Exploring Landscape of Drug-Target-Pathway-Side Effect Associations

Hansaim Lim\textsuperscript{1}, Aleksandar Poleksic, PhD\textsuperscript{2}, and Lei Xie, PhD\textsuperscript{1,3}
\textsuperscript{1}PhD program in Biochemistry, the City University of New York, New York, NY, United States; \textsuperscript{2}Department of Computer Science, University of Northern Iowa, Cedar Falls, IA, United States; \textsuperscript{3}Department of Computer Science, Hunter College, the City University of New York, New York, NY, United States

Abstract

Side effects are the second and the fourth leading causes of drug attrition and death in the US. Thus, accurate prediction of side effects and understanding their mechanism of action will significantly impact drug discovery and clinical practice. Here, we show REMAP, a neighborhood-regularized weighted and imputed one-class collaborative filtering algorithm, is effective in predicting drug-side effect associations from a drug-side effect association network, and significantly outperforms the state-of-the-art multi-target learning algorithm for predicting rare side effects. We also apply FASCINATE, an extension of REMAP for multi-layered networks, to infer associations among side effects and drug targets from drug-target-side effect networks. Then, using random permutation analysis and gene overrepresentation tests, we infer statistically significant side effect-pathway associations. The predicted drug-side effect associations and side effect-causing pathways are consistent with clinical evidences. We expect more novel drug-side effect associations and side effect-causing pathways to be identified when applying REMAP and FASCINATE to large-scale chemical-gene-side effect networks.

Introduction

Severe side effects are the second leading cause for drug attrition, and the fourth leading cause of death in the US. Severe side effects limit the use of the drugs, decrease their value, and negatively affect patients\textsuperscript{1,2}. Despite the importance of identifying potential side effects of a drug molecule in advance, it is daunting and prohibitive to test them experimentally. This results in biased, sparse and noisy knowledge about the biological and biochemical associations of side effect. To tackle the difficulty in studying drug side effects, systematic, large-scale methods have been developed to computationally predict drug-induced side effects\textsuperscript{3,4,5,6}. Although these approaches show acceptable accuracy for predicting common side effects of existing drugs, challenges remain to predict rare side effects as well as to systematically infer missing multi-scale drug-target-pathway-side effect associations. It is important to model drug actions on a multi-scale, since the drug response phenotypes result from complex interplay among biological pathways that are modulated by drug-target interactions.

It is not a trivial task for a machine learning method to infer novel drug-target-pathway-side effect associations based on incomplete, biased, and noisy data. Recently, we have developed a neighborhood-regularized weighted and imputed one-class collaborative filtering method REMAP to address this challenge\textsuperscript{7}. REMAP has several unique features, making it particularly suitable to infer missing relations from incomplete and noisy data sets such as drug side effects. First, REMAP does not require negative data for model training by utilizing the imputation. The drug-side effect associations in the existing database are mainly positive. The known negative associations are extremely sparse. These limitations impose hurdles for most classification methods. Second, REMAP can handle mislabeling problem by assigning a confidence score to each label. Mislabling is common in biological and clinical data sets due to systematic and random errors in experiments. Finally, by applying neighborhood regularization on drug, target, and side effect information, REMAP alleviates the cold start problem, where predicting new targets or side effects is difficult for chemicals without any known targets or side effects. In our earlier study, we have showed that REMAP can be successfully applied to predict unknown drug-target associations\textsuperscript{7}. In this paper, we extend its application to drug side effect prediction.

While REMAP shows high prediction accuracy and potential in understanding drug actions, it has limitations. One of the most important issues is that REMAP can take only two types of biological entities (e.g. drugs and targets) and their relationship, and model them as nodes and edges in a bipartite graph. As mentioned above, however, drug actions involve multiple biological entities that are linked with each other on a multi-scale. Thus, integrating information from more than two types of biological entities may be crucial for predicting drug action. For example, a drug interacts with an off-target. The off-target is involved in a biological pathway. The pathway is associated with a side effect. These biological entities (e.g. drug, target, pathway, and side effect) and their relationships can be modeled as a multi-layered...
network (Figure 1). To infer missing relations from the multi-layered network, most of conventional methods model multiple pairwise relations independently, and integrate these binary relations subsequently. Such an approach ignores the inter-dependency among binary relations. FASCINATE has been developed to infer novel missing relationships from multi-layered networks by jointly optimizing multiple bipartite graphs\(^8\). In the benchmark studies, FASCINATE outperforms other state-of-the-art methods in inferring multiple relations\(^8\).

