REPRODUCIBLE DRUG REPURPOSING: WHEN SIMILARITY DOES NOT SUFFICE

EMRE GUNEY*

Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine
c/ Baldiri Reixac 10-12, Barcelona, 08028, Spain
**E-mail: emre.guney@irbbarcelona.org*

Repurposing existing drugs for new uses has attracted considerable attention over the past years. To identify potential candidates that could be repositioned for a new indication, many studies make use of chemical, target, and side effect similarity between drugs to train classifiers. Despite promising prediction accuracies of these supervised computational models, their use in practice, such as for rare diseases, is hindered by the assumption that there are already known and similar drugs for a given condition of interest. In this study, using publicly available data sets, we question the prediction accuracies of supervised approaches based on drug similarity when the drugs in the training and the test set are completely disjoint. We first build a Python platform to generate reproducible similarity-based drug repurposing models. Next, we show that, while a simple chemical, target, and side effect similarity based machine learning method can achieve good performance on the benchmark data set, the prediction performance drops sharply when the drugs in the folds of the cross validation are not overlapping and the similarity information within the training and test sets are used independently. These intriguing results suggest revisiting the assumptions underlying the validation scenarios of similarity-based methods and underline the need for unsupervised approaches to identify novel drug uses inside the unexplored pharmacological space. We make the digital notebook containing the Python code to replicate our analysis that involves the drug repurposing platform based on machine learning models and the proposed disjoint cross fold generation method freely available at `github.com/emreg00/repurpose`.

Keywords: Drug repurposing; Machine learning; Drug similarity; Stratified disjoint cross validation.

1. Introduction

Computational drug repurposing has gained popularity over the past decade, offering a possibility to counteract the increasing costs associated with the conventional drug development pipelines. Several studies have focused on training similarity-based predictors (also known as knowledge-based or guilt-by-association-based methods) using drug chemical, target and side effect similarity between drugs (see Refs. 1–3 for recent reviews). These studies often combine various features including but not limited to chemical 2D fingerprint similarity, overlap or interaction network closeness of drug targets and correlation between drug side effects and build a machine learning model based on different algorithms, such as support vector machines, random forests and logistic regression classifiers.[4–11] The proposed models are then compared in a cross validation setting, in which a portion of the known drug-disease associations are hidden during training and used for the validation afterwards. The areas under reciver operating characteristic (ROC) curves in the cross validation analysis reported for these models range between 75-95%, suggesting that some of these models can accurately identify novel drug-disease associations. Nevertheless, in reality, the applicability of these methods for discovery of novel drug-disease associations has been limited due to "the reliance on data existing nearby in pharmacological space" as highlighted by Hodos et *al.*[2] Moreover, Vilar and colleagues alert the community about the potential "upstream bias introduced with the information provided in the construction of the similarity measurement" in similarity-based predictors.[12] Yet, since many studies do not provide the data and code used to build the models for repurposing, it is often cumbersome to validate, reproduce and reuse the underlying methodology.

In this study, first, we provide a Python-based platform for reproducible similarity-based drug repurposing and then seek to quantify the effect of the assumptions on the existing data nearby in pharmacological space. Following similar works evaluating various cross validation approaches for drug-target and protein-protein interaction prediction,[13,14] we adopt a stratified disjoint cross validation strategy for splitting drug-disease pairs, where none of the drugs in the training set appear in the test set. We show that, although a simple logistic regression classifier can achieve good performance on the data set under a conventional cross validation setting, it performs poorly when it faces with drugs it has never seen before.

Overall, our results suggest that the prediction accuracies reported by existing supervised methods are optimistic, failing to represent what one would expect in a real-world setting. We believe that the platform provided in this study could be useful for prospective studies to perform benchmarking in a unified manner.

2. Results

2.1. *A Python platform for reproducible similarity-based drug repurposing*

To incentivize reproducibility in computational drug repurposing research, we provide a Python-based platform[a] encapsulating several machine learning algorithms available in Python Scikit-learn package[15] available both as stand alone code and Jupyter notebook. The platform consists of methods to (*i*) parse a publicly available data set containing drug chemical substrucuture, target, side effect information, (*ii*) calculate drug similarity using a combination of the three features provided in the data set, (*iii*) balance data such that the drug-disease pairs have the same proportion of positive and negative instances, (*iv*) apply cross validation, and (*v*) build classifiers (Fig. 1).

