–gene pairs that are not known to be associated (Fig. 5a). For gene expression similarity features, we discovered that when taking maximum, mean or median as the aggregation function, genes have significantly more similar expression profiles to known targets of their targeters than to known targets of other drugs (Fig. 5b). Using minimum as the aggregation function rendered the opposite trend, which could be explained by drugs in drug–gene pairs that are known to be associated having a significantly more diverse target set than drugs in other drug–gene pairs (Supplementary Fig. S2c). Intriguingly, expression profile features, especially when aggregated with maximum, mean or median, exhibit almost no correlation with other groups of features, bringing in complementary information that other features do not capture (Fig. 5c).

It is worth noticing that drug–gene phenotype similarity and gene expression profile similarity features can be calculated even if the gene in question has no known drugs that target it. This potentializes us to make predictions for currently undrugged genes, thereby expanding the druggable genome. Our consideration of drugs that are known to target protein–protein interaction partners of the gene in question for both chemical similarity and side effect similarity (Fig. 1b) extends this advantage to drug–drug similarity features. In addition, we considered protein–protein interactors of the gene in

question for drug–gene phenotype similarity (Fig. 2b). These groups of features also possess distinguishing power in separating drug–target pairs and other drug–gene pairs (Supplementary Fig. S3a–d), and more importantly, the fact that they do not require our gene of interest to be targeted by any known drug enhances the ability of our model to make predictions for genes that remain undrugged.

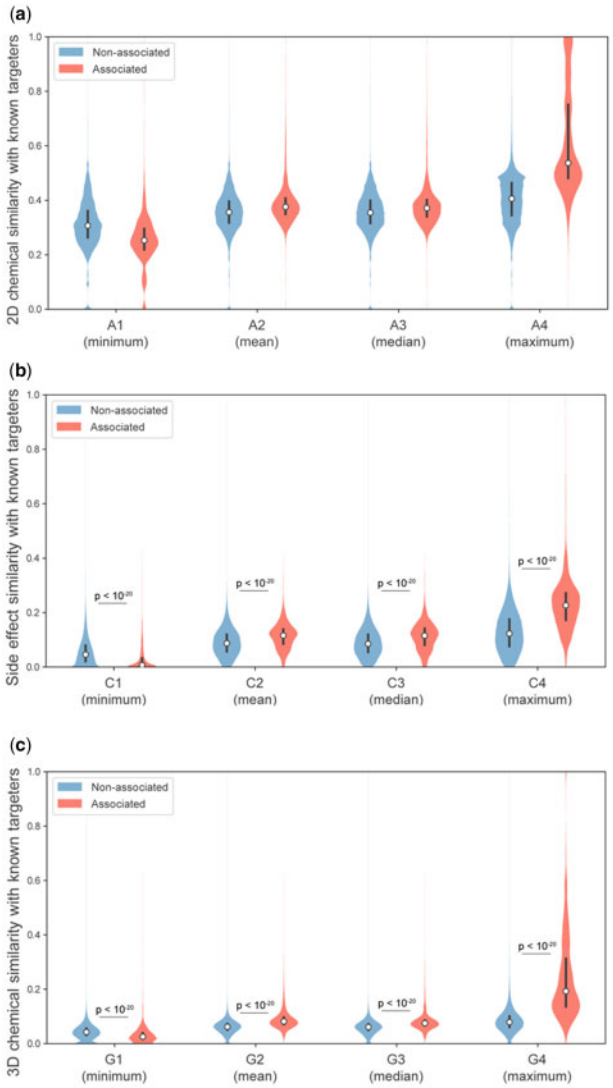

Fig. 4. Drug–drug similarity feature distributions. (**a**) Distribution of 2D chemical similarity features (feature group A). (**b**) Distribution of side effect similarity features (feature group C). (**c**) Distribution of 3D chemical similarity features (feature group G) (statistical significance determined by the two-sided Mann–Whitney *U*-test)