![Multi-layered network view of drugs causing side effects. Drugs may bind targets that are associated with side effects or relevant biological pathways. Thus, drugs may cause side effects through the interplay of biological networks. Solid lines: known associations used as training sets in this study. Dashed lines: no known associations used.](image)

Here, we apply REMAP and FASCINANTE to the prediction of drug side effects and identification of pathways associated with side effects, respectively. We first show that our network-regularized weighted and imputed one-class collaborative filtering method, REMAP, outperforms state-of-the-art multi-target learning methods for drug-side effect associations. Then, we perform random permutation analysis using the FASCINATE algorithm to predict statistically significant gene-side effect associations based on known associations among drug-gene, drug-side effect, and gene-side effect associations. The gene overrepresentation analysis is performed to infer associations among biological pathways and side effects. The predicted results suggest biological pathways associated with cataracts, glaucoma, depression, and other side effects. These putative associations are consistent with existing clinical evidence. Thus, FASCINATE has demonstrated great potential for identifying novel associations between side effects and biological pathways systematically from drug-gene-side effect networks. Our method relies on the known associations and node similarity information, such as drug-drug similarity by chemical substructure comparison. It is expected that a more accurate chemical similarity and side effect similarity measurement as well as more complete chemical-gene-side effect association datasets can discover more novel associations that may be directly applied to clinical studies.

**Contribution**

The contributions of this paper are both methodological and translational. Methodologically, we demonstrate that a network-regularized and imputed one-class collaborative filtering method when applied to bi-layered and multi-layered drug-target-side effect network is a potentially powerful tool to predict drug side effects and elucidate the molecular and cellular mechanisms of side effects. The translational application is that we predict side effects for several drugs in the database as well as biological pathways associated with these side effects. Majority of them are consistent with existing clinical evidences.

**Related work**

This section is a review of the existing methods for side effect-pathway prediction, and the core statistical and machine learning methods that have been applied to drug side effect predictions, followed by applications of nonnegative matrix factorization methods for association prediction problems.

To the best of our knowledge, only a few studies have been done to predict the associations between side effects and biological pathways. Fukuzaki et al. proposed a method, Cooperative Pathway Enumerator (CREPE) to predict associations between biological conditions (e.g. heat shock condition) and pathways using the frequent subgraph discovery method\(^9\). Wallach et al. presented a method to predict side effect-pathway associations by a combination of virtual docking and logistic regression methods, assuming that drugs affecting the same pathway may induce the same side effects\(^10\). More recently, Shaked et al. developed a prediction method, called Array of Model-based Phenotype Predictors (AMPP) to predict drug-target associations using Support Vector Machine (SVM), and extended it to connect key metabolic reactions and biomarkers to side effects, by applying Random Forest-based feature selection on the biomarker features\(^11\),\(^12\).
Similarity Ensemble Approach (SEA) is a method developed to predict whether a drug will bind to a target protein based on the chemical structure similarity \(^{13, 14, 15}\). Lounkine et al. applied SEA to drug-induced side effect prediction by connecting the predicted drug-target associations to the known target-side effect associations\(^3\). The strength of SEA is that it can capture information from sets of weak similarities.

Multi-label Learning (ML) is a class of machine learning methods suitable for predicting multiple labels that are not necessarily mutually exclusive, and ML applications on several domains have been developed\(^{16}\). Zhang and Zhou developed k-nearest neighbor-based ML (MLKNN)\(^{17}\) and it refined by applying mutual information-based feature selection on top of MLKNN (FS-MLKNN) to predict drug-induced side effects\(^6\). FS-MLKNN can be classified as an ensemble learning method, which combines weak base learners to improve prediction.

Canonical Correlation Analysis (CCA) is used to find the weights that maximize the correlation between the feature vectors\(^{18}\). Pauwels et al. developed a chemical substructure-based prediction method that applies CCA and refined it to Sparse CCA (SCCA) for easier interpretation\(^{15, 19}\). In addition to the predicted outcome, SCCA can also return to the optimized weights that represent the importance of each feature. Liu and Altman applied CCA and found correlations between binding affinities of drugs to essential genes and side effects of drugs. They combined CCA with singular value decomposition to extract features representing biological signals, excluding the frequency of information in the dataset\(^{20}\).

While these sophisticated methods are reasonably accurate, each class of methods has drawbacks. First, SEA, FS-MLKNN and SCCA only use chemical features. Therefore, these methods cannot utilize other side information, such as target-target similarity without integrating other methods. Second, these methods may not be able to handle the activity cliff, where a small change in the chemical structure results in a dramatic change in the biological outcome\(^{21}\). Third, the methods mentioned above generate models based on the assumption that the unobserved associations are negative. However, the unobserved associations may be positive, the idea which the weighted imputation-based method tries to adopt.

Nonnegative matrix factorization (NMF), first introduced by Paatero and Tapper in 1994, became a popular choice for recommendation problems after the development of the multiplicative update algorithm by Lee and Seung\(^{22, 23}\). One of the most successful NMF applications is the Netflix competition, where the user-video preferences are predicted by NMF\(^{24}\). To better predict the user-item preferences, advanced NMF algorithms take side information, such as user-user and item-item similarities to reflect the idea that similar users tend to choose similar items. Advanced NMF methods such as wiZAN-dual can also take weights and imputation values to account the fact that some of the unknown associations may be positive\(^{25}\). Based on the mathematical framework of wiZAN-dual, REMAP and its variation COSINE were developed to make predictions on unobserved chemical-protein associations\(^7, 26\). FASCINATE is an extension of REMAP on a multi-layered network, and has been applied to study drug-gene-disease networks\(^8\). In this paper, for the first time, we utilize REMAP and FASCINATE to infer novel drug-gene-pathway-side effect associations.