The platform facilitates access to several machine learning algorithms and cross validation utilities available in Scikit-learn. By changing the configuration values, the user can build a classifier using default parameters based on logistic regression, k-nearest neighbor classifier, support vector machine, random forest, and gradient boosting classifier. We note, however, these methods are provided as is and the user still has to take the necessary steps for parameter optimization for these methods. The user can also adjust the proportion of the positive and negative pairs within each fold by changing the parameter file. Furthermore, the platform is easily customizable, allowing the user to define her own data balancing, cross validation and classifier building methods.

2.2. *Evaluating similarity-based drug repurposing via cross validation*

Next, we show the utility of the platform by building a logistic regression based drug repurposing classifier that incorporates drug chemical, target, and side effect similarity, a simplified

[a] Available at `github.com/emreg00/repurpose`

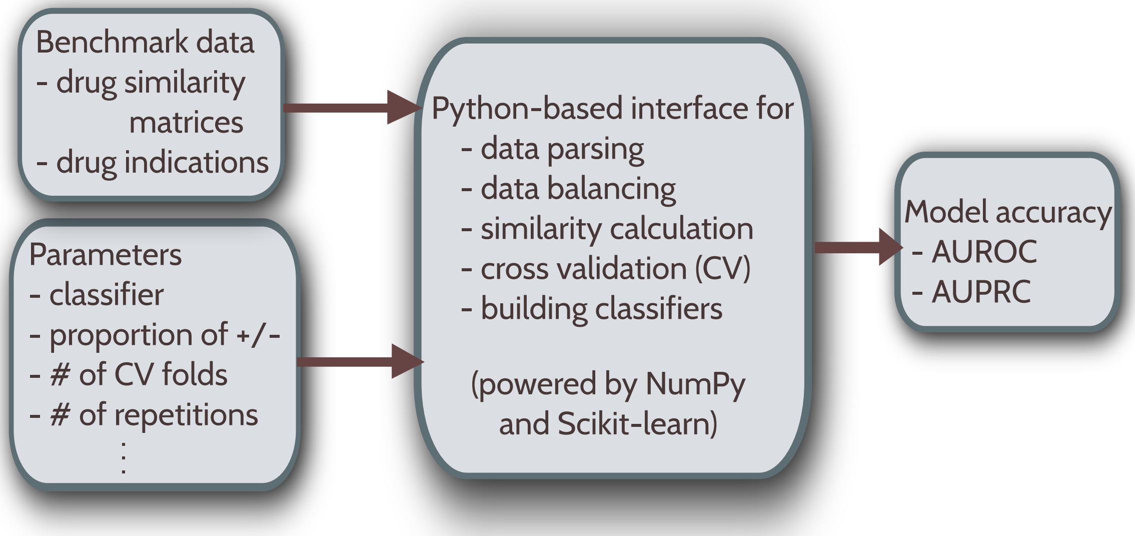

Fig. 1. Overview of the reproducible similarity-based repurposing platform.

version of the classifier suggested in a seminal paper by Gottlieb and colleagues.[4] Our model uses three drug-drug similarity based features compared to the combination of five drug-drug similarity (similarity of targets in terms of gene ontology functions and protein interaction network closeness in addition to the drug chemical, target, and side effect similarity) and two disease-disease similarity-based features (ten in total) proposed by Gottlieb and colleagues. We also incorporate the k-nearest-neighbor approach used by Zhang and coworkers,[7] who recently, built a classifier based on similarity to the 20 most similar drugs and compared it to Gottlieb and colleagues. We build our model on the same data set[b] used by Zhang and coworkers. We calculate the Pearson correlation between drugs using each of the three features mentioned above. For each feature, we assign a score corresponding to the likelihood of a given drug to be indicated for a disease based on the similarity scores and labels of the most similar 20 drugs. These scores are then combined in a logistic regression model and coefficients of the model is derived using a cross validation scheme (see Methods).