3.2 Performance of the novel training scheme

After feature selection (see Section 2), the final set of features comprised of 14 features, with every type of feature present, including the newly proposed drug–gene phenotype similarity and expression profile similarity features. Using a Bayesian TPE approach (Bergstra *et al.*, 2011), which has recently been shown to improve classifier performance drastically (Meyer *et al.*, 2018), we optimized hyperparameters for all classifiers in the ensemble under the new training scheme that we proposed. This resulted in an average training AUROC of 0.924 across all classifiers in the ensemble and an average training AUPR of 0.273 (Table 1). When evaluated on the hold-out test set which has no overlap with the training data in terms of either drugs or genes, we obtained an AUROC of 0.928 (Fig. 6a) and an AUPR of 0.268 (Fig. 6b). This similar performance between training and evaluation on a test set that does not overlap with the training set in terms of either drugs or genes demonstrates that our model is not subject to overfitting and illustrates the robustness of our training scheme. Notably, our model attained a precision of 78% on the top 50 predictions and a precision of 48.2% when examining the top 500 predictions. Considering the fact that these are lower bound estimates since drug–gene pairs labeled as non-associated could actually be undiscovered drug–target pairs, our model achieves accurate drug target prediction in the most general scenario where the drug and the gene could have no previously known drug–gene associations.

To demonstrate the effectiveness and necessity of our novel training scheme in preventing overfitting, we conducted an experiment with the same training subsets and test set, using conventional 5-fold cross-validation for hyperparameter optimization rather than the hold-out validation with no overlap splitting applied in our training scheme. Although average training AUROC was as high as 0.966, AUROC and AUPR on the same test set only reached 0.912 (Fig. 6a) and 0.239 (Fig. 6b), respectively, substantially lower than those obtained using our new training scheme. Furthermore, the large discrepancy between training and test AUROC values indicate that conventional cross-validation is prone to overfitting in this pair-input setting and that our training scheme is robust to this problem.

3.3 Model evaluation

To assess the contribution of each type of features to model performance, we dropped each type of features and retrained our classifier ensemble using the same number of TPE iterations. As expected, model performance declined when any of the five groups of features was taken out. Interestingly, exclusion of expression profile similarity resulted in the largest performance drop, followed by exclusion of 3D chemical similarity (Table 1). These results demonstrate that every type of feature contributes to classifier performance, and that

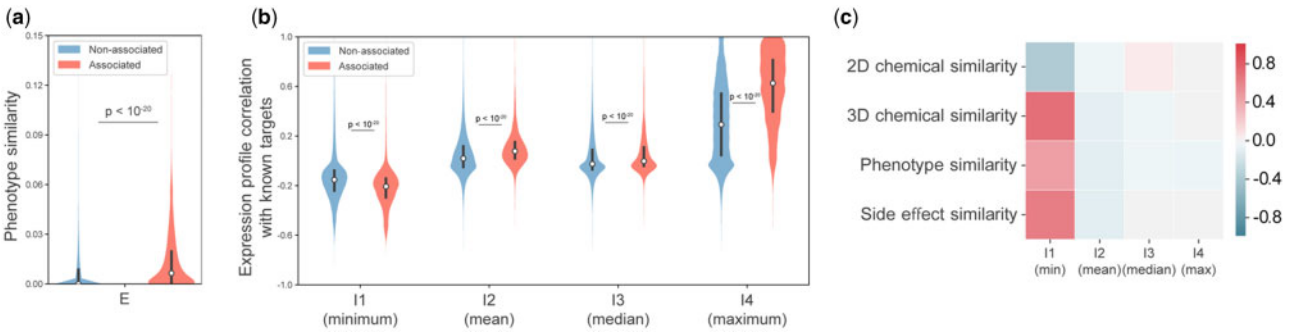

Fig. 5. Distribution of phenotype similarity and gene expression similarity features. (**a**) Distribution of the drug–gene phenotype similarity feature (feature group E). (**b**) Distribution of gene expression profile similarity features (feature group I). (**c**) Spearman correlation coefficients of gene expression similarity features with other types of features. Only the correlation coefficient corresponding to the most correlated or anti-correlated feature in each type of features is shown (statistical significance determined by the two-sided Mann–Whitney *U*-test)

Table 1. Classifier performance after dropping each feature type

Features	AUROC	AUPR
All features selected	**0.924**	**0.273**
Drop 2D chemical similarity	0.922	0.261
Drop side effect similarity	0.921	0.260
Drop phenotype similarity	0.923	0.268
Drop 3D chemical similarity	0.909	0.226
Drop expression profile similarity	0.903	0.209

Notes: AUROC and AUPR values indicate the average over all classifiers in the ensemble on the training set. $p < 0.001$ by two-sample t-test for all comparisons between the bold row with each row below (for both columns).