**Method**

As illustrated in Figure 2, the hypothesis of REMAP and its extension FASCINATE algorithms is that the observed sparse and noisy drug-gene profiles and drug-side effect profiles rise from the latent features (hidden variables) of the observed sparse and noisy drug-gene profiles and drug-side effect profiles rise from the latent features (hidden variables) of
chemicals, genes, and side effects. Similar chemicals, genes, or side effects will have similar latent features. The operation between chemical and gene latent features (e.g., inner product) will restore the complete chemical-gene associations; similarly for chemical-side effect and gene-side effect associations.

B. Dataset description

Our method has three types of nodes: drugs, genes, and side effects. We collected known associations from publicly available databases. Drug-side effect associations for small molecule drugs were obtained from SIDER4.1 database. Drug-target associations are obtained from the combination of three databases: ZINC, ChEMBL, and DrugBank. Drug-target associations were excluded if the drug was not included in the SIDER database. Gene-side effect associations were obtained from Lounkine et al. The known associations were treated as binary (i.e., 1 for observed associations, and 0 otherwise). The drug-drug similarity scores were calculated based on chemical 2D structure similarity as described in REMAP paper. Protein-protein interactions were obtained from BioGRID database and used as gene-gene similarity matrix. We have tested semantic similarity-based side effect-side effect similarity scores using UMLS-Similarity software. However, the semantic similarity scores did not help gene-side effect or drug-side effect prediction accuracy. The best performance with semantic similarity was within one standard deviation of the performance without using it (data not shown). Thus, an identity matrix was used as side effect-side effect similarity matrix. The data statistics are listed in Table 1.

Table 1. Data statistics of drug-gene-side effect network.

| Type                  | Count     |
|-----------------------|-----------|
| **Nodes**             |           |
| Drugs                 | 1549      |
| Genes                 | 19116     |
| Side effects (diseases)| 4251      |
| **Associations**      |           |
| Drug-gene             | 4727      |
| Drug-side effect       | 127595    |
| Gene-side effect       | 2209      |
| **Similarities**      |           |
| Drug-drug             | 5303      |
| Gene-gene             | 458135    |
| Side effect-side effect| 4251      |

C. Description of prediction method

REMAP is a variation of nonnegative matrix factorization algorithm for inferring novel binary relations. FASCINATE is an extension of REMAP on multi-layered network. The three layers used in this study are drugs, genes, and side effects, and the aim is to find low-rank matrices of latent features for each layer that minimizes the following equation:

$$\min_{F_i \geq 0 \ (i=1,\ldots,g)} \left\| \sum_{i=1}^{g} \| W_{ij} \odot (D_{ij} + P_{ij} - F_i F_j') \|_F^2 + \alpha \sum_{i=1}^{g} \text{tr}(F_i'(T_i - A_i) F_i) + \beta \sum_{i=1}^{g} \| F_i \|_F^2 \right\|_F^2$$  \(1\)

Here, \(g\) is the number of total layers (3 in this study), and \(F_i\) (e.g., \(F_1, F_2, \text{or } F_3\)) is the low-rank matrix for the \(i\)th layer, \(D_{ij}\) is the cross-layer association matrix between the \(i\)th and \(j\)th layers (e.g., drug-gene), \(A_i\) is the within-layer similarity matrix (e.g., drug-drug), \(T_i\) is the diagonal degree matrix of \(A_i\), \(W_{ij}\) is the weight matrix, \(P_{ij}\) is the imputation matrix, \(\odot\) is an element-wise product operator (Hadamard product), and \(\|M\|_F\) is the Frobenius norm of matrix \(M\). For the algorithmic efficiency, the weight matrix was replaced with a scalar, global weight for all layers, and the imputation value was uniformly set to a constant for all unobserved associations. Once the low-rank matrices for each layer are calculated by FASCINATE, the prediction matrix between the \(i\)th and \(j\)th layer can be obtained by the matrix product of \(F_i\) and \(F_j\). If there are only two layers, the objective function in Eq. (1) is identical to that of REMAP. More details about the algorithm can be found in the reference.

We evaluated the effectiveness of REMAP using multiple metrics. For varying degree of density in the gene-side effect network, we performed 10-fold cross-validation to measure four different performance metrics that are commonly used to evaluate a recommender system. The four metrics are the standard area under the receiver operating characteristic curve (AUC), Half-Life Utility (HLU), Mean Average Precision (MAP), and Mean Percentage
Ranking (MPR). HLU estimates the likelihood that a user accepts the recommended items under the assumption that it exponentially decays with the item’s ranking. MAP measures the average precision over all users in the test dataset at different recall levels. MPR measures the average of percentile-ranks of positive associations in the test dataset. For example, an MPR of 0% is obtained when all positive associations are top-ranked. On the other hand, 100% MPR is obtained when they are ranked to the least. The higher AUC, HLU, MAP and the lower MPR, the better performance. During the cross-validation, we categorized side effects by the number of known drug-side effect associations and measured performances from rare side effects to frequently observed ones. For comparison, we tested feature selection-based multi-label k-nearest neighbor method (MLKNN) under the same cross-validation and rarity conditions as those for REMAP.

