8

A partial canonical correlation model for identification of drug features specific to side effect classes

Drugs are known to cause adverse reactions in addition to their intended therapeutic effects. These 'side effects' of drugs have origin in a range of factors associated with chemical properties of drugs and the nature of their interactions with cellular molecules [Demetri, 2007]. Accurate prediction of side effects is of key value to drug discovery process, as unintended effects of drugs is a major cause of their rejection due to interaction with a web of molecular entities [C. P. Adams and Brantner, 2006; D. J. Adams, 2013; Demetri, 2007; Kola and Landis, 2004; Kramer, 2003]. Various computational techniques have been developed to use empirical data of chemical properties of drugs as well as their interaction with macromolecular entities in an attempt to predict adverse drug reactions so as to capture specific causative mechanisms. Such techniques include machine learning [Ammad-Ud-Din et al., 2014; M. Liu et al., 2012], artificial neural networks [Menden et al., 2013], support vector machine [Chen, Huang, et al., 2013; L.-C. Huang, and Wu, and Chen, 2011; L. C. Huang et al., 2013] and canonical correlation analysis (CCA) [Atias and Sharan, 2011; Mizutani et al., 2012; Pauwels et al., 2011]. CCA represents a strategy for identification of parameters that maximize two vectors representing two interdependent processes [Weenink, 2003]. In the case of drugs, for example, the chemical properties profile of drugs may constitute one vector and the other vector could be represented by its side effects profile vector. CCA has been demonstrated to be a good strategy for prediction of drug side effects [Atias and Sharan, 2011; Mizutani et al., 2012; Pauwels et al., 2011]. A large number of drug features could be causally linked to their side effects: 2D and 3D chemical features of drugs, clinical observations, drug-target interactions, gene ontology annotations, chemical-chemical interactions, protein-chemical interactions, pathways through which drugs act etc. Previous studies have used some of these features towards prediction of side effects using CCA model [Atias and Sharan, 2011; Mizutani et al., 2012; Pauwels et al., 2011]. While these reports highlight the effectiveness of CCA, given the complex nature of intertwined factors that lead to observed adverse drug reactions, it is also clear that CCA itself is of limited value.

Generalized Canonical Correlation Analysis (GCCA) facilitates integration of multiple drug features to create a composite model. GCCA has been demonstrated to be of better value in prediction of side effects. Earlier studies have integrated drug various features such as properties of chemical structures and details of their targets. After combining 3D and 2D chemical properties with target profile, Kanji et.al. reported that 3D chemical properties, especially Jurs descriptors, are crucial for specifying side effects [Kanji et al., 2015]. Other integrative approaches have combined chemical substructures and target profiles. Yaminishi et.al. demonstrated that such integration of features outperforms the use of individual feature using a kernel-based method [Yoshihiro Yamanishi et al., 2012b]. Also, in another study Bresso et.al. used linear and ring sub-structure of drugs, that were extracted from SMILES, along with protein-protein interactions for predicting drug side effects [Bresso et al., 2013].

The analytical treatment of GCCA model is heavily simplified by assuming that drug features integrated in it are independent. While the model has served well, it is evidently based on an unreasonable assumption. Drug features such as chemical properties, targets and side effects are intricately linked with each other contrary to this assumption. For example, structural property of a drug has a strong bearing on the class of proteins it may interact with. Further the

pathways with which a drug interferes, is linked to its phenotypic side effects. Hence, while it is luring to pool large number of drug features in a computational framework (such as artificial neural network or machine learning algorithms), the prediction accuracy of such a model may not necessarily scale with number of features used. Hence, to be able to build an effective computational model, it is extremely important to identify contribution of individual features towards accurate prediction of side effects. One of the challenges in this direction is to disentangle interdependence of features to identify contribution of individual features that specify side effects.

Selection of key features from a set of interdependent descriptors is an important problem in biomedical data analysis with applications for drug design and disease diagnosis [Y Liu, 2004; Perlman, 2004; Weston et al., 2003; Xiong et al., 2001]. Feature selection methods can be classified into following three categories each with their pros and cons [Bolón-Canedo, Sánchez-Maroño, and Alonso-Betanzos, 2013; Guyon, Isabelle and Gunn, Steve and Nikravesh, Masoud and Zadeh, 2008; Saeys, Inza, and Larranaga, 2007]: filter, embedded and wrapper. Among these, 'filter' method is the simplest which selects subset of features on the basis of mutual independence without using a classifier. Few of the popular procedures used as part of this method are mutual information, minimum redundancy maximum relevance, and fast correlation [Ding and Peng, 2003; Hanchuan Peng, Fuhui Long, 2005; L. Yu and Liu, 2004]. But this is a 'model free' method and does not allow class-specific feature selection. In contrast to 'filter' method, 'embedded' methods use classifiers (such as LASSO and Bolasso) to altogether ignore irrelevant features [Bach, 2008; Tibshirani, 1996]. In a more nuanced manner, soft methods such as weighted-SVM, weighted-logistic regression, and artificial neural network attribute weights for selection of best features [R. P. Li, Mukaidono, and Turksen, 2002; Ma and Huang, 2005; X. Wang and Tian, 2012]. 'Wrapper' methods follow a strategy of sampling random subset of features which are further prioritized with a classifier. The inevitable complexity associated with sampling large number of subsets is mitigated with procedures like sequential forward selection and sequential backward selection [Iñaki Inza et al., 2004; Xiong et al., 2001]. Heuristic strategies such as genetic algorithm and estimation of distribution algorithms start with a random set of features to converge onto an approximated set of best features[I Inza et al., 2000; Jirapech-Umpai and Aitken, 2005].

