5
Phenotypic side effects prediction by optimizing correlation with chemical and target profiles of drugs

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 as 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, 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 [M. Liu et al., 2012]. 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 [L.-C. Huang, Wu, et al., 2011]. 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 [Chen, Huang, et al., 2013]. 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 neighbours of drug targets [Brouwers et al., 2011]. 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 [Fukuzaki et al., 2009]. 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 [Campillos et al., 2008]. Prediction of drug off-targets was implemented for renal disorders through an *in silico* framework [Chang, Xie, Xie, Bourne, and Palsson, 2010].

One of the successful approaches towards prediction of side effects is that of canonical correlation analysis (CCA) [Weenink, 2003; Witten, Tibshirani, and Hastie, 2009]. 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 [Atias and Sharan, 2011]. 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 with chemical substructures (with the help of CCA) are better at predicting side effects than machine learning methods. Mizutani et al. proposed the first target-based feature extraction method using CCA that yielded better results than that based on chemical structures [Mizutani et al., 2012]. Yaminishi et al. reported a kernel regression model that integrates chemical space (chemical structures) and genomic space (drug-target interactions) with good prediction accuracy [Yoshihiro Yamanishi et al., 2012b]. From these studies we reckoned that the

interaction of drug with protein interactome is one of the key specifiers of side effects. While chemical space (representing chemical similarities) has been used for predictive models [Atias and Sharan, 2011; Chen, Huang, et al., 2013; Pauwels et al., 2011; Yoshihiro Yamanishi et al., 2012b], chemical properties of drugs have not been 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 drug's quantitative 2D and 3D chemical features. We predicted the side effects of 830 drugs, that are common to DrugBank [Knox et al., 2011] and SIDER2 [Kuhn, Campillos, Letunic, Jensen, and Bork, 2010], by using only target profiles, only chemical profiles, and by using both. We find that models based on chemical profiles have more consistent accuracy than those based on target profiles. With increasing number of features used in the model, the chemical profile-based model fares better than that based on target profile. We found that a few chemical features are critical in driving the accuracy of our model.

5.1 SIMPLIFIED GCCA

The GCCA model, which is described in Chapter 4, was implemented for analysis presented in this chapter. Since computing Lagrange multipliers involves exponential order complexity, we simplified the procedure as described in the verification section (Section 5.1.3).

5.1.1 Generalized Ordinary Canonical Correlation analysis (GCCA)

Lets say we have d drugs having t targets, c chemical features and s side effects. Each drug $x_{i=1,2,3\cdots d}$ 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 (Figure 3.5). Hence, drug-target matrix (D_r), 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 [Myers and Myers, 2007].

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

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,

$$f = \max_{\alpha,\beta} \alpha^T Z \beta$$

While the above expression can be utilized for predicting side effects using single feature, the expression for multiple features is as follows:

$$f = \sum_{i=1}^{2} \max_{\alpha,\beta} \alpha_i^T Z \beta$$

Such that, $\|\alpha_i\|_2 = \|\beta\|_2 = 1$.

Differentiating f w.r.t. β and α_i by setting $\frac{\partial f}{\partial \alpha_i} = \frac{\partial f}{\partial \beta} = 0$, yields $P\beta = \mu\beta$ and $Z_i\beta = \lambda_i\alpha_i$. β is the eigenvector of $P = \sum_i p_i$, 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 $\|\alpha_i\|_2 = \|\beta\|_2 = 1$, respectively.

5.1.2 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) [Witten et al., 2009].

$$Y = (B^T)^{-1}[\sum_{i=1}^{2} D_i A_i^T X_{p_i}]$$

Note that, $(B^T)^{-1} = B$ when all eigenvectors of P matrix are considered. Here, $B = [\beta_1 \beta_2 \ldots \beta_k \ldots]$ and $A = [\alpha_1 \alpha_2 \ldots \alpha_k \ldots]$, and Y is the predicted side effects profile. Anyway, $(B^T)^{-1}$ is unique as B consists of orthogonal vectors. Every entry D_i of matrix D is given by $\frac{\lambda_i}{\sum_{k=1}^{n}|\lambda_k|}$.