Since FASCINATE takes a few tuning parameters, we found the optimal parameters by using a grid search and 10-fold cross-validation. For each fold, we randomly selected approximately 10% of the associations from both drug-gene and drug-side effect layers as test sets, and measured the AUC for both layers separately. Then, we evaluated the performance by the average of the test AUC for 10 folds weighted by the number of associations in each layer. In this way, we found the optimal low-rank and maximum iteration number to be 770 and 100, respectively, while fixing the other parameters to be the default values. Using the optimal rank and iteration number, we found the optimal parameters to be $\alpha=0.5$, $\beta=1.0$, and weight=0.25.

### D. Random permutation and overrepresentation analysis

Next, we performed random permutation and overrepresentation analysis on the gene-side effect layer using the optimal parameters above. Briefly, we estimated the distribution of prediction scores from random permutations, and using the distribution, we filtered predicted pairs from FASCINATE without permutation. The filtered pairs were then used to search for significant pathway-side effect associations by an overrepresentation test. For each permutation test, we randomly permuted the drug-gene and drug-side effect associations while keeping the total number of associations in each layer. Using the permuted layers and the intact gene-side effect layer as inputs, we ran FASCINATE to obtain the prediction scores for gene-side effect pairs. We performed 200 permutation iterations and recorded the prediction scores for each gene-side effect pair. Then, we collected raw prediction scores from randomly selected 1,000 gene-side effect pairs from unknown associations. With the 200,000 raw scores (1,000 pairs with 200 scores each), Epanechnikov kernel was used to fit the distribution (Figure 3). Except those for known gene-side effect associations, the raw prediction scores were converted to $p$-values by the Epanechnikov kernel probability density function obtained above, followed by the False Discovery Rate adjustment. To infer associations among biological pathways and the side effects, the gene sets for each side effect were used for KEGG overrepresentation test provided in the cluster Profiler package. The enriched pathway-side effect associations were filtered out if the pathway is associated with more than 5 side effects, or the side effect is associated with more than 5 pathways. Also, predicted pairs with $q$-value (from hypergeometric test) higher than 0.05 were filtered out.

![Figure 3](image)

**Figure 3.** (A) Epanechnikov kernel fitting on FASCINATE prediction scores from random permutation analysis on gene-side effect associations. (B) Inset of (A) for smaller ranges of both axes.

### Result

A. Significantly improved performance of REMAP in predicting rare side effects

REMAP performs better than MLKNN for all four metrics in all tested rarity categories (Table 2). The performance gain of REMAP over MLKNN is more significant when the side effect is rare. The AUC of REMAP drops from 0.95
for all side effects to 0.84 for the rarest side effect class, while that of MLKNN drops from 0.92 to 0.68. This suggests that REMAP can be effectively applied for drug-side effect predictions from rare to common side effects. It is noted that while REMAP shows better AUC and MPR for more common side effects, REMAP’s performance is not necessarily improving as the rarity decreases, based on HLU and MAP. The same trend is observed for MLKNN, suggesting that presenting the four metrics provides more objective evaluations. In summary, our extensive benchmark studies show that REMAP has a great potential to predict novel associations.

Table 2. Performances of REMAP and MLKNN predicting drug-side effect associations.

| Number of drugs per SE* | Avg. AUC | Avg. HLU | Avg. MAP | Avg. MPR |
|-------------------------|----------|----------|----------|----------|
| REMAP                   | MLKNN    | REMAP    | MLKNN    | REMAP    | MLKNN    |
| < 12                    | 0.84     | 0.68     | 5.6      | 3.6      | .048     | .031     | 0.26     | 0.33     |
| < 25                    | 0.88     | 0.76     | 4.9      | 2.4      | .045     | .026     | 0.20     | 0.24     |
| < 50                    | 0.90     | 0.82     | 4.1      | 1.6      | .044     | .026     | 0.16     | 0.19     |
| < 100                   | 0.92     | 0.85     | 3.9      | 1.1      | .043     | .024     | 0.13     | 0.16     |
| < 200                   | 0.93     | 0.88     | 3.5      | 0.7      | .040     | .023     | 0.11     | 0.13     |
| < 400                   | 0.94     | 0.90     | 3.4      | 0.4      | .042     | .024     | 0.10     | 0.11     |
| < 800                   | 0.95     | 0.91     | 3.6      | 0.2      | .047     | .027     | 0.08     | 0.10     |
| All                     | 0.95     | 0.92     | 4.0      | 0.3      | .052     | .031     | 0.08     | 0.09     |

* Side effects having a certain number of associated drugs per side effect.