We test the prediction accuracy of the classifier under a ten fold cross validation scheme, where we split the available data set into ten groups, leave one group for testing the accuracy of the classifier and use the remaining groups to train the classifier. We repeat the cross validation analysis ten times to get estimates on the mean and standard deviation of the areas under ROC curves (AUC) and report these values in Table 1. We find that the AUC of the classifier is 84%, comparable to 87% reported by Zhang and coworkers. The slight discrepancy between the values can be explained by *(i)* the original study using imputation on the feature set and/or *(ii)* the authors reporting the AUC value from a single run instead of the mean

[b] Made publicly available by the Zhang *et al.* at `http://astro.temple.edu/~tua87106/drugreposition.html`

over multiple cross validation runs (due to the random subsampling of the data, the AUC values in consequent runs might vary slightly).

Table 1. Areas under ROC and Precision-Recall curves (AUC and AUPRC) under various validation schemes averaged over ten runs of ten-fold cross validation (S.d.: Standard deviation).

Disjoint folds	Mean AUC (%)	S.d. AUC (%)	Mean AUPRC (%)	S.d. AUPRC (%)
No	84.1	0.3	83.7	0.3
Yes	65.6	0.5	62.8	0.5

2.3. *Revisiting cross validation using disjoint folds*

Existing studies often assume that the drugs that are in the test set will also appear in the training set, a rather counter-intuitive assumption as, in practice, one is often interested in predicting truly novel drug-disease associations (i.e. for drugs that have no known indications previously). We challenge this assumption by evaluating the effect of having training and test sets in which none of the drugs in one overlaps with the drugs in the other. Accordingly, we implement a disjoint cross validation fold generation method that ensures that the drug-disease pairs are split such that none of the drugs in the training set appear in the test set (Fig. 2, see Methods for details).

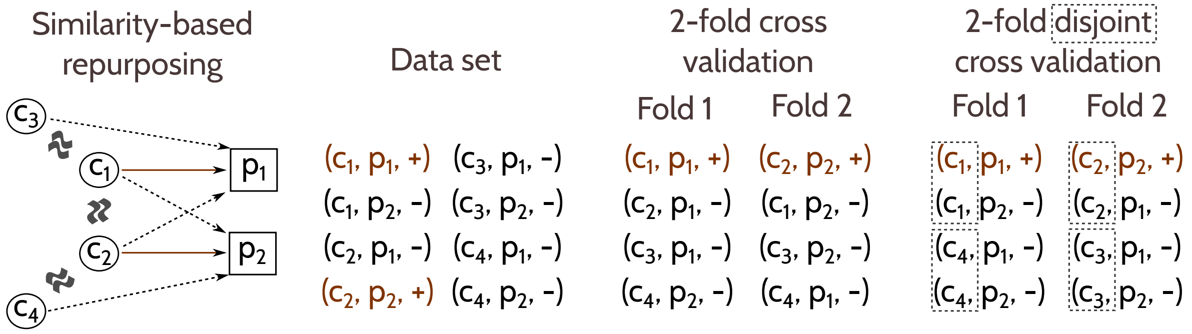

Fig. 2. Schematic representation of similarity-based repurposing and cross validation strategy. On a toy data set consisting of four compounds c_1, c_2, c_3, c_4 and two phenotypes p_1, p_2, the similarity-based drug repurposing approach is illustrated. c_1 and c_2 are indicated for p_1 and p_2, respectively. For instance, c_3 can be inferred to be useful for p_1 due to its similarity to c_1. Conventionally, k-fold cross validation randomly splits the data into k groups preserving the overall proportion of the labels in the data. We propose a disjoint cross validation scheme for paired data, such as drug-disease pairs in drug repurposing studies, that does not only preserve the proportion of the labels but also ensures that none of the drugs from the pairs in one fold are in the other folds. We demonstrate this on the toy data for k = 2 (two-fold cross validation).

In fact, several studies aim to investigate the prediction performance when the drugs in the test set are dissimilar to those in the training data set. Nonetheless, they usually do not guarantee that the trained models are unbiased with respect to unseen data. For instance, Luo *et al.*[11] use an independent set of drug-disease associations, yet, 95% of the drugs in the independent set are also in the original data set (109 out of 115). On the other hand, Gottlieb *et al.*[4] create the folds such that 10% of the drugs are hidden instead of 10% of the drug-disease pairs, but they do not ensure that the drugs used to train the model are disjoint from the drugs in the test set.

