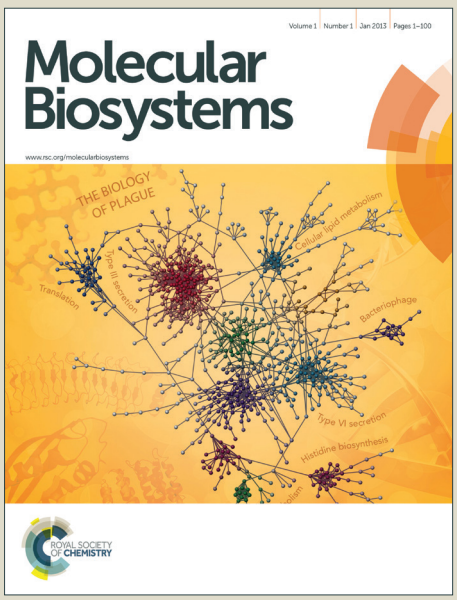
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

ARTICLE TYPE www. rsc.org/xxxxxx | XXXXXXXX

Phenotypic side effects prediction by optimizing correlation with chemical and target profiles of drugs†

Rakesh Kanji,[a] Abhinav Sharma ,[a] and Ganesh Bagler*[a]

Received 4th May 2015, Accepted Xth July 2015
First published on the web Xth July 2005
DOI: 10.1039/b000000x

Despite technological progresses and improved understanding of biological systems, discovery of novel drugs is an inefficient, arduous and expensive process. Research and development cost of drugs is unreasonably high, largely attributed to high attrition rate of candidate drugs due to adverse drug reactions. Computational methods for accurate prediction of drug side effects, rooted in empirical data of drugs, have the potential to enhance the efficacy of the drug discovery process. Identification of features critical for specifying side effects would facilitate efficient computational procedures for their prediction. We devised a generalized ordinary canonical correlation model for prediction of drug side effects based on their chemical properties as well as their target profiles. While the former is based on 2D and 3D chemical features, the latter enumerates a systems-level property of drugs. We find that the model incorporating chemical features outperforms that incorporating target profiles. Further we identified the 2D and 3D chemical properties that yield best results, thereby implying their relevance in specifying adverse drug reactions.

1 Introduction

While drugs are intended for therapeutic effect, they lead to side effects through unintended interactions with cellular processes. Accurate prediction of phenotypic side effects is extremely important so as to assess the effectiveness of candidate molecules as potential drugs. Various methods have been developed to model relevant aspects of drugs' interaction with cellular milieu leading to intended therapeutic effects as well adverse drug reactions (side effects). A living cell acts as a complex dynamical system of molecules interacting at different hierarchies. This web of molecular interactions comprises of interactions among genes, proteins (enzymes), metabolites and small molecules. To model mechanisms of side effects, it is important to consider this intertwined structure of cellular processes in which a drug is presented as an agent of molecular control.

Studies aimed at prediction of side effects have incorporated various data such as those of drug-drug similarity, drug-target interactions[9], protein-protein interactions, pathway activation and ontological correlates. In one such study, Liu et. al. compared performance of various machine learning methods by integrating information of side effects, chemical structures, targets and pathways to conclude that support vector machine (SVM) approach yields best results[1]. In another study, a computational strategy was developed by combining data of clinical observations, drug-targets, protein interactions and gene ontology (GO) annotations, and was demonstrated for the prediction of cardiotoxicity[2]. Chen et. al. developed a computational method for prediction and ranking of side effects with the help of chemical-chemical as well as protein-chemical interactions[3]. Drugs with common targets are expected to share side effects due to overlapping molecular mechanisms. Interestingly, a proportion of shared side effects between drugs are caused by network neighbors of drug targets[4]. Side effects could be seen as the result of inadvertent activation of unintended pathways. With this premise, a method was developed to enumerate 'cooperative pathways' that function together under identical conditions by combining pathway networks with the help of gene expression data[5]. It has been suggested that the similarity of drugs by virtue of shared targets correlates better with their side effects than that based on their chemical structures[6]. Prediction of drug off-targets was implemented for renal disorders through an *in silico* framework[7].

One of the successful approaches towards prediction of side effects is that of canonical correlation analysis (CCA)[11,12]. In contrast to earlier methods, in which side effects were treated individually, Atias and Sharan presented a novel approach for side effects prediction by considering an integrated side effects profile[13].

