7
Predicting hidden drug reactions using partial information of side effects classes

The side effects profile of a drug represents comprehensive information of its phenotypic responses. It is often not possible to acquire the complete side effects profile due to limitations on protocols. Obtaining the full profile is of key value in drug discovery process. While a range of features, from chemical substructures, chemical properties targets to interactome properties are reported to be of value for prediction of side effects, we probed the possibility of using partial information of side effects for predicting the rest of the (unknown) side effects. Thus we propose that partial side effects profile as a possible descriptor of the remaining side effects. Our proposal is based on the idea that phenotypic effects of drugs are interdependent. To probe our proposition, we segregated drugs presenting with side effects specific to an organ/system and used canonical correlation analysis to predict 'unknown' side effects starting with the partial knowledge of side effects. The notion of known and unknown side effects as well as details of each organ class studied are illustrated in Figure 7.1.

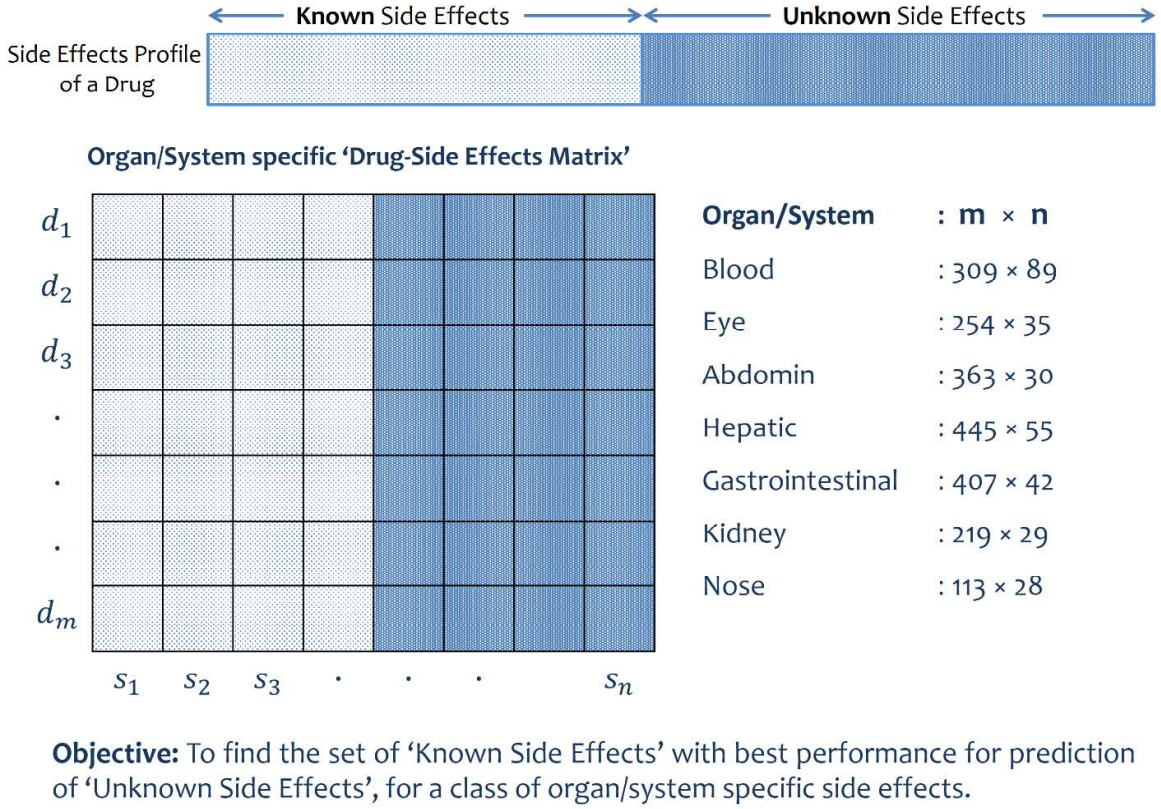

Figure 7.1 : Schematic depiction of concepts used and objective of this study. The notion of 'Known Side Effects' as well as 'Unknown Side Effects' of a drug profile is graphically illustrated. Data structure of drug-side effects matrix and dimensions of the matrix for each organ/system investigated in this study are also depicted.

7.1 ORGAN/SYSTEM-BASED CLASSES OF SIDE EFFECTS

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 around 97, the drug with worst adverse reactions profile was reported to have 757 side effects.

Starting with the data of drug side effects from SIDER4, we segregated drugs based on the organ/system to which they are linked through the side effects they cause. We identified side effects classes associated with various organs/systems such that each of these classes has more than 200 drugs (with an exception of Nose-related side effects class), and each drug is reported to present at least two side effects of the class (to reduce noise). The seven side effects classes for which studies were performed are as follow: Blood, Eye, Abdomen, Hepatic, Gastrointestinal, Kidney and Nose. For the purpose analysis we created the 'drug-side effects matrix' for each class. The list of side effects belonging to each class are presented in Table 7.1.

Table 7.1 : List of side effects for each of the side effects class linked to an organ/system.

Blood	'Blood albumin decreased' (1), 'Blood albumin increased' (2), 'Blood alkaline phosphatase increased' (3), 'Blood amylase increased' (4), 'Blood and lymphatic system disorders' (5), 'Blood bicarbonate decreased' (6), 'Blood bicarbonate increased' (7), 'Blood bilirubin abnormal' (8), 'Blood bilirubin decreased' (9), 'Blood bilirubin increased' (10), 'Blood bilirubin unconjugated increased' (11), 'Blood calcium abnormal' (12), 'Blood calcium decreased' (13), 'Blood chloride decreased' (14), 'Blood chloride increased' (15), 'Blood cholesterol abnormal' (16), 'Blood cholesterol decreased' (17), 'Blood cholesterol increased' (18), 'Blood cortisol decreased' (19), 'Blood cortisol increased' (20), 'Blood count abnormal' (21), 'Blood creatine increased' (22), 'Blood creatine phosphokinase MB increased' (23), 'Blood creatine phosphokinase MM increased' (24), 'Blood creatine phosphokinase abnormal' (25), 'Blood creatine phosphokinase increased' (26), 'Blood creatinine abnormal' (27), 'Blood creatinine increased' (28), 'Blood culture positive' (29), 'Blood disorder' (30), 'Blood electrolytes abnormal' (31), 'Blood follicle stimulating hormone increased' (32), 'Blood gases abnormal' (33), 'Blood gastrin increased' (34), 'Blood glucose abnormal' (35), 'Blood glucose decreased' (36), 'Blood glucose fluctuation' (37), 'Blood glucose increased' (38), 'Blood gonadotrophin increased' (39), 'Blood in stool' (40), 'Blood insulin decreased' (41), 'Blood insulin increased' (42), 'Blood iron decreased' (43), 'Blood iron increased' (44), 'Blood lactate dehydrogenase increased' (45), 'Blood luteinising hormone increased' (46), 'Blood magnesium decreased' (47), 'Blood methaemoglobin present' (48), 'Blood monocytes increased' (49), 'Blood neutrophil count decreased' (50), 'Blood oestrogen decreased' (51), 'Blood oestrogen increased' (52), 'Blood osmolarity decreased' (53), 'Blood pancreatic amylase increased' (54), 'Blood parathyroid hormone increased' (55), 'Blood phosphate increased' (56), 'Blood phosphorus decreased' (57), 'Blood phosphorus increased' (58), 'Blood potassium abnormal' (59), 'Blood potassium decreased' (60), 'Blood potassium increased' (61), 'Blood pressure abnormal' (62), 'Blood pressure diastolic decreased' (63), 'Blood pressure diastolic increased' (64), 'Blood pressure fluctuation' (65), 'Blood pressure inadequately controlled' (66), 'Blood pressure increased' (67), 'Blood pressure normal' (68), 'Blood pressure systolic decreased' (69), 'Blood pressure