B. Literature supports of top ranked predictions

Furthermore, the top ranked predictions from REMAP are supported by existing clinical evidences, as shown in Table 3. Newton et al. reported that patient groups receiving diazepam showed decreased libido compared to other groups including placebo in reviewing total of 984 patients with generalized anxiety disorder48. An ovarian carcinoma patient treated with paclitaxel and carboplatin was reported to develop acute myeloid leukemia (AML) in 2006 while the main culprit was unclear49. In 2016, another group reported a case of a patient who developed AML after endometrial chemotherapy including paclitaxel, where paclitaxel was the likely cause of AML40. Although indirectly related to ichthyosis and less severe, patients treated with lenalidomide showed significant dermatologic side effects41. It was reported that one of the most common side effect of erlotinib, when treated with emibetuzumab for non-small cell lung cancer (NSCLC), is hypocalcemic42. In addition, erlotinib alone was reported to cause hypocalcemia43. Two clinical case reports and human exposure studies suggest that propranolol may cause cardiac arrest if overdosed44,45. Although rare and the cases were limited to the elderly population, ranolazine treatment was reported to cause visual hallucinations, which resolved after discontinuation of treatment46. In a study of 5,194 males, approximately 35% of patients experienced erectile dysfunction, and it was associated with cumulative exposure to zalcitabine or enfuvirtide47. Ejaculatory failure was reported from oxcarbazepine treatment48. In three cases of long-term corticosteroid treatment, including two cases with prednisolone, the patients suffered from mild to severe impairment of vision, which could have resulted in permanent loss of vision49. In a study analyzing more than a million cases from September 1999 to April 2012, levofloxacin was found to be the greatest risk for patients. Patients who took levofloxacin had higher hazard ratios for serious cardiac arrhythmia and death compared to amoxicillin and azithromycin50. In a case report, Mahendran and Liew suggested that when prescribing alprazolam, symptoms of depression need to be carefully monitored. They described a healthy woman in her 20s who developed severe depression with suicidal thoughts after taking the prescribed doses of alprazolam three times51.

Table 3. Predicted top ranked drug-side effect associations with literature support.
C. Identification of pathway-side effect associations

A total of 7,572 gene-side effect pairs for 409 unique genes and 1,150 unique side effects were found with q-values lower than 0.001 by FASCINATE and subsequent random permutation analysis. Then these genes are associated with biological pathways using gene set overrepresentation analysis. The final predicted pathway-side effect associations with their q-value are listed in Table 4. The literature support for predicted top-ranked side-effect pathways suggest that our method can infer missing one type of binary relations by jointly learning other types of binary relations from a multi-layered network (Table 4). Wirtz and Keller identified mutations and significant up or down regulation in different kinds of an interleukin family of cytokines in mice with glaucoma, suggesting that disruption of interleukin-20 signaling in the eye may be associated with glaucoma. More recently, Gupta et al. showed that the concentration of tear film cytokines were significantly lower in patients with primary open-angle glaucoma. Liu and Neufeld showed that the epidermal growth factor receptor (EGFR) is significantly upregulated and tyrosine phosphorylated in the human glaucomatous optic nerve head in vivo. They demonstrated that it induces nitric oxide synthase to generate excessive nitric oxide, leading to increased intraocular pressure and glaucoma. In another study, the authors suggested that EGFR is a common regulatory pathway for neural injuries in the optic nerve, including glaucoma. Kwon and Tomarev showed that several components in integrin-focal adhesion kinase – serine/threonine kinase signaling pathway are activated in the eyes of mice expressing high level of myocilin, a protein known to be associated with glaucoma. Piotrowska et al. surveyed patients treated with a selective EGFR tyrosine kinase inhibitor, where they found a significant proportion of patients developed drug-induced cataracts, including those who had to undergo surgical repair. Although the association between cataract and focal adhesion pathway is subtle, Kokkinos et al. showed that the expression and phosphorylation pattern of focal adhesion kinase changes as the lens develops in mice. As one may simply guess, sex steroid hormone deficiency is known to be associated with erectile dysfunction. Although we could not find pertinent references, it may be trivial to validate the predicted association between steroid hormone synthesis and decreased libido. A more indirect association was described between erectile dysfunction and gastric acid secretion, which is clinically regulated by histamine receptor antagonists. In a study involving 55 patients with stage IV NSCLC, patients having EGFR mutations showed less severe depression compared to those without EGFR mutation, suggesting the association between depression and EGFR. In a review paper, evidence was found for omega-6 polyunsaturated fatty acid (e.g. linoleic acid) affecting the function of leukocytes. Several interesting points rise from the predicted side effect-pathway associations. First, glaucoma and cataract share two common pathways: EGFR tyrosine kinase inhibitor resistance and focal adhesion pathways. Second, the results suggest that erectile dysfunction and decreased libido are relevant to the steroid hormone biosynthesis pathway. Further studies with more details and experimental validations are necessary to refine the FASCINATE method for gene-side effect-biological pathway association prediction.