Towards our goal of obtaining features that contribute the most to side effects prediction, we present a partial canonical correlation analysis (PCCA) model that facilitates enumeration of contribution from individual drug features. We used this model in conjunction with 'wrapper' for enumerating contribution of an individual drug feature, irrespective of interdependence on other features. The generalized PCCA presented in this study allows us to measure contribution of individual features using a combination of analytical and numerical techniques. The solution to this model could be achieved by starting with approximation of one of the three constraint parameters, and to obtain its value through an iterative procedure that is complemented by the analytical solution. We demonstrate the utility of our model by identification of 2D features with most significant contribution to Nose, Eye, and Abdomen related side effects. Figure 8.1 provides an overview of the strategy implemented for application of PCCA model in identification of key drug features.

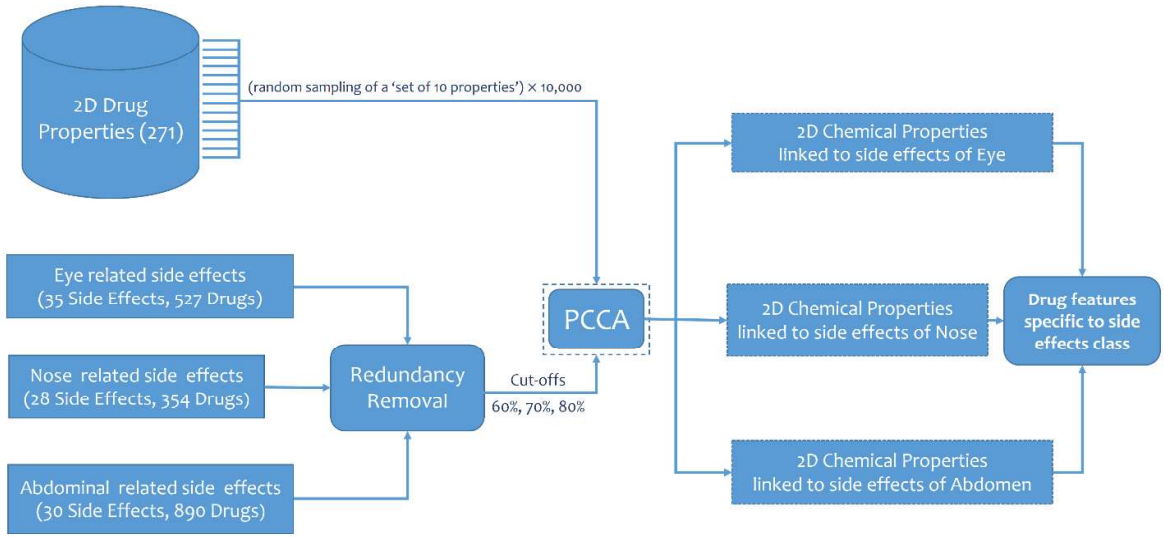

Figure 8.1 : The PCCA model facilitates enumeration of contribution of individual features using a combination of analytical and numerical techniques. This schematic depicts utility of PCCA for identification of 2D drug features that are linked with Eye, Nose, and Abdomen related side effects. This model is of value in drug discovery process as it allows to deal with multidimensional feature space towards identification of key features from an array of interdependent drug descriptors. Please see Methods section for mathematical description of PCCA. Also see figures in Annexure B and Figure 8.3 for Venn diagram depiction of process used post PCCA.

8.1 MATERIALS

Drug features were compiled from different data sources and by using tools available, such as SIDER [Kuhn et al., 2016] and Discovery Studio 4.0. We obtained data of side effects for 1430 drugs using SIDER 4.1 (January 2017). SIDER is a one of the most comprehensive open resources available and, beyond side effects prediction, has been used towards various other objectives [Campillos, 2016; Deghou et al., 2016; Wooden, Goossens, Hoshida, and Friedman, 2016]. The number of side effects for these drug range from 1 and 757. Starting from STITCH IDs obtained from SIDER, using corresponding PubChem IDs, SMILES and SDF files were compiled from PubChem website. For each of these drugs, chemical descriptors were computed using Discovery Studio 4.0 (CHARMm force field). A total of 271 2D and 56 3D chemical properties were obtained (Listed in Chapter 3). However the 3D descriptors were not used for further analysis due to excessive redundancy (Please see Section 8.4.1 for details). 2D features represent various structural aspects of drugs such as molecular properties, electrostatic properties, and molecular surface area. As part of preprocessing, each of the chemical feature vector was scaled between 0 to 10 to remove biases across features. Further for every chemical feature, the entries with value less than a cut-off (1.25 × mean) value (of the vector across all drugs) were nullified (0) as outliers. Subsequent to this preprocessing, each drug was assigned with three column vectors of size 271×1 and 5868×1 corresponding to 2D chemical profiles and side effects profile, respectively.