	systolic increased' (70), 'Blood prolactin decreased' (71), 'Blood prolactin increased' (72), 'Blood sodium increased' (73), 'Blood test abnormal' (74), 'Blood testosterone decreased' (75), 'Blood testosterone free increased' (76), 'Blood testosterone increased' (77), 'Blood thyroid stimulating hormone increased' (78), 'Blood triglycerides decreased' (79), 'Blood triglycerides increased' (80), 'Blood urea abnormal' (81), 'Blood urea increased' (82), 'Blood urea normal' (83), 'Blood uric acid decreased' (84), 'Blood uric acid increased' (85), 'Blood urine' (86), 'Blood urine present' (87), 'Bloodshot eye' (88), 'Bloody discharge' (89)
Eye	'Eye allergy' (1), 'Eye bleeding' (2), 'Eye burns' (3), 'Eye discharge' (4), 'Eye disorder' (5), 'Eye haemorrhage' (6), 'Eye infection' (7), 'Eye infection viral' (8), 'Eye injury' (9), 'Eye irritation' (10), 'Eye movement disorder' (11), 'Eye oedema' (12), 'Eye pain' (13), 'Eye penetration' (14), 'Eye pruritus' (15), 'Eye redness' (16), 'Eye rolling' (17), 'Eye swelling' (18), 'Eye ulcer' (19), 'Eyebrow loss of' (20), 'Eyelash darkening' (21), 'Eyelash discolouration' (22), 'Eyelash thickening' (23), 'Eyelid bleeding' (24), 'Eyelid disorder' (25), 'Eyelid exfoliation' (26), 'Eyelid function disorder' (27), 'Eyelid infection' (28), 'Eyelid irritation' (29), 'Eyelid margin crusting' (30), 'Eyelid oedema' (31), 'Eyelid pain' (32), 'Eyelid ptosis' (33), 'Eyelid retraction' (34), 'Eyelids pruritus' (35)
Abdominal	'Abdominal abscess' (1), 'Abdominal adhesions' (2), 'Abdominal aortic aneurysm' (3), 'Abdominal aortic bruit' (4), 'Abdominal aortic rupture' (5), 'Abdominal bloating' (6), 'Abdominal colic' (7), 'Abdominal cramps' (8), 'Abdominal discomfort' (9), 'Abdominal disorder' (10), 'Abdominal distension' (11), 'Abdominal distension gaseous' (12), 'Abdominal distress' (13), 'Abdominal hernia' (14), 'Abdominal infection' (15), 'Abdominal injury' (16), 'Abdominal neoplasm' (17), 'Abdominal pain' (18), 'Abdominal pain aggravated' (19), 'Abdominal pain generalised' (20), 'Abdominal pain localised' (21), 'Abdominal pain lower' (22), 'Abdominal pain upper' (23), 'Abdominal rigidity' (24), 'Abdominal sepsis' (25), 'Abdominal spasm' (26), 'Abdominal strangulated hernia' (27), 'Abdominal symptom' (28), 'Abdominal tenderness' (29), 'Abdominal wall abscess' (30)
Hepatic	'Hepatic adenoma' (1), 'Hepatic angiosarcoma' (2), 'Hepatic artery thrombosis' (3), 'Hepatic atrophy' (4), 'Hepatic cancer' (5), 'Hepatic cirrhosis' (6), 'Hepatic congestion' (7), 'Hepatic cyst' (8), 'Hepatic encephalopathy' (9), 'Hepatic enzyme abnormal' (10), 'Hepatic enzyme decreased' (11), 'Hepatic enzyme increased' (12), 'Hepatic failure' (13), 'Hepatic fibrosis' (14), 'Hepatic function abnormal' (15), 'Hepatic granuloma' (16), 'Hepatic haemorrhage' (17), 'Hepatic impairment' (18), 'Hepatic infiltration eosinophilic' (19), 'Hepatic insufficiency' (20), 'Hepatic lymphoma' (21), 'Hepatic necrosis' (22), 'Hepatic neoplasm' (23), 'Hepatic pain' (24), 'Hepatic steatosis' (25), 'Hepatic vein thrombosis' (26), 'Hepatitis' (27), 'Hepatitis A' (28), 'Hepatitis B' (29), 'Hepatitis B e antigen positive' (30), 'Hepatitis B reactivation' (31), 'Hepatitis B surface antigen positive' (32), 'Hepatitis C' (33), 'Hepatitis C antibody positive' (34), 'Hepatitis F' (35), 'Hepatitis acute' (36), 'Hepatitis aggravated' (37), 'Hepatitis alcoholic' (38), 'Hepatitis cholestatic' (39), 'Hepatitis chronic active' (40), 'Hepatitis fulminant' (41), 'Hepatitis granulomatous' (42), 'Hepatitis symptom' (43), 'Hepatitis toxic'

	(44), 'Hepatitis viral' (45), 'Hepatobiliary disease' (46), 'Hepatoblastoma' (47), 'Hepatocellular adenoma' (48), 'Hepatocellular carcinoma' (49), 'Hepatocellular injury' (50), 'Hepatomegaly' (51), 'Hepatorenal syndrome' (52), 'Hepatosplenic T-cell lymphoma' (53), 'Hepatosplenomegaly' (54), 'Hepatotoxicity' (55)
Gastrointstinal	'Gastroduodenitis' (1), 'Gastroenteritis' (2), 'Gastroenteritis adenovirus' (3), 'Gastroenteritis bacterial' (4), 'Gastroenteritis salmonella' (5), 'Gastroenteritis yersinia' (6), 'Gastrointestinal candidiasis' (7), 'Gastrointestinal carcinoma' (8), 'Gastrointestinal cramps' (9), 'Gastrointestinal discomfort' (10), 'Gastrointestinal disorder' (11), 'Gastrointestinal fistula' (12), 'Gastrointestinal fullness' (13), 'Gastrointestinal fungal infection' (14), 'Gastrointestinal haemorrhage' (15), 'Gastrointestinal hypermotility' (16), 'Gastrointestinal hypomotility' (17), 'Gastrointestinal infection' (18), 'Gastrointestinal inflammation' (19), 'Gastrointestinal inflammatory conditions' (20), 'Gastrointestinal malformation' (21), 'Gastrointestinal motility disorder' (22), 'Gastrointestinal necrosis' (23), 'Gastrointestinal neoplasm' (24), 'Gastrointestinal neoplasm malignant' (25), 'Gastrointestinal obstruction' (26), 'Gastrointestinal pain' (27), 'Gastrointestinal perforation' (28), 'Gastrointestinal signs and symptoms' (29), 'Gastrointestinal signs and symptoms NEC' (30), 'Gastrointestinal sounds abnormal' (31), 'Gastrointestinal stoma complication' (32), 'Gastrointestinal stromal tumour' (33), 'Gastrointestinal symptom NOS' (34), 'Gastrointestinal toxicity' (35), 'Gastrointestinal tract irritation' (36), 'Gastrointestinal ulcer' (37), 'Gastrointestinal ulcer haemorrhage' (38), 'Gastrointestinal ulcer perforation' (39), 'Gastrooesophageal reflux disease' (40), 'Gastrooesophageal sphincter insufficiency' (41), 'Gastrooesophagitis' (42)
Renal/Kidney	'Renal aplasia' (1), 'Renal artery stenosis' (2), 'Renal artery thrombosis' (3), 'Renal cancer' (4), 'Renal cell carcinoma' (5), 'Renal colic' (6), 'Renal cortical necrosis' (7), 'Renal cyst' (8), 'Renal dysgenesis' (9), 'Renal failure' (10), 'Renal failure acute' (11), 'Renal failure aggravated' (12), 'Renal failure chronic' (13), 'Renal function test abnormal' (14), 'Renal graft loss' (15), 'Renal haemorrhage' (16), 'Renal impairment' (17), 'Renal infarct' (18), 'Renal insufficiency aggravated' (19), 'Renal mass' (20), 'Renal neoplasm' (21), 'Renal pain' (22), 'Renal papillary necrosis' (23), 'Renal rickets' (24), 'Renal tubular acidosis' (25), 'Renal tubular disorder' (26), 'Renal tubular necrosis' (27), 'Renal vasculitis' (28), 'Renal vein thrombosis' (29)
Nasal/Nose	'Nasal discomfort' (1), 'Nasal disorder' (2), 'Nasal dryness' (3), 'Nasal inflammation' (4), 'Nasal irritation' (5), 'Nasal itching' (6), 'Nasal mucosal disorder' (7), 'Nasal mucosal erythema' (8), 'Nasal mucus blood tinged' (9), 'Nasal obstruction' (10), 'Nasal odour' (11), 'Nasal oedema' (12), 'Nasal pain' (13), 'Nasal passage irritation' (14), 'Nasal polyps' (15), 'Nasal septum deviation' (16), 'Nasal septum disorder' (17), 'Nasal septum perforation' (18), 'Nasal septum ulceration' (19), 'Nasal sinus drainage' (20), 'Nasal soreness' (21), 'Nasal ulcer' (22), 'Nasopharyngeal carcinoma' (23), 'Nasopharyngeal disorder' (24), 'Nasopharyngitis' (25), 'Nose infection NOS' (26), 'Nose infection NOS' (27), 'Nosocomial pneumonia' (28)