5.1.3 Verification of model

Herein we present verification of our model. By substituting $Z_i\beta = \lambda_i\alpha_i$, we obtained the objective function $f = \sum_i \lambda_i$, which is always positive since $\lambda_i = \|Z_i\beta\|_2$ (norm can never be negative). Next we show that choosing β as the largest eigenvector maximizes the objective function. By substituting α_i in the objective function we obtain $f = \beta^T(\sum_i \frac{1}{\lambda_i} p_i)\beta$. To achieve maximization of this objective function one needs to know β, as λ_i is dependent on β. Therefore, to obtain the objective function we relax the equation by assuming $\lambda_1 = \lambda_2$. With this assumption, the objective function changes to $f = \mu\beta^T\beta = \mu$. Hence, maximization of f can be achieved by choosing the largest eigenvector with μ as its largest eigenvalue.

5.2 RESULTS

We measured the performance of CCA and GCCA model using unconstrained optimization, as described in Section 4.2.1 and Section 4.3.1 in Chapter 4. Table 5.1 shows results for different features used for prediction of side effects.

Table 5. 1 : Features used and AUC for prediction of side effects using 'unconstrained canonical correlation analysis'.

Feature used	AUC
Target profile	0.1468
2D chemical profile	0.1637
Target & 2D chemical profile	0.2189
3D chemical profile	0.1549
Target & 3D chemical profile	0.2602

The procedure of unconstrained approach is simple, but yields poor performance. Hence, we implemented constrained optimization which is based on eigenvectors calculations. All the following results were obtained using constrained optimization. Interestingly, we observed that integration of target and chemical profile improves performance. This highlights the relevance of system-level perspective, which was the key motivation to develop GCCA procedure.

5.2.1 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 5.1 depicts the AUC using first eigenvector of drug-target matrix and that of integrated matrices with 3D 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 eigenvectors of drug-target matrix used for prediction (Figure 5.2).

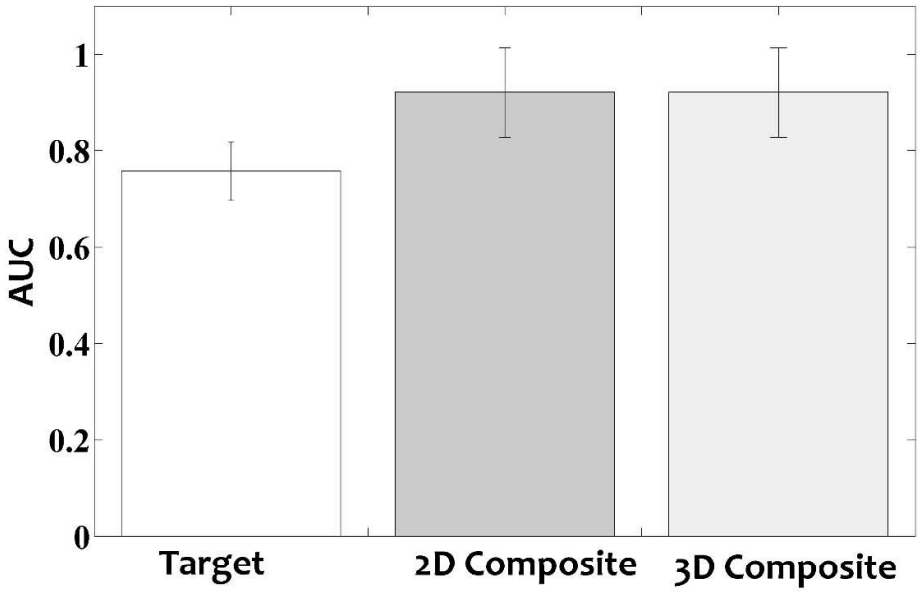

Figure 5.1 : AUC measured using the largest eigenvector of drug-target profiles matrix only and including that of 2D and 3D drug-chemical matrix. The error bars indicate standard error of data from 10-fold cross-validation experiments. Inclusion of chemical profiles data significantly improves the side effects prediction efficacy.