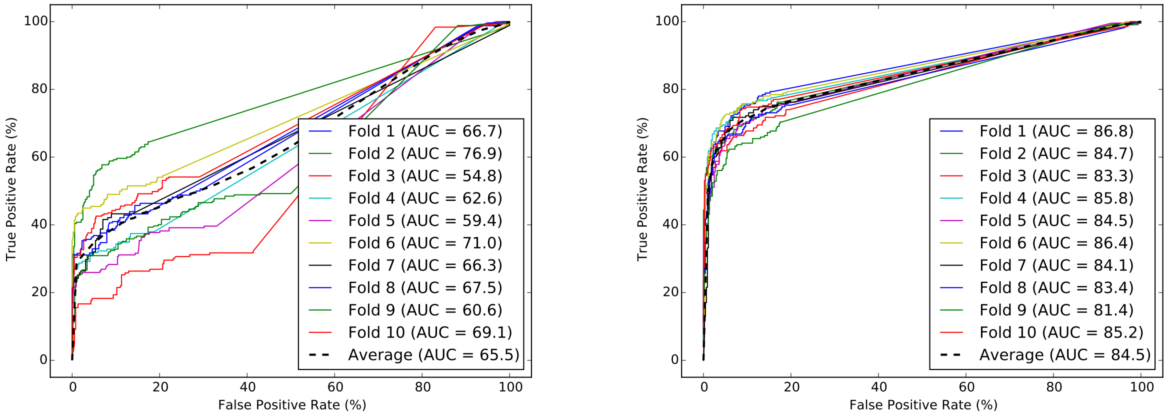

Fig. 3. ROC curves for each fold with and without disjoint cross validation (in a single run).

2.4. *Effect of the cross validation strategy on classifier performance*

We use the drug-wise disjoint cross validation strategy to study its effect on the classifier performance. We observe that the AUC drops significantly from 84% to 66% ($P = 6.9 \times 10^{-23}$, assessed by two-sided t-test) when the classifier is trained with drug-disease associations coming from the drugs that do not exist in the test data set (Table 1).

We suspect that this is due to the limited information within the test set from which the similarity-based drug-disease associations are calculated (using 20 most similar drugs) before they are fed to the classifier. To verify this, we repeat the analysis using two-fold, five-fold and 20-fold cross validation and show that the number of folds does indeed have an effect on the classifier performance (Table 2). In the two-fold disjoint cross validation scheme, the classifier accuracy is almost as good as the ten-fold cross validation accuracy without using disjoint folds, probably due to the number of drug-disease pairs within the test fold being large enough to capture the similarity relationships between drugs. Conversely, in the 20-fold disjoint cross validation scheme, the AUC drops to 59%, emphasizing the effect of the test set size due to the increased number of folds.

We next turn to the ROC curve of each cross validation fold under the two different strategies to examine the consistency among different folds (Fig. 3). We recognize that the variance between the ROC curves is higher when the folds are drug-wise disjoint compared to when drugs are shared among folds. As a result, the standard deviation over the corresponding AUC values is larger in the drug-wise disjoint case (6.0% in disjoint vs 1.5% in non-disjoint),

Table 2. Areas under ROC and Precision-Recall curves under disjoint k-fold cross validation scheme for $k = 2$, 5, 10 and 20 averaged over ten runs.

Number of folds	Mean AUC (%)	S.d. AUC (%)	Mean AUPRC (%)	S.d. AUPRC (%)
2	80.7	0.3	79.3	0.3
5	73.6	0.7	71.9	0.7
10	65.6	0.5	62.8	0.5
20	59.1	0.6	57.0	0.3

suggesting that the predictions are less robust against the partitioning of the drugs in disjoint cross validation.

Compiled mainly via text mining, the drug side effect information in SIDER is prone to a high number of false positives. Given the reduced number of drugs with high similarity, the effect of false positive associations might be more pronounced in the disjoint cross validation than the non-disjoint scenario. Thus, to inspect whether the observed decline in the AUC can be attributed to one of the features used in the classifier –such as side effect based similarity–, we check the contribution of each feature under the disjoint cross validation scheme (Fig. 4). We confirm that this is not the case. In fact, the feature based on side effect similarity is slightly more predictive than the rest (AUC=65% for side effect similarity vs 62% and 61% for chemical and target similarity, respectively), corroborating the promise of side effect profiles to describe similarities between drugs,[4,7,16,17] despite potential noise in the annotations.