Entries in the side effects profile vector indicates presence (1) or absence (0) of a particular side effect with the corresponding drug. The drug-side effects matrix had 139,756 associations among drugs and their side effects, and was found to be sparse with 1.67% of the maximum possible associations 8,391,240. This is a reflection on the fact that from a large number of specific phenotypic side effects, each drug is associated with a small fraction of them. While the average number of side effects for drugs was ~97, the drug with worst adverse reactions profile was reported to have 757 side effects.

8.1.1 Compilation of Eye related Drug-Side Effects matrix

Starting with the comprehensive drug-side effects matrix compiled, 'Eye' related subspace of drugs and adverse reactions were obtained by filtering for side effects that contained word 'Eye' in them. Every drug in this list had at least one Eye related side effect. This protocol yielded 35 side effects and 527 drugs. For the list of these Eye related side effects please refer to Table 7.1. The sparseness of this matrix is 6% average and maximum number of side effects for these drugs is 2.12 and 15, respectively. 5-fold cross-validation was performed over 160, 400 and 505 drugs for redundancy cut-off of 60%, 70% and 80%, respectively (Table 8.1).

Table 8.1 : Statistics for adverse reactions related to 'Eye' for different cut-offs of redundancy.

	≤0.6	≤0.7	≤0.8
Drugs	160	400	505
Sparse-ness	9%	6.64%	6%
Mean side effects	2	2.19	2.11
Maximum side effects	12	15	15

8.1.2 Compilation of Nose-related Drug-Side Effects matrix

The same protocol was used to compile the drug-side effects matrix for Nose-related adverse reactions using keyword 'Nasal'. It yielded 28 side effects and 354 associated drugs. For the list of these Nose-related side effects please refer to Table 7.1. The matrix had 6% sparse-ness, with on an average 1.6 side effects and maximally having 15 adverse reactions. 5-fold cross-validation was performed over 60, 230 and 305 drugs for redundancy cut-off of 0.6, 0.7 and 0.8, respectively (Table 8.2).

Table 8.2 : Statistics for adverse reactions related to 'Nose' for different cut-offs of redundancy.

	≤0.6	≤0.7	≤0.8
Drugs	60	230	305
Sparse-ness	13%	7%	5.8%
Mean side effects	1.93	1.72	1.62
Maximum side effects	9	9	9

8.1.3 Compilation of Abdomen-related Drug-Side Effects matrix

Using the same protocol the drug-side effects matrix for Nose-related adverse reactions was compiled with the keyword 'Nasal'. It yielded 30 side effects and 890 associated drugs. For the list of these Nose-related side effects please refer to Table 7.1. The matrix had 6% sparse-ness, with on an average 1.6 side effects and maximally having 15 adverse reactions. 5-fold cross-validation was performed over 260, 640 and 850 drugs for redundancy cut-off of 0.6, 0.7 and 0.8, respectively (Table 8.3).

Table 8.3 : Statistics for adverse reactions related to 'Abdomen' for different cut-offs of redundancy.

	≤0.6	≤0.7	≤0.8
Drugs	260	640	850
Sparse-ness	12%	7%	6%
Mean side effects	1.85	1.79	1.77
Maximum side effects	7	7	7

8.2 METHOD

Let's say a given dataset comprises of d drugs for which we would like to find out best of the chemical features that are most independently associated with s side effects. From the available features (2D properties), we select r features randomly to know their contribution towards prediction of side effects, after eliminating interdependence with rest of the c features. These randomly selected and rest of the features were assumed to be linearly correlated with side effects. Every drug $p_{i=1,2,3\cdots,d}$ was assigned with chemical profile vector (2D) and a side-effects profile vector having dimension $1 \times r$ (for randomly selected features), $1 \times c$ (remaining properties) and $1 \times s$, respectively. Thus, drug-chemical features matrix for random features (D_r), rest of features (D_c) and drug-side effect matrix (D_s) have dimension of $d \times r$, $d \times c$ and $d \times s$ respectively. The objective function, f that encodes correlation between two variables, was developed according to the definition of partial correlation $\rho_{xy,w}$ [Anderson, 1958].

$$\rho_{xy,w} = \frac{\rho_{xy} - \rho_{xw}\rho_{yw}}{\sqrt{(1-\rho_{xw}^2)}\sqrt{(1-\rho_{yw}^2)}} \tag{8.1}$$

Thus, $\rho_{xy,w}$ defines partial correlation between x and y such that the effect of third controlling variable w is removed. So as to obtain partial correlation of D_r matrix with D_s matrix by eliminating the effect of third controlling variable D_c, these matrices are required to be of the same dimension. This could be achieved through transformation of these matrices: $U_r = D_r\alpha_1$, $U_c = D_c\alpha_2$, and $V = D_s\beta$. These transformations implicitly entail following assumptions:

$$\rho_{xy} = \frac{U_r^T V}{\|U_r\|_2 \|V\|_2}$$

$$\rho_{xw} = \frac{U_r^T U_c}{\|U_r\|_2 \|U_c\|_2}$$

$$\rho_{yw} = \frac{U_c^T V}{\|U_c\|_2 \|V\|_2}$$

Here $Z_1 = D_r^T D_s$, $Z_2 = D_r^T D_c$ and $Z_3 = D_c^T D_s$. The unknown vector parameters α_1, β and α_2, corresponding to randomly selected features, rest of the chemical features and side effects, could be obtained for maximizing the objective function (f) that represents the numerator of Eq. (8.1) Therefore, the objective function f takes the following form:

$$f = \alpha_1^T Z_1 \beta - \alpha_1^T Z_1 \beta \alpha_2^T Z_2 \beta$$

Such that, $\alpha_1^T \alpha_1 = \alpha_2^T \alpha_2 = \beta^T \beta = 1$.