Using a model based on a similar method, Yaminishi and coauthors predicted side effects of drugs using their target profiles and chemical fingerprints [Mizutani et al., 2012]. 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 [Günther et al., 2008].

In Table 5.2, prediction performance of every pair of Lagrange multipliers are shown according to the analytical derivation provided in Section 4.3.2. No significant difference was observed in prediction performance for varying values of theta (θ). This points to the limitation of hyper-sphere technique that was used for creating an analytical relationship among Lagrange multipliers. Therefore, we introduced the simplified GCCA procedure to perform experiments involving use of eigenvectors beyond the first eigenvector. Interestingly, simplified GCCA showed performance (Table 5.2) comparable to that of GCCA ('2D Composite' and '3D Composite' in Figure 5.1).

Table 5.2 : AUC measured of 2D and 3D chemical profile along with target profile. Integration of target profile with 2D and 3D chemical profiles have been referred as 2D+ and 3D+. θ was varied between 0 to 90 degrees.

	$\theta = 1^{o}$	10^{o}	20^{o}	30^{o}	40^{o}	50^{o}	60^{o}	70^{o}	80^{o}	89^{o}
2D +	0.9049	0.9178	0.9186	0.9187	0.9188	0.9188	0.9188	0.9188	0.9188	0.9188
3D +	0.9044	0.9171	0.9183	0.9185	0.9186	0.9187	0.9187	0.9187	0.9187	0.9187

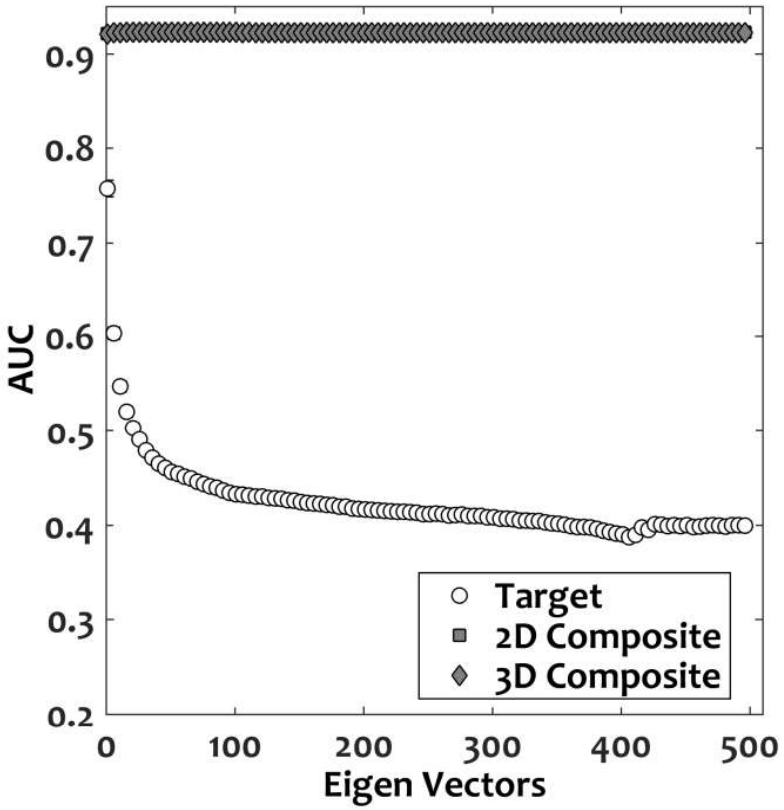

Figure 5.2 : AUC measured using eigenvectors of drug-target profiles matrix only and including that of 2D and 3D drug-chemical matrix. The error bars indicate standard error of data from 10-fold cross-validation experiments. Inclusion of chemical profiles data, significantly improves the side effects prediction efficacy.