This study provided a breakthrough approach through an algorithmic framework that combined CCA and network-based diffusion. It has been demonstrated that drug profiles created

† Electronic Supplementary Information (ESI) available: [details of any supplementary information available should be included here]. See DOI: 10.1039/b000000x/

[a] *Indian Institute of Technology Jodhpur, Ratanada, Jodhpur, India. Tel: 91 77938 20447; E-mail: ganesh.bagler@gmail.com*

with chemical substructures (with the help of CCA) are better at predicting side effects than machine learning methods [14]. Mizutani et. al. proposed the first target-based feature extraction method using CCA that yielded better results than that based on chemical structures [15]. Yaminishi et. al. reported a kernel regression model that integrates chemical space (chemical structures) and genomic space (drug-target interactions) with good prediction accuracy [16]. Integration based studies done for prediction of drug-drug interactions have shed light on side effects and have led to drug repositioning. Some of the studies have focused on classification of drug-drug interactions using similarity measures based on structural, therapeutic, phenotypic properties, and also by integration of these measures [8,10]. These studies highlight the fact that integration based models work better. From these studies we reckoned that the interaction of drug with protein interactome is one of the key specifier of side effects. While chemical space (representing chemical similarities) has been used for predictive models [3,13,14,16], chemical properties of drugs have not explored enough through a generalized model. We propose that chemical profiles of drugs embody relevant therapeutic correlates that have a strong bearing on the side effects. A generalized canonical correlation analysis (GCCA) model facilitates consolidation of various aspects of drugs providing a systems-level perspective. Such a generalized approach would also allow one to identify drug features that are critical in specification of phenotypic side effects.

In this study, we used GCCA model based on drugs' target profiles as well as their chemical profiles. The former represents a binary profile of the drug indicating reported interactions with targets, whereas the latter represents the drugs quantitative 2D and 3D chemical features. We predicted the side effects of 830 drugs, that are common to Drugbank [17] and SIDER2 [18], by using only target profiles, only chemical profiles, and by using both. We find that model based on chemical profiles have more consistent accuracy than those based on target profiles. With increasing number of features used in the model, chemical profile-based model fare better than that based on target profile. We found that a few chemical features are critical in driving the accuracy of our model.

2 Materials

2.1 Compilation of drug-target and drug-side effects datasets

Drug-Target interaction and Drug-Side effect relations data were collected from Drugbank 3.0 [17] and SIDER2 [18] respectively. The former presents target profiles of drugs, whereas the latter presents side effect profiles. From Drug Bank, 3520 targets having gene sequence information were found to be associated with 5789 drugs. From SIDER2, 4192 side effects were found to be associated with 996 drugs.

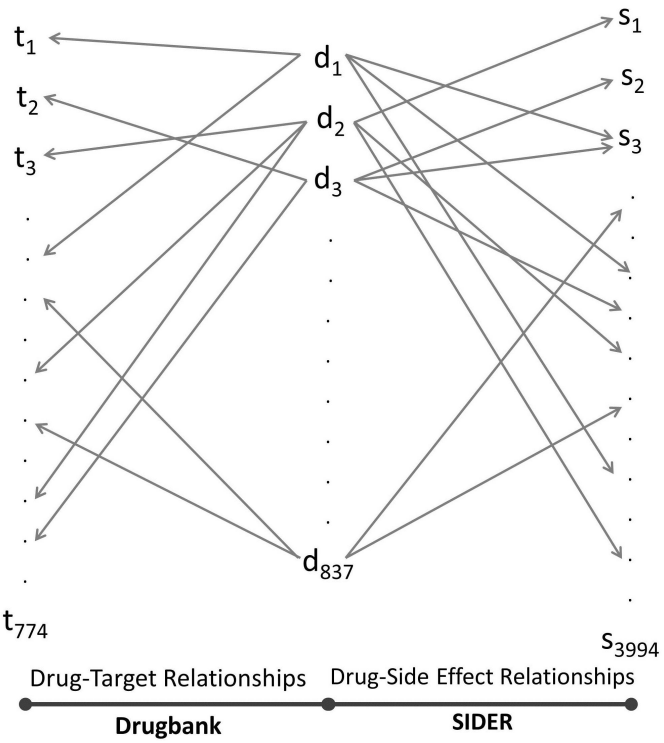


Fig. 1 Illustration of the Drug-Target-Side tri-partite network. The drug-target regulatory associations (Drugbank) and drug-side effect data (SIDER) was merged to create a composite dataset.

2.2 Construction of composite dataset

With the intention of associating side effect profiles of drugs with those of target and chemical features, we further created a composite dataset of drugs that are common between Drugbank as well as SIDER2. This composite dataset comprises of 837 drugs having both, target and side effect profiles. This subset of drugs were found to be associated with 774 targets and 3994 side effects. Following procedure was used for identification of this composite dataset. Out of 996 drugs from SIDER2, 774 were found to be having a direct match between Drug Bank using their generic drug names. Among the remaining 222 drugs from SIDER2, 118 drugs were mapped to Drug Bank by resolving their aliases through PubChem. Out of these 892 common drugs, target profiles of only 837 were available (since not all the drugs listed in Drug Bank have associated target information).

Accordingly, each of these 837 drugs was assigned with two binary column vectors of size 774×1 and 3994×1 corresponding to target binding profile and side effects profile, respectively. Entries in these profile vectors indicates pres-

ence (1) or absence (0) of association with the corresponding drug. Figure 1 illustrates interrelationships in the composite tri-partite data.