7.2 IDENTIFICATION OF SET OF CRUCIAL 'KNOWN SIDE EFFECTS'

Our proposition is that side effects of drugs are interdependent, and hence with the knowledge of the 'Known Side Effects', the remainder of the side effects could be predicted using canonical correlation analysis. Having grouped the side effects based on their presentation in an organ/system as described in Section 7.1, the task is to find out the best set of 'Known Side Effects' for each class. Finding such a best set in general would require exhaustive search for all possible fraction of known side effects, making the problem intractable. Hence, we reduced the problem to three cases such that percentage of known side effects is either 30%, 40% or 50%. The fraction of known side effects was also chosen in this manner, so that knowing smaller subset remaining side effects could be predicted. Further for a given class and a chosen percentage of known side effects in that class, we sampled large number of sets of side effects and computed the prediction performance for each set. We defined the 'weighted performance' for every set as per Eq. (7.1) to create a balanced metric.

$$\textbf{Weighted Performance} = 0.2 \times AUC + 0.4 \times F1_score + 0.4 \times MCC \tag{7.1}$$

For each percentage of know side effects (30%, 40%, 50%), from 1000 randomly sampled sets of side effects, we obtained three sets with best 'Weighted Performance'. Using a total of nine sets of known side effects (three each from 30%, 40% and 50%), every side effect that occurs in these nine sets was ranked (κ_i) as per Eq. (7.2).

$$\kappa_i = \sum_{car=30.40,50} w_{car} \sum_{j=1}^{3} f(\phi_{car}^{j}) \tag{7.2}$$

Here, ϕ_j^{car} is a randomly chosen set with cardinality 'car' and j represents its rank among in terms of its Prediction Performance. Top three sets from each category were considered while ranking the sets. Eq. (7.3) and Eq. (7.4) represent weight of each category and occurrence of side effect i, respectively.

$$w_{car} = \frac{100 - car}{\sum_{car=30,40,50} car} \tag{7.3}$$

$$f(\phi_{car}) = \begin{cases} 1 & if \quad i \in \phi_{car} \\ 0 & if \quad i \notin \phi_{car} \end{cases} \tag{7.4}$$

In Table 7.2– 7.16, groups of known and missing side effects with prediction performance have been shown with the indexes as given by respective organs related side effects in Table 7.1.

7.2.1 Justification for sample size

Before we present the results for each class of side effects, using the side effects class of 'Blood' (the class with largest number of side effects, among the organ/system based classes considered for our study), we justify that set size of 1000 for random sampling is reasonable. For 30%, 40% and 50% known side effects of Blood class, we randomly sampled 500, 1000 and 1500 side effects and compared the mean as well as standard deviation of Weighted Performances (Table 7.2).

Table 7.2 : 'Blood' Class: Mean and standard deviation of Weighted Performance corresponding to different sample size and proportion of known side effects.

	30% known side effects		40% known side effects		50% known side effects	
Sample size	**Mean**	**Standard deviation**	**Mean**	**Standard deviation**	**Mean**	**Standard deviation**
500	0.31	0.11	0.35	0.09	0.39	0.09
1000	0.30	0.10	0.36	0.09	0.39	0.09
1500	0.31	0.10	0.35	0.09	0.39	0.09

Figure 7.2—7.4 depict the data of Weighted Performance for each case. As evident from these figures as well as statistics presented in Table 7.2, regardless of the size of sampling the Weighted Performance does not vary significantly. Also, we find that while the best Weighted Performance obtained improves with increased sampling size (0.58 for 500 sampling size, 0.64 for 1000 sampling size and 0.64 for 1500 sampling size), with sample size of 1000 we could achieve reasonably good performance.

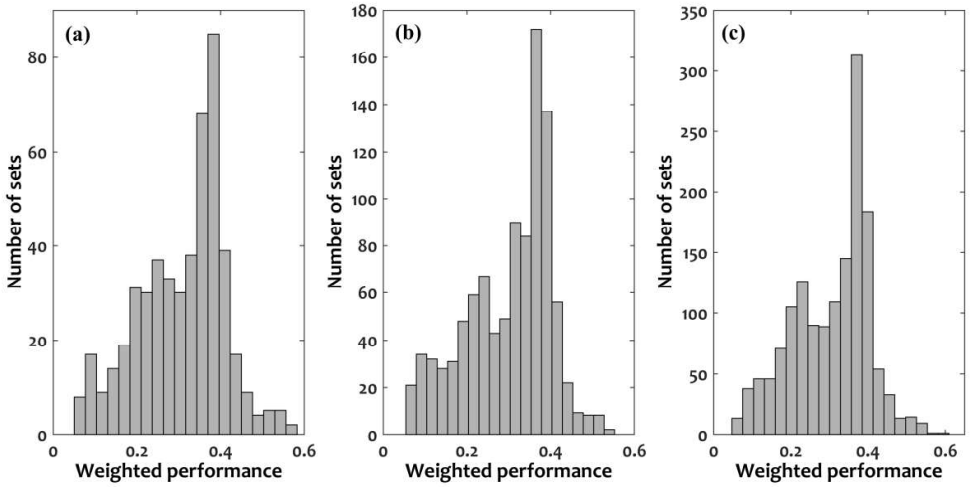

Figure 7.2 : 'Blood' Class: Distribution of Weighted Performance with 30% of known side effects for varying size of randomly chosen side effects sets. (a) 500 (b) 1000 and (c) 1500.

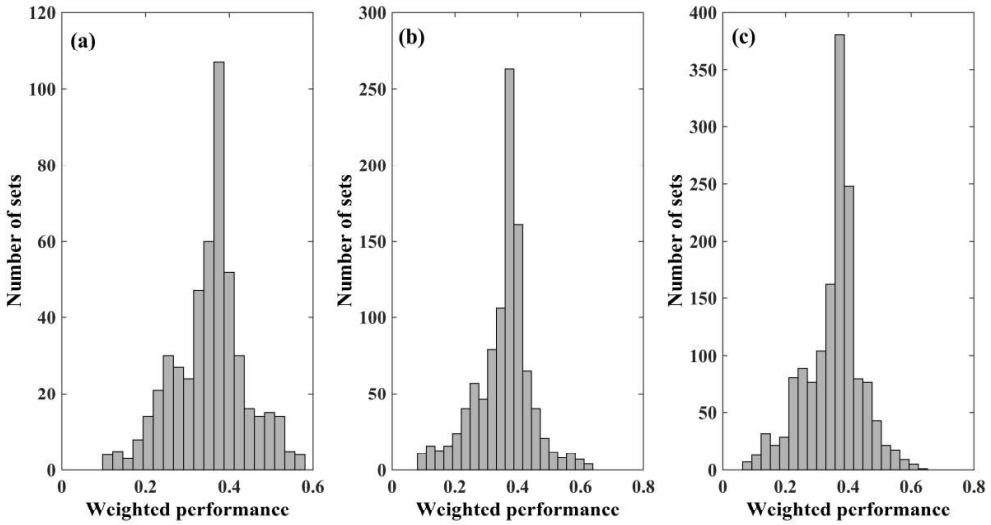

Figure 7.3 : 'Blood' Class: Distribution of Weighted Performance with 40% of known side effects for varying size of randomly chosen side effects sets. (a) 500 (b) 1000 and (c) 1500.