To find the unknown parameters α_1, β and α_2, the objective function is rewritten in its Lagrange form as,

$$f = \alpha_1^T Z_1 \beta - \alpha_1^T Z_1 \alpha_2 \alpha_2^T Z_2 \beta + \lambda_1(1-\alpha_1^T \alpha_1) + \lambda_2(1-\alpha_2^T \alpha_2) + \mu(1-\beta^T \beta)$$

Here, λ_1, λ_2 and μ are Lagrange multiplier for constraints equation associated with unknown parameters. By setting equation of differentiating f w.r.t α_1, β and α_2 to be 0 will give us three equation, which are given below.

$$[Z_1 - Z_2\alpha_1\alpha_2^T Z_3][Z_1 - Z_2\alpha\ \alpha_2^T Z_3]^T \alpha_1 = \lambda_1\mu\alpha_1 \quad (8.2)$$

$$[Z_1 - Z_2\alpha\ \alpha_2^T Z_3]^T[Z_1 - Z_2\alpha\ \alpha_2^T Z_3]\beta = \lambda_1\mu\beta \quad (8.3)$$

$$[Z_2^T\alpha_1\beta^T Z_3^T - Z_3\beta\alpha_1^T Z_2]\alpha_2 = \lambda_2\alpha_2 \quad (8.4)$$

Therefore, an iterative approach was applied further to compute unknown parameters. Please find the PCCA strategy depicted as a flowchart (Figure 8.2) and the pseudocode, below. Refer to Section 4.3 for detailed derivation.

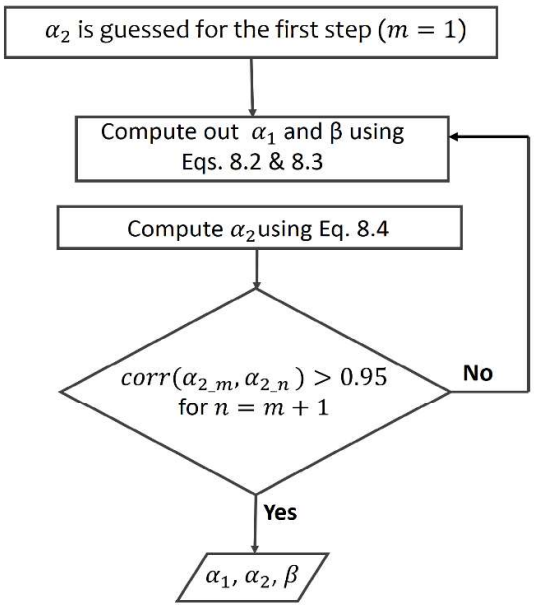

Figure 8.2 : Flowchart of Partial Canonical Correlation Analysis for three parameter system.

8.2.1 Pseudocode

Step 1: α_1 is initially assigned with a vector of every element 1.
Step2: Estimate α_1 and β by putting α_2 into Eqs. (8.2) and (8.3).
Step3: Updating the old value of $\alpha_2$with new α_2 which is estimated by putting α_1 and β into Eq. (8.4).
Step4: Check correlation of new and old α_2 with predefined threshold (0.95).
Step5: If correlation value exceeds the threshold; then stop this process with estimated α_1, β and α_2. Otherwise, go to Step2.

8.2.3 Prediction model

The objective function of PCCA depicted in Eq. (8.1) was maximized, aligned with the argument presented in Section 4.4.1. We set unknown parameters α_1, β and α_2 as highest eigenvector of Eq. (8.2), Eq. (8.3) and Eq. (8.4) equations respectively. In this study, PCCA quantifies contribution of 2D chemical profile. Following formula was used for predicting the drug side effect profile (Y).

$$Y = (\beta^T)^{-1}[\alpha_1^T x_1 - \frac{\alpha_1^T x_1 x_2^T \alpha_2}{\sqrt{1-(\alpha_1^T x_1 x_2^T \alpha_2)^2}}\alpha_2^T x_2] \quad (8.5)$$

Here, x_1 refers to the randomly chosen set of features and x_2 refers to the remainder of the features.