2.3 Compilation of chemical profiles

Starting from the InChI (Textual chemical identifier) of these 837 drugs, their SMILES were identified using Open Babel software. Further various 2D and 3D chemical properties of these drugs were computed using the Calculate Molecular Properties module of Discovery Studio 4.0. These chemical properties represent 61 three-dimensional and 145 two-dimensional chemical properties. The 3D properties enumerate various aspects related to dipole, energy, Jurs descriptors, principal moments, shadow indices, surface area, volume and molecular counts. The 2D properties enumerate molecular aspects related to surface area, volume, molecular counts and electrostatic properties. The complete list of 3D and 2D chemical properties is provided in the Supplementary Information.

2.4 Statistics of drug profile matrices

For performing 10-fold cross-validation 830 of the total 837 drugs were used. These 830 drugs, partitioned in to 10 groups, were associated with 3994 side effects and 774 targets. The drug-side effects matrix was sparse comprising of only 2.64% of the maximum possible associations (87,537). The number of side effects for these drugs was in the range of 1 to 724 with an average of 105.46. The drug-target association matrix was much sparser with only 2877 (less than 1%) of the maximum interactions possible (642,420). The 3D and 2D chemical profile matrices of the drugs had sparsity of around 36% and 23%, respectively, after thresholding with mean value as a cut-off of each parameter.

3 Method

3.1 Generalized Ordinary Canonical Correlation

Lets say we have d drugs having t targets, c chemical features and s side effects. Each drug $x_{i=1,2,3.....d}$ is assigned with a target profile vector, a side-effects profile vector, and chemical profiles vector having dimension $1\times t$, $1\times s$ and $1\times c$, respectively. Hence, drug-target matrix (D_t), drug-side effect matrix (D_s) and drug-chemical features matrix (D_c) have dimension of $d\times t$, $d\times s$ and $d\times c$, respectively. We considered cosine similarity(ρ) for developing an objective function[19].

$$\rho = \frac{U^T V}{\|U\|_2 \|V\|_2}, -1 \le \rho \le 1. \quad (1)$$

To find cosine similarity, for enumerating correlation, between matrices A and B having dimension $n\times p$ and $n\times q$, one needs to vectorize these matrices, $U = A\alpha$ and $V = B\beta$. The maximization of objective function f can be written as,

$$f = \max_{\alpha,\beta} \alpha^T Z \beta, \;\; Z = A^T B \quad (2)$$

In general,

$$f = \sum_i \max_{\alpha_i,\beta} \alpha_i^T Z_i \beta, \;\; Z_i = A_i^T B \quad (3)$$

such that $\|\alpha_i\|_2 = \|\beta\|_2 = 1$.
Differentiating f w.r.t. β and α_i by setting $\frac{\partial f}{\partial \beta} = \frac{\partial f}{\partial \alpha_i} = 0$, yields $P\beta = \mu\beta$ and $Z_i\beta = \lambda_i\alpha_i$.
β is the eigen vector of $P = \sum P_i$, where $P_i = Z_i^T Z_i$ and α_i can be solved by assuming $\lambda_i = \|Z_i\beta\|_2$.
Here, μ and λ_i are Lagrange multipliers with constraints of $\|\beta\|_2 = 1$ and $\|\alpha_i\|_2 = 1$, respectively.

3.2 Prediction model

For a drug with target profile X_{P_1} and chemical profile X_{P_2}, following formula was used for predicting its side effect profile (Y)[12].

$$Y = (B^T)^{-1}[\sum_i D_i A_i^T X_{P_i}]$$

Note that $(B^T)^{-1} = B$ when all eigen vectors of P matrix are considered. Here, $B = [\beta_1, \beta_2, ...\beta_i....]$ and $A = [\alpha_1, \alpha_2, ...\alpha_i....]$, and Y is the predicted side effects profile. Anyway, $(B^T)^{-1}$ is unique as B consists of independent orthogonal vectors. Every entry D_i of matrix D is given by $\frac{\lambda_i}{\sum_j |\lambda_j|}$.

3.3 Verification of the model

Herein we present verification of our model. By substituting $Z_i\beta = \lambda_i\alpha_i$ we obtain the objective function f by using equation 3. The substitution yields $f = \sum_i \lambda_i$ which is always positive since $\lambda_i = \|Z_i\beta\|_2$, and norm can not be negative.

Next we show that choosing β as the largest Eigen vector maximizes the objective function. By substituting α_i in the objective function we obtain $f = \beta^T(\sum_i \frac{1}{\lambda_i} P_i)\beta$. Since λ_i is a scaling factor we get $f = \mu\beta^T\beta = \mu$. Therefore, to maximize the objective function the largest Eigen vector needs to chosen as μ.

