MACHINE LEARNING APPROACHES IN PHARMACOKINETIC AND TOXICITY PREDICTION

YAP CHUN WEI

(B. Sc (Pharm)(Hons), NUS)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF PHARMACY

NATIONAL UNIVERSITY OF SINGAPORE

2006

Acknowledgements

I would like to dedicate this thesis to my wife, who has been very patient in listening to my project ideas throughout these years, even though she is busy with her own PhD study.

I wish to express my heartfelt appreciation to my supervisor, Associate Professor Chen Yu Zong, who has provided me with excellent guidance and instilled upon me the necessary skills for scientific research.

Many thanks to Dr Cai Cong Zhong for introducing support vector machine to our group and Dr Li Ze Rong and Dr Xue Ying for programming the molecular descriptors used in this work.

Finally, I wish to thank all members of the BIDD group for their insightful discussions and help in one way or another.

Table of Contents

Summary

Drug development aims at finding therapeutic compounds that possess desirable pharmacodynamic and pharmacokinetic properties and low toxicological profiles. Historically, inappropriate pharmacokinetic properties and side-effects have been the primary reasons for the failure of drug candidates in later stages of development. Thus tools for predicting pharmacokinetic and toxicological properties in early design stages are needed for fast elimination of compounds with undesirable properties so that development effort can be focused on the most promising candidates. As part of the effort for developing such tools, computational methods have been explored for predicting various pharmacokinetic and toxicological properties of pharmaceutical compounds. In particular, quantitative structure pharmacokinetic relationship (QSPkR) and qualitative structure pharmacokinetic relationship (qSPkR) methods have shown promising potential for performing these tasks by statistically analyzing the correlation between chemical structures and a specific pharmacokinetic, or toxicological (ADMET) property to derive statistical models or rules for predicting whether a drug candidate possesses a specific property or for predicting the activity level of the drug candidate.

Previously, QSPkR/qSPkR models were frequently built using datasets with a limited number of related compounds and by using linear statistical methods. Hence they may not be suitable for the prediction of ADMET properties of diverse groups of compounds and also ADMET properties that are controlled by multiple mechanisms. Thus it is of interest to examine the potential of using a larger number and more diverse groups of compounds and non-linear machine learning methods in improving the quality of QSPkR/qSPkR models. In this work, machine learning methods, such as support vector machines, support vector regression, and general regression neural

network, consensus modeling methods, larger number and more diverse groups of compounds, as well as compounds with known human ADMET data were used to develop QSPkR/qSPkR models for various ADMET properties. A novel method for identification of relevant physicochemical and structural properties of a compound from non-linear QSPkR/qSPkR models, which are traditionally regarded as black boxes, is also introduced.

The results show that the quality of QSPkR/qSPkR models can be improved by using the methods discussed in this work. The prediction capabilities of QSPkR/qSPkR models developed in this work for human intestinal absorption, p-glycoprotein substrates, blood-brain barrier penetration, human serum albumin binding, milk-plasma ratio, cytochrome isoenzymes substrates and inhibitors, total body clearance, and genotoxicity are higher than those developed in earlier studies. In addition, machine learning methods were found to be useful for developing qSPkR models for torsade de pointes, a rare but serious adverse drug reaction, which has not been sufficiently explored in earlier studies.

List of Tables

List of Figures

List of Abbreviations

ADMET – Absorption, distribution, metabolism, excretion, toxicity

ADR – Adverse drug reaction

ANN – Artificial neural network

BBB – Blood-brain barrier

C4.5 DT – C4.5 decision tree

CL_{tot} – Total clearance

cQSPkR – Consensus quantitative structure pharmacokinetics relationship

CSVM – Consensus support vector machine

CYP – Cytochrome

DI – Diversity index

FN – False negatives

FP – False positives

GA – Genetic algorithm

GRNN – General regression neural network

HIA – Human intestinal absorption

HSA – Human serum albumin

kNN – *k* nearest neighbour

LDA – Linear discriminant analysis

LOO – Leave-one-out

LSER – Linear solvation energy relationship

MCC – Matthews correlation coefficient

MDR – Multidrug resistant

MLFN – Multilayer feedforward neural network

MLR – Multiple linear regression

MSE – Mean square error

PC – Principal component

PCA – Principal component analysis

PLS – Partial least squares

PNN – Probabilistic neural network

Q – Overall accuracy

QSAR – Quantitative structure activity relationship

QSPkR – Quantitative structure pharmacokinetics relationship

qSPkR – Qualitative structure pharmacokinetics relationship

QSPR – Quantitative structure property relationship

QSTR – Quantitative structure toxicity relationship

RFE – Recursive feature elimination

RI – Representativity index

SAR – Structure activity relationship

SE – Sensitivity

SP – Specificity

SVM – Support vector machine

SVR – Support vector regression

TdP – Torsade de pointes

TN – True negatives

TP – True positives

List of Publications

A. Publications relating to research work from the current thesis

1. **Yap CW**, Li ZR and Chen YZ (2006). Quantitative structure-pharmacokinetic relationships for drug clearance by using statistical learning methods. *Journal of Molecular Graphics and Modelling* **24**(5): 383-395.
2. **Yap CW** and Chen YZ (2005). Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. *Journal of Chemical Information and Modeling* **45**(4): 982-992.
3. Li H, Ung CY, **Yap CW**, Xue Y, Li ZR, Cao ZW and Chen YZ (2005). Prediction of genotoxicity of chemical compounds by statistical learning methods. *Chemical Research in Toxicology* **18**(6): 1071-1080.
4. **Yap CW** and Chen YZ (2005). Quantitative structure-pharmacokinetic relationships for drug distribution properties by using general regression neural network. *Journal of Pharmaceutical Sciences* **94**(1): 153-168.
5. Xue Y, Li ZR, **Yap CW**, Sun LZ, Chen X and Chen YZ (2004). Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. *Journal of Chemical Information and Computer Sciences* **44**(5): 1630-1638.
6. Xue Y, **Yap CW**, Sun LZ, Cao ZW, Wang JF and Chen YZ (2004). Prediction of p-glycoprotein substrates by support vector machine approach. *Journal of Chemical Information and Computer Sciences* **44**(4): 1497-1505.
7. **Yap CW**, Cai CZ, Xue Y and Chen YZ (2004). Prediction of torsade-causing potential of drugs by support vector machine approach. *Toxicological Sciences* **79**(1): 170-177.

B. Publications from other projects not included in the current thesis

1. Xue Y, Li H, Ung CY, **Yap CW** and Chen YZ (2006). Classification of a diverse set of Tetrahymena Pyriformis toxicity chemical compounds from molecular descriptors by statistical learning methods. *Chemical Research in Toxicology* **19**(8): 1030-1039.
2. **Yap CW**, Xue Y, Li ZR and Chen YZ (2006). Application of support vector machines to in silico prediction of cytochrome P450 enzyme substrates and inhibitors. *Current Topics in Medicinal Chemistry* **6**(15): 1593-1607.
3. **Yap CW**, Xue Y, Li H, Li ZR, Ung CY, Han LY, Zheng CJ, Cao ZW and Chen YZ (2006). Prediction of compounds with specific pharmacodynamic, pharmacokinetic or toxicological property by statistical learning methods. *Mini Reviews in Medicinal Chemistry* **6**(4): 449-459.
4. Li H, **Yap CW**, Xue Y, Li ZR, Ung CY, Han LY and Chen YZ (2006). Statistical learning approach for predicting specific pharmacodynamic, pharmacokinetic or toxicological properties of pharmaceutical agents. *Drug Development Research* **66**(4): 245-259.
5. Li H, Ung CY, **Yap CW**, Xue Y, Li ZR and Chen YZ (2006). Prediction of estrogen receptor agonists and characterization of associated molecular descriptors by statistical learning methods. *Journal of Molecular Graphics and Modelling* **25**(3): 313-323.
6. Zheng CJ, Han LY, **Yap CW**, Ji ZL, Cao ZW and Chen YZ (2006). Therapeutic targets: Progress of their exploration and investigation of their characteristics. *Pharmacological Reviews* **58**(2): 259-279.

7. Zheng CJ, Han LY, **Yap CW**, Xie B and Chen YZ (2006). Progress and difficulties in the exploration of therapeutic targets. *Drug Discovery Today* **11**(9-10): 412-420.

8. Li H, **Yap CW**, Ung CY, Xue Y, Cao ZW and Chen YZ (2005). Effect of selection of molecular descriptors on the prediction of blood-brain barrier penetrating and non-penetrating agents by statistical learning methods. *Journal of Chemical Information and Modeling* **45**(5): 1376-1384.

9. Zheng CJ, Han LY, **Yap CW**, Xie B and Chen YZ (2005). Trends in exploration of therapeutic targets. *Drug News and Perspectives* **18**(2): 109-127.

10. Zheng CJ, Zhou H, Xie B, Han LY, **Yap CW** and Chen YZ (2004). TRMP: A Database of Therapeutically Relevant Multiple-Pathways. *Bioinformatics* **20**: 2236-2241.

11. Ji ZL, Han LY, **Yap CW**, Sun LZ, Chen X and Chen YZ (2003). Drug adverse reaction target database (DART): Proteins related to adverse drug reactions. *Drug Safety* **26**(10): 685-690.

Chapter 1

Introduction

In Silico methods are increasingly employed to reduce the time and cost needed for evaluating the pharmacokinetics and toxicity of drug candidates. The most common In Silico methods are traditional linear statistical methods such as multiple linear regression. Recently, non-linear machine learning methods such as artificial neural networks and support vector machines have been evaluated for their usefulness for the prediction of pharmacokinetics and toxicological properties because of their success in many diverse fields such as data mining, image and speech recognition, and process control. The first section (section 1.1) of this chapter gives an overview of the application of in silico methods for pharmacokinetics and toxicity prediction. The motivation for this work and an outline of the structure of this document is given in the next two sections of this chapter (sections 1.2, 1.3).

1.1 Application of *in silico* methods for pharmacokinetics and toxicity prediction

1.1.1 Drug discovery process

Modern drug discovery efforts have primarily been based on the search and optimization of compounds that possess specific pharmacodynamic and pharmacokinetic properties, and on the test of their potential toxicological and side effects (Caldwell *et al.* 1995; Drews 2000; Park *et al.* 2000). Pharmacodynamics is the study of the biochemical and physiological effects of drugs and their mechanisms

of action (Hardman *et al.* 2002). For a drug to be effective, it must have optimal pharmacodynamic properties so that it can inhibit a disease process, correct the imbalances and brings about the normal functioning of the body. Pharmacokinetics is the study of the time course of a drug within the body and incorporates the processes of absorption, distribution, metabolism and excretion, which together with toxicological properties are referred to as ADMET properties (Smith *et al.* 2001b). A drug must have optimal pharmacokinetic properties so as to achieve sufficient concentration at target site while possibly limiting its distribution elsewhere so as to produce desired therapeutic action with minimum side effects.

The drug discovery process is typically a lengthy and costly process. The average time required for a drug to proceed from initial design effort to market approval is 13 years and the estimated average development cost of a new drug is US$802 million, with the preclinical phase and clinical phase costing US$335 million and US$467 million respectively (DiMasi *et al.* 2003). Traditionally, pharmacokinetic and toxicological properties of drug candidates have primarily been evaluated during later design stages, particularly in the expensive animal tests and clinical trials (van de Waterbeemd *et al.* 2003). According to a recent report, approximately 40% of all drug failures during the clinical phase, excluding failures of anti-infectives, is due to poor pharmacokinetics (7%) or unacceptable toxicity (33%). If anti-infectives are considered, the percentage increases to approximately 60% with 39% and 21% due to poor pharmacokinetics and unacceptable toxicity respectively (Kubinyi 2003). To reduce the cost and time of drug development, there has been a paradigm shift such that ADMET properties are now considered and evaluated in increasingly earlier stages of drug discovery process. Thus methods for predicting these ADMET properties, particularly in the early design stages, are useful for facilitating drug

development and drug safety evaluation (Drews 2000; Ekins *et al.* 2000b; White 2000).