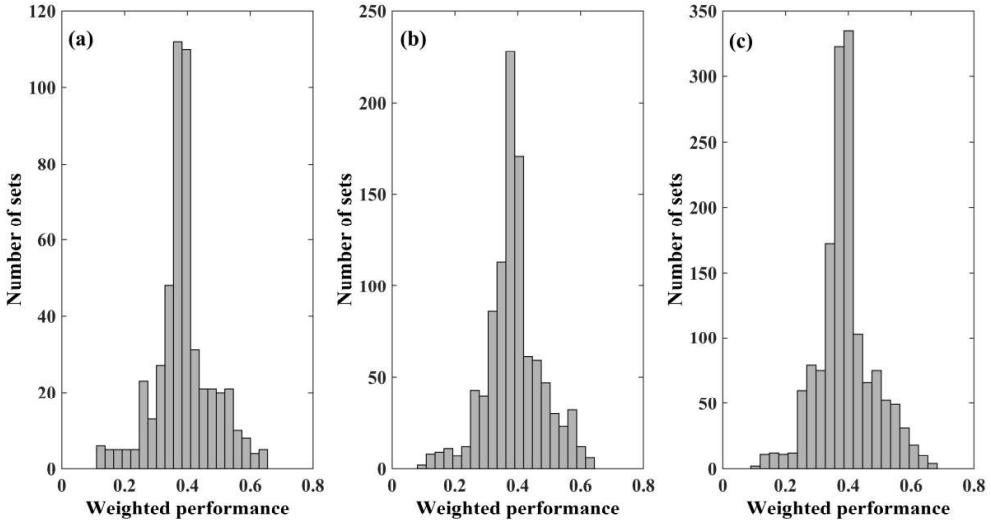

Figure 7.4 : 'Blood' Class: Distribution of Weighted Performance with 50% of known side effects for varying size of randomly chosen side effects sets. (a) 500 (b) 1000 and (c) 1500.

Further we obtained the 'overall best set of known side effects' for different sampling sizes, using the protocol described in Section 7.2. We find that the results are comparable with many shared side effects such as 'Blood bilirubin increased' (10), 'Blood creatinine increased' (28), 'Blood disorder' (30), 'Blood pressure increased' (68), 'Blood cholesterol abnormal' (16), 'Blood culture positive' (29), 'Blood glucose increased' (38), and 'Blood urea increased' (84). Thus, we infer that random sampling of side effect sets of size 1000 yield good result. Below, we present results for each of the side effects classes with 'sample size of 1000'.

7.3 PREDICTING HIDDEN SIDE EFFECTS

In subsequent sections, we have presented our studies with different side effects classes (Blood, Eye, Abdominal, Hepatic, Gastrointestinal, Kidney and Nose) using the above-described protocol, to identify best set of 'known side effects' so as to predict the remainder of side effects.

7.3.1 Blood

Blood carries life-sustaining substances like nutrients and oxygen to cells and also transports metabolic waste out of cells. It is composed many key components (such as albumin, bilirubin, cholesterol) whose imbalance could lead to adverse reactions. For example, Cytrabine, a drug used in chemotherapy for treatment of acute lymphocytic leukemia [Jabbour et al., 2007], is known to cause side effects such as increase in bilirubin, increase in uric acid, lymphatic system disorder etc. We compiled the drug-side effects matrix for Blood-related adverse reactions using key word 'Blood'. 309 drugs associated with 89 relevant side effects were obtained. The list of these Blood-related side effects and drugs is provided in Table 7.1. The matrix had 4% sparseness, with an average of 3.8 side effects per drug and maximally having 26 adverse reactions. 10-fold cross-validation was performed using 300 drugs after removing 9 drugs. The training set and test set consisted of 270 and 30 drugs, respectively.

We obtained ranked list of known side effects using the protocol described in Section 7.2. Here, we computed the Prediction Performance of 1000 randomly sampled sets starting with the prior knowledge of 30%, 40% or 50% of side effects (Figure 7.2(b), Figure 7.3(b) and Figure 7.4(b)). As expected we find marginal improvement in performance with increasing prior knowledge of side effects. Table 7.3 lists top 3 sets with best performance, which were further used for obtaining a ranked list of side effects. The final ranked list represents 27 (30%) overall best Eye related side effects (Table 7.4). Our method suggests that for any drug, prior knowledge of these side effects allows us to predict the rest of the side effects with high confidence. We find that the prediction performance with these side effects is (0.6365) better than the best of the randomly chosen 27 side effects (0.5528).

Table 7.3 : 'Blood' Class: Three best groups with various proportion of known side effects and their performance.

Proportion of given side effects	Known side effects profile	Performance
30%	3, 6, 9, 10, 11, 16, 19, 20, 21, 26, 28, 30, 32, 40, 44, 48, 50, 51, 61, 63, 67, 71, 72, 75, 77, 81, 87	0.5528
	1, 3, 8, 10, 16, 20, 21, 24, 28, 29, 30, 37, 38, 41, 42, 47, 50, 53, 55, 63, 67, 70, 76, 79, 84, 87, 88	0.5303
	1, 7, 8, 10, 12, 15, 17, 19, 26, 28, 30, 43, 44, 46, 47, 56, 58, 60, 61, 62, 63, 67, 68, 86, 87, 88, 89	0.5278
40%	3, 8, 10, 17, 18, 21, 22, 27, 28, 30, 34, 37, 38, 40, 42, 44, 47, 48, 52, 58, 60, 63, 66, 67, 68, 70, 71, 72, 75, 76, 78, 80, 82, 83, 85, 87	0.6390
	4, 7, 9, 10, 13, 17, 19, 20, 23, 26, 28, 29, 30, 35, 37, 38, 40, 44, 49, 56, 57, 58, 59, 61, 63, 66, 67, 71, 75, 78, 80, 81, 82, 84, 87, 89	0.6303
	2, 3, 4, 10, 11, 13, 16, 17, 18, 26, 27, 28, 29, 30, 32, 34, 35, 41, 48, 49, 50, 52, 54, 57, 59, 66, 67, 68, 72, 74, 76, 78, 80, 82, 87, 89	0.6229

50%	2, 4, 7, 9, 10, 11, 12, 16, 18, 19, 20, 21, 22, 26, 27, 28, 30, 31, 33, 34, 36, 39, 43, 45, 46, 47, 48, 49, 54, 55, 58, 60, 63, 67, 69, 71, 74, 77, 78, 79, 80, 82, 83, 86, 87	0.6440
	1, 2, 4, 6, 8, 10, 12, 13, 14, 16, 18, 20, 23, 24, 25, 27, 28, 29, 30, 36, 38, 39, 40, 41, 43, 45, 52, 53, 55, 56, 57, 58, 62, 65, 67, 68, 69, 70, 72, 75, 78, 79, 82, 83, 85,89	0.6360
	1, 3, 6, 7, 8, 10, 12, 13, 15, 17, 18, 20, 21, 24, 27, 28, 29, 30, 32, 33, 36, 38, 39, 40, 42, 45, 46, 47, 50, 51, 55, 58, 66, 67, 68, 71, 72, 73, 74, 76, 77, 79, 82, 84, 89	0.6352

Table 7.4 : 'Blood' Class: A ranked list of critical side effects obtained with sample size 1000.

Rank	Side effects	Rank	Side effects	Rank	Side effects
1	'Blood bilirubin increased' (10)	10	'Bloody discharge' (89)	19	'Blood culture positive' (29)
2	'Blood creatinine increased' (28)	11	'Blood alkaline phosphatase increased' (3)	20	'Blood glucose increased' (38)
3	'Blood disorder' (30)	12	'Blood bilirubin abnormal' (8)	21	'Blood in stool' (40)
4	'Blood pressure increased' (67)	13	'Blood cholesterol abnormal' (16)	22	'Blood magnesium decreased' (47)
5	'Blood urea increased' (82)	14	'Blood cholesterol decreased' (17)	23	'Blood pressure normal' (68)
6	'Blood urine present' (87)	15	'Blood cholesterol increased' (18)	24	'Blood prolactin decreased' (71)
7	'Blood cortisol increased' (20)	16	'Blood count abnormal' (21)	25	'Blood prolactin increased' (72)
8	'Blood phosphorus increased' (58)	17	'Blood creatine phosphokinase increased' (26)	26	'Blood thyroid stimulating hormone increased' (78)
9	'Blood pressure diastolic decreased' (63)	18	'Blood creatinine abnormal' (27)	27	'Blood albumin decreased' (1)

7.3.2 Eye

Eye is an important organ responsible for critical ability for vision. Some of the drugs cause eye related side effects. For example Chloramphenicol, which is used against Cholera and Meningitis, is known to cause various eye problems such as irritation, dryness, and infection etc. It would be valuable if knowing some of these side effects, one could reliably ascertain other eye-specific reactions. The subset of drugs with 'Eye' related adverse reactions was obtained by filtering for side effects that contained word 'Eye' in them. This protocol yielded 244 drugs linked to 35 side effects. Complete list of these Eye related side effects is provided in Table 7.1. The drug-side effects matrix had sparseness of 9.8%. The average number of side effects was 3.4 and the maximum was 15. To perform 10-fold cross-validation, we eliminated 4 drugs to obtain a final matrix of 240 drugs which was used for analysis. The training and test set consisted of 216 and 24 drugs, respectively.