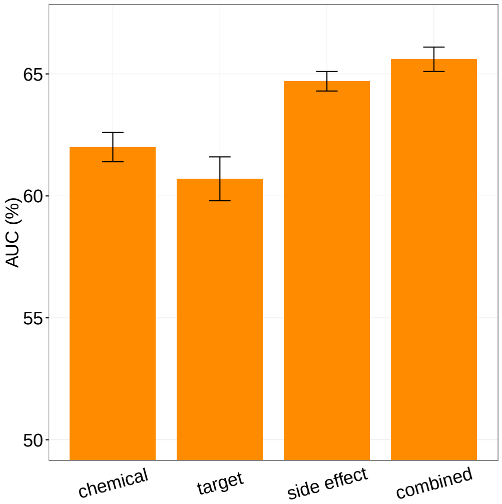

Fig. 4. Prediction accuracy (AUC) when each similarity feature used individually in disjoint cross validation. Error bars show standard deviation of AUC over ten runs of ten-fold cross validation.

2.5. *When similarity does not suffice*

The drop in the AUC confirms that many drug-disease associations are missed when the drugs in the test set have not been seen while training the classifier. For instance, the gold standard data contains several lipid lowering agents indicated for hypercholesterolemia: cholesterol absorption inhibitors (ezetimibe); fibrates (clofibrate, fenofibrate, gemfibrozil); and statins (atorvastatin, fluvastatin, lovastatin, pravastatin, simvastatin). We observe that most of these drugs can be predicted for hypercholesterolemia due to their chemical, target, and side effect based similarity to the other drugs within the same family when drugs are allowed to overlap across cross validation folds. However, when the classifier is trained using disjoint cross validation, most of these drug-disease associations can not be predicted correctly. Likewise, the drugs used for juvenile rheumatoid arthritis (diclofenac, ibuprofen, methotrexate, naproxen, oxaprozin, sulfasalazine, toletin) fail to manifest similarity to other drugs in the cross validation fold, hence missed by the classifier. We also note a similar trend for acute myeloid leukemia drugs (cyclophosphamide, daunorubicin, etoposide, idarubicin, mitoxantrone). In Table 3, we highlight the similarity-based scores of the drug to the other drugs and the probability calculated by the logistic regression classifier in a cross validation fold for several of these drug-disease associations.

Table 3. Similarity scores and logistic regression based probabilities for several known drug-disease associations missed using disjoint cross validation.

	Non-disjoint cross validation				Disjoint cross validation			
Drug	Chemical score	Target score	Side effect score	Probability	Chemical score	Target score	Side effect score	Probability
Hypercholesterolemia drugs								
fenofibrate	0.76	0.71	1.10	0.82	0.57	0	0.46	0.36
lovastatin	1.93	1.97	2.92	0.99	0	0	0	0.14
Juvenile rheumatoid arthritis drugs								
ibuprofen	0.82	3.50	1.08	1.00	0	0.50	0.43	0.43
sulfasalazine	1.39	1.99	0.43	0.96	0	0.50	0.43	0.43
Acute myeloid leukemia drugs								
daunorubicin	1.77	1.50	0	0.87	0	0	0	0.15
idarubicin	0.78	2.00	0.81	0.97	0	0	0	0.14

3. Methods

3.1. *Data sets*

We have retrieved the data set Zhang *et al.* curated for the analysis of the drug repurposing classifier they proposed.[7] They collected 1,007 approved drugs and their targets from DrugBank,[18] the chemical structure information of these drugs from PubChem[19] and the side effect information from SIDER.[20] The drugs were represented by a combination of 775 targets extracted from DrugBank and 881 substructures in PubChem. They were able to map side effects of 888 out of 1,007 drugs using SIDER, covering 61,102 drug-side effect associations coming from 1,385 side effects. The known drug-disease indications span 3,250 associations between 799 drugs and 719 diseases and were extracted from the National Drug File - Reference Terminology (NDF-RT) as suggested in a previous study by Li and Lu.[21] The data set is publicly available online at `http://astro.temple.edu/~tua87106/drugreposition.html`. We used the 536 drugs that were common among chemical, target, side effect, and indication data, corresponding to 2,229 drug-disease associations covering 578 diseases and 40,455 drug-side effect associations covering 1,252 side effects.