1.1.2 Application of quantitative structure pharmacokinetics relationship (QSPkR) and qualitative structure pharmacokinetics relationship (qSPkR) models in ADMET prediction

As part of an effort to accelerate and reduce the cost of drug discovery processes, computational methods have been explored for predicting compounds that possess specific pharmacodynamic, pharmacokinetic or toxicological property (Katritzky *et al.* 1997; Manallack *et al.* 1999; van de Waterbeemd *et al.* 2003; Hansch *et al.* 2004). In particular, statistical learning methods have shown promising potential for performing these tasks by statistically analyzing the structural and physicochemical features of the compounds known to possess a particular property to derive explicit or hidden statistical models or rules for predicting the activity or property of new compounds (Manallack *et al.* 1999; Burbidge *et al.* 2001; Trotter *et al.* 2003).

The development of QSPkR models have been instrumental for the early testing of ADMET properties of drug candidates. Hansch is one of the pioneers in exploring the usefulness of QSPkR models (Hansch 1972). His work on the use of the partition coefficient, log P, to model drug metabolism has generated a significant interest in applying QSPkR models for prediction of other ADMET properties. The initial QSPkR models were usually built from small congeneric groups of compounds with known *in vivo* ADMET data (Hansch 1972; Seydel *et al.* 1981; Toon *et al.* 1983; Markin *et al.* 1988). The results of these studies suggested that QSPkR models are potentially useful for the prediction of ADMET properties. However, the small

amount of available *in vivo* ADMET data limits the widespread development of QSPkR models. Subsequently, the development of combinatorial chemistry and high-throughput screening using *in vitro* assays enable large numbers of closely related compounds to be rapidly synthesized and screened for their ADMET properties. This creates a wealth of *in vitro* ADMET data, which enables the evaluation of *in silico* methods, thereby increasing the confidence in the results obtained when these methods are applied to scarce human data (Clark *et al.* 2003).

QSPkR/qSPkR models have now been built for a number of ADMET properties. These include cellular permeability (van de Waterbeembd *et al.* 1996), intestinal absorption (Stenberg *et al.* 2000), bioavailability (Mandagere *et al.* 2003), active transport processes (Ekins *et al.* 2000c) and skin permeability (Abraham *et al.* 1999), blood-brain barrier penetration (Ecker *et al.* 2004), milk-plasma ratio (Meskin *et al.* 1985), serum protein binding (Toon *et al.* 1983), volume of distribution (Toon *et al.* 1983), P450 isoenzyme substrates and inhibitors (Koymans *et al.* 1992; Ekins *et al.* 1999a), first pass (Watari *et al.* 1988), total clearance (Toon *et al.* 1983), renal clearance (Toon *et al.* 1983), half-life (Markin *et al.* 1988), genotoxicity (Mosier *et al.* 2003), carcinogenicity (Benigni *et al.* 2000), mutagenicity (Benigni *et al.* 2000), and QT prolongation (Muzikant *et al.* 2002). Table 1.1 and Table 1.2 give a list of some of these QSPkR/qSPkR models. There are many applications of these QSPkR/qSPkR models. Some qSPkR models, such as the Lipinski's rule of five (Lipinski *et al.* 1997), are useful as computational filters for the high-throughput screening of chemical libraries for potential drug leads with acceptable ADMET properties. QSPkR/qSPkR models that identify pharmacophoric models of metabolic enzymes are useful in the rational design of drug candidates to avoid potential drug-drug interactions (Ekins *et al.* 2000a). Those models that estimate the pharmacokinetics behavior in humans,

such as the bioavailability (Mandagere *et al.* 2003) and milk-plasma ratio (Agatonovic-Kustrin *et al.* 2002), are useful for determining the appropriate starting dose during the clinical phase or to evaluate the potential risk to the infant.

Table 1.1 Performance of classification-based statistical learning methods for predicting compounds of specific pharmacokinetic or toxicological property.

Property	Method	Molecular descriptors	Number of compounds in training set	Validation method [a]	Reported prediction accuracy			Reference
					SE (%)	SP (%)	Q (%)	
Human intestinal absorption (HIA)	LDA	TOPS-MODE	82	Validation set (127)	95.5	76.5	92.9	(Pérez *et al.* 2004)
	C-SAR	Simple physicochemical parameters	977	Training set (977)	97.0	81.7	95.7	(Zmuidinavicius *et al.* 2003)
	PNN	Log P, MR, TOP	76	Validation set (10)	100.0	50.0	80.0	(Niwa 2003)
	SVM	Simple molecular properties, molecular connectivity and shape, E-state, Q-C, GEO	196	5 fold CV (196)	90.0	80.7	86.7	(Xue *et al.* 2004b)
Bioavailability	ORMUCS	Log P, structural	232	Validation set (40)	-	-	60.0	(Yoshida *et al.* 2000)
	Adaptive fuzzy partition	CON, information, TOP, E-state, physicochemical, ELE	352	Validation set (75)	-	-	64.0	(Pintore *et al.* 2003)
P-gp substrate	SVM	Simple molecular properties, molecular connectivity and shape, E-state, Q-C, GEO	142	Validation set (25)	84.2	66.7	80.0	(Xue *et al.* 2004c)

BBB penetration	MLR	Daylight, thermodynamic, spatial, structural, TOP, charge	48	Validation set (150)	81.0	95.8	88.0	(Lobell *et al.* 2003a)
	Discrimination function analysis	TOP, substructures, GEO, Q-C	28	LOO (28)	100.0	91.7	96.4	(Basak *et al.* 1996)
	PLS	Log P, PSA, E-state	58	Validation set (181)	85.7	46.7	66.3	(Subramanian *et al.* 2003)
	PLS-DA	ADME screen, geometry, topology, VAMP electronic parameters, VAMP energy parameters, Sybyl surface areas	1696	Validation set (82)	90.0	92.0	91.0	(Adenot *et al.* 2004)
	SUBSTRUCT	Substructures	8678	10 fold CV (8678)	83.3	71.2	76.3	(Engkvist *et al.* 2003)
	Bayesian neural network	CON, log P, ISIS fingerprint	>73000	Validation set (84)	94.7	73.9	83.3	(Ajay *et al.* 1999)
	PCA	VolSurf	110	Validation set (120)	90.9	64.8	71.7	(Crivori *et al.* 2000)
	SVM	Structural	172	Validation set (304)	78.9	60.4	76.0	(Trotter *et al.* 2001)
		VolSurf	238	Validation set (238)	91.8	68.5	86.6	(Trotter *et al.* 2003)
		MW, lipophilicity, H-bond	274	Validation set (50)	82.7	80.2	81.5	(Doniger *et al.* 2002)
CYP3A4 inhibitor	PLS	CATS, TOP, ELE, count, structural, atom types	311	Validation set 1 (50) Validation set 2 (10)	93.1 100.0	85.7 66.7	90.0 90.0	(Zuegge *et al.* 2002)

	ANN	Unity fingerprint	218	Validation set (72)	91.7	88.9	90.3	(Molnar *et al.* 2002)
	Consensus SVM	DRAGON	602	Validation set (100)	92.0	97.3	96.0	(Yap *et al.* 2005a)
CYP2D6 inhibitor	Consensus recursive partitioning	TOP, E-state, physicochemical, fragment keys, 1D similarity scores	100	Validation set (51)	100	76.0	80.0	(Susnow *et al.* 2003)
	Consensus SVM	DRAGON	602	Validation set (100)	90.0	95.0	94.0	(Yap *et al.* 2005a)
CYP2C9 inhibitor	Consensus SVM	DRAGON	602	Validation set (100)	88.9	96.3	95.0	(Yap *et al.* 2005a)
CYP2D6 substrate	Consensus SVM	DRAGON	602	Validation set (100)	98.2	90.9	95.0	(Yap *et al.* 2005a)
CYP3A4 substrate	Consensus SVM	DRAGON	602	Validation set (100)	96.6	94.4	95.0	(Yap *et al.* 2005a)
CYP2C9 substrate	Consensus SVM	DRAGON	602	Validation set (100)	85.7	98.8	97.0	(Yap *et al.* 2005a)
Genotoxic	KNN	TOP, GEO, ELE, PSA	120	Validation set (20)	66.7	92.9	85.0	(Mosier *et al.* 2003)
	Consensus KNN	TOP, GEO, ELE, Q-C, CPSA, H-bond, nitrogen-specific	334	3 fold CV (334)	69.3	74.1	72.2	(Mattioni *et al.* 2003)

	Consensus model (KNN, LDA, PNN)	TOP, GEO, ELE, CPSA, H-bond	227	3 fold CV (227)	73.8	84.3	81.2	(He *et al.* 2003)
	SVM	Simple molecular properties, molecular connectivity and shape, E-state, Q-C, GEO	577	Validation set (123)	77.8	92.7	89.4	(Li *et al.* 2005a)
Torsade de pointes causing compound	SVM	LSER	271	Validation set (78)	97.4	84.6	91.0	(Yap *et al.* 2004)

Abbreviations: **C-SAR** - classification structure-activity relations; **ORMUCS** – ordered multicategorical classification method using the simplex technique; **PLS-DA** – partial least squares-discriminant analysis; **TOPS-MODE** – topological substructural molecular design; **MR** – molar refractivity; **TOP** – topological; **E-state** – electrotopological state indices; **Q-C** – quantum-chemical; **GEO** – geometrical; **CON** – constitutional; **ELE** – electronic; **PSA** – polar surface area; **MW** – molecular weight; **H-bond** – hydrogen bonding capabilities; **CPSA** – charged polar surface area; **LSER** – linear solvation energy relationship; **CV** – cross validation

[a] – number in parenthesis denotes the number of compounds used for model validation.

Table 1.2 Performance of regression-based statistical learning methods for predicting compounds of specific pharmacokinetic or toxicological property.

Property	Activity	Method	Molecular descriptors	Validation method[a]	Reported prediction statistics	Reference
HIA	%FA	MLR	LSER	Training set (38) Validation set (131)	r^2=0.82, q^2=0.77, SE=15, F=53 RMSE=14, MAE=11	(Zhao *et al.* 2001)
			Physicochemical, structural fragment	Training set (417) Validation set (50)	r^2=0.79, SE=12.34, F=38.83 r^2=0.79, SE=12.32	(Klopman *et al.* 2002)
		Sigmoidal	PSA	Training set (20)	r^2=0.94, RMSE=9.2%	(Palm *et al.* 1997)
		PLS	Log P, molecular size, H-bond, counts	Training set (16) Validation set (63)	r^2=0.55, q^2=0.45 RMSE=28.6	(Oprea *et al.* 1999)
			Atom type	Training set (169)	r^2=0.921, q^2=0.787	(Sun 2004)
		ANN	TOP, ELE, GEO, CPSA, H-bond	Training set (67) Validation set (10)	RMSE=0.4, MAE=6.7 RMSE=16.0, MAE=11.0	(Wessel *et al.* 1998)
			CON, TOP, chemical, GEO, Q-C	Training set (67) Validation set (10)	RMSE=0.590 r^2=0.802, RMSE=0.425	(Agatonovic-Kustrin *et al.* 2001)
			TOP	Training set (396) Validation set (185)	r^2=0.92, RMSE=9.1, MAE=7.3 r^2=0.80, RMSE=11.8, MAE=9.8	(Votano *et al.* 2004)

		GRNN	Log P, MR, TOP	Training set (67)	RMSE=6.5	(Niwa 2003)
				Validation set (10)	RMSE=22.8	
	FA	CART	Structural	Training set (899)	AAE=0.120	(Bai *et al.* 2004)
				Validation set 1 (362)	AAE=0.169	
				Validation set 2 (67)	AAE=0.170	
				Validation set 3 (90)	AAE=0.200	
				Validation set 4 (37)	AAE=0.140	
	logit(%FA)	PLS	MolSurf	Training set (13)	r^2=0.903, q^2=0.685, RMSE=0.523	(Norinder *et al.*
				Validation set (7)	RMSE=0.488	1999)
			TOP	Training set (13)	r^2=0.903, q^2=0.818, RMSE=0.523	(Norinder *et al.*
				Validation set (7)	RMSE=0.413	2001)
		SVR	Log P, MR, E-state	Training set	RMSE=0.445, MAE=0.404	(Norinder 2003)
				Validation set	RMSE=0.372, MAE=0.290	
Bioavailability	%F	Regression	Substructure counts	Training set (591)	r^2=0.71, q^2=0.63, RMSE=17.92	(Andrews *et al.*
				2000 runs of 80/20 splits (591)	r^2=0.58, RMSE=20.40	2000)
		MLR	Bulk properties, solubility parameters,	Training set (159)	r^2=0.352, q^2=0.254	(Turner *et al.*