Using the protocol described in Section 7.2, using the prior knowledge of 30%, 40% or 50% of side effects, we computed the Prediction Performance of 1000 random sampled sets (Figure 7.5). As observed in the figure, we find marginal improvement in performance with increasing prior knowledge (from 30% to 50%) of side effects. Table 7.5 lists top 3 sets with best performance, which were further used for obtaining a ranked list of side effects. The final ranked list represents the 11 overall best Eye related side effects (Table 7.6). Our methods suggest that for any drug, prior knowledge of these side effects allows us to predict the rest of the side effects with high confidence. We find that the prediction performance with these side effects is (0.7035) better than the best of the randomly chosen 11 (30%) side effects (0.6752).

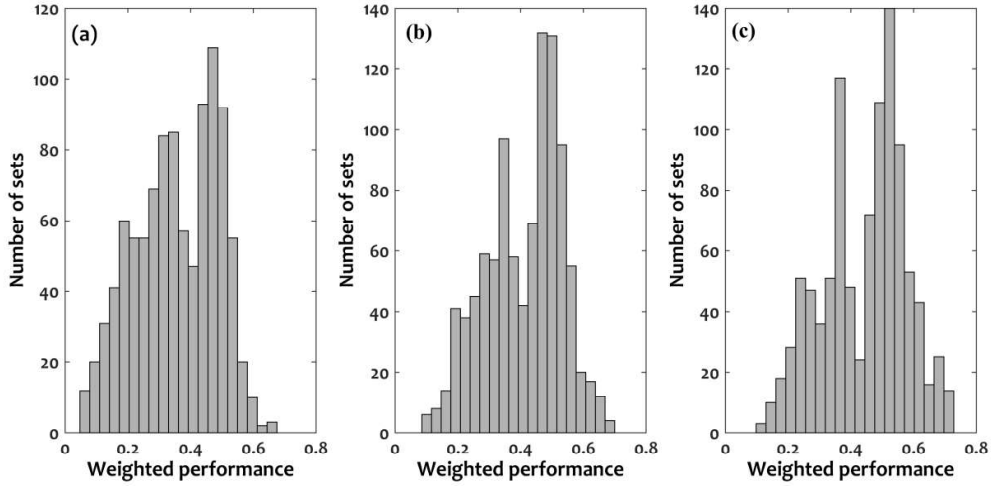

Figure 7.5 : 'Eye' class: Relation between weighted performance and their corresponding number of sets with different proportion of known side effects. (a) 30% (b) 40% and (c) 50%.

Table 7.5 : 'Eye' Class: Three best groups with various proportion of known side effects and their performance.

Fraction of known side effects	Known side effects profile	Performance
30%	4, 6, 10, 13, 16, 18, 22, 25, 31, 33, 34	0.6752
	4, 6, 10, 12, 13, 18, 27, 28, 30, 31, 34	0.6668
	1, 2, 6, 10, 11, 13, 14, 18, 24, 31, 33	0.6524

40%	1, 3, 6, 7, 10, 13, 15, 17, 18, 20, 22, 26, 27, 31	0.6999
	6, 7, 10, 12, 13, 15, 17, 19, 20, 21, 23, 25, 31, 34	0.6864
	3, 6, 7, 10, 13, 17, 18, 19, 24, 25, 27, 31, 33, 34	0.6800
50%	3, 4, 6, 8, 9, 10, 11, 12, 13, 14, 15, 18, 23, 27, 30, 31, 33, 34	0.7281
	4, 6, 8, 10, 11, 13, 15, 16, 18, 20, 23, 25, 27, 29, 30, 31, 32, 33	0.7244
	1, 2, 4, 6, 7, 8, 9, 10, 13, 14, 15, 16, 18, 27, 29, 31, 34, 35	0.7200

Table 7.6 : 'Eye' Class: A ranked list of critical side effects obtained with sample size 1000.

Rank	Side effect	Rank	Side effect	Rank	Side effect
1	'Eye haemorrhage' (6)	6	'Eyelid function disorder' (27)	11	'Eyelid disorder' (25)
2	'Eye irritation' (10)	7	'Eyelid retraction' (34)		
3	'Eye pain' (13)	8	'Eye discharge' (4)		
4	'Eyelid oedema' (31)	9	'Eye pruritus' (15)		
5	'Eye swelling' (18)	10	'Eye infection' (7)		

7.3.3 Abdominal

Abdomen is known to be associated with various adverse reactions such as abdominal pain, abdominal cramp, abdominal blotting. Lamivudine, used against HIV-1, is known to cause many abdomen related side effects. Similarly, Metronidazole, used for treatment of pelvic inflammatory disease, shows a range of abdominal related side effects [Ursing and Kamme, 1975; Zapolanski, 1980]. Few of the most common abdomen-associated side effects are 'Abdominal aortic bruit', 'Abdominal distension', 'Abdominal injury'. The drug-side effects matrix for Abdomen-related adverse reactions was compiled using keyword 'Abdominal'. It yielded 30 side effects and 363 associated drugs. List of these Abdomen-related side effects is provided in Table 7.1. The matrix had 9.7% sparse-ness, with drugs having an average of 2.9 and maximally 7 adverse reactions. 10-fold cross-validation was performed using 360 drugs after removing 4 drugs. The training set and test set consisted of 324 and 36 drugs, respectively.

Table 7.7 lists the best of 1000 randomly sampled sets of (30%, 40% or 50%) side effects, knowing which one could predict the remainder of side effects. Figure 7.6 depicts statistics of Prediction Performance of these random sets. As expected, increasing the fraction of known side effects improves the prediction performance, along with enhanced performance among the best sets. Further, by ranking side effects thus obtained in their relevance for prediction, we identified nine best side effects (Table 7.8) with a prediction performance (0.8763) better than what is expected from best of the randomly sampled sets of same size (0.8502; 30%).

Table 7.7 : 'Abdominal' Class: Three best groups with various proportion of known side effects and their performance.

Fraction of known side effects	Known side effects profile	Performance
30%	4, 6, 8, 9, 10, 11, 20, 23, 27	0.8502
	1, 6, 8, 9, 11, 14, 23, 25, 26	0.8456
	3, 8, 9, 11, 13, 20, 21, 22, 23	0.8261
40%	1, 6, 8, 9, 11, 12, 15, 21, 22, 23, 24, 28	0.8794
	2, 6, 8, 9, 11, 14, 17, 20, 23, 24, 27, 30	0.8514
	1, 5, 7, 8, 9, 11, 14, 15, 20, 23, 27, 29	0.8252
50%	1, 2, 6, 7, 8, 9, 11, 15, 16, 21, 22, 23, 27, 28, 29	0.9150
	1, 5, 6, 8, 9, 10, 11, 15, 19, 22, 23, 24, 26, 28, 29	0.9056
	1, 5, 6, 8, 9, 10, 11, 16, 17, 20, 22, 23, 25, 27, 29	0.9054