Knowing that the distribution of number of targets as well as number of side effects that drugs have is heterogeneous (thick-tailed; Figure 3.2 and Figure 3.3), we evaluated the role of such hub drugs characterized with exceptionally large number of targets and side effects (Table 5.3). We removed dominant hub drugs from the drug-target and drug-side effect matrix to observe their contribution to 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 were removed (chosen based on the number targets they regulate), AUC decreased significantly.

Table 5.3 : Relevance of drug hubs. While generally neither the drugs causing most side effects nor 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.

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

Similarly, we removed the profiles of most promiscuous side effects (and targets, independently) to enumerate their contribution towards prediction efficacy (Table 5.4). 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.

Table 5.4 : Relevance of promiscuous profile hubs. While removal of profiles with most prevalent side effects does not affect the prediction efficacy significantly, the profiles of most promiscuous targets are critical.

	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

5.2.2 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 [Norinder and Bergström, 2006]. Structural properties of drugs have been reported to be critical in specifying toxicity of drugs [Sherhod, Gillet, Judson, and Vessey, 2012]. Integration of genomic features and chemical properties have been systematically used to test potential efficacy of drugs as anti-tumor agents [Menden et al., 2013]. We extended our studies to include chemical profiles of drugs to predict their side effects.

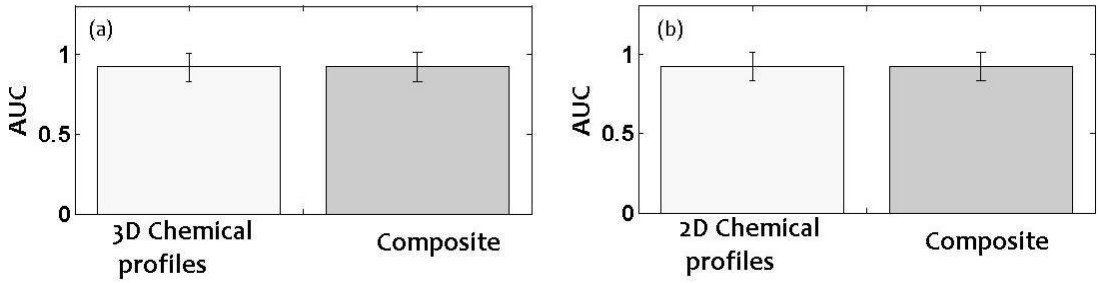

Figure 5.3 : AUC measured using the largest eigenvector for 3-dimensional (a) and 2-dimensional (b) chemical profiles, and including that of drug-target profiles matrix. The error bars indicate standard error of data from 10-fold cross-validation experiments.

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 5.3(a) depicts performance with 3D chemical profiles and that with composite data for first eigenvector. 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 eigenvectors, as shown in Figure 5.4.

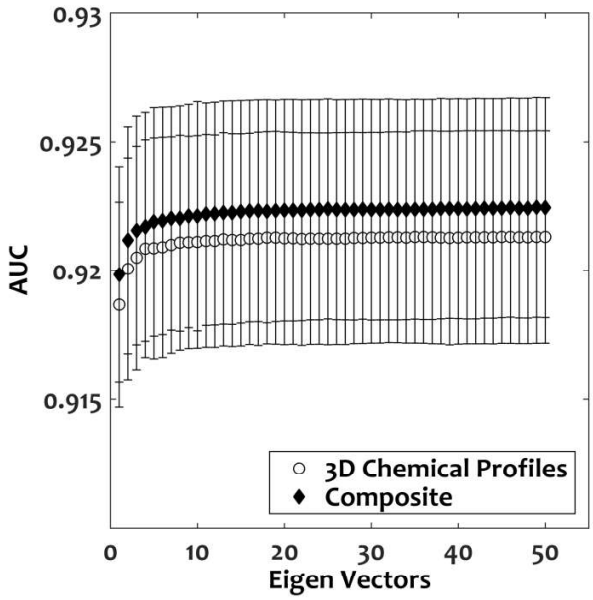

Figure 5.4 : AUC measured using eigenvectors of drug-3D chemical profiles matrix only and including that of drug-target matrix. The error bars indicate standard error of data from 10-fold cross-validation experiments. Inclusion of chemical profiles data significantly improves the side effects prediction efficacy.