			Q-C, CON, TOP	Validation set (10)	r^2=0.72	2003a)
		ANN	CON, TOP, chemical, GEO, Q-C, bulk properties, solubility parameters	Training set (137) Validation set (15)	r^2=0.736, RMSE=19.21 r^2=0.680, RMSE=20.47	(Turner *et al.* 2004a)
		CODES neural network	CODES	Training set (28)	q^2=0.90	(Dorronsoro *et al.* 2004)
P-gp inhibitor	log(1/EC_{50})	PLS	SIBAR	Training set (100)	r^2=0.731, q^2=0.661	(Klein *et al.* 2002)
BBB penetration	logBB	MLR	MW, log P	Training set (20)	r^2=0.691, SE=0.439, F=40.23	(Young *et al.* 1988)
			LSER	Training set (57)	r^2=0.907, SE=0.197, F=99.2	(Abraham *et al.* 1994)
			Solvation energy	Training set (55)	r^2=0.672, SE=0.41, F=108.3	(Lombardo *et al.* 1996)
			MW, log P	Training set (33)	r^2=0.897, SE=0.126, F=131.1	(Kaliszan *et al.* 1996)
			H-bond	Training set (20)	r^2=0.723, SE=0.0012, F=46.93	(Segarra *et al.* 1999)
			PSA	Training set (45)	r^2=0.841, F=229	(Kelder *et al.*

			1999)
PSA, log P	Training set (55) Validation set 1 (5) Validation set 2 (5)	r^2=0.787, SE=0.354, F=95.8 MAE=0.14 MAE=0.24	(Clark 1999)
PSA	Training set (45)	r^2=0.95	(Ertl *et al.* 2000)
Solvation free energy	Training set (55) Validation set 1 (7) Validation set 2 (5) Validation set 3 (25)	r^2=0.72, SE=0.37 MAE=0.16 MAE=0.14 MAE=0.37	(Keserü *et al.* 2001)
MW, molecular lipoaffinity	Training set (55) Validation set (11)	r^2=0.790, q^2=0.763, SE=0.35, F=97.7 r^2=0.838, SE=0.30	(Liu *et al.* 2001)
LSER	Training set 1 (148) 2 fold CV (148) 5 runs of 80/20 splits (148)	r^2=0.745, q^2=0.711, SE=0.343, F=69 r^2=0.718, SE=0.381 r^2=0.733, SE=0.356	(Platts *et al.* 2001)
Hydrogen bonding, molecular volume, solvent-accessible surface area	Training set (76)	r^2=0.94, SE=0.173, F=311.307	(Kaznessis *et al.* 2001)

	Spatial, structural, thermodynamic	Training set (59) Validation set (12) Validation set (21)	r^2=0.757, q^2=0.701, SE=0.408, F=42.135 RMSE=0.29 RMSE=0.50	(Hou *et al.* 2002)
	E-state	Training set (102) Validation set (20) 5 fold CV (102)	r^2=0.66, q^2=0.62, SE=0.45, F=62.4 RMSE=0.38, MAE=0.32 RMSE=0.47, MAE=0.38	(Rose *et al.* 2002)
	Solute aqueous dissolution and solvation, solute-membrane interaction, general intramolecular solute	Training set (56) Validation set (7)	r^2=0.845, q^2=0.795 RMSE=0.449, MAE=0.398	(Iyer *et al.* 2002)
	Daylight, thermodynamic, spatial, structural, TOP, charge	Training set (48) Validation set (17)	r^2=0.837, q^2=0.786, MAE=0.26, SE=0.19 r^2=0.68, MAE=0.41	(Lobell *et al.* 2003a)
	Hydrophobicity, hydrophilicity, molecular bulkiness	Training set (78) Validation set 1 (13) Validation set 2 (22)	r^2=0.767, q^2=0.736, SE=0.364, F=81.5 r^2=0.88, RMSE=0.26, MAE=0.16 r^2=0.61, RMSE=0.48, MAE=0.39	(Hou *et al.* 2003)
	4D molecular similarity measures	Training set (104) Validation set (46)	r^2=0.69, q^2=0.64 r^2=0.56	(Pan *et al.* 2004)
	Physicochemical, GEO, structural, TOP	Training set (88)	r^2=0.864, q^2=0.847, SE=0.392, F=60.98	(Narayanan *et al.*

			Validation set 1 (13) Validation set 2 (15)	RMSE=0.558, MAE=0.407 RMSE=0.533, MAE=0.437	2005)
	Least-median-of-squares regression		Training set (86)	r^2=0.89, RMSE=0.31	(Cheng *et al.* 2002)
	PCR	Log P, H-bond, PSA	Training set (61) Validation set 1 (14) Validation set 2 (25)	r^2=0.730, q^2=0.688, RMSE=0.424 r^2=0.576, RMSE=0.628 r^2=0.616, RMSE=0.789	(Feher *et al.* 2000)
		Atomic contributions to van der Waals surface area, log P, MR, partial charge	Training set (75)	r^2=0.83, q^2=0.73, RMSE=0.32	(Labute 2000)
	PLS	MolSurf	Training set (28) Validation set 1 (28) Validation set 2 (6)	r^2=0.862, q^2=0.782, RMSE=0.288 RMSE=0.353 RMSE=0.473	(Norinder *et al.* 1998)
		TOP, molecular volume, MW, CON, H-bond	Training set (58) Validation set 1 (12) Validation set 2 (22)	r^2=0.850, q^2=0.752, SE=0.318, F=102 RMSE=0.235 RMSE=0.408	(Luco 1999)
		TOP	Training set (28)	r^2=0.751, q^2=0.696, RMSE=0.368	(Norinder *et al.*

				Validation set (30)	RMSE=0.375	2001)
			Log P, MW, MR, molar volume, H-bond	Training set (19) Validation set (37)	r^2=0.905, q^2=0.791, RMSE=0.287 RMSE=0.338	(Osterberg *et al.* 2001)
			VolSurf	Training set (79)	r^2=0.78, q^2=0.65	(Ooms *et al.* 2002)
			Log P, PSA, E-state	Training set (58) Validation set (39)	r^2=0.846, RMSE=0.308, MAE=0.232 r^2=0.617, RMSE=0.413, MAE=0.499	(Subramanian *et al.* 2003)
			Atom type	Training set (57) Validation set (13)	r^2=0.910, RMSE=0.502 RMSE=0.326	(Sun 2004)
		CODES neural network	CODES	Training set (36)	q^2=0.88	(Dorronsoro *et al.* 2004)
		Bayesian neural net	Property-based, TOP indices, CIMI, atomic charges	Training set (106)	r^2=0.76, q^2=0.65, SE=0.54	(Winkler *et al.* 2004)
		GRNN	DRAGON	Validation set (30)	r^2=0.701, RMSE=0.361	(Yap *et al.* 2005b)
		SVR	Log P, MR, E-state	Training set Validation set	RMSE=0.242, MAE=0.200 RMSE=0.439, MAE=0.298	(Norinder 2003)
HSA binding	log Khsa	MLR	E-state	Training set (84)	r^2=0.77, q^2=0.70, SE=0.29, F=43	(Hall *et al.* 2004)

				10% CV (84)	r^2=0.68	
				Validation set (10)	r^2=0.74, RMSE=0.32, MAE=0.31	
			ELE, TOP, information-content, spatial, structural, thermodynamic	Training set (84)	r^2=0.78, q^2=0.73	(Colmenarejo *et al.* 2001)
				Validation set (10)	r^2=0.88	
		GRNN	DRAGON	Validation set (18)	r^2=0.851, RMSE=0.202	(Yap *et al.* 2005b)
		SVR	CON, TOP, GEO, electrostatic, Q-C	Training set (84)	r^2=0.94, RMSE=0.124	(Xue *et al.* 2004a)
				Validation set (10)	r^2=0.89, RMSE=0.222	
Protein binding	log((1-fu)/fu)	MLR	Log P	Training set (226)	r^2=0.68, MAE=0.45	(Lobell *et al.* 2003b)
				Validation set (94)	r^2=0.51, MAE=0.53	
	%fb	Non-linear regression	Log P	Training set 1 (84)	r^2=0.803, MAE=0.104	(Yamazaki *et al.* 2004)
				Training set 2 (44)	r^2=0.786, MAE=0.055	
				Validation set (23)	r^2=0.830	
	fb	ANN	Atom and functional group counts, connectivity index differences, connectivity index quotients, charge indices, vertex counts, ramifications, Wiener number, MW, Log P	Validation set (6)	r^2=0.745	(Turner *et al.* 2004b)

Milk-plasma ratio	M/P	ANN	CON, TOP, molecular connectivity, GEO, Q-C, physicochemical, liquid properties	Training set (123)	r^2=0.61, RMSE=0.781	(Agatonovic-Kustrin *et al.* 2002)
		GRNN	DRAGON	Validation set (20)	r^2=0.677, RMSE=0.454	(Yap *et al.* 2005b)
Total clearance	CL_{tot}	KNN	TOP, physical properties, partial charge, pharmacophore feature, potential energy	Training set (32) Validation set (6)	q^2=0.77 r^2=0.94	(Ng *et al.* 2004)
		ANN	Atom and functional group counts, connectivity index differences, connectivity index quotients, charge indices, vertex counts, ramifications, Wiener number, MW, Log P	Validation set (6)	r^2=0.731	(Turner *et al.* 2004b)
		GRNN	Lipophilicity, ionization, molecular size, H-bond	Training set (23)	r^2=0.775, q^2=0.731	(Karalis *et al.* 2003)

Abbreviations: **FA** – fraction absorbed; **F** – bioavailability; **BB** – ratio of concentration of drug in brain to concentration of drug in blood; **Khsa** – binding affinity of drug to human serum albumin; **fu** – fraction of drug unbound in plasma; **fb** – fraction of drug bound in plasma; **CART** – classification regression tree; **PCR** – principal component regression; **SIBAR** – similarity based structure activity relationship; **CIMI** – chemically intuitive molecular index; **3DMoRSE** – 3D molecule representation of structures based on electron diffraction; **ATS** – Moreau-Broto autocorrelation; **GETAWAY** - geometry, topology, and atom-weights assembly; **RDF** – radial distribution function; **WHIM** – weighted holistic invariant molecular descriptors

[a] – number in parenthesis denotes the number of compounds used for model validation.

1.1.3 *In silico* methods

There are a number of *in silico* methods that have been used to develop QSPkR/qSPkR models. Traditional statistical methods, such multiple linear regression and partial least squares, have been widely adopted for the development of QSPkR/qSPkR models because they can be easily used and the derived models can be easily interpreted. These methods are highly successful in developing QSPkR/qSPkR models by using small groups of congeneric compounds, which can be used in the modification of drug leads by identifying important physicochemical and structural properties which affect the ADMET properties. Studies have been conducted to apply these methods to develop QSPkR/qSPkR models by using larger and more diverse groups of compounds. The derived QSPkR/qSPkR models usually have lower prediction accuracies than those of QSPkR/qSPkR models developed by using small groups of congeneric compounds (Herman *et al.* 1994). This suggests that multiple mechanisms are involved in determining the ADMET properties of diverse groups of compounds and thus recent studies have explored methods based on non-linear relationships, such as machine learning methods, for constructing QSPkR/qSPkR models (Smith *et al.* 2001a).

Machine learning is the study of computer prediction, classification or analysis algorithms that improve automatically through experience (Mitchell 1997). Machine learning methods have been successfully used in many diverse fields with numerous applications such as pharmacodynamic properties prediction (Czerminski *et al.* 2001; Livingstone *et al.* 2003), protein function prediction (Cai *et al.* 2003), medical decision making (Veropoulos 2001), spam categorization (Drucker *et al.* 1999), detection of oil spills (Kubat *et al.* 1998), and speech recognition (Nuttakorn *et al.* 2001). A reason for the widespread adoption of machine learning methods in different

fields is that they do not make any assumption about the nature of the relationship between the property to be predicted and the factors affecting that property. This enables complex relationships to be modeled accurately and thus improves the prediction accuracies of these models.