All computations were performed on Dell Precision T5610 workstations (*Charaka, Sushruta*) of the Complex Systems Laboratory, IIT Jodhpur.

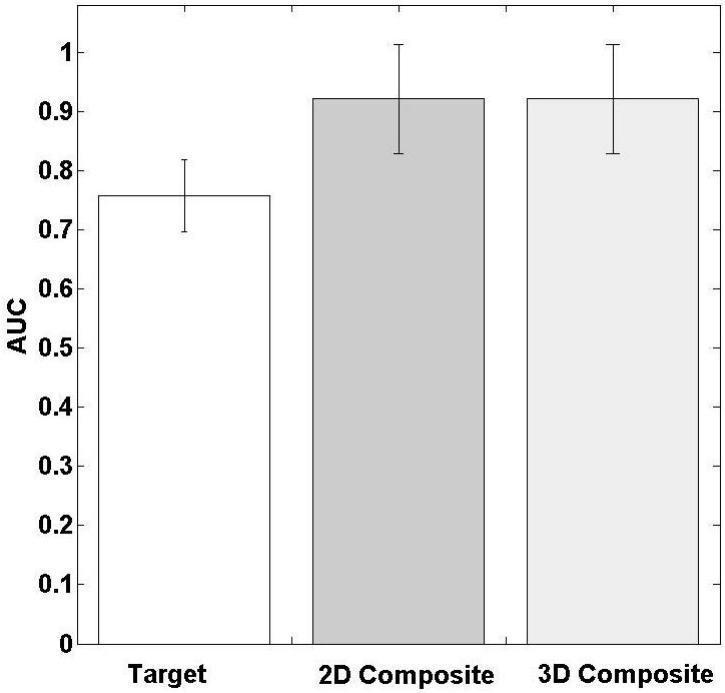


Fig. 2 AUC measured using the largest eigen vector of drug-target profiles matrix only and including that of 2D and 3D drug-chemical matrix. The error bars indicate standard error of data from 10-fold cross-validation experiments. Inclusion of chemical profiles data significantly improves the side effects prediction efficacy.

4 Results

Using the generalized ordinary canonical correlation method, we predicted side effects with 10-fold cross-validation, with each set containing 83 drugs. The predictions were made using the largest eigen vector of the drug-target matrix (Section 4.1), that of chemical profile matrix (Section 4.2) and that of composite matrix (Section 4.3). For each of these experiments, Area Under Curve (AUC) was computed to asses the performance of the method. The effect of increase in the number of eigen vectors on the performance was also tested (Supplementary Information). When we refer to either drug-target matrix or chemical profiles matrix in connection with computation of eigen vectors, we are referring to their corresponding P matrices (For more details, please see to Section 3.1 in Methods).

4.1 Drug-target profiles

Using the empirical data of drug-target interactions and side effects, we maximize the objective function f and obtain corresponding α and β. With the help of optimized parameters, we further map the test drugs to their predicted side effect profiles, based on their known target profiles. Figure 2 depicts the AUC using first eigen vector of drug-target matrix and that of integrated matrices with 3D chemical and 2D chemical profiles, respectively. After integration of 2D and 3D chemical profiles data, the AUC obtained from only the drug-target matrix significantly improved from 0.76 to 0.92 for both ($p < 10^{-5}$). Thus the chemical profiles of drugs add to the predictive ability of our model. We observed that AUC decreases from 0.76 to 0.40 with increasing number of eigen vectors of drug-target matrix used for prediction (Supplementary Fig. 1).

Using a model based on similar method, Yaminishi and coauthors predicted side effects of drugs using their target profiles and chemical fingerprints [15]. They observed that prediction based on target profiles is better than that based on chemical fingerprints (AUC of 0.8850 as opposed to 0.8355). We surmise that the superior AUC returned by their method is largely due to inclusion of large number indirect drug-target interactions (obtained by text-mining) from MATADOR [20].

Knowing that the distribution of number of targets as well as number of side effects that drugs have is heterogeneous (thick tailed) we evaluated the role of such hub drugs with exceptionally large number of targets and side effects (Table 2). We removed dominant hub drugs from the drug-target and drug-side effect matrix to observe their contribution to prediction