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 eigenvector for the prediction. Figure 5.7(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 5.5 and Figure 5.7(a)). 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.

2D chemical profiles

Knowing that 3D 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 2D 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 5.3(b), we find that prediction performance, as measured in terms of AUC was close to that returned with 3D chemical features. The AUC did not improve much when additional eigenvectors were included (Figure 5.5). Interestingly, this implies that the contribution of 2D chemical features considered for these experiments is comparable to that of 3D descriptors (AUC in Figure 5.4).

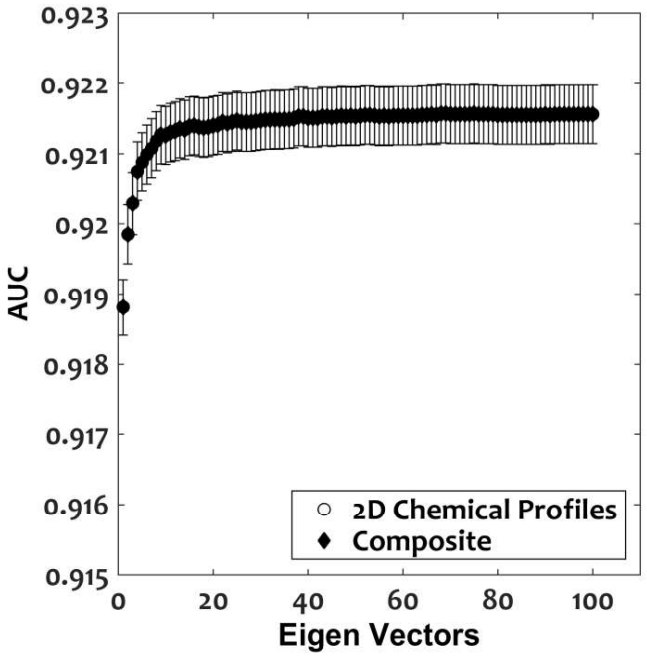

Figure 5.5 : AUC measured using eigenvectors of drug-2D chemical profiles matrix only and including that of drug-target matrix. The error bars indicate standard error of data from 10-fold cross-validation experiments. Inclusion of chemical profiles data significantly improves the side effects prediction efficacy.

Performance of 3D and 2D chemical profiles

We performed an experiment by using certain number of 3D and 2D chemical properties randomly. Each of this experiments was repeated 10 times for statistical significance. Figure 5.6 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. Furthermore, 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 2D chemical properties, we performed 10855 experiments with pairwise combinations of 145 2D chemical properties. Figure 5.7(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 performed best regardless of the parameter it is paired with (Table 5.6 and Figure 5.7(b)). The 2D 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.