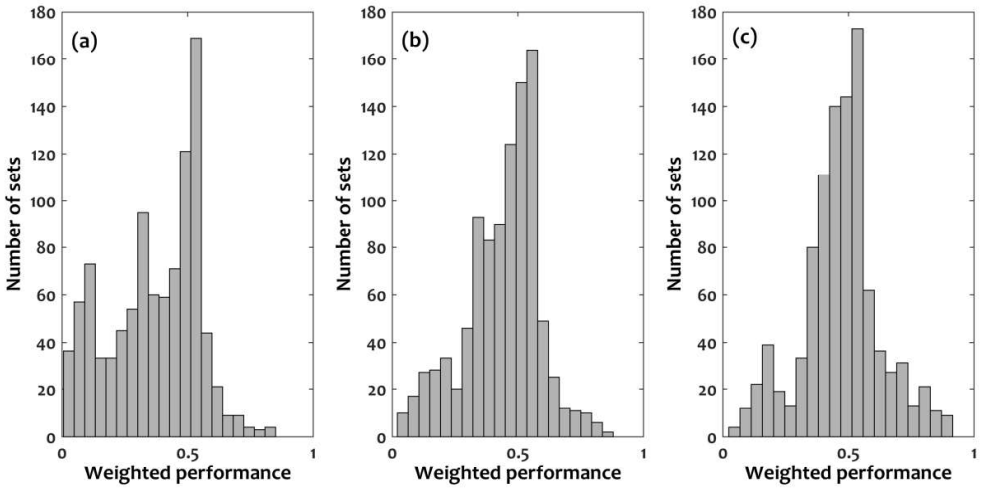

Figure 7.6 : 'Abdomen' class: Relation between weighted performance and their corresponding number of sets with different proportion of known side effects. (a) 30% (b) 40% and (c) 50%.

Table 7.8 : 'Abdominal' Class: A ranked list of critical side effects obtained with sample size 1000.

Rank	Side effect	Rank	Side effect
1	'Abdominal cramps' (8)	6	'Abdominal abscess' (1)
2	'Abdominal discomfort' (9)	7	'Abdominal pain generalised' (20)
3	'Abdominal distension' (11)	8	'Abdominal pain lower' (22)

4	'Abdominal pain upper' (23)	9	'Abdominal strangulated hernia' (27)
5	'Abdominal bloating' (6)		

7.3.4 Hepatic

The word Hepatic is derived from the Greek word 'Hepato' which means liver. It is one of the most vital organs, known to be involved in metabolism and regulation of different functions such as storage of glycogen, red blood decomposition, plasma protein. Finding unknown side effects of consequence to liver with the partial knowledge of hepatic adverse reactions could be of value for existing drugs as well as for candidate drugs under screening. The drug-side effects matrix for hepatic (liver-associated) side effects was compiled using key word 'Hepatic'. It yielded 55 side effects and 449 associated drugs. List of these side effects is provided in Table 7.1. The matrix had 6% sparse-ness, with on an average 3.8 and maximally having 13 adverse reactions. 10-fold cross-validation was performed using 440 drugs after removing 9 drugs. The training set and test set consisted of 400 and 44 drugs, respectively.

Using the protocol described in Section 7.2, using the prior knowledge of 30%, 40% or 50% of all possible side effects, we computed the Prediction Performance of 1000 randomly sampled sets (Figure 7.7). As seen in the figure, while the overall nature of distribution remains unaffected, the average performance as well as best performance improves with increased prior knowledge. Table 7.9 lists top 3 sets with best performance obtained, which were further used for obtaining a ranked list of 17 side effects (Table 7.10). The results suggest that for any drug, prior knowledge of these side effects can facilitate prediction of the remainder of the side effects with good confidence (0.6230). Our strategy of segregating side effects with best possible predictive connection with rest of the adverse reactions, yields performance that is far better than that achieved from screening random sets of side effects of the same size (30%; 0.5512).

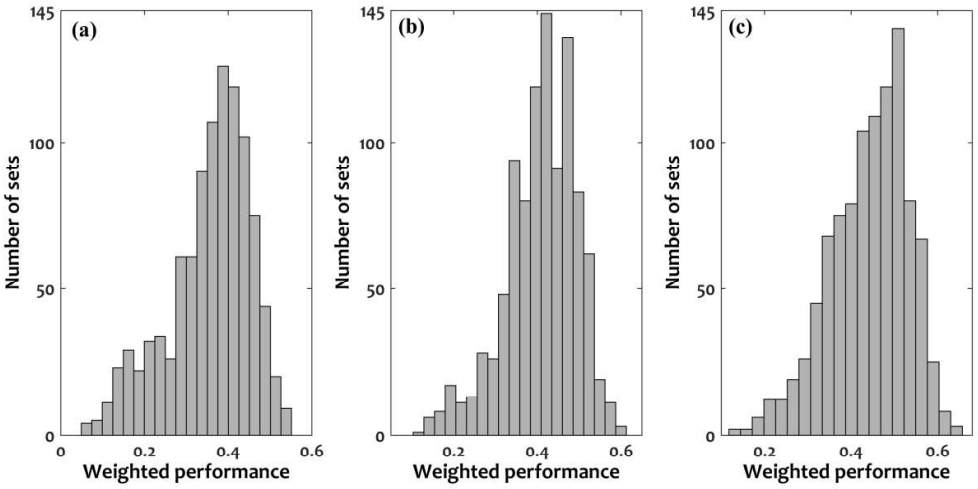

Figure 7.7 : 'Hepatic' class: Relation between weighted performance and their corresponding number of sets with different proportion of known side effects. (a) 30% (b) 40% and (c) 50%.

Table 7.9 : 'Hepatic' Class: Three best groups with various proportion of known side effects and their performance.

Fraction of known side effects	Known side effects profile	Performance
30%	3, 6, 9, 12, 13, 15, 20, 21, 22, 32, 34, 36, 39, 42, 46, 53, 54	0.5512
	2, 3, 6, 8, 10, 11, 12, 15, 18, 27, 36, 39, 41, 44, 47, 51, 55	0.5412
	2, 3, 10, 11, 12, 15, 22, 30, 32, 35, 39, 40, 44, 46, 48, 50, 51	0.5397
40%	2, 3, 4, 5, 8, 9, 12, 13, 15, 18, 29, 31, 36, 39, 40, 41, 42, 46, 49, 50, 51, 55	0.6127
	4, 10, 11, 12, 13, 15, 18, 20, 22, 28, 29, 31, 32, 35, 38, 44, 48, 49, 50, 51, 54, 55	0.6118
	5, 8, 9, 10, 12, 13, 15, 17, 19, 20, 22, 24, 31, 33, 35, 39, 41, 44, 48, 50, 53, 55	0.6016
50%	2, 3, 8, 10, 12, 13, 14, 15, 18, 19, 20, 22, 23, 25, 28, 33, 34, 38, 39, 40, 41, 42, 46, 49, 50, 51, 54, 55	0.6572
	1, 2, 6, 7, 9, 10, 12, 13, 14, 15, 17, 20, 22, 24, 26, 28, 31, 33, 38, 39, 40, 43, 45, 47, 50, 51, 54, 55	0.6378
	1, 4, 8, 9, 12, 13, 15, 16, 18, 20, 21, 23, 25, 30, 33, 34, 39, 41, 43, 45, 47, 49, 50, 51, 52, 53, 54, 55	0.6328

Table 7.10 : 'Hepatic' Class: A ranked list of critical side effects obtained with sample size 1000.

Rank	Side effect	Rank	Side effect	Rank	Side effect
1	'Hepatic enzyme increased' (12)	7	'Hepatotoxicity' (55)	13	'Hepatic cyst' (8)
2	'Hepatic function abnormal' (15)	8	'Hepatic enzyme abnormal' (10)	14	'Hepatic encephalopathy' (9)
3	'Hepatitis cholestatic' (39)	9	'Hepatic insufficiency' (20)	15	'Hepatic impairment' (18)

4	'Hepatic failure' (13)	10	'Hepatic necrosis' (22)	16	'Hepatitis fulminant' (41)
5	'Hepatocellular injury' (50)	11	'Hepatic angiosarcoma' (2)	17	'Hepatosplenomegaly' (54)
6	'Hepatomegaly' (51)	12	'Hepatic artery thrombosis' (3)		

7.3.5 Gastrointestinal

This class refers to side effects associated with stomach and intestine. These organs are responsible for extracting energy and nutrients from food, and are critical for metabolism of drugs. Drugs such as Bupropion, Citalopram are known to cause many gastrointestinal side effects such as hemorrhage, gastroenteritis, gastric pH fluctuations, gastro intestinal obstruction and discomfort [Fouladi and Mortazavi, 2012; Hibbard, Propst, Frank, and Wyse, 2009]. The protocol, when used to compile the drug-side effects matrix for Gastroduodentitis-related adverse reactions using the key word 'Gastroduodentitis' yielded 42 side effects and 407 associated drugs. The list of these side effects is provided in Table 7.1. The matrix had 6% sparseness, with an average 2.8 side effects per drug and maximally having 10 adverse reactions. 10-fold cross-validation was performed using 400 drugs after removing 7 drugs. The training set and test set consisted of 360 and 40 drugs, respectively.