Table 4. Predicted pathway-side effect associations by random permutation and overrepresentation analysis.

| Side Effect (disease)       | KEGG Pathway                        | Q-value | Reference |
|-----------------------------|-------------------------------------|---------|-----------|
| Glaucoma                    | Cytokine-cytokine receptor interaction | 1.1E-2  | 52, 53    |
| Glaucoma                    | EGFR tyrosine kinase inhibitor resistance | 9.2E-4  | 54, 55    |
| Glaucoma                    | Focal adhesion                      | 6.9E-3  | 56        |
| Cataract                    | EGFR tyrosine kinase inhibitor resistance | 7.3E-3  | 57        |
| Cataract                    | Focal adhesion                      | 2.2E-2  | 58        |
| Erectile dysfunction        | Steroid hormone biosynthesis         | 5.6E-3  | 59        |
| Libido decreased            | Steroid hormone biosynthesis         | 9.3E-6  | *         |
| Erectile dysfunction        | Gastric acid secretion               | 8.1E-3  | 60, 61    |
| Depression                  | EGFR tyrosine kinase inhibitor resistance | 1.2E-2  | 62        |
| Leukopenia                  | Linoleic acid metabolism             | 1.4E-2  | 63        |

*No pertinent references found.

Conclusion

In this study, REMAP outperformed state-of-the-art MLKNN on drug-side effect prediction as measured by the four metrics (AUC, HLU, MAP, and MPR), especially for rare side effects. Furthermore, FASCINATE, a multi-layer extension of REMAP, was applied to jointly predict the drug-target-side effect network and to infer missing pathway-side effect associations in combination with random permutation analysis and the overrepresentation test. Many of the predicted drug-side effect and pathway-side effect associations are supported by existing experimental and clinical
evidences (Table 3 and Table 4). These results suggest that network-regularized weighted imputed one-class collaborative filtering is a powerful tool for inferring missing biological relations on a multi-scale. There are many ways to improve the performance of REMAP and FASCINATE. Their performances are dependent on the similarity measure of drug-drug, target-target, and side effect-side effect pairs as well as the coverage of reconstructed drug-gene-side effect networks. The side effect-side effect semantic similarity used in this study does not help much with the prediction accuracy. Thus, new side effect-side effect similarity measurement is needed. In the earlier studies, it was shown that the structure-based ligand binding site similarity improves the prediction accuracy of side effect\textsuperscript{16}. In addition, the datasets used in this study are not comprehensive. Only 1,549 unique FDA approved drugs and 4,727 drug-gene interactions are included in the current network. It is expected that more novel relations can be identified if the reliable chemical-gene network which include approximately 500,000 chemicals (including drugs) and thousands of targets genes\textsuperscript{7} and improved drug-side effect database are applied\textsuperscript{65}.

Acknowledgment
We appreciate the constructive comments from reviewers. We acknowledge Russ B. Altman, Ph. D. and Tianyun Liu, Ph. D. for providing the gene-side effect association data. This work was supported by Grant Number R01GM122845 from the National Institute of General Medical Sciences (NIGMS) of the National Institute of Health (NIH), and R01LM011986 from the National Library of Medicine (NLM) of the NIH.

To whom correspondence should be addressed: Lei Xie, Email: lei.xie@hunter.cuny.edu