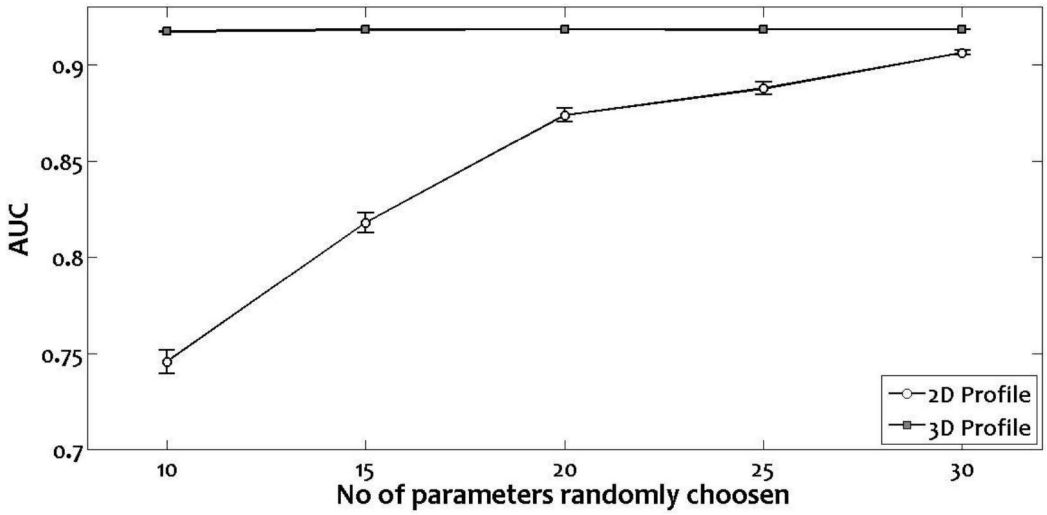

Figure 5.6 : AUC measured using the largest eigenvector 3-dimensional (a) and 2-dimensional (b) chemical profiles. The error bars indicate standard error of data from 10-fold cross-validation experiments.

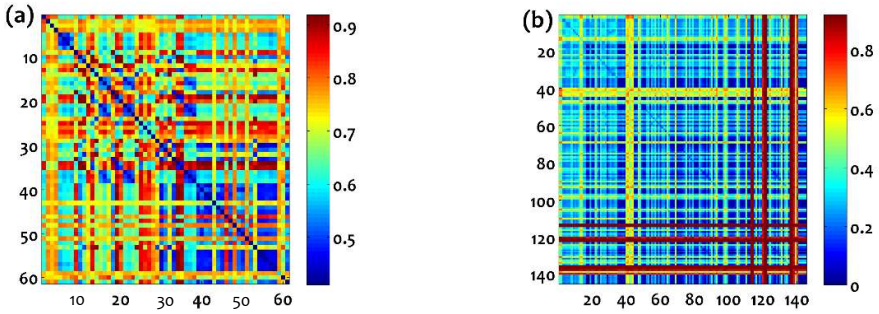

Figure 5.7 : The AUC matrices for pairwise combinations of (a) 3-dimensional and (b) 2-dimensional chemical properties. Starting from pairwise results a few parameters from each of the two categories were obtained that were critical for side effects prediction.

Table 5.5 : 3D Chemical parameter pairs with best correlation with side effects.

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 5.6 : 2D 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

5.2.3 Effect of increasing number of eigenvectors

After predicting drug side effects using only the largest eigenvector of feature matrices independently as well as together (Figure 5.1 and 5.3), we assessed the effect of increasing number of eigenvectors included in the GCCA.

Drug-Target profiles extended to include chemical profiles

Generalizing the CCA model built with drug-target profile to include 3D and 2D chemical profiles led to dramatic improvement in the performance (AUC improved from 0.76 to 0.92, Figure 5.1). The model enriched by including increasing number of eigenvectors from drug-target profiles matrix led to improved and consistent AUC (Figure. 5.2). This result suggests that the composite models are capable of optimizing the objective function when the largest eigenvector is included. We observed that including first eigenvector led to good optimization, which can be explained by spectral analysis of profile matrices. For these matrices, the eigenvalues were observed to decrease sharply from first to second value, and decayed marginally onward. This composite model returns the best AUC of 0.93.

3D and 2D chemical profiles extended to include Drug-Target profiles

In a similar manner, we enriched generalized CCA models built with (3D and 2D) chemical properties and target profiles, to include increasing number of eigenvectors from chemical profile matrices (Figure 5.2, Figure 5.4 and Figure 5.5). 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.

5.3 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 with the highest eigenvector was considered, 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, the proposed model is relatively faster and easy to operate with multivariate data than neural network method [Menden et al., 2013]. Also, this method produces unique solution without using kernel functions which are used in other generalized canonical analysis methods [Yoshihiro Yamanishi et al., 2012b].

5.4 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 [Pauwels et al., 2011] or using chemical structures [Atias and Sharan, 2011]. While the utility of chemical features for assessing drug toxicity has been demonstrated earlier [Norinder and Bergström, 2006; Sherhod et al., 2012], 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 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.

…