'Traditional' machine learning methods, such as artificial neural networks and decision trees, have been explored for the development of QSPkR/qSPkR models for a number of ADMET properties. These include human intestinal absorption (Wessel *et al.* 1998; Bai *et al.* 2004; Wegner *et al.* 2004), bioavailability (Yoshida *et al.* 2000; Pintore *et al.* 2003; Turner *et al.* 2004a), blood-brain barrier penetration (Ajay *et al.* 1999; Winkler *et al.* 2004), milk-plasma ratio (Agatonovic-Kustrin *et al.* 2002), serum protein binding (Gobburu *et al.* 1995; Turner *et al.* 2004b), volume of distribution (Gobburu *et al.* 1995; Turner *et al.* 2004b), P450 isoenzyme substrates and inhibitors (Molnar *et al.* 2002; Susnow *et al.* 2003; Balakin *et al.* 2004), total clearance (Turner *et al.* 2004b), and genotoxicity (Maran *et al.* 2003; Mattioni *et al.* 2003). The prediction or classification accuracies of these QSPkR/qSPkR models are usually better than those of QSPkR/qSPkR models developed by using traditional statistical methods (Manallack *et al.* 1999; Svetnik *et al.* 2005).

1.2 Motivation

There are three main objectives of this work. The first is to improve the quality of previous QSPkR/qSPkR models for ADMET prediction. In this work, four strategies will be used to achieve this objective. The first strategy is to apply newer machine learning methods, such as support vector machine (SVM), support vector regression (SVR), and general regression neural network (GRNN), for the development of QSPkR/qSPkR models. These methods have shown promising potential for predicting pharmacodynamic properties of drugs (Burbidge *et al.* 2001; Czerminski *et al.* 2001; Mosier *et al.* 2002; Huang *et al.* 2003) and it is of interest to compare these newer methods with 'traditional' machine learning methods, such as artificial neural networks and decision trees, for the prediction of ADMET properties. The second strategy is to employ consensus modeling to combine different QSPkR/qSPkR models. There may be several QSPkR/qSPkR models for prediction of a single ADMET property that are developed by using different *in silico* methods or different descriptors. Thus it is of interest to determine if combining these models into a consensus model will improve the overall prediction accuracies for the ADMET property. The third strategy is to use a larger number and more diverse groups of compounds for developing QSPkR/qSPkR models. Some of the previous QSPkR/qSPkR models were built using datasets with a limited number of related compounds and thus may are not suitable for prediction of ADMET properties of diverse groups of compounds. The last strategy is to use compounds with known human ADMET data for developing QSPkR/qSPkR models. The large number of QSPkR/qSPkR models developed by using *in vitro* or animal ADMET data has helped to improve the *in silico* methods and descriptors for developing QSPkR/qSPkR models. However, there are large, non-systemic variations of ADMET properties

across species for individual compounds and thus these previous QSPkR/qSPkR models may not be suitable for prediction of human ADMET properties.

The second objective is to improve on the interpretability of QSPkR models developed by machine learning methods. A common problem with these QSPkR models is that they are often complex with multiple parameters and weights. Thus it is difficult to determine which physicochemical and structural properties of a compound are important in determining its ADMET properties. Hence it will be useful to have a method which can identify the important physicochemical and structural properties.

The last objective of this work is to construct qSPkR models for important ADMET properties which have not received sufficient attention. An example of such ADMET properties is the potential of drug candidates to cause torsade de pointes, which is a rare but serious adverse drug reaction.

1.3 Thesis structure

A QSPkR/qSPkR model consists of three main components: (1) ADMET data, (2) physicochemical and structural descriptions of a compound, and (3) a statistical learning technique to correlate the first two components. In chapter 2, these three components are described and the methods used in this work for developing QSPkR/qSPkR models are given. Methods that are used for checking the validity and usefulness of QSPkR/qSPkR models are also described.

A new machine learning library, YMLL, and a Microsoft Windows software, PHAKISO, is introduced in chapter 3. YMLL contains algorithms that are essential for performing a QSPkR/qSPkR experiment. PHAKISO provides a graphical user interface to the algorithms in YMLL so that a QSPkR/qSPkR model can be developed and validated easily with just a few mouse clicks. Both YMLL and PHAKISO are available freely on the PHAKISO website (http://www.phakiso.com) for non-commercial uses.

The prediction of absorption-related processes, in particular, human intestinal absorption, and p-glycoprotein substrates, is presented in chapter 4. SVM was used to develop classification systems for identifying compounds that are absorbable by human intestine and compounds that are substrates of the p-glycoprotein transporter. The effect of recursive feature elimination (RFE), a method for identifying relevant descriptors, on the classification accuracies of the SVM classification systems is discussed. Analysis of the RFE-selected descriptors and comparison with other classification studies are also presented.

Chapter 5 describes the prediction of a few important distribution processes, such as blood-brain barrier penetration, human serum albumin binding and milk-plasma ratio by using GRNN. The prediction accuracies of the GRNN-developed

models were compared with those of QSPkR models developed by using MLR and MLFN. A new method for interpreting GRNN-developed QSPkR models, which enables relevant physicochemical and structural properties of a compound to be identified, is also introduced.

The use of consensus SVM model strategy to improve the prediction accuracies of substrates and inhibitors of three cytochrome P450 isoenzymes, 3A4, 2D6 and 2C9 is presented in chapter 6. Physicochemical and structural properties of compounds that are important for the identification of substrates and inhibitors and factors that may affect the prediction accuracies are discussed.

Chapter 7 describes three machine learning approaches for the prediction of total clearance. Several different sets of descriptors are compared for their usefulness in modeling total clearance. Important physicochemical and structural properties of a compound are also identified by using the new method that is introduced in Chapter 5.

Chapter 8 describes two important drug toxicities: genotoxicity and torsade de pointes. The classification accuracies of the qSPkR models for prediction of genotoxic potential and torsade-causing potential of compounds developed by using SVM and other classification methods are presented. The possible reasons for misclassification of some compounds are also discussed.

Chapter 9 summarizes the major findings and contributions of this work to the progress of using machine learning approaches for pharmacokinetics and toxicity predictions. Limitations of the present work and possible areas for future studies are also discussed.

Chapter 2

Quantitative/Qualitative Structure Pharmacokinetics Relationship (QSPkR/qSPkR)

A QSPkR/qSPkR model consists of three main components: (1) ADMET data (section 2.2), (2) physicochemical and structural descriptions of a compound (section 2.3), and (3) a statistical learning technique to correlate the first two components (section 2.4). In this chapter, these three components are described and the methods used in this work for developing QSPkR/qSPkR models are given. Methods that are used for checking the validity and usefulness of QSPkR/qSPkR models are also described (section 2.5).

2.1 Introduction

A QSPkR/qSPkR model is a mathematical model which approximates the relationship between an ADMET property of a compound and its structure-derived physicochemical and structural features (Johnson *et al.* 1990). The two main objectives of QSPkR/qSPkR modeling are to allow prediction of the ADMET properties of a not yet biologically tested, but chemically characterized compound and to extract clues as to which molecular characteristics of a compound are important for the ADMET properties. In this work, the term "QSPkR model" is used to refer to quantitative models (regression problems), and the term "qSPkR model" will be used to refer to qualitative models (classification problems).

Figure 2.1 Flowchart showing the various processes during the development of a QSPkR/qSPkR model.

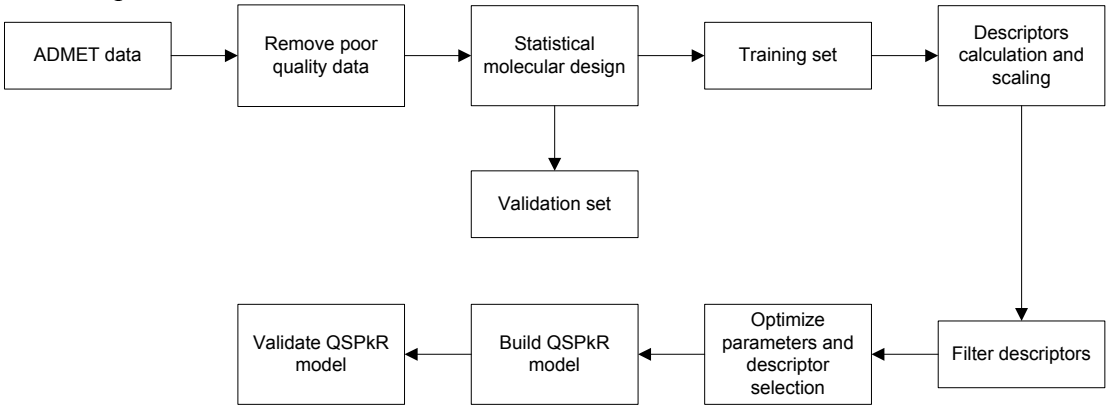


Figure 2.1 shows a basic scheme for developing a QSPkR/qSPkR model. The initial step is the collection of relevant ADMET data and the elimination of low quality data that are likely to affect the quality of the model. The next step is the selection of representative compounds into a training set and a validation set to calibrate and evaluate the QSPkR/qSPkR model respectively. Molecular descriptors are then computed for representing the physicochemical and structural properties of the compounds studied and those that are redundant or contained little information are removed prior to the modeling process. A machine learning method is then used to develop a model that relates the ADMET property to the physicochemical and structural properties of the compounds. During a modeling process, optimization of the essential parameters of the machine learning methods and the selection of relevant descriptor subsets are conducted simultaneously. The optimum set of parameters and descriptor subset are used to construct a final QSPkR/qSPkR model, which is subsequently subjected to various validation methods to ensure that it is valid and useful.

2.2 Dataset

2.2.1 Quality analysis

The development of reliable QSPkR/qSPkR models depends on the availability of high quality ADMET data that have low experimental errors (Cronin *et al.* 2003). Ideally, these ADMET properties should be measured by a single protocol so that different compounds can be reliably compared with each other. However, human ADMET data have been determined only for a limited number of compounds and these data are rarely determined by the same protocol. Thus data selection has been primarily based on such considerations as the comparison of the data of the compounds commonly studied by different protocols, and the incorporation of additional experimental information.

In this work, several methods are adopted to ensure that interlaboratory variations in experimental protocols do not significantly affect the quality of the training sets. The sources for the ADMET data for each compound were investigated to ensure that there were no wide variations in experimental protocol from those of the majority of the compounds in the training set. Compounds that were investigated in more than one source are used to estimate the quality of each source. It is assumed that sources which give ADMET data that are closer to the median of the values from the different sources are more accurate. In classification problems, the most common range of the ADMET data for the compounds investigated in more than one source was used to select compounds for the different classes (Susnow *et al.* 2003).

2.2.2 Statistical molecular design

2.2.2.1 Introduction

The use of an external independent validation set, which has been collected independently of the training set, is widely regarded as the best way to assess the quality of a QSPkR/qSPkR model (Wold *et al.* 1995) (details on model validation will be described in section 2.5). However, it is usually difficult to find additional sources of ADMET data to construct an independent validation set and thus the typical method is to split the original dataset into two different sets, a training set for developing the QSPkR/qSPkR model and a validation set for evaluating the model performance (Gramatica *et al.* 2004). The training set should contain compounds of diverse structures that can adequately represent all of the compounds that possess a particular ADMET property (Rajer-Kanduc *et al.* 2003; Schultz *et al.* 2003). The validation set also needs to be sufficiently diverse and representative of the compounds studied in order to accurately assess the accuracies of the QSPkR/qSPkR models (Rajer-Kanduc *et al.* 2003; Schultz *et al.* 2003).

There are a number of approaches for creating diverse training sets and representative validation sets from the datasets, which are given in Table 2.1. These include random selection, cluster-based methods, dissimilarity-based methods, cell-based methods, stochastic techniques, statistical experimental designs and neural networks (Daszykowski *et al.* 2002; Leach *et al.* 2003). Studies have shown that dissimilarity-based methods, such as Kennard and Stone algorithm and removal-until-done algorithm, are more effective than other algorithms in selecting diverse training sets and representative validation sets for developing and validating QSPkR/qSPkR models (Snarey *et al.* 1997; Rajer-Kanduc *et al.* 2003). Thus these two methods are used in this work to select training and validation sets.