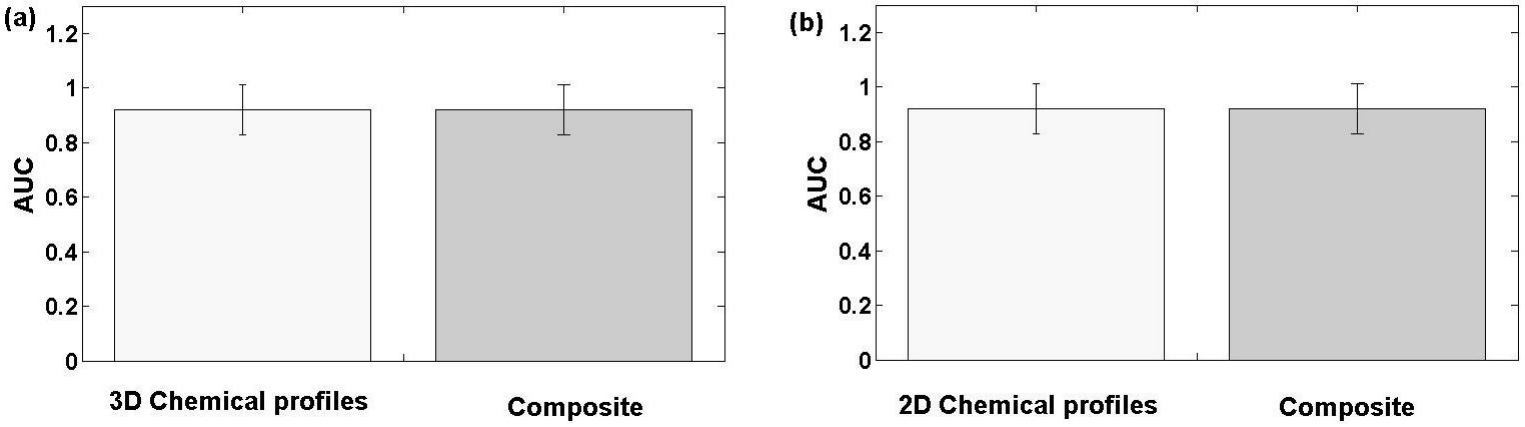


Fig. 3 AUC measured using the largest eigen vector 3-dimensional (a) and 2-dimensional (b) chemical profiles, and including that of drug-target profiles matrix. The error bars indicate standard error of data from 10-fold cross-validation experiments.

	Number of profiles				
Promiscuous profile hubs	1	2	3	4	5
Side effect hubs	0.756	0.755	0.754	0.753	0.752
Random side effects	0.755 ± 0.0083	0.753 ± 0.0081	0.753 ± 0.0081	0.753, ± 0.0081	0.753, ± 0.0081
Target hubs	**0.702**	**0.700**	**0.6936**	**0.692**	**0.686**
Random targets	0.755 ± 0.0083	0.753 ± 0.0081	0.752 ± 0.0081	0.752 ± 0.0081	0.752 ± 0.0081

Table 1 Relevance of promiscuous profile hubs. While removal of profiles most prevalent side effects does not affect the prediction efficacy significantly, the profiles of most promiscuous targets are critical.

	Number of Drug Hubs				
Drug hubs	50	100	150	200	250
Side effect based	0.747	0.735	0.720	0.713	0.705
Target based	**0.747**	**0.722**	**0.703**	**0.6564**	**0.641**
Random	0.748 ± 0.0083	0.740 ± 0.0082	0.730 ± 0.0088	0.721, ± 0.0087	0.708, ± 0.0089

Table 2 Relevance of drug hubs. While generally neither the drugs causing most side effects or those regulating large number of drugs seem to be critical for prediction efficacy, with increasing number of drugs removed, the latter seem to be relevant for prediction.

efficacy. We found that the contribution of such hub drugs towards AUC was not significantly different from that of drugs chosen randomly. Although when increasing number of drug hubs, chosen based on the number targets they regulate, were removed, AUC decreased significantly.

Similarly, we removed the profiles of most promiscuous side effects (and targets, independently) to enumerate their contribution towards prediction efficacy (Table 1). We found that efficacy depended more on target profiles of proteins regulated by most number of drugs than side effect profiles of most frequent adverse reactions.

4.2 Drug-chemical profiles

Chemical descriptors of drugs embody relevant therapeutic correlates that have a strong bearing on their side effects. Toxicity response of drugs are specified by their chemical properties and have been widely used in QSAR models[21]. Structural properties of drugs have been reported to be critical in specifying toxicity of drugs [22]. Integration of genomic features and chemical properties have been systematically used to test potential efficacy of drugs as anti-tumor agents[23]. We extended our studies to include chemical profiles of drugs to predict their side effects.

4.2.1 3D chemical profiles We created chemical profiles of drugs with the help of 61 3D chemical properties (enumerating dipole, energy, Jurs descriptors, principal moments, shadow indices, surface area, volume and molecular count). We expect chemical feature matrix composed of these properties to meaningfully represent their therapeutic aspects. We repeated the experiment using drug-chemical profile matrix and by implementing the generalized model by adding the drug-target interactions data (Composite). Figure 3(a) depicts performance with 3D chemical profiles and that with compos-