References

1. Giacomini KM, Krauss RM, Roden DM, Eichelbaum M, Hayden MR, Nakamura Y. When good drugs go bad. Nature. 2007;446(7139):975-7.
2. Arrowsmith J. Trial watch: phase III and submission failures: 2007-2010. Nat Rev Drug Discov. 2011;10(2):87.
3. Loukine E, Keiser MJ, Whitebread S, Mikhailov D, Hamon J, Jenkins JL, et al. Large-scale prediction and testing of drug activity on side-effect targets. Nature. 2012;486(7403):361-7.
4. Jahid MJ, Ruan J. An ensemble approach for drug side effect prediction. Proceedings (IEEE Int Conf Bioinformatics Biomed). 2013:440-5.
5. Pauwels E, Stoven V, Yamanishi Y. Predicting drug side-effect profiles: a chemical fragment-based approach. BMC Bioinformatics. 2011;12:169.
6. Zhang W, Liu F, Luo L, Zhang J. Predicting drug side effects by multi-label learning and ensemble learning. BMC Bioinformatics. 2015;16:365.
7. Lim H, Poleksic A, Yao Y, Tong H, He D, Zhuang L, et al. Large-scale off-target identification using fast and accurate dual regularized one-class collaborative filtering and its application to drug repurposing. PLoS Comput Biol. 2016;12(10):e1005135.
8. Chen C, Tong H, Xie L, Ying L, He Q. FASCINATE: fast cross-layer dependency inference on multi-layered networks. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; San Francisco, California, USA. 2939784: ACM; 2016. p. 765-74.
9. Fukuzaki M, Seki M, Kashima H, Sese J, editors. Side effect prediction using cooperative pathways. 2009 IEEE International Conference on Bioinformatics and Biomedicine; 2009 1-4 Nov. 2009.
10. Wallach I, Jaitly N, Lilien R. A structure-based approach for mapping adverse drug reactions to the perturbation of underlying biological pathways. PLoS One. 2010;5(8):e12063.
11. Breiman L. Random forests. Machine learning. 2001;45(1):5-32.
12. Shaked I, Oberhardt MA, Altas N, Sharan R, Ruppin E. metabolic network prediction of drug side effects. Cell Syst. 2016;2(3):209-13.
13. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. Relating protein pharmacology by ligand chemistry. Nat Biotechnol. 2007;25(2):197-206.
14. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, et al. Predicting new molecular targets for known drugs. Nature. 2009;462(7270):175-81.
15. Hert J, Keiser MJ, Irwin JJ, Oprea TI, Shoichet BK. Quantifying the relationships among drug classes. J Chem Inf Model. 2008;48(4):755-65.
16. Madjarov G, Kocev D, Gjorgjevikj D, Džeroski S. An extensive experimental comparison of methods for multi-label learning. Pattern Recognition. 2012;45(9):3084-104.
17. Zhang M-L, Zhou Z-H. ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition. 2007;40(7):2038-48.
18. Hotelling H. Relations between two sets of variates. Biometrika. 1936;28(3/4):321-77.
19. Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics. 2009;10(3):515-34.
20. Liu T, Altman RB. Relating essential proteins to drug side-effects using canonical component analysis: a structure-based approach. J Chem Inf Model. 2015;55(7):1483-94.
21. Cruz-Monteagudo M, Medina-Franco JL, Perez-Castillo Y, Nicolotti O, Cordeiro MN, Borges F. Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde? Drug Discov Today. 2014;19(8):1069-80.
22. Paatero P, Tapper U. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. Environmetrics. 1994;5(2):111-26.
23. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999;401(6755):788-91.
24. Bennett J, Lanning S, editors. The netflix prize. Proceedings of KDD cup and workshop; 2007: New York, NY, USA.
25. Yao Y, Tong H, Yan G, Xu F, Zhang X, Szymanski BK, et al., editors. Dual-regularized one-class collaborative filtering. Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management; 2014: ACM.
26. Lim H, Gray P, Xie L, Poleksic A. Improved genome-scale multi-target virtual screening via a novel collaborative filtering approach to cold-start problem. Sci Rep. 2016;6:38860.
27. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res. 2016;44(D1):D1075-9.
28. Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG. ZINC: a free tool to discover chemistry for biology. J Chem Inf Model. 2012;52(7):1757-68.
29. Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, et al. The ChEMBL bioactivity database: an update. Nucleic acids research. 2013:gkt1031.
30. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. 2008;36(Database issue):D901-6.
31. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 2006;34(Database issue):D535-9.
32. McInnes BT, Pedersen T, Pakhomov SVS. UMLS-interface and UMLS-similarity: open source software for measuring paths and semantic similarity. AMIA Annual Symposium Proceedings. 2009;2009:431-5.
33. Breese JS, Heckerman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering. Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence; Madison, Wisconsin. 2004; Morgan Kaufmann Publishers Inc.; 1998. p. 43-52.
34. Li Y, Hu J, Zhai C, Chen Y, editors. Improving one-class collaborative filtering by incorporating rich user information. Proceedings of the 19th ACM international conference on Information and knowledge management; 2010: ACM.
35. Hu Y, Koren Y, Volinsky C, editors. Collaborative filtering for implicit feedback datasets. Data Mining, 2008 ICDM'08 Eighth IEEE International Conference on; 2008: IEEE.