8.3 PERFORMANCE EVALUATION

For a given number of drug features (properties), the performance of PCCA model was computed through 10000 unique experiments by choosing a set of 10 features randomly for every experiment. For each class of side effects ('Eye', 'Nose' and 'Abdomen'), 5-fold cross-validation statistics was implemented after obtaining non-redundant drugs with different cut-offs. The statistics for these groups are provided in Table 8.1, Table 8.2 and Table 8.3. Accordingly, for the least redundant sets (cut-off≤0.6) of Eye, Nose and Abdomen related side effects the number of drugs used for testing were 32, 12, and 52, and those used for training were 128, 48 and 208, respectively. Prediction of side effect profile was made by using the highest eigenvector, as it carries sufficient information according to spectral theory, as demonstrated in our earlier work [Kanji et al., 2015]. The side effects profile was normalized so as to limit their range between 0 and 1. The performance of the model was evaluated using True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR) and False Negative Rate (FNR). These were computed by binarizing the profile vector at intervals of 0.001 between 0 and 1. In the drug-side effects matrix, the proportion 'side effects present (1)' was significantly lower than the 'side effect absent (0)'. Therefore, for prediction performance evaluation of such an unbalanced dataset, we considered F1-score [Powers, 2011] and Mathews Correlation Coefficient (MCC) [Matthews, 1975]. In the presence of unbalanced data, these measures are preferred over AUC (Area under curve) and Accuracy [Bhasin and Raghava, 2004; Chen, Feng, Cai, Chou, and Li, 2010; B. Q. Li et al., 2013; Petersen, Lundegaard, and Petersen, 2010].

$$F1\_score = \frac{2TP}{2TP + FP + FN}$$

$$MCC = \frac{TP \times TN}{\sqrt{(TP + FP)(TP + FP)(TN + FP)(TN + FP)}}$$

$F1\_score$ varies between 0 to 1, which represent 'as good as random' and 'perfect prediction performance', respectively. Different forms of $F1\_score$ can be implemented by varying the significance (weight) attached to precision and sensitivity/recall [Van Rijsbergen, 1986]. In our implementation, we are counting contribution from both the measures. Unlike $F1\_score$, MCC accounts for true negatives and hence was incorporated along with $F1\_score$. MCC ranges between -1 and 1, with 0 standing for quality of prediction comparable to that from random sampling. Values of MCC greater and lesser than 0 indicate better or worse quality of prediction as compared to random sampling.

Using different weights for $F1\_score$ and MCC ((0.5,0.5), (0.6,0.4), (0.4,0.6)), we arrived at three different ranked lists of features. We used a weighted performance (WP) metric to rank randomly sampled sets. The ranking of sets was practically invariant regardless of the weights used. For our analysis, we have used the sets ranked with equal weights ($a = 0.5$) given to both $F1\_score$ and MCC.

$$WP = a \times MCC + (1 - a) F1\_score, \text{ Here } a \in \{0.6, 0.5, 0.4\}$$

8.4 RESULTS

Given the availability of large number of drug features, it is important to integrate them towards prediction of side effects. One of the problems in this approach is that drug features tend to be interdependent. The PCCA model, presented in this article, facilitates identification of most independent features that are critical for specifying the side effects. This model, represented by Eq. (8.5), could be segmented into two parts. The first part, analogous to CCA, measures contribution of a given feature in the absence of other features. The second part represents interdependence with other features. We implemented the PCCA model, along with a 'wrapper' class of feature selection strategy, on three groups of side effects (Eye, Nose and Abdomen).

Section 8.4.1 describes the strategy used for quantification of key drug features linked to for Eye (Section 8.4.2), Nose (Section 8.4.3) and Abdomen (Section 8.4.4) related side effects. The schematic of strategy implemented in this Chapter is depicted in Figure 8.3.