Table 2.1 Methods for selecting training and validation sets

Hierarchical	*Non-hierarchical*
Cluster-based methods	
Single linkage (Leach *et al.* 2003)	K-means (Forgy 1965)
Complete linkage (Leach *et al.* 2003)	Jarvis-Patrick clustering (Jarvis *et al.* 1973)
Group average (Leach *et al.* 2003)	DBSCAN (Ester *et al.* 1996)
Wards method (Leach *et al.* 2003)	OPTICS (Ankrest *et al.* 1999)
Centroid method (Leach *et al.* 2003)	DENCLUE (Han *et al.* 2001)
Median method (Leach *et al.* 2003)	
Dissimilarity-based methods	
MaxSum (Snarey *et al.* 1997)	OptiSim (Clark 1997)
Kennard and Stone algorithm (Kennard *et al.* 1969)	IcePick (Mount *et al.* 1999)
Removal-until-done (Hobohm *et al.* 1992)	Minimum spanning tree error function (Waldman *et al.* 2000)
Sphere exclusion (Hudson *et al.* 1996)	
Cell-based methods	
Cummins algorithm (Cummins *et al.* 1996)	
Menard algorithm (Menard *et al.* 1998)	
Uniform cell coverage (Lam *et al.* 2002)	
Stochastic techniques	
Techniques using Monte Carlo sampling (Agrafiotis 1996; Hassan *et al.* 1996)	
Techniques using genetic algorithms (Sheridan *et al.* 2000; Gillet *et al.* 2002)	
Statistical experimental designs	
D-optimal design (Mitchell 1974)	
Factorial design (Box *et al.* 1978)	
Others	
Random selection	
Kohonen's self-organizing map	
Informative design (Miller *et al.* 2002)	

2.2.2.2 Kennard and Stone algorithm

Two compounds with the largest Euclidean distance apart were initially selected for the training set. The remaining compounds for the training set were selected by maximizing the minimum distances between the compounds in the training set and the rest of the compounds in the dataset. This selection process continues until the desired number of compounds was selected for the training set. The remaining compounds in the dataset will be used as the validation set (Kennard *et al.* 1969).

2.2.2.3 Removal-until-done algorithm

Compounds are sequentially removed from the dataset in pairs and placed in the training and validation sets until a defined similarity threshold or desired number of compounds was selected for the validation set. The selection of the compounds to be removed was based on their distribution in the chemical space. Here, chemical space is defined by the structural and chemical descriptors used to represent a compound and each descriptor value is a point in a multidimensional space. Each compound occupies a particular location in this chemical space. All possible pairs of the compounds in the dataset were generated and a similarity score was computed for each pair. These pairs were then ranked in terms of their similarity scores, based on which compounds of similar structural and chemical features were evenly assigned into the training and validation sets. For those compounds without enough structurally and chemically similar counterparts, they were assigned to the training set.

2.2.3 Diversity and representativity of datasets

The diversity of a dataset can be estimated by a diversity index (DI) which is the average value of the similarity between all of the pairs of compounds in that dataset (Perez 2005):

$$DI = \frac{\sum_{i=1}^{n} \sum_{j=1, i \neq j}^{n} sim(i,j)}{n(n-1)} \quad (2.1)$$

where *sim(i,j)* is a measure of the similarity between compound *i* and *j*, and *n* is the number of compounds in a dataset. The diversity of a dataset increases with decreasing DI. The similarity between two compound *i* and *j* is commonly described by the Tanimoto coefficient (Potter *et al.* 1998; Willett *et al.* 1998; Molnar *et al.* 2002):

$$sim(i,j) = \frac{\sum_{d=1}^{p} x_{di} x_{dj}}{\sum_{d=1}^{p} (x_{di})^2 + \sum_{d=1}^{p} (x_{dj})^2 - \sum_{d=1}^{p} x_{di} x_{dj}} \quad (2.2)$$

where *p* is the number of descriptors of the compounds in the dataset. The mean maximum Tanimoto coefficient of the compounds in dataset A and those in dataset B can be used as a representativity index (RI) to measure the extent to which dataset B is representative of dataset A. Dataset B is more representative of dataset A if the RI value between dataset A and B is higher.

2.3 Molecular descriptors

2.3.1 Types

A descriptor is “the final result of a logical and mathematical procedure which transforms chemical information encoded within a symbolic representation of a

compound into an useful number or the result of some standardized experiment" (Todeschini *et al.* 2000). There are currently over 3,700 types of descriptors, which are classified into three broad categories: 1-, 2- and 3-D descriptors that encode chemical composition, topology, and 3D shape and functionality respectively (Todeschini *et al.* 2000; Farnum *et al.* 2003). A descriptor can be simple, like molecular volume, which encode only one feature of a compound, or can be complex, like 3D-MoRSE, which encode multiple physicochemical and structural properties of a compound. Several computer programs have been developed for deriving molecular descriptors of a compound. Examples of the most popularly used and internet accessible programs are DRAGON (Todeschini *et al.* 2005), Molconn-Z (Hall *et al.*), JOELib (Wegner 2005), and MODEL (Li *et al.* 2005b). Table 2.2 below lists some of the common types of descriptors used in QSPkR/qSPkR studies.

Table 2.2 Common descriptors used in QSPkR/qSPkR studies

Constitutional	**Hydrophobic**
Functional groups	Aromaticity indices (Randic 1975)
Molecular weight	Hansch substituent constant (Fujita *et al.* 1964)
Simple counts e.g. number of atoms, bonds, rings	Log D
Topological	Log P
Atom-pairs (Carhart *et al.* 1985)	**Steric**
Balaban index (Balaban 1986)	Charton steric parameter (Charton 1975)
BCUT (Pearlman *et al.* 1999)	Molar refractivity (Pauling *et al.* 1945)
Information content indices (Basak *et al.* 1983)	Parachor (McGowan 1963)
Kappa shape indices (Kier 1997)	Taft steric parameter (Taft 1952)
Kier and Hall connectivity indices (Kier *et al.* 1986)	**Quantum chemical (Karelson *et al.* 1996)**
Kier flexibility index (Kier 1990)	Charges
Kier shape indices (Kier 1990)	HOMO and LUMO energies

Molecular walk counts (Rücker *et al.* 1993)	Orbital electron densities
Randic indices (Randic 1991)	Superdelocalizabilities
Wiener index (Nikolic *et al.* 1995)	Atom-atom polarizabilities
Geometric	Molecular polarizabilites
Gravitation index (Katritzky *et al.* 1996)	Dipole moments and polarity indices
Molecular surface area	Energies
Molecular volume (Higo *et al.* 1989)	**Combination**
Shadow indices (Rohrbaugh *et al.* 1987)	3D-MoRSE (Schuur *et al.* 1996)
Solvent accessible molecular surface area	Electrotopological state indices (Kier *et al.* 1999)
Electrostatic	GETAWAY (Consonni *et al.* 2002)
Charged polar surface area (Stanton *et al.* 1990)	LSER (Platts *et al.* 1999)
Galvez topological charge indices (Galvez *et al.* 1994)	MolSurf (Sjoberg 1997)
Hydrogen bonding capacities	Moreau-Broto topological autocorrelation (Moreau *et al.* 1980)
Maximum and minimum partial charges (Kirpichenok *et al.* 1987)	Randic molecular profiles (Randic 1995)
Molecular polarizabilities (Dewar *et al.* 1984)	RDF (Hemmer *et al.* 1999)
Fingerprints	VolSurf (Cruciani *et al.* 2000b)
Daylight (Craig *et al.* 2005)	WHIM (Bravi *et al.* 1997)
MDL keys (Durant *et al.* 2002)	
UNITY (Patterson *et al.* 1996)	

In this work, descriptors were computed from the 3D structure of the compounds. The 2D structure of each of the compounds studied was generated by using DS ViewerPro 5.0 (Accelrys 2005), which was subsequently converted into 3D structure by using CONCORD (Pearlman). The 3D structure of each compound was manually inspected to ensure that the chirality of each chiral compound is properly represented after the CONCORD's transformation. All salts and elements, such as sodium or calcium, were removed prior to descriptor calculation.

2.3.2 Scaling

Molecular descriptors are usually scaled before they are used for QSPkR/qSPkR modeling. This is to ensure that all descriptors have equal potential to affect the QSPkR/qSPkR model (Livingstone 1995b). There are four main types of descriptor scaling: autoscaling (Livingstone 1995a), range scaling (Livingstone 1995a), feature weighting (Livingstone 1995a) and Pareto scaling (Eriksson *et al.* 2001a). Autoscaling and range scaling are the two most common types of descriptor scaling methods used in QSPkR/qSPkR modeling.

2.3.2.1 Autoscaling

In autoscaling, the mean is subtracted from the descriptor values and the resultant values are divided by the standard deviation:

$$X'_{ij} = \frac{X_{ij} - \bar{X}_j}{\sigma_j} \qquad (2.3)$$

where X'_{ij} is the new scaled value for descriptor *j* of compound *i* and $\bar{X}_j$ and σ_j are the mean and standard deviation of descriptor *j* respectively. The autoscaled descriptors have a mean of zero and a standard deviation of one. The advantage of autoscaling is that it is less susceptible to effects of compounds with extreme values because they are mean centred. In addition, variance of one is useful in variance-related methods since they each contribute one unit of variance to the overall variance of a dataset.

2.3.2.2 Range scaling (Normalization)

In range scaling, the minimum value of the descriptor is subtracted from the descriptor values and the resultant values are divided by the range:

$$X'_{ij} = \frac{2(X_{ij} - X_{j,\min})}{X_{j,\max} - X_{j,\min}} - 1 \qquad (2.4)$$

where $X_{j,min}$ and $X_{j,max}$ are the minimum and maximum value of descriptor j respectively. The range-scaled descriptors have a minimum and maximum value of -1 and 1 respectively. Range scaling can be carried out over any preferred range by multiplication of the range-scaled values by a factor. The disadvantage of range scaling is that it is dependent on the minimum and maximum values of the descriptors, which makes it very sensitive to outliers.

2.3.3 Selection

The purpose of descriptor selection is to remove descriptors irrelevant or negligible to an ADMET property of the compounds, so as to improve computation speed, performance and interpretability of predictive models. Irrelevant and redundant descriptors are removed either by using a filter or a wrapper approach or a combination of these approaches. The filter approach is independent of the *in silico* method and is frequently used to remove redundant descriptors or descriptors of low information content. Descriptors are chosen or removed based on one or more of the following considerations: prior knowledge of factors affecting a particular ADMET property, the properties of the descriptors (e.g. variance), the correlation between different descriptors, and the distribution of the descriptor values in different data classes. In the wrapper approach, a descriptor selection algorithm is incorporated into an *in silico* classification method (Guyon *et al.* 2003).

In many cases, it is difficult to uniquely select an optimum set of descriptors due to the high redundancy and overlapping of many descriptors (Gramatica *et al.* 2004). Separate sets of descriptors containing different members of redundant descriptor classes have been found to give similar prediction accuracies (Izrailev *et al.* 2004). The interpretation of the prediction results in these cases should be more appropriately conducted at the descriptor class level where redundant and overlapping descriptors are grouped into one class. Table 2.3 gives a list of the common descriptor selection methods used in QSPkR/qSPkR studies.