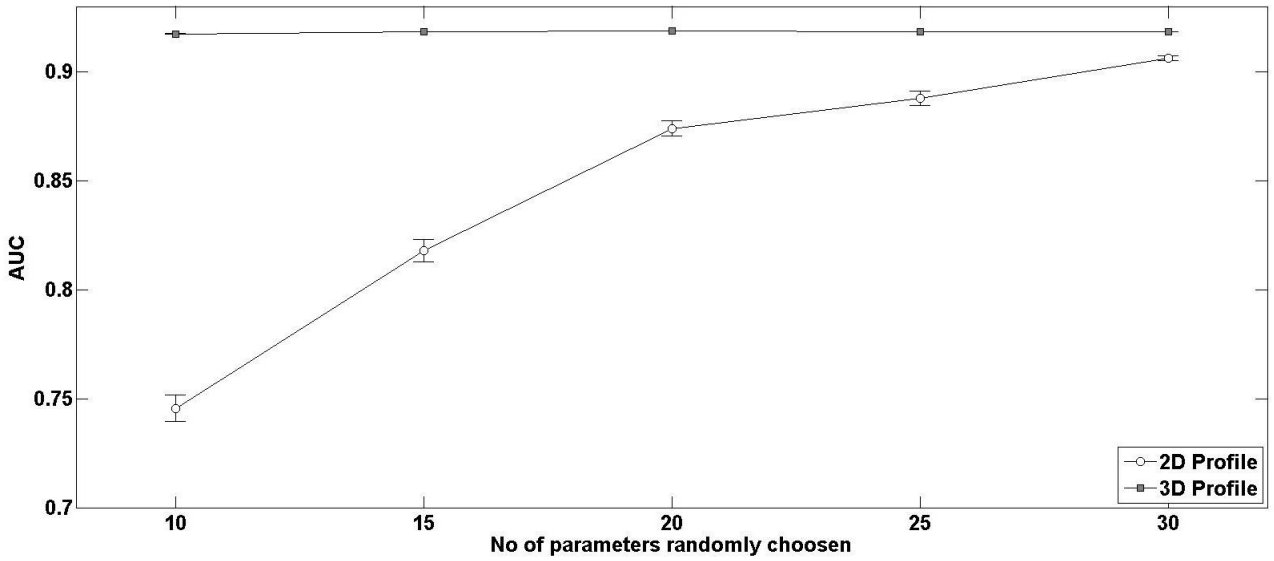


Fig. 4 AUC measured using the largest eigen vector 3-dimensional (a) and 2-dimensional (b) chemical profiles. The error bars indicate standard error of data from 10-fold cross-validation experiments.

ite data for first eigen vector. Including drug-targets profiles data did not improve the prediction performance (AUC, 0.92; Kolmogorov-Smirnov test). We found that the AUC did not change significantly with inclusion of more number of eigen vectors, as present Supplementary Fig. 3.

To probe the relevance of individual chemical properties in prediction of side effects, we further implemented the model by using two properties at a time. Thus, we performed 1830 experiments with pairwise combinations of 61 3D chemical properties. For these experiments we used only the first eigen vector for the prediction. Figure 5(a) depicts the AUC computed for each of the pairwise combinations of chemical properties.

Among the five chemical property pairs that yield best values of AUC, while parameters 11 and 20 together yield the best AUC (0.9188), parameters 10 and 34 have best performance regardless of the parameter they are paired with (Table 3 and Figure 5(a)). Thus we find that best pair constitutes of Jurs_DPSA_3 (11) and Jurs_PNSA_3 (20). Jurs descriptors reflect electronic information present in surface area of individual atoms in the chemicals. Broadly, we find that Jurs chemical features had best correlation for prediction of side effects.

4.2.2 2D chemical profiles Knowing that 3*D* chemical properties of drugs could serve as a critical feature for specification of their side effects, we further created chemical profiles of drugs with the help of 145 chemical properties (enumerating surface area, molecular count and electrostatic properties). We intended to explore the relevance of 2-dimensional chemical properties for specifying adverse drug reactions. As depicted in Figure 3(b), we find that prediction performance, as measured in terms of AUC was close to that returned with 3*D* chemical features. Interestingly, this implies that the contribution of 2*D* chemical features considered for these experiments is comparable to that of 3*D* descriptors (AUC in Supplementary Fig. 5).

4.2.3 Performance of 3D and 2D chemical profiles We performed an experiment of using certain number of 3D and 2D chemical properties randomly. Each of this experiments was repeated 10 times for statistical significance. Figure 4 clearly depicts that 3D chemical properties outperform 2D chemical parameters. With increasing number of (2D and 3D) parameters used for the prediction the accuracy of prediction, as measured in terms of AUC, tend to match. This indicates that 3D features are robust parameters for prediction of side effects with our method. Further to obtain composite set of parameters that could be effectively used together for prediction of side effects, we used 2D and 3D features in a pairwise manner.

In our studies of pair-wise 2*D* chemical properties we performed 10585 experiments with pairwise combinations of 145 2D chemical properties. Figure 5(b) depicts the AUC computed for each of the pairwise combinations of chemical properties. Among the five chemical property pairs that yield best values of AUC, parameter 75 performs best regardless of the parameter it is paired with (Table 4 and Figure 5(b)). The 2*D*

Chemical parameter pair	AUC
11 & 20	0.9188
10 & 19	0.9187
34 & 58	0.9186
10 & 34	0.9185
34 & 41	0.9183

Table 3 3D Chemical parameter pairs with best correlation with side effects.