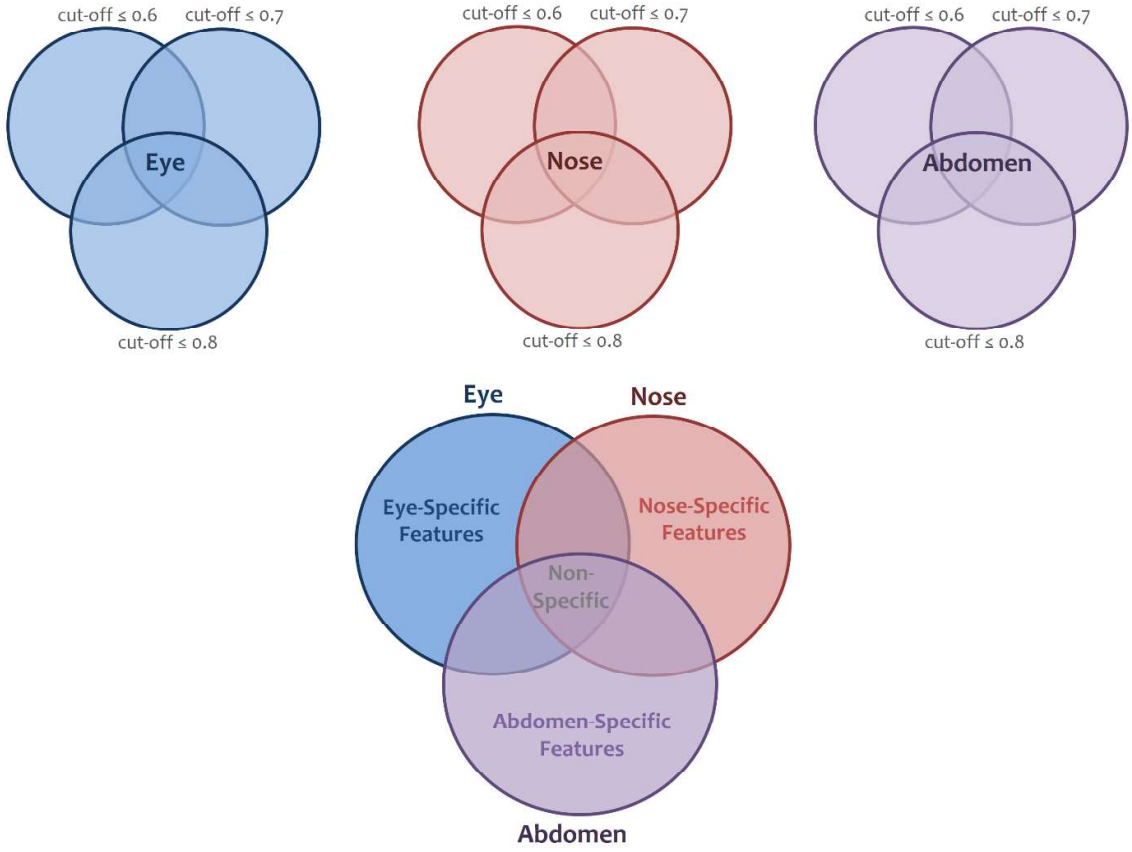

Figure 8.3 : Identification of 2D features specific to side effects of Eye, Nose, and Abdomen. Starting from features that were consistently found among the top sets at different cut-offs of redundancy using PCCA for each class (Refer to Table 8.4, Table 8.5 and Table 8.6, and Annexure B), we obtained features that were specific to Eye, Nose, and Abdomen. Features that are linked to each class were found to be highly specific with no overlap across all three classes (Only one feature was found in the intersection of Eye and Nose linked features), highlighting the ability of PCCA and the protocol presented for arriving at class-specific features.

8.4.1 Strategy for identification of 2D features specific to organ/class

Figure 8.1 depicts the outline of the strategy used for identification drug features that are closely linked to side effects of Eye, Nose, and Abdomen. To assess redundancy in each of these classes drug similarity was computed across 2D and 3D chemical properties using Kendall Tau correlation [Daniel, 2000]. The drugs were found to be highly similar (~90%) with respect to 3D properties, thereby rendering them of limited utility. When compared against 2D properties, on the other hand, the drugs in each of these classes were fairly non-redundant. Hence we randomly sampled 10000 sets (of size 10) of 2D properties (3D properties were rejected due to high redundancy) for different cut-offs of redundancy. These 2D chemical properties broadly belong to following classes: 'Molecular property', 'Electrostatic property', and 'Molecular surface area'. We selected three levels of redundancy cut-offs (60%, 70%, and 80%) to identify drugs that were increasingly redundant. Further, we used PCCA for computing prediction performance of each of these 10000 sets. The Prediction Performance was measured in terms of MCC and $F1\_score$ with 10-fold cross-validation (Please see 'Performance Evaluation' section for more details). For each organ-specific class, we identified a set of features that were consistently present in top-ranked sets at all three similarity cut-offs, which were deemed as closely linked to the side effect class (See data in Annexure B as well as results in Table 4, Table 5 and Table 6, for Eye, Nose, and

Abdomen class, respectively). In this procedure, Top 5 best ranked sets were used at lower similarity (0.6 and 0.7) and Top 25 best ranked sets were used at higher similarity, to account for redundancy. Further, using these organ-linked 2D features thus obtained, features specific (not overlapping with other organs classes) to each class were identified (Figure 8.3)

8.4.2 Chemical features contributing to Eye related side effects

Following the strategy described above, we obtained 10000 random sets of 2D properties. Figure 8.4 depicts the statistics of their prediction performance computed using PCCA in conjunction with the eye-related drug-side effects matrix. The sets with higher performance contribute more to organ-specific side effects than the sets with poor performance. The range of performance across random sets became narrower with increasing redundancy. For ≤60% similarity, the performance varied between 0 and 0.69 (with mean performance of 0.64 ± 0.03; Figure 8.4(a)). For ≤70% similarity, the performance ranged between 0.21 and 0.57 with an average of 0.54± 0.02 (Figure 8.4(b)). And, for ≤80% similarity, the performance ranged between 0.4 and 0.62 with an average of 0.6± 0.02 (Figure 8.4(c)). Overall, this suggests that among randomly sampled chemical properties the prediction performance is highly variable, and thus our strategy can facilitate identification of best features closely linked to the Eye related side effects.