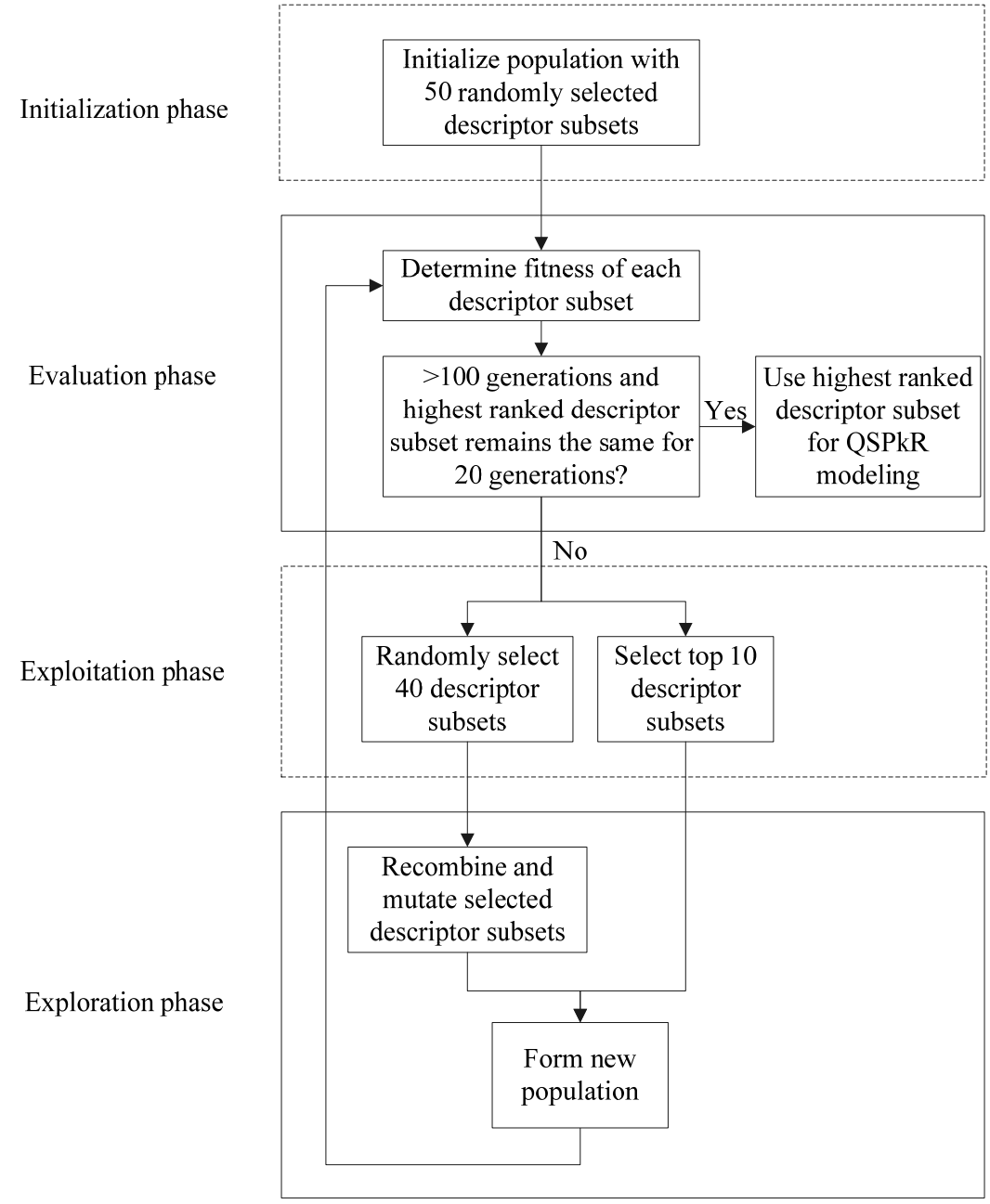
Table 2.3 Common descriptor selection methods used in QSPkR studies

Filter methods	**Wrapper methods**
Remove descriptors with low variance	Forward selection (Xu *et al.* 2001)
Remove highly correlated descriptors	Backward elimination (Xu *et al.* 2001)
CORCHOP (Livingstone *et al.* 1989)	Stepwise regression (Xu *et al.* 2001)
Decision tree (Cardie 1993)	Branch and bound (Narendra *et al.* 1977)
FOCUS (Almuallim *et al.* 1994)	Floating search (Pudil *et al.* 1994)
LVF (Brassard *et al.* 1996)	Adaptive floating search (Somol *et al.* 1999)
RELIEF (Kononenko 1994)	Oscillating search (Somol *et al.* 2000)
Discrimination scores (Guyon *et al.* 2002)	Tabu search (Glover 1989)
Information gain (Liu 2004)	Simulated annealing (Sutter *et al.* 1993)
Mutual information (Liu 2004)	Genetic algorithm (Siedlecki *et al.* 1989)
χ^2-test (Liu 2004)	Recursive feature elimination (Guyon *et al.* 2002)
Odds ratio (Liu 2004)	
GSS coefficient (Liu 2004)	

2.3.3.2 Genetic algorithm-based descriptor selection

The scheme for the genetic algorithm-based descriptor selection method used in this work is shown in Figure 2.2. It comprises of four phases: initialization, evaluation, exploitation and exploration. The initialization phase involves constructing an initial population of 50 randomly selected descriptor subsets. During the evaluation phase, each descriptor subset is evaluated by calculating its fitness score, which indicates the relevance of a descriptor subset to the ADMET property. In the exploitation phase, the descriptor subsets were first ranked by their fitness value. The higher ranked descriptor subsets were given a higher probability of being chosen for reproduction. The top 40 selected descriptor subsets were then used to replace the 40 lowest ranking descriptor subsets in the population. These 40 new descriptor subsets, together with the 10 highest ranked descriptor subsets in the current generation, form a new generation of descriptor subsets. In the last phase, which is the exploration phase, the 40 new descriptor subsets were subjected to one point crossover and mutation to increase the diversity of the population. In the mutation process, descriptors might be randomly added to or deleted from a descriptor subset. After the exploration phase, the genetic algorithm returns to the evaluation phase and the cycle repeats until at least 100 generations have passed and the highest ranked descriptor subset remains the same for 20 generations. The highest ranked descriptor subset was used to construct the final QSPkR/qSPkR model.

Figure 2.2 Schematic diagram of the genetic algorithm-based descriptor selection method



2.3.3.3 Recursive feature elimination (RFE)

It has been suggested that the ranking criterion for descriptor selection can be formulated from the variation in an objective function upon removing each descriptor (Kohavi *et al.* 1997). In order to improve the efficiency of support vector machine (SVM) training, this objective function is represented by a cost function *J* for

the *i*-th descriptor and it is computed by using the training set only. When the *i*-th descriptor is removed or its weight w_i is reduced to zero, the variation of the cost function *DJ(i)* is given by

$$DJ(i) = \frac{1}{2}\frac{\partial^2 J}{\partial w_i^2}(Dw_i)^2 \qquad (2.5)$$

The case of $Dw_i = w_i - 0$ corresponds to the removal of descriptor *i*.

Guyon *et al* have used RFE to reduce the descriptors of a linear SVM classification system for cancer detection from gene selection data (Guyon *et al.* 2002). In the corresponding linear SVM classifier, the cost function is

$$J = \frac{1}{2}\|\mathbf{w}\|^2 - \boldsymbol{\alpha}^T\mathbf{l} \qquad (2.6)$$

where **l** is an *m* dimensional identity vector (*m* is the number of compounds in the training set). Therefore *DJ(i)* = (1/2) w_i^2 and w_i^2 can be used as a descriptor ranking criterion. Yu *et al* have used RFE to reduce the descriptors of a non-linear SVM classification system of polynomial kernels for prediction of drug activity (Yu *et al.* 2003). However, because of the diversity and complexity of the compounds to be classed, the use of linear and polynomial kernels may not always be sufficient for accurate prediction of various pharmaceutical and biological properties. Thus, in this work, SVM classification systems of Gaussian kernels were used. In this case, the cost function to be minimized, under the constraints $0 \le \alpha_k \le C$ and $\sum_k \alpha_k y_k = 0$, is

$$J = \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{H}\boldsymbol{\alpha} - \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1} \qquad (2.7)$$

where **H** is the matrix with elements $y_i y_j \exp(-\|x_i - x_j\|^2/(2\sigma^2))$.

To compute the variation in the cost function upon removal of input component *i*, the parameters αs were kept unchanged and the matrix **H** was re-computed. The resulting ranking coefficient is

$$DJ(i) = \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{H}\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{H}(-i)\boldsymbol{\alpha} \quad (2.8)$$

where **H**(-*i*) is the matrix computed by using the same method as that of matrix **H** but with its *i*-th component removed. One or more of the descriptors with the smallest *DJ(i)* can thus be eliminated.

2.4 Machine learning methods

2.4.1 Methods for classification problems

2.4.1.1 Support vector machine (SVM)

SVM is based on the structural risk minimization principle from statistical learning theory (Vapnik 1995; Burges 1998; Evgeniou *et al.* 2001). A compound is represented by a vector $\mathbf{x}_i$ which is its molecular descriptors. In linearly separable cases, SVM constructs a hyperplane which separates two data classes of compounds with a maximum margin. This is accomplished by finding another vector **w** and a parameter *b* that minimizes $\|\mathbf{w}\|^2$ and satisfies the following conditions:

$$\mathbf{w}\cdot\mathbf{x}_i + b \geq +1, \text{ for } y_i = +1 \qquad \text{Class 1 } (D+) \quad (2.9)$$

$$\mathbf{w}\cdot\mathbf{x}_i + b \leq -1, \text{ for } y_i = -1 \qquad \text{Class 2 } (D-) \quad (2.10)$$

where y_i is the data class index of compound *i*, **w** is a vector normal to the hyperplane, $|b|/\|\mathbf{w}\|$ is the perpendicular distance from the hyperplane to the origin and $\|\mathbf{w}\|^2$ is the Euclidean norm of **w**. After the determination of **w** and *b*, a given compound with vector **x** can be classified by:

$$\hat{y} = sign[(\mathbf{w}\cdot\mathbf{x}) + b] \quad (2.11)$$

In non-linearly separable cases, SVM maps the vectors into a higher dimensional feature space using a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$. Table 2.4 below lists three different types of kernel functions which are commonly used. The Gaussian radial basis function kernel has been extensively used in a number of different studies with good results (Burbidge *et al.* 2001; Czerminski *et al.* 2001; Trotter *et al.* 2001).

Table 2.4 Commonly used kernel functions

Kernel	Equation
Polynomial	$K(\mathbf{x}_i, \mathbf{x}_j) = \left(\mathbf{x}_i \cdot \mathbf{x}_j + 1\right)^p$
Gaussian radial basis function	$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\left\|\mathbf{x}_j - \mathbf{x}_i\right\|^2 / 2\sigma^2}$
Sigmoidal	$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh\left(\kappa \mathbf{x}_i \cdot \mathbf{x}_j - \delta\right)$

Linear support vector machine is applied to this feature space and then the decision function is given by:

$$\hat{y} = sign(\sum_{i=1}^{l} \alpha_i^{\,0} y_i K(\mathbf{x}, \mathbf{x}_i) + b) \tag{2.12}$$

where l is the number of support vectors and the coefficients $\alpha_i^{\,0}$ and b are determined by maximizing the following Langrangian expression:

$$\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \tag{2.13}$$

under the following conditions:

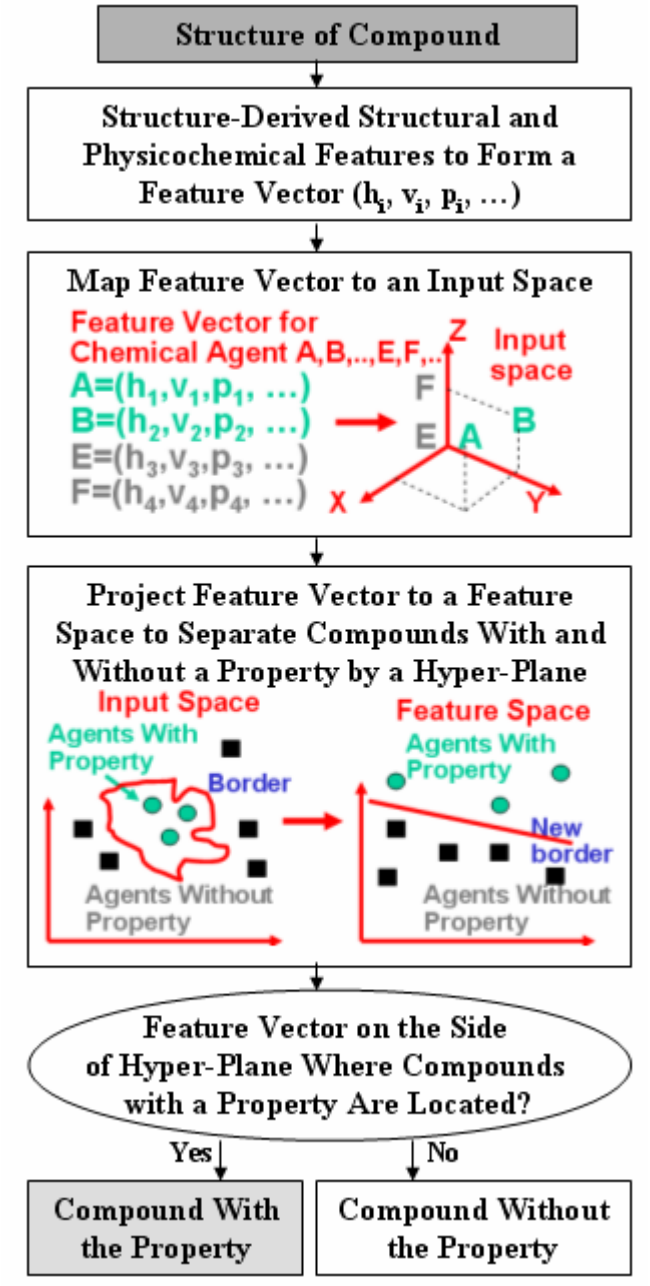
$$0 \le \alpha_i \le C \tag{2.14}$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \tag{2.15}$$

where C is a penalty for training errors. A positive or negative value from equation (2.12) indicates that the compound with vector $\mathbf{x}$ belongs to the positive ($D+$) or

negative data class (*D*–) respectively. Figure 2.3 below shows a schematic diagram illustrating the process of the prediction of compounds with a particular ADMET property from its structure by using SVM.