Knowing that side effects of a given organ/system are expected to be closely linked with each other, using canonical correlation analysis we computed the prediction performance through a model that uses the prior knowledge of 30%, 40% or 50% of side effects (protocol described in Section 7.2). Figure 7.8 depicts the statistics of prediction performance from 1000 randomly sampled sets of side effects. With increasing fraction of known side effects, more number of sets had high performance. Table 7.11 lists top three random side effect sets for each category, which were further ranked (Table 7.12). We obtained the prediction performance of 0.7430 using the top 13 (30%) side effects, compared to the best performance from randomly sampled sets of the same size (0.7732). This is a result contrary to expectation and needs to be probed further.

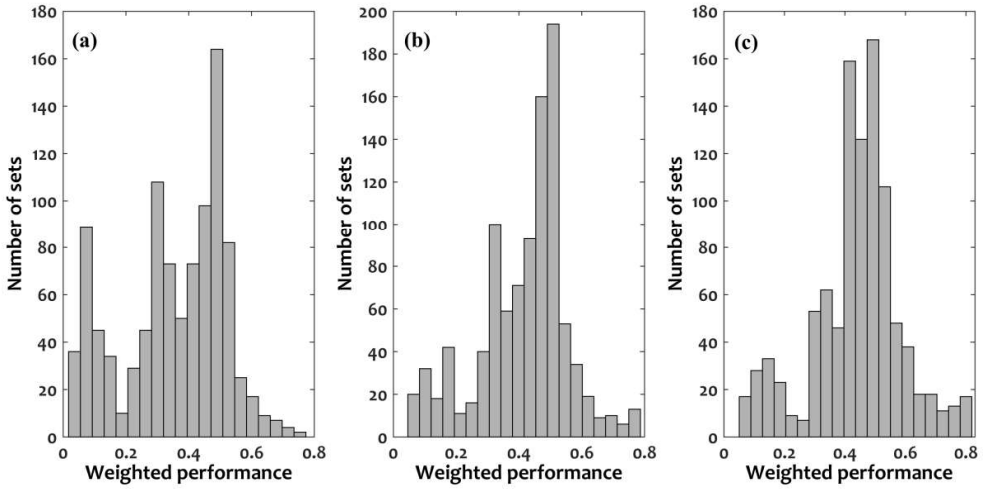

Figure 7.8 : 'Gastrointestinal' class: Relation between weighted performance and their corresponding number of sets with different proportion of known side effects. (a) 30% (b) 40% and (c) 50%.

Table 7.11 : 'Gastrointestinal' Class: Three best groups with various proportion of known side effects and their performance.

Fraction of known side effects	Known side effects profile	Performance
30%	2, 4, 9, 12, 15, 17, 26, 28, 30, 34, 35, 36, 40	0.7732
	2, 8, 9, 15, 25, 30, 34, 36, 37, 38, 40, 41, 42	0.7479
	2, 3, 5, 6, 8, 15, 16, 17, 20, 21, 34, 37, 40	0.7314
40%	1, 2, 3, 6, 7, 9, 10, 15, 22, 30, 31, 34, 35, 36, 37, 39, 40	0.7869
	2, 4, 6, 7, 15, 19, 21, 24, 26, 27, 29, 30, 31, 34, 35, 39, 40	0.7768
	2, 3, 4, 6, 7, 10, 15, 17, 19, 21, 26, 32, 34, 36, 37, 40, 41	0.7762
50%	2, 3, 4, 10, 12, 13, 14, 15, 17, 19, 24, 26, 27, 28, 31, 33, 34, 35, 37, 38, 40	0.8194
	2, 6, 9, 10, 14, 15, 16, 17, 19, 21, 22, 23, 27, 28, 30, 31, 32, 34, 35, 37, 40	0.8172
	2, 3, 6, 7, 15, 17, 19, 22, 24, 25, 27, 28, 29, 30, 34, 35, 36, 38, 39, 40, 41,	0.8153

Table 7.12 : 'Gastrointestinal' Class: A ranked list of critical side effects obtained with sample size 1000.

Rank	Side effect	Rank	Side effect	Rank	Side effect
1	'Gastroenteritis' (2)	6	'Gastroenteritis yersinia' (6)	11	'Gastroenteritis adenovirus' (3)
2	'Gastrointestinal haemorrhage' (15)	7	'Gastrointestinal hypomotility' (17)	12	'Gastrointestinal inflammation' (19)
3	'Gastrointestinal symptom NOS' (34)	8	'Gastrointestinal signs and symptoms NEC' (30)	13	'Gastroenteritis bacterial' (4)
4	'Gastrointestinal toxicity' (35)	9	'Gastrointestinal ulcer' (37)		
5	'Gastrooesophageal reflux disease' (40)	10	'Gastrointestinal ulcer haemorrhage' (38)		

7.3.6 Kidney

Adverse reactions associated to kidney are also referred to as renal side effects. This organ, extremely critical in drug excretion, purifies blood and also balances the acid-base level of blood. Drugs such as pramipexole dihydrochloride, which is used as a Dopamine agonistic against Parkinson's disease, are known to cause side effects such as renal failure, renal artery stenosis and renal cyst [Dooley and Markham, 1998]. This class is among the smallest in our list with only 29 side effects. For obtaining drug-side effects matrix associated with Kidney (Renal) problems, we used the protocol with the keyword 'Renal'. This yielded 29 side effects and 219 associated drugs. The list of these Renal-related side effects is provided in Table 7.1. The matrix had 9% sparse-ness, with an average of 2.7 and maximally having 9 adverse reactions. 10-fold cross-validation was performed using 210 drugs after removing 9 drugs. The training set and test set consisted of 189 and 21 drugs, respectively.

Figure 7.9 depicts the prediction performance of 1000 sets of (30%, 40% or 50%) randomly selected sets of drugs. The data is divided into two distance groups: poor performers and good performers. The overall performance as well as best performing sets improve with increasing prior knowledge. Table 7.13 lists results of our experiments that were performed using the protocol described in Section 7.2. These inputs were further used to create a ranked list of side effects (Table 7.14). While the procedure of sampling and ranking does help in improving the performance from random sets (0.6896) to ranked sets (0.7026) for 30% known side effects, the improvement is marginal.

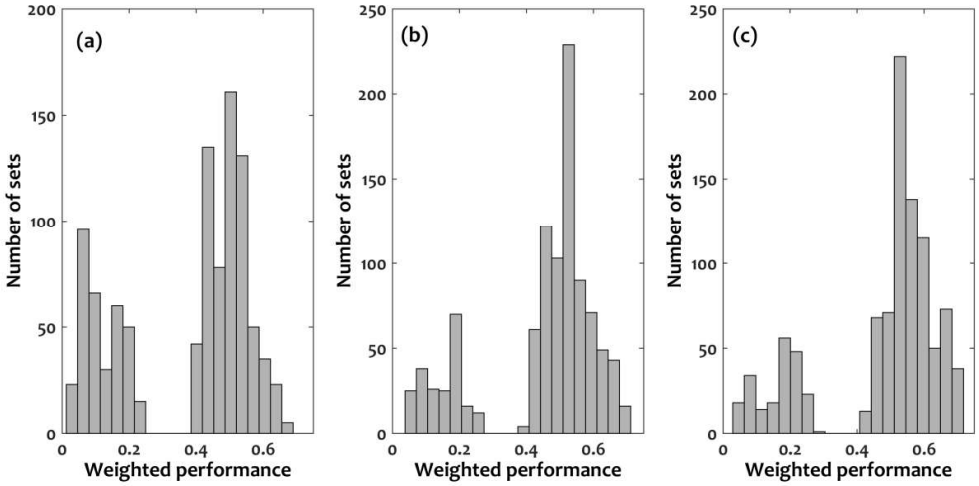

Figure 7.9 : 'Kidney' class: Relation between weighted performance and their corresponding number of sets with different proportion of known side effects. (a) 30% (b) 40% and (c) 50%.