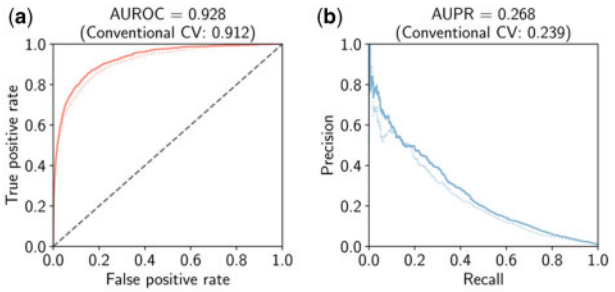

Fig. 6. Performance evaluation. (**a**) Receiving operating characteristic (ROC) curve when evaluating the model on the hold-out test set. Dashed line indicates ROC curve of the classifier tuned with conventional cross-validation. (**b**) Precision–recall curve when evaluating the model on the hold-out test set. Dashed line indicates precision–recall curve of the classifier tuned with conventional cross-validation

Table 2. Performance comparison on the test set

Method	AUROC	AUPR
Ezzat *et al.* (2016)	0.685	0.035
Wen *et al.* (2017)	0.521	0.009
Our method	**0.928**	**0.268**

Notes: statistical significance does not apply, since this is only a one-shot evaluation performed on the independent test set.

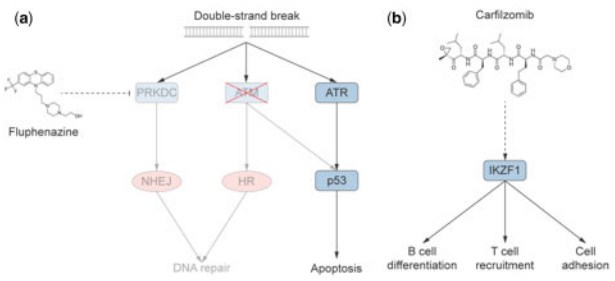

Fig. 7. Novel therapeutic opportunities for cancer uncovered by drug target predictions. (**a**) Fluphenazine is predicted to target DNA-PKc encoded by the PRKDC gene. DNA-PKc is a key mediator of non-homologous end joining (NHEJ), which is an alternative mechanism for DNA double-strand break (DSB) repair to homologous recombination (HR). Fluphenazine can potentially be repurposed to treat ATM-deficient cancer by disabling NHEJ. (**b**) Carfilzomib is predicted to target the transcription factor IKZF1, which is a modulator of immune responses. Activation of IKZF1 enhances cell adhesion and promotes B cell differentiation and T cell recruitment. This renders IKZF1 a potential drug target that might enhance the efficacy of immunotherapy when activated

our newly proposed features are crucial components in our model for prediction of drug targets.

To demonstrate that our method achieves superior performance on a highly imbalanced test set with no drugs and genes overlapping with the training set, we compared the performance of two previously published feature-based methods trained on our training set and tested on our test set (Ezzat *et al.*, 2016; Wen *et al.*, 2017). As shown in Table 2, our classifier substantially outperforms both drug–target prediction methods.