Figure 2.3 Schematic diagram illustrating the process of the prediction of compounds with a particular ADMET property from its structure by using SVM method. A,B: feature vectors of compounds with the property; E,F: feature vectors of compounds without the property; feature vector (hj, pj, vj,…) represents such structural and physicochemical properties as hydrophobicity, volume, polarizability, etc.



2.4.1.2 Probabilistic neural network (PNN)

PNN was introduced by Specht in 1990 (Specht 1990) and is a form of neural network designed for classification through the use of Bayes' optimal decision rule:

$$h_i c_i f_i(\mathbf{x}) > h_j c_j f_j(\mathbf{x}) \tag{2.16}$$

where h_i and h_j are the prior probabilities, c_i and c_j are the costs of misclassification and $f_i(\mathbf{x})$ and $f_j(\mathbf{x})$ are the probability density function for data class i and j respectively. A given compound with vector **x** is classified into data class i if the product of all the three terms is greater for data class i than for any other data class j not equal to i. In most applications, the prior probabilities and costs of misclassifications are treated as being equal. The probability density function for each data class for a univariate case can be estimated by the Parzen's nonparametric estimator (Parzen 1962):

$$g(x) = \frac{1}{n\sigma}\sum_{i=1}^{n} W(\frac{x - x_i}{\sigma}) \tag{2.17}$$

where n is the sample size, σ is a scaling parameter which defines the width of the bell curve that surrounds each compound, $W(d)$ is a weight function which has its largest value at $d = 0$ and $(x - x_i)$ is the distance between a given compound and a compound in the training set. The Parzen's nonparametric estimator was later expanded by Cacoullos (Cacoullos 1966) for the multivariate case.

$$g(x_1,\ldots,x_p) = \frac{1}{n\sigma_1 \ldots \sigma_p}\sum_{i=1}^{n} W(\frac{x_1 - x_{1,i}}{\sigma_1},\ldots,\frac{x_p - x_{p,i}}{\sigma_p}) \tag{2.18}$$

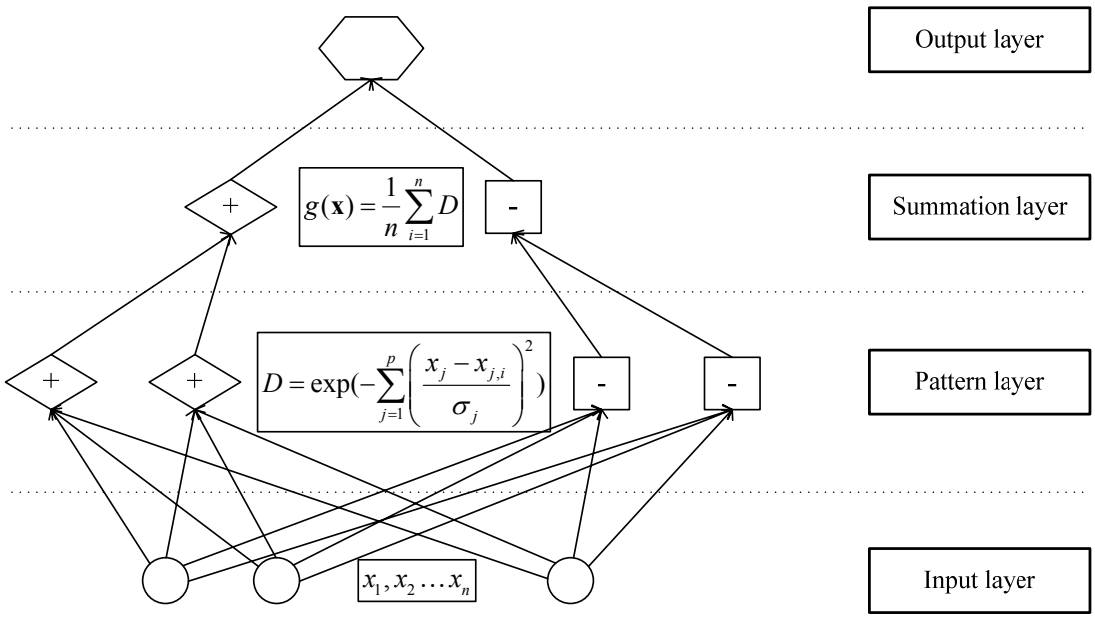
The Gaussian function is frequently used as the weight function because it is well behaved, easily calculated and satisfies the conditions required by Parzen's estimator. Thus the probability density function for the multivariate case becomes

$$g(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \exp(-\sum_{j=1}^{p}\left(\frac{x_j - x_{j,i}}{\sigma_j}\right)^2) \tag{2.19}$$

To simplify the equation, a single σ that is common to all the descriptors (single-sigma model) can be used instead of an individual σ for each descriptor (multi-sigma model). Single-sigma models could be computed faster and can produce reasonable models when all the descriptors are of approximately equal importance. However, multi-sigma models are more general than single-sigma model and are useful when descriptors are of different nature and importance (Masters 1995).

PNN can be implemented as a neural network (Masters 1995), which is shown in Figure 2.4. The network architecture of a PNN is determined by the number of compounds and descriptors in the training set. There are 4 layers in a PNN. The input layer provides input values to all neurons in the pattern layer and has as many neurons as the number of descriptors in the training set. The number of pattern neurons is determined by the total number of compounds in the training set. Each pattern neuron computes a distance measure between the input compound and the training compound represented by that neuron and then subjects the distance measure to the Parzen's nonparameteric estimator. The summation layer has a neuron for each data class and the neurons sum all the pattern neurons' output corresponding to members of that summation neuron's data class to obtain the estimated probability density function for that data class. The single neuron in the output layer then determines the final data class of the input compound by comparing all the probability density functions from the summation neurons and choosing the data class with the highest value for the probability density function.

Figure 2.4 PNN architecture.



2.4.1.3 k nearest neighbour (kNN)

kNN is a basic instance-based method and was introduced by Fix and Hodges (Fix *et al.* 1951). kNN measures the Euclidean distance between a given compound with vector **x** and each compound in the training set with individual vector $\mathbf{x}_i$ (Fix *et al.* 1951; Johnson *et al.* 1982). The Euclidean distances for the vector pairs are calculated using the following formula:

$$D = \sqrt{\left\|\mathbf{x} - \mathbf{x}_i\right\|^2} \tag{2.20}$$

A total of *k* number of training compounds nearest to the given compound is used to determine its data class:

$$\hat{y} = \arg\max_{v \in V} \sum_{i=1}^{k} \delta(v, y_i) \tag{2.21}$$

where $\delta(a,b)$=1 if a=b and $\delta(a,b)$=0 if $a \neq b$, argmax is the maximum of the function, V is a finite set of data classes. k is usually an odd number to prevent ambiguity in the estimation of $\hat{y}$.

2.4.1.4 C4.5 decision tree (DT)

C4.5 DT is a branch-test-based classifier (Quinlan 1993). A branch in a decision tree corresponds to a group of data classes and a leaf represents a specific data class. A decision node specifies a test to be conducted on a single descriptor value, with one branch and its subsequent data classes as possible outcomes of the test. A given compound with vector **x** is classified by starting at the root of the tree and moving through the tree until a leaf is encountered. At each non-leaf decision node, a test is conducted and the classification process proceeds to the branch selected by the test. Upon reaching the destination leaf, the data class of the given compound is predicted to be that associated with the leaf.

The algorithm is a recursive greedy heuristic that selects descriptors for membership within the tree. It uses recursive partitioning to examine every descriptor of the compounds in the training set and rank them according to their ability to partition the remaining compounds, thereby constructing a decision tree. Whether or not a descriptor is included within the tree is based on the value of its information gain. As a statistical property, information gain measures how well the descriptor separate training cases into subsets in which the data class is homogeneous. For descriptors with continuous values, a threshold value had to be established within each descriptor so that it could partition the training cases into subsets. These threshold values for each descriptor were established by rank ordering the values within each descriptor from lowest to highest and repeatedly calculating the information gain using the arithmetical midpoint between all successive values within

the rank order. The midpoint value with the highest information gain was selected as the threshold value for the descriptor. That descriptor with the highest information gain (information being the most useful for classification) was then selected for inclusion in the DT. The algorithm continued to build the tree in this manner until it accounted for all training cases. Ties between descriptors that were equal in terms of information gain were broken randomly (Carnahan *et al.* 2003).

2.4.2 Methods for regression problems

2.4.2.1 Support vector regression (SVR)

The theoretical background of SVR is similar to that of SVM (Smola *et al.*; Vapnik 1995; Yuan *et al.* 2004). In SVR, the kernel function is used to map the vectors into a higher dimensional feature space and linear regression is then conducted in this space. The optimal regression function can be represented by:

$$\hat{y} = \sum_{i=1}^{l} (\alpha^* - \alpha) K(\mathbf{x}, \mathbf{x}_i) + b \qquad (2.22)$$

where $\hat{y}$ represents the predicted value of an ADMET property, and the coefficients α, α^* and bias b are determined by maximizing the following Langrangian expression:

$$-\varepsilon \sum_{i=1}^{n} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{n} (\alpha_i + \alpha_i^*) y_i - \frac{1}{2} \sum_{i,j=1}^{n} (\alpha_i + \alpha_i^*)(\alpha_j + \alpha_j^*)(x_i, x_j) \qquad (2.23)$$

under the following conditions:

$$0 \le \alpha_i, \alpha_i^* \le C \qquad (2.24)$$

$$\sum_{i=1}^{n} (\alpha_i + \alpha_i^*) = 0 \qquad (2.25)$$

2.4.2.2 General regression neural network (GRNN)

GRNN is a modification of PNN for regression problems (Specht 1991). For GRNN, the predicted value of the ADMET property is the most probable value, which is given by

$$\hat{y} = \frac{\int_{-\infty}^{\infty} yf(\mathbf{x}, y)dy}{\int_{-\infty}^{\infty} f(\mathbf{x}, y)dy} \tag{2.26}$$

where *f(***x***,y)* is the joint density and can be estimated by using Parzen's nonparametric estimator (equation (2.17) or (2.18)). Substituting Parzen's nonparametric estimator for *f(***x***,y)* and performing the integrations leads to the fundamental equation of GRNN.

$$\hat{y} = \frac{\sum_{i=1}^{n} y_i \exp(-D(\mathbf{x}, \mathbf{x}_i))}{\sum_{i=1}^{n} \exp(-D(\mathbf{x}, \mathbf{x}_i))} \tag{2.27}$$

where

$$D(\mathbf{x}, \mathbf{x}_i) = \sum_{j=1}^{p} \left(\frac{x_j - x_{j,i}}{\sigma_j} \right)^2 \tag{2.28}$$

The network architecture of a GRNN is similar to that of a PNN except that its summation layer has two neurons that calculate the numerator and denominator of equation (2.27). The single neuron in the output layer then performs a division of the two summation neurons to obtain the predicted ADMET value of the given compound.

2.4.2.3 k nearest neighbour (kNN)

kNN can be modified for regression problems by replacing equation (2.21) with the following equation:

$$\hat{y} = \frac{\sum_{i=1}^{k} y_i}{k} \tag{2.29}$$

The predicted ADMET value of the given compound is the average of the ADMET values of its *k* nearest neighbours. Unlike kNN that is used for classification problems, *k* need not be an odd number in this case.

2.4.3 Optimization of the parameters of machine learning methods

Different machine learning methods have different types of parameters that must be optimized. In this work, SVM and SVR are trained by using a Gaussian kernel function which has an adjustable parameter σ. For PNN and GRNN, the only parameter to be optimized is a scaling parameter, σ. In kNN, the optimum number of nearest neighbours, *k*, needs to be derived for each training set.