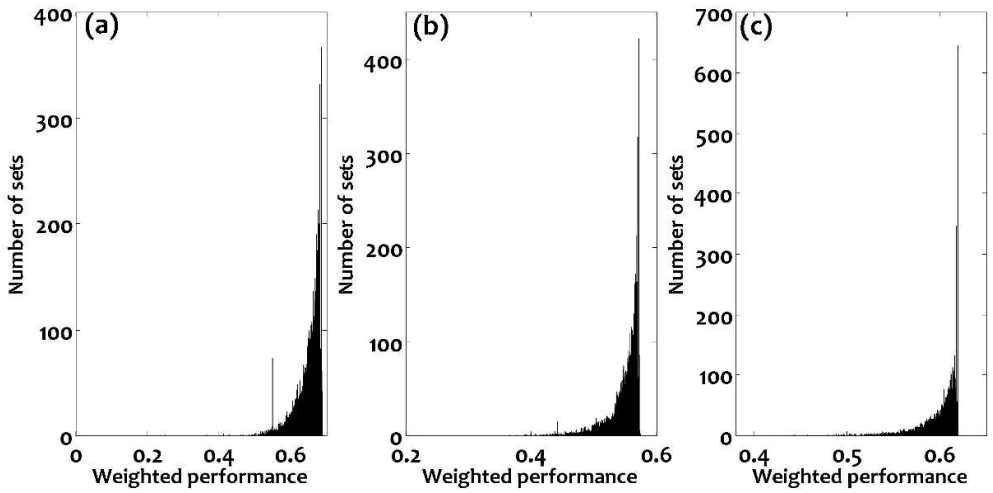

Figure 8.4 : Distribution of Weighted Performance of 10,000 randomly sets of size 10 sampled from all 2D features, for Eye related side effects, with increasing redundancy. (a) ≤60% (b) ≤70% and (c) ≤80%.

Top features closely linked to Eye, that were consistently present across varying redundancy levels, are listed in Table 8.4. Please refer to Figure B1, Table B1, Table B2, and Table B3 in Annexure B for more details.

Table 8.4 : 2D features closely linked with 'Eye' related side effects, obtained using PCCA.

Chemical feature family	Class-linked features
Electrostatic property	ES Count aaNH
Electrostatic property	ES Count dS
Electrostatic property	ES Count sSH
Electrostatic property	ES Count tCH

Molecular property	Log D

8.4.3 Chemical features contributing to Nose related side effects

Using the same randomly sampled sets of 2D properties, we computed their prediction performance using PCCA in conjunction with the Nose-related drug-side effects matrix (Figure 8.5). The figure depicts statistics of PCCA prediction performance with different levels of redundancy. For drugs with 60% similarity, the performance varied between 0 and 0.65, with mean performance of 0.62±0.03 (Figure 8.5(a)). At the cut-off of 70%, the performance ranged between 0.31 and 0.61 with an average of 0.56±0.03 (Figure 8.5(b)). And, for 80% similarity, the performance varied between 0.33 and 0.59, with mean performance of 0.56±0.02 (Figure 8.5(c)). With the observed variability in prediction performance our strategy allows identification of Abdomen-linked best features. Right tails of these distributions hold feature sets with best performance contributing most to organ-specific side effects.

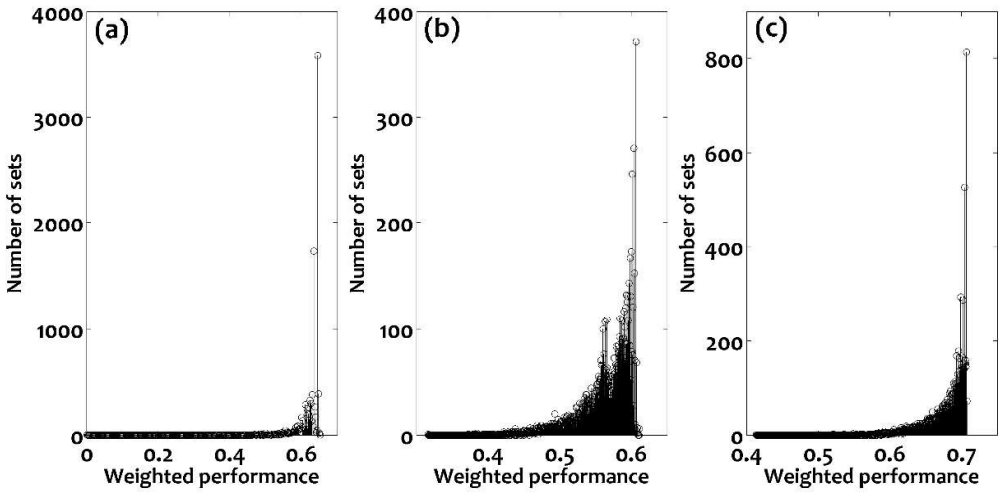

Figure 8.5 : Distribution of Weighted Performance of 10,000 randomly sets of size 10 sampled from all 2D features, for Nose related side effects, with increasing redundancy. (a) ≤60% (b) ≤70% and (c) ≤80%.

Top features closely linked to Nose, that were consistently present across varying redundancy levels, are listed in Table 8.5. Please refer to Figure B2, Table B4, Table B5, and Table B6 in Annexure B for more details.

Table 8.5 : 2D features closely linked with 'Nose' related side effects, obtained using PCCA.

Chemical feature family	Class-linked features
Electrostatic property	ES Count dS
Electrostatic property	ES Count sF
Electrostatic property	ES Sum sssNH
Molecular property	HBA Count

8.4.4 Chemical features contributing to Abdomen related side effects

Following the strategy described in earlier sections, we used the set of randomly sampled sets of 2D properties for arriving at features linked to Abdomen. Figure 8.6 depicts their prediction performance using PCCA in conjunction with the Abdomen-related drug-side effects matrix. With increasing levels of redundancy, the figure depicts the statistics of PCCA prediction performance. For drugs with 60% similarity, the performance varied between 0.24 and 0.61, with mean performance of 0.58±0.02 (Figure 8.6(a)). At the cut-off of 70%, the performance ranged between 0.42 and 0.62 with an average of 0.59±0.02 (Figure 8.6b)). And, for 80% similarity, the performance varied between 0.40 and 0.61, with mean performance of 0.59±0.02 (Figure 8.6(c)). With the observed variability in prediction performance our strategy allows identification of Abdomen-linked best features. Right tails of these distributions hold feature sets with best performance contributing most to organ-specific side effects.