3.4 Cancer treatment opportunities revealed by newly predicted drug–target associations

We applied our trained model on drug–gene pairs that were not previously known to be associated and predicted novel drug–target associations. By examining newly predicted associations between known drugs and genes that are already known to be druggable (Supplementary Table S3), we discovered new drug repurposing opportunities. As an example, the antipsychotic drug, fluphenazine, is predicted to target the *PRKDC* gene with high probability. Commonly prescribed for treatment of schizophrenia, fluphenazine primarily acts on dopamine receptors and G-protein coupled receptors (Bisson *et al.*, 2007; Seeman, 2002). The *PRKDC* gene encodes a DNA-dependent protein kinase (DNA-PKc) that mediates non-homologous end joining (NHEJ) (Ma *et al.*, 2004), which is an important mechanism by which cells can repair double-strand breaks (DSB) in DNA without a homologous template (Davis and Chen, 2013). Cells deficient of the *ATM* gene, which is commonly mutated in various types of cancers (Choi *et al.*, 2016), can evade p53-mediated apoptosis but become reliant on NHEJ for DSB repair (Jiang *et al.*, 2009). Therefore, while *ATM*-deficient cancers are largely resistant to genotoxic chemotherapy, it has been reported that exposure to a DNA-PKc inhibitor diminishes their ability to repair DSBs and prolongs survival using a *ATM*-deficient mouse lymphoma model (Riabinska *et al.*, 2013). This, along with studies in other cancer types (Muraki *et al.*, 2013; Tanori *et al.*, 2019), establishes the *PRKDC* gene as a potential target for treatment of *ATM*-deficient cancer and opens up the possibility of repurposing fluphenazine for cancer treatment (Fig. 7a).

In addition to predicting associations between known drugs and genes that are already known as drug targets, we also predicted potential associations between drugs and genes that are not yet known to be druggable, in an effort to expand the druggable genome (Supplementary Table S4). For instance, carfilzomib and *IKZF1* is the drug–target pair with the highest probability to be associated where the gene is not yet known to be druggable. *IKZF1* encodes a transcription factor and has been shown to have tumor suppressive function during leukemia development (Payne and Dovat, 2011). It is a modulator of immune responses and its activation facilitates recruitment of T cells (Fig. 7b). Furthermore, according to a recent publication, overexpression of *IKZF1* in tumors results in significantly improved responses to immunotherapy, including anti-PD1 and anti-CTLA4 treatment (Chen *et al.*, 2018). These discoveries have rendered *IKZF1* an ideal candidate for drug development and showcase the power of our drug target prediction method in expanding the druggable genome.

4 Discussion

In this article, we introduce a machine learning method for drug–target prediction that leverages newly proposed features and a novel training scheme. We demonstrate that the new features, including drug–gene phenotype similarity and gene expression profile similarity features, provide complementary information that other features do not capture and enhance the predictive power of our model. In addition, we show that our novel training scheme warrants robust prediction by preventing overfitting and our model achieves accurate prediction while possessing the ability to predict associations for new drugs or currently "undrugged" genes. By doing so, we predict new drug–gene associations and reveal previously unexplored opportunities for drug repurposing and expansion of the druggable genome.

In a conventional train–test split setting, drugs and genes that show up in both the training set and the test set may cause overfitting through data leakage. Although several papers have proposed

alternate splitting schemes and experimental settings, including splitting based on drugs or genes (Yu *et al.*, 2012), few of them adjust their cross-validation strategy, leading to biased performance estimates. Mayr *et al.* introduced a cluster-cross-validation strategy where drugs undergo clustering before being split into training and test sets (Mayr *et al.*, 2018). The resulting training and test sets contain drugs from different clusters and thus have no drugs that are similar. However, although this strategy works in the scenario where only single drug-based descriptors are used as features, it decreases performance in our settings where similarity-based metrics between drug pairs are important features for our predictions (Supplementary Table S5). Pahikkala *et al.* proposed a nested cross-validation technique for the same splitting scheme used here (Pahikkala *et al.*, 2015), yet its complexity renders it computationally expensive considering the fact that in this drug–target prediction problem the training and test feature matrices need to be re-calculated for each fold (Supplementary Fig. S1a and b). Our newly proposed hold-out validation scheme achieves the same effect of avoiding overfitting with less computation while integrating well with the hyperparameter optimization method. Along with the ensemble approach to solving the class imbalance problem, it can be readily applied to other biological prediction problems, especially those involving biological networks. While other methods specifically designed for predicting drug–gene associations within the explored drug and gene space might become more favorable as more drug targets are discovered, they too would benefit from incorporating novel features introduced in this study, such as drug–gene phenotype similarity and gene expression profile similarity, as well as adapting the ensemble method in the scenario of extreme class imbalance.

