

## List of Figures

Figures	Title	page
3.1	The procedure implemented for integration of drugs, their targets and side effects. Figure adapted from [Sharma, 2015].	14
3.2	(a) Number of targets that are bound by $\geq d$ drugs. (b) Number of drugs that bind $\geq t$ targets.	15
3.3	(a) Number of side effects that are caused by per $\geq d$ drugs. (b) Number of drugs are associated with $\geq s$ side effects. This plots show that presence of certain number of drugs and side effects with more than certain number of side effects and drugs respectively.	16
3.4	Illustration of the Drug-Target-Side tri-partite relationships. The drug-target regulatory associations (DrugBank) and drug-side effect data (SIDER) was merged to create a composite dataset. This integrated dataset was used for investigations presented in Chapter 5 and Chapter 6.	16
3.5	(a) 2D and (b) 3D structure representation of Levobunolol. Images were obtained from PubChem.	18
4.1	Flowchart of Canonical correlation analysis.	25
4.2	Flowchart of Generalized partial canonical correlation analysis.	30
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.	34
5.2	AUC measured using the largest eigenvector 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.	35
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.	35
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.	37
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.	37
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.	39
5.7	The AUC matrices for pair wise combinations of (a) 3-dimensional and (b) 2-dimensional chemical properties. Starting from pair wise results a few parameters from each of the two categories were obtained that were critical for side effects prediction.	39
6.1	Symbolic depiction of how drugs interact with cellular mechanisms leading to an array of reactions. Cell's pathways are shown as a network of molecular interactions. Starting with a therapeutic target, rational drug discovery aims at identification of chemicals that are specific regulators of such targets. In reality, drugs interact with cellular mechanisms in an unintended manner giving rise to adverse reactions. In this study, we created Genomic Space representing molecular network in cell, to assess the value of information derived from various graph theoretical features towards prediction of side effects.	43
6.2	Illustration of Genomic Space, and drug target interactions. The primary targets and their secondary as well as higher order interactions with the cellular milieu are depicted.	44
6.3	Statistics of 'number of target pairs' in the Genomic Space and number of shared drugs.	45
6.4	Relation between 'number of drugs shared' by a pair of targets and their 'sequence similarity'. The sequence similarity cut-off used for creating the genomic space (0.35) is indicated with an arrow.	45
6.5	(a) AUC and (b) AUPR measured using the largest eigenvector. The error bars indicate standard error of data from 10-fold cross-validation.	48

6.6	(a) MCC (b) F1 score and (c) Accuracy measured using the largest eigenvector. The error bars indicate standard error of data from 10-fold cross validation.	48
7.1	Schematic depiction of concepts used and objective of this study. The notion of ‘Known Side Effects’ as well as ‘Unknown Side Effects’ of a drug profile is graphically illustrated. Data structure of drug-side effects matrix and dimensions of the matrix for each organ/system investigated in this study are also depicted.	49
7.2	‘Blood’ Class: Distribution of Weighted Performance with 30% of known side effects for varying size of randomly chosen side effects sets. (a) 500 (b) 1000 and (c) 1500.	54
7.3	‘Blood’ Class: Distribution of Weighted Performance with 40% of known side effects for varying size of randomly chosen side effects sets. (a) 500 (b) 1000 and (c) 1500.	55
7.4	‘Blood’ Class: Distribution of Weighted Performance with 50% of known side effects for varying size of randomly chosen side effects sets. (a) 500 (b) 1000 and (c) 1500.	55
7.5	‘Eye’ class: Relation between weighted performance and their corresponding number of sets with different proportion of known side effects. (a) 30% (b) 40% and (c) 50%.	56
7.6	‘Abdominal’ class: Relation between weighted performance and their corresponding number of sets with different proportion of known side effects. (a) 30% (b) 40% and (c) 50%.	60
7.7	‘Hepatic’ class: Relation between weighted performance and their corresponding number of sets with different proportion of known side effects. (a) 30% (b) 40% and (c) 50%.	61
7.8	‘Gastrointestinal’ class: Relation between weighted performance and their corresponding number of sets with different proportion of known side effects. (a) 30% (b) 40% and (c) 50%.	63
7.9	‘Kidney’ class: Relation between weighted performance and their corresponding number of sets with different proportion of known side effects. (a) 30% (b) 40% and (c) 50%.	65
7.10	‘Nose’ class: Relation between weighted performance and their corresponding number of sets with different proportion of known side effects. (a) 30% (b) 40% and (c) 50%.	67
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.	71
8.2	Flowchart of Partial canonical correlation analysis for three parameter system.	74
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.	76
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) $\leq 60\%$ (b) $\leq 70\%$ and (c) $\leq 80\%$ .	77
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) $\leq 60\%$ (b) $\leq 70\%$ and (c) $\leq 80\%$ .	78
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) $\leq 60\%$ (b) $\leq 70\%$ and (c) $\leq 80\%$ .	79
B.1	Identification of 2D features linked to Eye-related side effects class that were consistently found among the top sets at different cut-offs of redundancy using PCCA. Sets with Top 5 best performance were shortlisted for redundancy cut-off of $\leq 0.6$ and $\leq 0.7$ , whereas those with Top25 best performances were used for the cut-off of $\leq 0.8$ . This yielded a total of 5 features consistently associated with Eye-related side effects (See Table 8.4).	85
B.2	Identification of 2D features linked to Nose-related side effects class that were consistently found among the top sets at different cut-offs of redundancy using PCCA.	87

Top 5 sets with best performance were shortlisted for redundancy cut-off of  $\leq 0.6$  and  $\leq 0.7$ , whereas Top25 sets were used for the cut-off of  $\leq 0.8$ . This yielded a total of 4 features consistently associated with Nose-related side effects (See Table 8.5).

- B.3 Identification of 2D features linked to Abdomen-related side effects class that were consistently found among the top sets at different cut-offs of redundancy using PCCA. Top 5 sets with best performance were shortlisted for redundancy cut-off of  $\leq 0.6$  and  $\leq 0.7$ , whereas Top 25 sets were used for the cut-off of  $\leq 0.8$ . This yielded a total of 14 features consistently associated with Abdomen-related side effects (See Table 8.6).

89