Chemical parameter pair	AUC
75 & 113	0.9193
113 & 116	0.9193
57 & 113	0.9189
75 & 122	0.9186
75 & 114	0.9186

Table 4 2D Chemical parameter pairs with best correlation with side effects.

chemical properties critically important for specifying side effects include ES_Count_dO (75) and other parameters enumerating Electrostatic property, which reflects stationary or slow moving electric charges of the chemicals.

4.3 Effect of increasing number of eigen vectors

After predicting drug side effects using only the largest eigen vector of feature matrices independently as well as together (Fig. 2 and Fig. 3), we assessed the effect of increasing number of eigen vectors included in the GCCA.

4.3.1 Drug-Target profiles extended to include chemical profiles Generalizing the CCA model built with drug-target profile to include 3*D* and 2*D* chemical profiles led to dramatic improvement in the performance (AUC improved from 0.76 to 0.92, Fig. 2). The model enriched by including increasing number of eigen vectors from drug-target profiles matrix, led to improved and consistent AUC (Supplementary Fig. 1). This result suggests that the composite models are capable of optimizing the objective function when the largest eigen vector is included. We observed that including first eigen vector led to good optimization, which can be explained by spectral analysis of profile matrices. For these matrices, the eigen values were observed to decrease sharply from first to second value, and decayed marginally onward. This composite model returns the best AUC of 0.93.

4.3.2 3*D* **amd** 2*D* **chemical profiles extended to include Drug-Target profiles**In a similar manner, we enriched generalized CCA models built with (3*D* and 2*D*) chemical properties and target profiles, to include increasing number of eigen vectors from chemical profile matrices (Supplementary Figs. 4-7). These composite models enhance the predictive value of chemical property-based models, albeit only marginally. The overall predictive value of composite models is dictated by the chemical profiles, thence emerging as a critical aspect for side effects prediction.

5 Performance evaluation

We repeated each experiment by 10-fold cross-validation with 83×10 drugs. For every experiment, the predicted side effect profile of each drug was normalized with its largest value, and was further binarized with varying thresholds ranging from 0 to 1 with an interval of 0.001. After the binarization, we computed AUC for each experiment. AUC (area under receiver operating curve) was computed from response curve of true positive rate (sensitivity) with increasing false positive rate (1-specificity), and its value reflects the quality of the model. AUC of 0.5 indicates that the model is indistinguishable from that of random sampling; the higher the AUC, the better is the model quality. AUC has been taken with highest eigen vector as this is sufficient to capture major variation present in dataset. This could be described in terms of spectral analysis of composite matrix as (12823, 992, 468,..), (15806,1060, 560,..) and (12863,285, 234,..) corresponding drug-chemical for 3D features, 2D features and drug target matrix respectively. Moreover, proposed model is relatively faster and easy to operate with multivariate data than neural network method[23] Also, this method produces unique solution without using kernel functions which are used in other Generalized cannonical analysis method[16].

6 CONCLUSIONS

We conclude that the performance of canonical correlation model is better using chemical profiles (2- and 3-dimensional properties) as compared to that using target profiles[15] or using chemical structures[14]. While the utility of chemical features for assessing drug toxicity has been demonstrated earlier[21,22], here we show their effectiveness on the basis of empirical data of therapeutic side effects by the application of CCA and GCCA models. Knowing the importance of identification of drug features that are critical for specifying their adverse effects, we propose a generalized ordinary canonical correlation analysis model that integrates the target profiles and chemical profiles of drugs. We anticipated that the target profiles, which encode the off-target aspects of drugs, may be