In addition to 2D and 3D fingerprints that we used for feature construction, recent works have explored various alternative ways of representing drugs, including SMILES strings, graph-based representation, image-based representation and molecular descriptors. SMILES strings represent the chemical structure of molecules as a sequence of characters, and they have been widely applied to prediction tasks including drug–target prediction (Ozturk *et al.*, 2016), chemical–chemical interaction prediction (Kwon and Yoon, 2017) and drug toxicity prediction (Cao *et al.*, 2012). Although a SMILES string-based chemical similarity metric has been developed, replacing either 2D or 3D chemical similarity with this SMILES string-based similarity resulted in a decrease in performance (Supplementary Table S6). Nevertheless, future studies may provide better SMILES string-based similarity metrics that could potentially improve drug–target prediction. Molecular graphs constructed from the chemical structure of compounds can be processed by graph neural networks, and this representation has been employed for predicting drug–target binding affinity (Nguyen *et al.*, 2019). Image-based representation of drugs, compatible with convolutional neural networks, has been exploited for drug–target prediction (Rifaioglu *et al.*, 2018) and drug function prediction (Meyer *et al.*, 2019). Finally, drugs can be characterized by molecular descriptors that encode their chemical information, and softwares for calculating molecular descriptors such as Dragon (Mauri *et al.*, 2006) and Mordred (Moriwaki *et al.*, 2018) have been extensively applied in drug-related prediction tasks. It is worth noticing that graph- and image-based drug representations each require specific algorithms that are compatible with them, and that molecular descriptors are features for drugs alone, and are therefore not easily integrable with our algorithm, which uses similarity-based features calculated from the network of drugs and genes. However, it would be interesting to see future work that integrate these alternative algorithms and the training scheme introduced in this paper.

One potential limitation of our approach lies in the construction of the non-interacting drug–gene set by taking all drug–gene pairs not known to be interacting. Although this has been the most common approach to representing the non-interaction space in a supervised learning framework (Ezzat *et al.*, 2016; Wen *et al.*, 2017; Yamanishi *et al.*, 2008; Yu *et al.*, 2012), the fact that possible drug–target associations could exist in the negative set may lead to inaccuracies in feature calculation and classifier fitting. Therefore, our method could potentially be improved in the future by applying a positive-unlabeled learning framework, which comprises of a number of techniques that avoid treating all unlabeled data as negatives.

As more chemogenomic and phenotypic information about compounds and genes becomes available, we expect that our method will reach even better performance. Since there is no universal method for drug repurposing and druggable gene discovery in place (Finan *et al.*, 2017; Makley and Gestwicki, 2013; Novac, 2013; Pushpakom *et al.*, 2019), our drug target prediction method can serve as an intermediate step in the drug discovery pipeline, generating reliable candidates which can then be tested by downstream experimental validation. This could greatly accelerate the drug development process and create new opportunities for disease treatment.

Acknowledgements

The authors would like to thank G. Hooker, J. F. Beltrán, D. Xiong, S. Qian and S. Chen for helpful discussions.

Funding

This work was supported by National Institute of General Medical Sciences grants [R01 GM124559, R01 GM125639]; and National Science Foundation [DBI-1661380 to H.Y.].

Conflict of Interest: none declared.

References

Axen,S.D. *et al.* (2017) A simple representation of three-dimensional molecular structure. *J. Med. Chem.*, **60**, 7393–7409.

Bergstra,J. *et al.* (2011) Algorithms for hyper-parameter optimization. In: *Proceedings of the 24th International Conference on Neural Information Processing Systems*, 2011, Granada, Spain, pp. 2546–2554.

Bisson,W.H. *et al.* (2007) Discovery of antiandrogen activity of nonsteroidal scaffolds of marketed drugs. *Proc. Natl. Acad. Sci. USA*, **104**, 11927–11932.

Bodenreider,O. (2004) The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, **32**, D267–D270.

Breheny,P. and Huang,J. (2011) Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Stat.*, **5**, 232–253.

Brown,E.G. *et al.* (1999) The medical dictionary for regulatory activities (MedDRA). *Drug Saf.*, **20**, 109–117.