36. Pan R, Zhou Y, Cao B, Liu NN, Lukose R, Scholz M, et al., editors. One-class collaborative filtering. Data Mining, 2008 ICDM'08 Eighth IEEE International Conference on; 2008: IEEE.
37. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16(5):284-7.
38. Newton RE, Marunycz JD, Alderdice MT, Napoletti MJ. Review of the side-effect profile of buspirone. The American journal of medicine. 1986;80(3):17-21.
39. See HT, Thomas DA, Bueso-Ramos C, Kavanagh J. Secondary leukemia after treatment with paclitaxel and carboplatin in a patient with recurrent ovarian cancer. Int J Gynecol Cancer. 2006;16 Suppl 1:236-40.
40. Bhatnagar UB, Singh D, Glazyrin A, Moormeier M, Jaakola T, Pauleit W. Paclitaxel induced MDS and AML: a case report and literature review. Case Rep Oncol Med. 2016;2016:8308179.
41. Sviggum HP, Davis MD, Rajkumar SV, Dispensieri A. Dermatologic adverse effects of lenalidomide therapy for amyloidosis and multiple myeloma. Arch Dermatol. 2006;142(10):1298-302.
42. Rosen LS, Goldman JW, Algazi AP, Turner PK, Moser B, Hu T, et al. A first-in-human phase I study of a bivalent MET antibody, emibetuzumab (LY2875358), as monotherapy and in combination with erlotinib in advanced cancer. Clin Cancer Res. 2017;23(8):1910-9.
43. Tridente G. Adverse events and oncotargeted kinase inhibitors: Academic Press; 2017.
44. Sandroni C, Cavallaro F, Caricato A, Scapigliati A, Fenici P, Antonelli M. Enoximone in cardiac arrest caused by propranolol: two case reports. Acta Anaesthesiol Scand. 2006;50(6):759-61.
45. McEvoy GKe. American hospital formulary service - drug information 2003. Americal Society of Health-System Pharmacists, Inc. 2003:1785.
46. Kalra SS, Arora Z, Nielsen C. Ranolazine and Hallucinations2017.
47. Wang Q, Young J, Bernasconi M, Cavassini M, Vernazza P, Hirschel B, et al. The prevalence of erectile dysfunction and its association with antiretroviral therapy in HIV-infected men: the Swiss HIV Cohort Study. Antivir Ther. 2013;18(3):337-44.
48. Boora K, Chiappone K, Dubovsky SL. Oxcarbazepine-induced reversible anorgasmia and ejaculatory failure: a case report. Prim Care Companion J Clin Psychiatry. 2009;11(4):173-4.
49. Loo JL, Lee SY, Ang CL. Can long-term corticosteroids lead to blindness? A case series of central serous chorioretinopathy induced by corticosteroids. Ann Acad Med Singapore. 2006;35(7):496-9.
50. Rao GA, Mann JR, Shoaibi A, Bennett CL, Nahhas G, Sutton SS, et al. Azithromycin and levofloxacin use and increased risk of cardiac arrhythmia and death. Ann Fam Med. 2014;12(2):121-7.
51. Mahendran R, Liew E. A case of suicidal thoughts with alprazolam. Prim Care Companion J Clin Psychiatry. 2010;12(3).
52. Wirtz MK, Keller KE. The role of the IL-20 subfamily in glaucoma. Mediators Inflamm. 2016;2016:4083735.
53. Gupta D, Wen JC, Hubsner JL, Stinnett S, Kraus VB, Tseng HC, et al. Cytokine biomarkers in tear film for primary open-angle glaucoma. Clin Ophthalmol. 2017;11:411-6.
54. Liu B, Neufeld AH. Activation of epidermal growth factor receptor signals induction of nitric oxide synthase-2 in human optic nerve head astrocytes in glaucomatous optic neuropathy. Neurobiology of disease. 2003;13(2):109-23.
55. Liu B, Chen H, Johns TG, Neufeld AH. Epidermal growth factor receptor activation: an upstream signal for transition of quiescent astrocytes to reactive astrocytes after neural injury. Journal of Neuroscience. 2006;26(28):7532-40.
56. Kwon HS, Tomarev SI. Myocilin, a glaucoma-associated protein, promotes cell migration through activation of integrin-focal adhesion kinase-serine/threonine kinase signaling pathway. J Cell Physiol. 2011;226(12):3392-402.
57. Pietrowska Z, Liu E, Varga A, Thakur M, Narayanan V, Liu SV, et al. Rociletinib-associated cataracts in EGFR-mutant NSCLC. Annals of Oncology. 2016;27(suppl_6):1239P-P.
58. Kokkinos MI, Brown HJ, de Jongh RU. Focal adhesion kinase (FAK) expression and activation during lens development. Molecular vision. 2007;13:418.
59. Traish AM, Abu-Zahra H, Guay AT. The brain, the penis and steroid hormones: clinical correlates with endothelial dysfunction. Current pharmaceutical design. 2008;14(35):3723-36.
60. Kim YC, Davies MG, Lee TH, Hagen P-O, Carson CC. Characterization and function of histamine receptors in corpus cavernosum. The Journal of urology. 1995;153(2):506-10.
61. Shamburek RD, Schubert ML. Control of gastric acid secretion. Histamine H2-receptor antagonists and H+K+-ATPase inhibitors. Gastroenterol Clin North Am. 1992;21(3):527-50.
62. Jacobs JM, Traeger L, Eusebio J, Simon NM, Sequist LV, Greer JA, et al. Depression, inflammation, and epidermal growth factor receptor (EGFR) status in metastatic non-small cell lung cancer: A pilot study. J Psychosom Res. 2017;99:28-33.
63. Pompeia C, Lopes M, Miyasaka C, Procopio J, Sannomiya P, Curi R. Effect of fatty acids on leukocyte function. Brazilian Journal of Medical and Biological Research. 2000;33(11):1255-68.
64. Poleksic A, Turner C, Dalal R, Gray P, Xie L. Mining FDA resources to compute population-specific frequencies of adverse drug reactions. IEEE International Conference on Bioinformatics and Biomedicine; November 13-16; Kansas City, MO, USA.