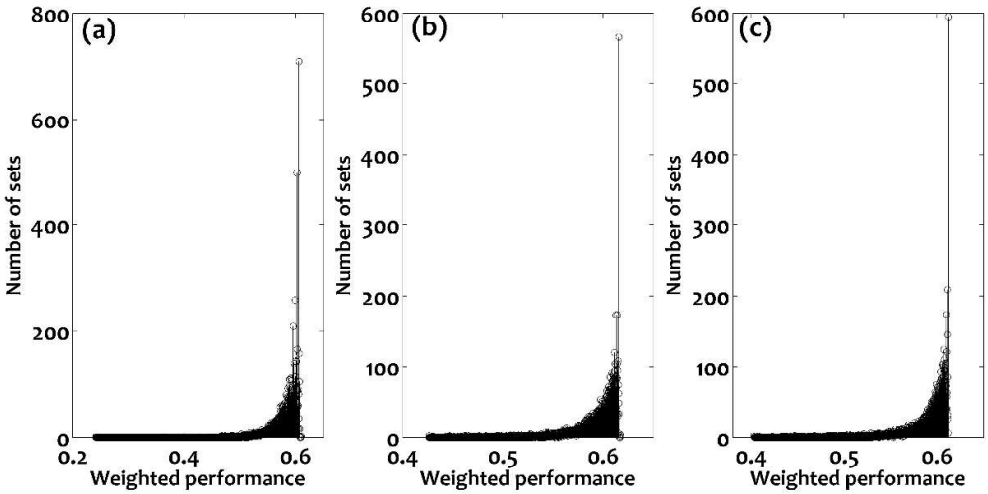

Figure 8.6 : Distribution of Weighted Performance of 10,000 randomly sets of size 10 sampled from all 2D features, for Abdomen related side effects, with increasing redundancy. (a) ≤60% (b) ≤70% and (c) ≤80%.

Table 8.6 : 2D features closely linked with 'Abdomen' related side effects, obtained using PCCA.

Chemical feature family	Class-linked features
Electrostatic property	ES Count aaN
Electrostatic property	ES Count ssNH2
Electrostatic property	ES Count sssP
Electrostatic property	ES Count ssssN
Electrostatic property	ES Sum dsN
Molecular property	QED MW
Molecular property	Num AromaticBonds
Molecular property	Num Chains
Molecular property	Num DoubleBonds

Molecular property	Num Fragments
Molecular property	Num H Donors
Molecular property	Num MesoStereoAtoms
Molecular property	Num RingAssemblies
Molecular surface area	Molecular FractionalPolarSurfaceArea

8.5 DISCUSSION

Herein, we present a generalized PCCA model for quantification of contribution from individual drug features towards side effects prediction while screening out interdependence of other features. The model is a combination of analytical and numerical strategies, and can be used to arrive at the most effective set of drug features starting from a range of available descriptors. The analytical treatment presented here, in conjunction with numerical strategy, can be used to identify most effective set of features to incorporate in a model that assumes orthogonality. With case studies of side effects presented in eye, nose and abdomen, we demonstrate the application of our model to identify chemical features that are specific to these class of drug reactions.

We started by considering 2D and 3D features of drugs compiled from SIDER4.1. After testing for their redundancy using Kendell Tau correlation, we rejected 3D features due to high redundancy. For each class of side effects, we retrieved the set of drugs with 60%, 70% and 80% redundancy. Randomly sampled features sets of 2D chemical properties were ranked with PCCA for each class of side effects at different level of redundancy (Figure 8.3). The most independent feature sets are expected to provide best performance.

For eye and nose classes, the performance of drugs with less redundancy (60%) was better compared with drugs with higher redundancy as expected theoretically. This pattern was not present for Abdomen, which could be attributed to its large drug-side effects matrix. To assess the impact of top sets that were used at different redundancy (Top 5 for 60% and 70%; Top 25 for 80%) for arriving at features that are linked to a class of side effects, we instead considered Top 10 sets for 60% and 70% redundancy and Top 50 sets for 80% redundancy. Interestingly, we observed that despite relaxing the stringency of performance the set of features arrived at, were again invariant. This points to the strength of the hypothesis used as a premise for the PCCA technique proposed in this study [Hall, 1999].

PCCA is evidently useful in drug discovery process as it helps in identification of most relevant drug features that potentially contribute to their adverse drug reactions. With the availability of large amount of data with an array of interdependent drug descriptors, this model is of value in drug discovery process as it enables in dealing with multidimensional drug features space. While we present our model in the context of drug features with applications to drug development, PCCA is of value to any real-life situation where one feature is jointly dependent on multiple other features simultaneously, and it is important to identify correlation among dependent features.

…