Campillos,M. *et al.* (2008) Drug target identification using side-effect similarity. *Science*, **321**, 263–266.

Cao,D.S. *et al.* (2012) In silico toxicity prediction by support vector machine and SMILES representation-based string kernel. *SAR QSAR Environ. Res.*, **23**, 141–153.

Chawla,N.V. *et al.* (2002) SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, **16**, 321–357.

Chen,C. *et al.* (2004) *Using Random Forest to Learn Imbalanced Data*. University of California, Berkeley, CA.

Chen,J.C. *et al.* (2018) IKZF1 enhances immune infiltrate recruitment in solid tumors and susceptibility to immunotherapy. *Cell Syst.*, **7**, 92–103.e104.

Chen,T. and Guestrin,C. (2016) XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 2016, San Francisco, CA, USA, pp. 785–794.

Chen,X. *et al.* (2016) Drug-target interaction prediction: databases, web servers and computational models. *Brief Bioinform.*, **17**, 696–712.

Choi,M. *et al.* (2016) ATM mutations in cancer: therapeutic implications. *Mol. Cancer Ther.*, **15**, 1781–1791.

Consortium,G.T. (2013) The genotype-tissue expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.

Davis,A.J. and Chen,D.J. (2013) DNA double strand break repair via non-homologous end-joining. *Transl. Cancer Res.*, **2**, 130–143.

DiMasi,J.A. *et al.* (2016) Innovation in the pharmaceutical industry: new estimates of R&D costs. *J. Health Econ.*, **47**, 20–33.

Ding,H. *et al.* (2014) Similarity-based machine learning methods for predicting drug-target interactions: a brief review. *Brief Bioinform.*, **15**, 734–747.

Ezzat,A. *et al*. (2016) Drug-target interaction prediction via class imbalance-aware ensemble learning. *BMC Bioinformatics*, **17**, 509.

Ezzat,A. *et al*. (2019) Computational prediction of drug-target interactions using chemogenomic approaches: an empirical survey. *Brief Bioinform*., **20**, 1337–1357.

Finan,C. *et al*. (2017) The druggable genome and support for target identification and validation in drug development. *Sci. Transl. Med*., **9**, eaag1166.

Ganganwar,V. (2012) An overview of classification algorithms for imbalanced datasets. *Int. J. Emerging Technol. Adv. Eng*., **2**, 42–47.

Jacob,L. and Vert,J.P. (2008) Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics*, **24**, 2149–2156.

Jiang,H. *et al*. (2009) The combined status of ATM and p53 link tumor development with therapeutic response. *Genes Dev*., **23**, 1895–1909.

Keiser,M.J. *et al*. (2007) Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol*., **25**, 197–206.

Keiser,M.J. *et al*. (2009) Predicting new molecular targets for known drugs. *Nature*, **462**, 175–181.

Kuhn,M. *et al*. (2016) The SIDER database of drugs and side effects. *Nucleic Acids Res*., **44**, D1075–D1079.

Kutuk,O. and Basaga,H. (2004) Aspirin inhibits TNFalpha- and IL-1-induced NF-kappaB activation and sensitizes HeLa cells to apoptosis. *Cytokine*, **25**, 229–237.

Kwon,S. and Yoon,S. (2017) DeepCCI. In: *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics - ACM-BCB '17*, 2017, Boston, MA, USA, pp. 203–212.

Li,H. *et al*. (2015) Aspirin prevents colorectal cancer by normalizing EGFR expression. *EBioMedicine*, **2**, 447–455.

Lounkine,E. *et al*. (2012) Large-scale prediction and testing of drug activity on side-effect targets. *Nature*, **486**, 361–367.

Ma,Y. *et al*. (2004) A biochemically defined system for mammalian nonhomologous DNA end joining. *Mol. Cell*, **16**, 701–713.

Makley,L.N. and Gestwicki,J.E. (2013) Expanding the number of 'druggable' targets: non-enzymes and protein-protein interactions. *Chem. Biol. Drug Des*., **81**, 22–32.