3.2. *Drug similarity definitions*

We used the data sets described above to build a drug-drug similarity matrix for each one of the three feature types: chemical substructures, targets, side effects. For each feature type, the drug i was defined by a binary vector $X_i = [x_1, x_2, \ldots, x_n]^T$, corresponding to the existence of the feature for that drug (1 if exists, 0 otherwise). The Pearson product-moment correlation coefficient between two drugs i and j was then calculated using $\rho_{ij} = \frac{C_{ij}}{\sqrt{C_{ii} * C_{jj}}}$, where C_{ij} given by

$$C_{ij} = \mathrm{cov}(X_i, X_j) = \mathrm{E}[(X_i - \mathrm{E}(X_i))(X_j - \mathrm{E}(X_j))]$$

The *corrcoef* function implemented in NumPy Python package was used to calculate correlation coefficients for each drug-drug pair.

3.3. *Similarity-based logistic regression classifier*

We trained a logistic regression model to predict the drug-disease associations based on the drug-drug similarities defined by the targets, chemical substructures, and side effects combined for the 20 most similar drugs to the drug in concern. Therefore, the probability of observing an association between the drug i and the disease j is

$$P(Y_{ij} = 1 | s_{ij}^{\text{chemical}}, s_{ij}^{\text{target}}, s_{ij}^{\text{side effect}}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 * s_{ij}^{\text{chemical}} + \beta_2 * s_{ij}^{\text{target}} + \beta_3 * s_{ij}^{\text{side effect}})}}$$

where for each feature $f \in \{$ chemical, target, and side effect $\}$, the similarity-based drug-disease score s_{ij}^f is defined as

$$s_{ij}^f = \sum_{k \in \mathrm{NN}(i)} \mathrm{sim}^f(i,k) * X(k,j)$$

with $sim^f(i,k)$ being the similarity between two drugs i and k (calculated via Pearson product-moment correlation coefficient as explained above), $\mathrm{NN}(i)$ is the set of 20 most similar drugs to drug i (nearest neighbors in the similarity space), and $X(k,j)$ being an indicator function with values 1 if drug k is a known indication for disease j, and 0 otherwise.

We used the *LogisticRegression* function in Scikit-learn Python package with the L2 regularization option and the default values (inverse regularization strength of 1 and stopping tolerance of 0.0001).

3.4. *Prediction accuracy evaluation*

To assess the prediction performance of the logistic regression classifier, we calculated the area under ROC curve (AUC) using k-fold cross validation scheme (e.g., $k = 2, 5, 10, 20$). We used 2,229 known drug-disease associations as the positive instances and marked all remaining possible associations between 536 drugs and 578 diseases ($536 \times 578 - 2,229 = 307,579$ associations) as negative instances. Following the previous studies, we balanced the data set such that it contained twice as many negative instances as positives.[4,7] Thus, in a k-fold cross validation run, we created k groups containing $2,229/k$ positive instances and $2 \times 2,229/k$ negative instances that were randomly chosen among all negative instances. Each fold was used as the test set once, in which all the remaining folds were used to train the classifier. In order to get robust estimates of the AUC, we repeated the cross validation procedure ten times and recorded the mean and the standard deviation of the AUC values over these runs. Note that, the classifier we built relies on both the similarity and the labels of the training drug-disease associations, as we calculate a drug-disease association score using the most similar 20 drugs and their indication information. We made sure not to use the training information in the test phase and calculated the drug-disease association scores within the training and test folds separately. We used the *roc_curve* and *auc* functions in Scikit-learn Python package to first get the true and false positive rates at various cutoffs and then to calculate the AUC using the trapezoidal rule.

3.5. *Stratified disjoint cross validation for defining non-overlapping drug groups*

To investigate the robustness of the drug-disease association classifier in the case of unseen data, we used a disjoint cross validation scheme, in which none of the drugs in one fold appear in another fold. We created cross validation folds such that all the drugs with the same name were in the same fold by first converting the drug's name into an integer value and then taking the modulo (k) of this value (for k-fold cross validation). To produce different groupings at each run, we added a random integer to the integer value of the drug calculated based on its name. The details of the algorithm are as follows:

D: data set containing drug-disease pairs, c: drug, p: disease,
l: label (1 if c is known to be indicated for p, 0 otherwise), k: number of cross validation folds,
$fold$: dictionary containing the fold index of each drug-disease pair

```
$i :=$ random($[0, 100]$)
$fold := \{\}$
for each $(c, p, l) \in D$ do
  $sum := 0$
  for each $x \in$ characters($c$) do
    $sum := sum +$ to_integer($x$)
  $fold(c, p) :=$ modulo($sum + i, k$)
return $fold$
```

To preserve the balance between positive and negative instances (stratified cross validation), we first grouped the data set into positive ($D^{l=1}$) and negative ($D^{l=0}$) pairs and applied the proposed disjoint fold generation algorithm above to each group.