Table 7.13 : 'Kidney' Class: Three best groups with various proportion of known side effects and their performance.

Fraction of known side effects	Known side effects profile	Performance
30%	1, 5, 6, 11, 14, 17, 22, 23, 28	0.6896
	4, 9, 11, 17, 22, 23, 25, 26, 27	0.6785
	6, 8, 9, 11, 14, 17, 19, 22, 29	0.6616

40%	3, 6, 7, 11, 14, 17, 21, 22, 23, 25, 27, 28	0.7112
	1, 5, 6, 11, 14, 17, 19, 22, 25, 26, 27, 28	0.7108
	1, 6, 9, 11, 13, 14, 16, 17, 19, 22, 23, 25	0.6979
50%	2, 4, 6, 9, 11, 12, 13, 14, 16, 17, 21, 22, 23, 25, 28	0.7180
	2, 6, 7, 8, 11, 12, 14, 15, 17, 18, 19, 22, 23, 26, 27	0.7091
	2, 6, 8, 9, 11, 12, 17, 21, 22, 23, 24, 25, 27, 28, 29	0.7077

Table 7.14 : 'Kidney' Class: A ranked list of critical side effects obtained with sample size 1000.

Rank	Side effects	Rank	Side effects	Rank	Side effects
1	'Renal failure acute' (11)	4	'Renal colic' (6)	7	'Renal tubular acidosis' (25)
2	'Renal impairment' (17)	5	'Renal function test abnormal' (14)	8	'Renal dysgenesis' (9)
3	'Renal pain' (22)	6	'Renal papillary necrosis' (23)	9	'Renal tubular necrosis' (27)

7.3.7 Nose

The nose or nasal system is important for olfactory responses (smell). Few of the most frequent nasal side effects are nasal dryness, nasal disorder and nasal irritation. Among the drugs that cause nasal reactions are fentanyl, fluticasone propionate. These drugs also cause other side effects such as nasal septum perforation, congestion, ulcer etc. Nose-related drug-side effects matrix was compiled using keyword 'Nasal'. It yielded 28 side effects and 113 associated drugs. The list of these Nasal-related side effects is provided in Table 7.1. The matrix had 10% sparseness, with an average 2.9 side effects per drug and maximally having 9 adverse reactions. 10-fold cross-validation was performed using 110 drugs after removing 3 drugs. The training set and test set consisted of 99 and 11 drugs, respectively.

Table 7.15 depicts the side effect sets obtained with best prediction performance after sampling 1000 times (Protocol described in Section 7.2). Figure 7.10 reflects a clear division among these random sets, thereby strengthening our idea that while prior knowledge of some side effects can be useful for prediction of the remainder of side effects, not all sets are as good. We further ranked these side effects (Table 7.16). The prediction performance was improved with ranked side effects (0.7921) compared to random sets (0.7798), for 30% of known side effects.

Table 7.15 : 'Nose' Class: Three best groups with various proportion of known side effects and their performance.

Fraction of known side effects	Known side effects profile	Performance
30%	3, 5, 7, 8, 12, 21, 25, 27	0.7798

	3, 4, 5, 7, 12, 18, 21, 27	0.7711
	3, 5, 7, 13, 18, 25, 26, 27	0.7677
40%	3, 5, 7, 8, 9, 10, 11, 14, 20, 27, 28	0.8019
	3, 5, 7, 8, 13, 14, 15, 24, 25, 27, 28	0.7949
	1, 3, 5, 6, 7, 14, 16, 19, 22, 27, 28	0.7937
50%	1, 3, 5, 7, 12, 13, 14, 15, 16, 19, 20, 23, 24, 27	0.8238
	3, 5, 7, 8, 13, 14, 15, 17, 18, 19, 20, 21, 22, 27	0.8210
	3, 5, 7, 9, 13, 14, 16, 17, 19, 20, 22, 24, 26, 27	0.8138

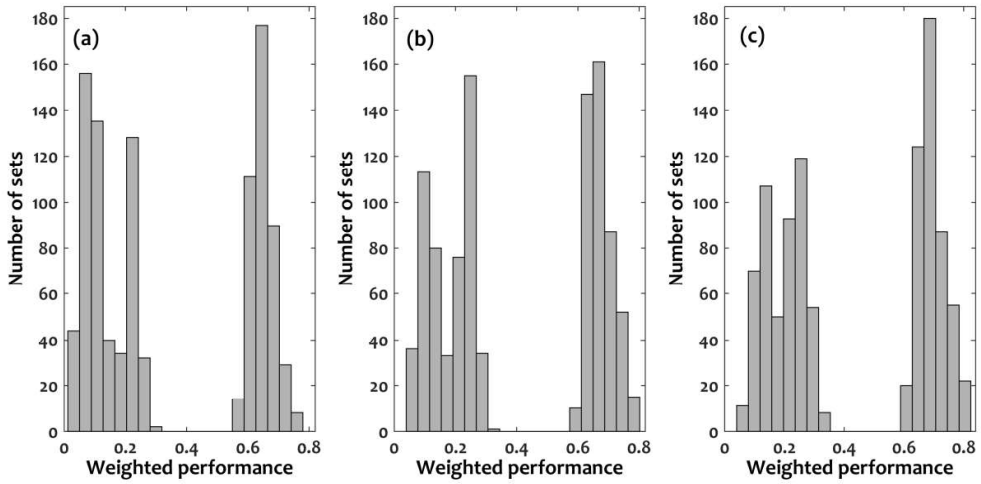

Figure 7.10 : 'Nose' class: Relation between weighted performance and their corresponding number of sets with different proportion of known side effects. (a) 30% (b) 40% and (c) 50%.

Table 7.16 : 'Nose' Class: A ranked list of critical side effects obtained with sample size 1000.

Rank	Side effects	Rank	Side effects	Rank	Side effects
1	'Nasal discomfort' (3)	4	'Nasopharyngitis' (27)	7	'Nasal itching' (8)
2	'Nasal dryness' (5)	5	'Nasal oedema' (14)	8	'Nasal septum disorder' (19)
3	'Nasal irritation' (7)	6	'Nasal odour' (13)		

7.4 DISCUSSION

In contrast to using chemical information such as their substructure, properties, and target profile, we implemented a strategy for predicting 'unknown side effects' starting with the knowledge of 'known side effects' (Figure 7.1). This approach could be useful for modeling chemical activity, especially in the absence of explicit information of chemistry or targets of drugs. The objective of our study was to identify a minimal set of known side effects which could predict

remaining side effects with good confidence. Towards this aim, seven classes of adverse reactions (blood, eye, abdomen, hepatic, gastrointestinal, kidney and nose) were considered. Unknown adverse reactions were predicted using CCA technique by considering different proportion of known side effects. For each of these classes, best known profiles of different proportion (30%, 40% and 50%) were further aggregated to assess the importance of individual known side effect to specify hidden side effects. Based on their importance, a ranked set was created for 30% known side effects. Across the classes, it was observed that such a ranked list was effective. This suggests that the strategy implemented in this study is useful for obtaining an optimum set.

Going beyond the above-mentioned strategy, we also implemented Ridge regression in order to verify the credibility of best ranked known side effects obtained for each class. It was observed that the performance of Ridge regression was comparable to that of CCA, with one of these methods outperforming the other depending on the class of side effects (Table 7.17). Thus we conclude that the best 'known side effects' obtained through our studies could be integrated into a suitable model to achieve better results than obtained with CCA or Ridge regression. Our studies with Restricted Boltzmann Machine and Diffusion Model yielded results that were inferior to that of CCA and Ridge regression.

Table 7.17 : Comparison of CCA and Ridge regression strategies in predicting 'unknown side effects', starting with the knowledge of 30% 'known side effects' for different organ/system.

	Blood	**Eye**	**Abdominal**	**Hepatic**	**Gastrointestinal**	**Kidney**	**Nose**
CCA	0.63	0.70	0.87	0.69	0.73	0.68	0.77
Ridge	0.63	0.71	0.85	0.68	0.73	0.69	0.77

…