Mauri,A. *et al*. (2006) Dragon software: an easy approach to molecular descriptor calculations. *MATCH Commun. Math. Comput. Chem*., **56**, 237–248.

Mayr,A. *et al*. (2018) Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chem. Sci*., **9**, 5441–5451.

Meyer,J.G. *et al*. (2019) Learning drug functions from chemical structures with convolutional neural networks and random forests. *J. Chem. Inf. Model*, **59**, 4438–4449.

Meyer,M.J. *et al*. (2018) Interactome INSIDER: a structural interactome browser for genomic studies. *Nat. Methods*, **15**, 107–114.

Moriwaki,H. *et al*. (2018) Mordred: a molecular descriptor calculator. *J. Cheminform*., **10**, 4.

Muraki,K. *et al*. (2013) The role of ATM in the deficiency in nonhomologous end-joining near telomeres in a human cancer cell line. *PLoS Genet*., **9**, e1003386.

Nguyen,T. *et al*. (2019) Prediction of drug–target binding affinity using graph neural networks. *bioRxiv*.

Novac,N. (2013) Challenges and opportunities of drug repositioning. *Trends Pharmacol. Sci*., **34**, 267–272.

Ozturk,H. *et al*. (2016) A comparative study of SMILES-based compound similarity functions for drug-target interaction prediction. *BMC Bioinformatics*, **17**, 128.

Pahikkala,T. *et al*. (2015) Toward more realistic drug-target interaction predictions. *Brief Bioinform*., **16**, 325–337.

Park,Y. and Marcotte,E.M. (2012) Flaws in evaluation schemes for pair-input computational predictions. *Nat. Methods*, **9**, 1134–1136.

Payne,K.J. and Dovat,S. (2011) Ikaros and tumor suppression in acute lymphoblastic leukemia. *Crit. Rev. Oncogene*, **16**, 3–12.

Pinero,J. *et al*. (2017) DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res*., **45**, D833–D839.

Pushpakom,S. *et al*. (2019) Drug repurposing: progress, challenges and recommendations. *Nat Rev. Drug Discov*., **18**, 41–58.

Rampe,D. *et al*. (1997) A mechanism for the proarrhythmic effects of cisapride (Propulsid): high affinity blockade of the human cardiac potassium channel HERG. *FEBS Lett*., **417**, 28–32.

Riabinska,A. *et al*. (2013) Therapeutic targeting of a robust non-oncogene addiction to PRKDC in ATM-defective tumors. *Sci. Transl. Med*., **5**, 178–189.

Rifaioglu,A.S. *et al*. (2018) DEEPScreen: high performance drug-target interaction prediction with convolutional neural networks using 2-D structural compound representations. *Chem. Sci*., **11**, 2531–2557.

Rogers,D. and Hahn,M. (2010) Extended-connectivity fingerprints. *J. Chem. Inf. Model*, **50**, 742–754.

Seeman,P. (2002) Atypical antipsychotics: mechanism of action. *Can. J. Psychiatry*, **47**, 29–38.

Skuta,C. *et al*. (2017) Probes and drugs portal: an interactive, open data resource for chemical biology. *Nat. Methods*, **14**, 759–760.

Tanori,M. *et al*. (2019) Cancer risk from low dose radiation in Ptch1(+)(/)(-) mice with inactive DNA repair systems: therapeutic implications for medulloblastoma. *DNA Repair (Amst.)*, **74**, 70–79.

Tatonetti,N.P. *et al*. (2012) Data-driven prediction of drug effects and interactions. *Sci. Transl. Med*., **4**, 125–131.

Tsujii,M. *et al*. (1998) Cyclooxygenase regulates angiogenesis induced by colon cancer cells. *Cell*, **93**, 705–716.

Wen,M. *et al*. (2017) Deep-learning-based drug-target interaction prediction. *J. Proteome Res*., **16**, 1401–1409.

Yamanishi,Y. *et al*. (2008) Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, **24**, i232–i240.

Yu,H. *et al*. (2012) A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data. *PLoS One*, **7**, e37608.

Zhao,S., and Li,S. (2010) Network-based relating pharmacological and genomic spaces for drug target identification. *PLoS One*, **5**, e11764."