4. Conclusions

Many recent similarity-based drug repurposing studies reported stunningly high prediction performances, suggesting that drugs can be predicted for novel uses almost with perfect accuracy. Yet, there has not been an observable improvement in the drug discovery in the pharma industry over the past years. We suspect this could be *(i)* because similarity-based methods do not provide insights on the mechanism of action of drugs, failing to explain clinical failures due to the lack of efficacy and safety and/or *(ii)* the reported accuracies being unrealistic due to the underlying validation scheme.

To look into various validation schemes and toward increasing the reproducibility in computational drug repurposing research, we provide a Python-based platform encapsulating machine learning algorithms available in Python Scikit-learn package and propose a disjoint cross fold generation method. This platform allows us to easily evaluate the prediction performance of a logistic regression classifier built using drug chemical, target, and side effect similarity under various experimental settings. Using this platform, we investigate the role of the experimental settings in similarity-based drug repurposing studies in producing optimistic prediction accuracies. In particular, we seek to validate the drug repurposing model when it has never seen the drug beforehand. To test this idea, we use a cross validation approach in which the data is split such that none of the drugs in the test set are in the training set. We show that the high success rate of the model drops sharply under such cross validation setting.

Indeed, in many computational biology problems dealing with paired data, such as predicting drug targets, side effects, drug-drug interactions, protein-protein interactions, functional annotations, and disease-genes, researches aim to leverage machine learning methods using similarity between biomolecules. Our findings suggest that failure to take into account the parity in such data sets can produce optimistic prediction accuracies, supporting earlier studies on drug-target and protein-protein interaction prediction.[13,14] We particularly point out the effect of the training set size when the drugs in the training and test sets do not overlap. Hence, we argue that, though useful in highlighting potential unknown drug-disease pairs, similarity-based methods are likely to be ineffective to explore drugs that are not in the nearby pharmacological space, i.e. the drugs with low chemical similarity or for which target and side effect data are not abundant.

Alternatively, systems-level drug discovery approaches can offer insights on the mechanism of action of the drugs by matching gene expression signatures upon drug treatment to compensate the genomic changes caused by the disease[22,23] or exploiting the paths from drug targets to the genes perturbed in the diseases to explain the efficacy of treatments given the interaction network.[24] Nonetheless, these approaches are still at their infancy and their accuracies remain modest,[24,25] leaving room for improvement.

Acknowledgments

The author is grateful to Dr. Patrick Aloy for hosting EG who is supported by EU-cofunded Beatriu de Pinós incoming fellowship from the Agency for Management of University and

Research Grants (AGAUR) of Government of Catalunya. I also would like to thank Zhang and colleagues for making the data used in their analysis online.