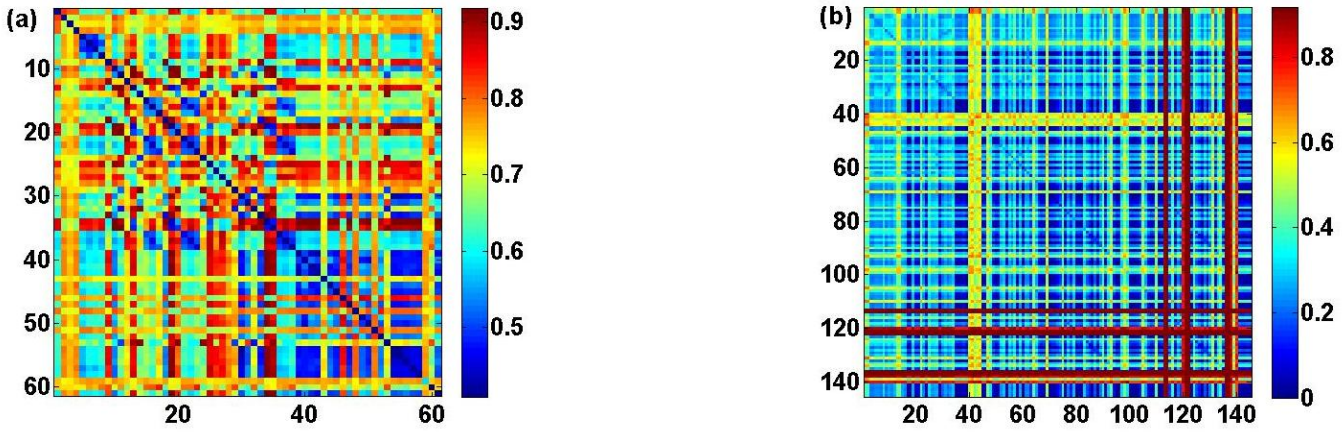


Fig. 5 The AUC matrices for pairwise combinations of (a) 3-dimensional and (b) 2-dimensional chemical properties. Starting from pairwise results a few parameters from each of the two categories were obtained that were critical for side effects prediction.

more relevant in predicting side effects. We found that while target profiles are useful for side effects prediction, chemical features-based predictions outperform it. This implies that low dimensional information is sufficient to achieve good efficacy; *less is more*. Individual chemical properties, that are key to side effects prediction, may provide further insights for improving side effects prediction models with reduced data.

References

1 M. Liu, Y. Wu, Y. Chen, J. Sun, Z. Zhao, X.-w. Chen, M. E. Matheny and H. Xu, *Journal of the American Medical Informatics Association : JAMIA*, 2012, **19**, e28–35.

2 L.-C. Huang, X. Wu and J. Y. Chen, *BMC genomics*, 2011, **12 Suppl 5**, S11.

3 L. Chen, T. Huang, J. Zhang, M.-Y. Zheng, K.-Y. Feng, Y.-D. Cai and K.-C. Chou, *BioMed research international*, 2013, **2013**, 485034.

4 L. Brouwers, M. Iskar, G. Zeller, V. van Noort and P. Bork, *PLOS ONE*, 2011, **6**, year.

5 M. Fukuzaki, M. Seki, H. Kashima and J. Sese, *2009 IEEE International Conference on Bioinformatics and Biomedicine*, 2009, 142–147.

6 M. Campillos, M. Kuhn, A.-C. Gavin, L. J. Jensen and P. Bork, *Science (New York, N.Y.)*, 2008, **321**, 263–6.

7 R. L. Chang, L. Xie, L. Xie, P. E. Bourne and B. O. Palsson, *PLoS computational biology*, 2010, **6**, e1000938.

8 F. Cheng and Z. Zhao, *Journal of the American Medical Informatics Association*, 2014.

9 F. Cheng, Y. Zhou, J. Li, W. Li, G. Liu and Y. Tang, *Molecular BioSystem*.

10 Y. Tang, *journal chemical information and modeling*, 2013, 753–762.

11 D. Weenink, 2003, **25**, 81–99.

12 D. M. Witten, R. Tibshirani and T. Hastie, *Biostatistics (Oxford, England)*, 2009, **10**, 515–34.

13 N. Atias and R. Sharan, *Journal of computational biology : a journal of computational molecular cell biology*, 2011, **18**, 207–18.

14 E. Pauwels, V. Stoven and Y. Yamanishi, *BMC bioinformatics*, 2011, **12**, 169.

15 S. Mizutani, E. Pauwels, V. Stoven, S. Goto and Y. Yamanishi, *Bioinformatics (Oxford, England)*, 2012, **28**, i522–i528.

16 Y. Yamanishi, E. Pauwels and M. Kotera, *Journal of chemical information and modeling*, 2012, **52**, 3284–92.

17 C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. C. Guo and D. S. Wishart, *Nucleic acids research*, 2011, **39**, D1035–41.

18 M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen and P. Bork, *Molecular systems biology*, 2010, **6**, 343.

19 Newmann, *Network An Introduction*, OXFORD UNIVERSITY PRESS, Great Clarendon Street,Oxford ox2 6DP, 2011.

20 S. Günther, M. Kuhn, M. Dunkel, M. Campillos, C. Senger, E. Petsalaki, J. Ahmed, E. G. Urdiales, A. Gewiess, L. J. Jensen, R. Schneider, R. Skoblo, R. B. Russell, P. E. Bourne, P. Bork and R. Preissner, *Nucleic acids research*, 2008, **36**, D919–22.

21 U. Norinder and C. a. S. Bergström, *ChemMedChem*, 2006, **1**, 920–37.

22 R. Sherhod, V. J. Gillet, P. N. Judson and J. D. Vessey, 2012.

23 M. P. Menden, F. Iorio, M. Garnett, U. McDermott, C. H. Benes, P. J. Ballester and J. Saez-Rodriguez, *PloS one*, 2013, **8**, e61318.