References

1. G. Jin and S. T. C. Wong, Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines, *Drug Discovery Today* **19**, 637 (May 2014).
2. R. A. Hodos, B. A. Kidd, K. Shameer, B. P. Readhead and J. T. Dudley, In silico methods for drug repurposing and pharmacology, *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* **8**, 186 (May 2016).
3. S. Vilar and G. Hripcsak, The role of drug profiles as similarity metrics: applications to repurposing, adverse effects detection and drug-drug interactions, *Briefings in Bioinformatics* , p. bbw048 (June 2016).
4. A. Gottlieb, G. Y. Stein, E. Ruppin and R. Sharan, PREDICT: a method for inferring novel drug indications with application to personalized medicine, *Mol Syst Biol* **7**, p. 496 (June 2011).
5. J. Li and Z. Lu, A new method for computational drug repositioning using drug pairwise similarity, in *2012 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, October 2012.
6. F. Napolitano, Y. Zhao, V. M. Moreira, R. Tagliaferri, J. Kere, M. D'Amato and D. Greco, Drug Repositioning: A Machine-Learning Approach through Data Integration, *Journal of Cheminformatics* **5**, p. 30 (2013).
7. P. Zhang, P. Agarwal and Z. Obradovic, Computational Drug Repositioning by Ranking and Integrating Multiple Data Sources, in *Machine Learning and Knowledge Discovery in Databases*, eds. H. Blockeel, K. Kersting, S. Nijssen and F. ZeleznyLecture Notes in Computer Science (Springer Berlin Heidelberg, September 2013) pp. 579–594. DOI: 10.1007/978-3-642-40994-3_37.
8. M. Oh, J. Ahn and Y. Yoon, A Network-Based Classification Model for Deriving Novel Drug-Disease Associations and Assessing Their Molecular Actions, *PLOS ONE* **9**, p. e111668 (October 2014).
9. Z. Liu, F. Guo, J. Gu, Y. Wang, Y. Li, D. Wang, L. Lu, D. Li and F. He, Similarity-based prediction for Anatomical Therapeutic Chemical classification of drugs by integrating multiple data sources, *Bioinformatics* **31**, 1788 (June 2015).
10. W. Dai, X. Liu, Y. Gao, L. Chen, J. Song, D. Chen, K. Gao, Y. Jiang, Y. Yang, J. Chen and P. Lu, Matrix Factorization-Based Prediction of Novel Drug Indications by Integrating Genomic Space, *Computational and Mathematical Methods in Medicine* **2015**, p. e275045 (May 2015).
11. H. Luo, J. Wang, M. Li, J. Luo, X. Peng, F.-X. Wu and Y. Pan, Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm, *Bioinformatics* , p. btw228 (May 2016).
12. S. Vilar, E. Uriarte, L. Santana, T. Lorberbaum, G. Hripcsak, C. Friedman and N. P. Tatonetti, Similarity-based modeling in large-scale prediction of drug-drug interactions, *Nature Protocols* **9**, 2147 (September 2014).
13. T. Pahikkala, A. Airola, S. Pietil, S. Shakyawar, A. Szwajda, J. Tang and T. Aittokallio, Toward more realistic drugtarget interaction predictions, *Briefings in Bioinformatics* **16**, 325 (March 2015).
14. Y. Park and E. M. Marcotte, A flaw in the typical evaluation scheme for pair-input computational predictions, *Nature methods* **9**, 1134 (December 2012).
15. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research* **12**, p. 28252830 (October 2011).

16. M. Campillos, M. Kuhn, A.-C. Gavin, L. J. Jensen and P. Bork, Drug Target Identification Using Side-Effect Similarity, *Science* **321**, 263 (July 2008).
17. L. Yang and P. Agarwal, Systematic Drug Repositioning Based on Clinical Side-Effects, *PLOS ONE* **6**, p. e28025 (December 2011).
18. D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam and M. Hassanali, DrugBank: a knowledgebase for drugs, drug actions and drug targets, *Nucleic Acids Research* **36**, D901 (January 2008).
19. Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang and S. H. Bryant, PubChem: a public information system for analyzing bioactivities of small molecules, *Nucleic Acids Research* **37**, W623 (July 2009).
20. M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen and P. Bork, A side effect resource to capture phenotypic effects of drugs, *Molecular Systems Biology* **6**, p. 343 (2010).
21. J. Li, X. Zhu and J. Y. Chen, Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts, *PLoS computational biology* **5**, p. e1000450 (July 2009).
22. J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross, M. Reich, H. Hieronymus, G. Wei, S. A. Armstrong, S. J. Haggarty, P. A. Clemons, R. Wei, S. A. Carr, E. S. Lander and T. R. Golub, The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease, *Science* **313**, 1929 (September 2006).
23. M. Sirota, J. T. Dudley, J. Kim, A. P. Chiang, A. A. Morgan, A. Sweet-Cordero, J. Sage and A. J. Butte, Discovery and preclinical validation of drug indications using compendia of public gene expression data, *Science Translational Medicine* **3**, p. 96ra77 (August 2011).
24. E. Guney, J. Menche, M. Vidal and A.-L. Barabási, Network-based in silico drug efficacy screening, *Nature Communications* **7**, p. 10331 (February 2016).
25. J. Cheng, L. Yang, V. Kumar and P. Agarwal, Systematic evaluation of connectivity map for disease indications, *Genome Medicine* **6** (December 2014).