Optimization of the parameter for each of these statistical learning methods is conducted by scanning the parameter through a range of values. The set of parameters that produces the best QSPkR/qSPkR model, which is determined by using cross-validation methods, such as 5-fold cross-validation, 10-fold cross-validation or a modeling testing set, is used to construct a final QSPkR/qSPkR model which is then further validated to ensure that it is valid and useful for the ADMET property (see section 2.5).

2.5 Model validation

2.5.1 Performance evaluation of a QSPkR/qSPkR model

One of the objectives of QSPkR/qSPkR modeling is to allow prediction of the ADMET properties of compounds which have not been biologically tested. Thus it is important to determine the ability of the developed QSPkR/qSPkR model to predict the ADMET properties of compounds that are not present in the training set. There are two methods which are commonly used to determine the predictive capability of a QSPkR/qSPkR model (Wold *et al.* 1995). The first method is the use of cross-validation, which includes leave-one-out (LOO) and *k*-fold cross-validation. In LOO, a compound is left out of the training set and the remaining compounds are used to train the machine learning method. The derived QSPkR/qSPkR model is then used to predict the ADMET property of the left-out compound. This process is repeated until every compound in the training set has been left out once. In *k*-fold cross-validation, the training set was randomly divided into *k* mutually exclusive subsets of approximately equal size. *k*-minus-one of the subsets were combined to form a modeling training set for developing a QSPkR/qSPkR model. The remaining subset was used as a modeling testing set to assess the predictive capability of the QSPkR/qSPkR model. This process was repeated until *k* QSPkR/qSPkR models were developed and each subset had been used as a modeling testing set once.

There are reports of the lack of correlation between cross-validation methods and the prediction capability of a QSPkR/qSPkR model (Golbraikh *et al.* 2002; Kozak *et al.* 2003; Reunanen 2003; Olsson *et al.* 2004). Moreover, cross-validation methods have a tendency of underestimating the prediction capability of a QSPkR/qSPkR model, especially if important molecular features are present in only a minority of the compounds in the training set (Mosier *et al.* 2002; Hawkins *et al.* 2004). Thus a model

having low cross-validation results can still be quite predictive (Mosier *et al.* 2002). This lead to some studies which suggests that an independent validation set may provide a more reliable estimate of the prediction capability of a QSPkR/qSPkR model (Wold *et al.* 1995; Golbraikh *et al.* 2002). Despite these disadvantages, cross-validation methods are still useful for assessing QSPkR/qSPkR models during optimization of parameters of machine learning methods and during descriptor selection.

A validation set should ideally be obtained independently of the training set. However, validation sets are usually constructed by using statistical molecular design (section 2.2.2) because of the limited availability of high-quality ADMET data. Regardless of the method used to obtain a validation set, a good validation set should be representative of the training set so that it can properly assess the prediction capabilities of the QSPkR/qSPkR model (Tropsha *et al.* 2003).

2.5.1.1 Methods for measuring predictive capability of qSPkR models

The following statistics are usually calculated to determine the predictive capability of a qSPkR model.

$$\text{Sensitivity (SE)} = \frac{\text{TP}}{\text{TP+FN}} \times 100\% \tag{2.30}$$

$$\text{Specificity (SP)} = \frac{\text{TN}}{\text{TN+FP}} \times 100\% \tag{2.31}$$

$$\text{Overall accuracy (Q)} = \frac{\text{TP+TN}}{\text{TP+FN+TN+FP}} \times 100\% \tag{2.32}$$

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \tag{2.33}$$

where MCC is the Matthews correlation coefficient (Matthews 1975), TP is number of the true positives, TN is the number of true negatives, FP is number of the false positives and FN is the number of false negatives. Sensitivity (SE) and specificity (SP) are the classification accuracies of a qSPkR model for the positive and negative data classes respectively. Overall accuracy (Q) is the classification accuracy of the qSPkR model for both positive and negative data classes. The shortcoming of the overall accuracy is that an imbalance in the data classes may result in a high overall accuracy even if either sensitivity or specificity is low. For example, a qSPkR model which has a sensitivity of 100% and specificity of 0% will have an overall accuracy of 90% for a validation set that have 9 times more compounds of the positive data class than compounds of the negative data class. Thus MCC, which is a weighted measure, is increasingly being used to measure the predictive capability of qSPkR models. A MCC value of 1 indicates that the qSPkR model can predict the data classes of unknown compounds perfectly, a MCC value of 0 is expected for a qSPkR model that is not better than random guessing, and a MCC value of -1 indicates total disagreement between the predicted data classes and the actual data classes. For the above example, MCC will give a value of 0, which is a more accurate representation of the predictive capability of the model.

2.5.1.2 Methods for measuring predictive capability of QSPkR models

The following statistics are commonly calculated to determine the predictive capability of a QSPkR model.

$$r^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \qquad (2.34)$$

$$\text{Mean square error (MSE)} = \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n} \quad (2.35)$$

$$\text{Mean absolute error (MAE)} = \frac{\sum_{i=1}^{n}|\hat{y} - y|}{n} \quad (2.36)$$

$$\text{fold-error of a compound} = \begin{cases} \frac{\hat{y}}{y} & \text{if } \hat{y} > y \\ \frac{y}{\hat{y}} & \text{all others} \end{cases} \quad (2.37)$$

$$\text{Average-fold error} = 10^{\frac{\sum\left|\log\frac{\hat{y}}{y}\right|}{n}} \quad (2.38)$$

The r^2 value measures the explained variance between the predicted and actual ADMET values. The fold-error of a compound measures the degree of overprediction or underprediction for a compound and is useful for identifying chemical structures which are not well-represented by the QSPkR model. The average-fold error avoids the cases in which poor overpredictions are cancelled by equally poor underpredictions. A QSPkR model that predicts a ADMET property perfectly gives an average-fold error of 1 and a model with an average-fold error of less than 2 is considered to be a successful one (Obach *et al.* 1997).

2.5.2 Overfitting

It is not sufficient for a QSPkR/qSPkR model to have good predictive capability. A second requirement for a good quality QSPkR/qSPkR model is that it must not suffer from overfitting. There are two main types of overfitting: (1) using a model that is more flexible than it needs to be and (2) using a model that includes

irrelevant descriptors (Hawkins 2004). There are various methods that can be used to prevent or to check for these two types of overfitting.

A number of different QSPkR/qSPkR models can be developed using machine learning methods of varying complexities. The QSPkR/qSPkR model with the best balance between complexity of the machine learning method used and its predictive capability is the one that is most suitable for predicting the ADMET property of interest. This method prevents the use of a QSPkR/qSPkR model that is more flexible than is necessary.

A frequently used method for checking whether a QSPkR/qSPkR model is overfitted is to compare its prediction capability determined by using cross-validation methods with those determined by using independent validation sets (Hawkins 2004). Even though cross-validation methods tend to give a pessimistic estimate of the predictive capability of a QSPkR/qSPkR model, a model that is not overfitted should not have large differences in the estimates of its predictive capability from cross-validation methods and independent validation sets.

Y-randomization is commonly used to determine the probability of chance correlation during descriptor selection (Manly 1997; Leardia *et al.* 1998). In classification problems, a portion of *D*+ compounds in the training set is randomly exchanged with *D*– compounds in the training set, creating new training sets with false *D*+ and *D*– compounds. For regression problems, the ADMET properties of all the compounds in the training set are randomly rearranged. The machine learning method is trained using this scrambled training set. The randomization is repeated a number of times and prediction capabilities of the new scrambled QSPkR/qSPkR model from each run are compared to that of the original QSPkR/qSPkR model. If the scrambled training set gives significantly lower prediction capabilities than the

original training set, it can be concluded that the original QSPkR/qSPkR model was relevant and unlikely to arise as a result of chance correlation.

In order to determine whether the selected descriptors of the original QSPkR/qSPkR model include those irrelevant for the prediction of an ADMET property, different groups of QSPkR/qSPkR models, each containing different number of descriptors, can be generated by using the descriptor selection method. Each group contains a fixed number of QSPkR/qSPkR models having the same number of descriptors. The prediction capabilities of the QSPkR/qSPkR models in each group are determined and the average prediction capabilities of all the groups are compared and used to determine the optimal number of descriptors for the particular ADMET property. If the optimal number of descriptors coincide with the number of descriptors in the original QSPkR/qSPkR model, the original model is unlikely to contain irrelevant descriptors.

2.5.3 Functional dependence study of QSPkR models

A functional dependence study can provide insights on the type of molecular characteristics that are important for a particular ADMET property and how changes in these molecular characteristics affect the ADMET property. This information is useful for guiding structural changes during computer-aided drug design so that the desired ADMET property can be obtained. It is also useful for validating a QSPkR model. A valid QSPkR model should be consistent with previous findings of important factors that affect the ADMET property.

For QSPkR models developed from linear modeling methods, the descriptors are either positively or negatively correlated to ADMET properties in a linear relationship. In contrast, descriptors in models developed by using machine learning

methods correlate to ADMET properties in a non-linear relationship. Thus these models can potentially provide more information about the relationships between descriptors and ADMET properties.

The relationships between descriptors and ADMET properties can be obtained by using functional dependence plots where the value of a single descriptor is varied through its range, while all other descriptors are held constant at a certain value (Wessel *et al.* 1998). However, QSPkR models usually contain descriptors that are correlated with one another and these intercorrelations can drastically alter the shape of a functional dependence plot if the values of the descriptors that are held constant are changed (Andrea *et al.* 1991). In addition, descriptors may encode multiple physicochemical and structural aspects of the molecule. This makes it difficult to determine the relationship between a specific molecular characteristic and an ADMET property.

In this work, principal component analysis (PCA) is used to overcome both problems. PCA can extract dominant patterns in the descriptor subsets and group similar descriptors under a single principal component (PC). Different PCs encode different molecular characteristics and the orthogonality among the PCs can be exploited to determine the correlation between a molecular characteristic and an ADMET property without the influence of other molecular characteristics. A descriptor may belong to multiple PCs and the explained variations of a descriptor in each PC can be used to determine its level of contribution in the PCs (Eriksson *et al.* 2001b). Artificial testing sets are created to determine the relationship between the PCs and ADMET property. Each artificial testing set contains 1000 artificial compounds and initially used PCs as descriptors. The PC to be evaluated is varied uniformly from -5 to 5 while all of the other PCs are assigned a value of zero. The

loadings derived from PCA are then used to transform the PCs back to the original molecular descriptors. Artificial compounds with molecular descriptors outside the range of the corresponding descriptor in the training set are removed to prevent extrapolation of the model. The values of the ADMET property of the remaining artificial compounds are predicted by using the developed QSPkR models. Functional dependence plots of the ADMET property against the PCs can then be used to find the trends between various molecular characteristics and the ADMET property. In this work, PCA and the transformation of the PCs back to the original molecular descriptors were carried out using the software PHAKISO.

Chapter 3

Machine Learning Library

A new machine learning library, YMLL (section 3.2), and a Microsoft Windows software, PHAKISO (section 3.3), is introduced in this chapter. YMLL contains algorithms that are essential for performing a QSPkR/qSPkR experiment. PHAKISO provides a graphical user interface to the algorithms in YMLL so that a QSPkR/qSPkR model can be developed and validated easily with just a few mouse clicks. Both YMLL and PHAKISO are available freely on the PHAKISO website (http://www.phakiso.com) for non-commercial uses.

3.1 Introduction

One of the fundamental requirements for the conduct of an *in silico* QSPkR/qSPkR experiment is the availability of appropriate software. A good software for QSPkR/qSPkR experiments should possess the following features:

1. Ease of data entry.
2. Containing several common statistical molecular design algorithms so that appropriate training and testing sets can be obtained from the original datasets.
3. Containing several common machine learning methods so that the best machine learning method for developing QSPkR/qSPkR model of a particular ADMET property can be determined.

4. Containing several common descriptor selection methods so that a relevant descriptor subset for a particular ADMET property can be determined.
5. Containing several common methods to validate QSPkR/qSPkR models to ensure that the models are valid and useful.

When this work was started, there is no freely available QSPkR/qSPkR software except a few machine learning freeware available. Two such machine learning software are Torch (Collobert *et al.* 2002) and Weka (Witten *et al.* 2005). Torch is a machine learning library, written in C++, which is under a Berkeley Software Distribution (BSD) licence. Its objective is to apply machine learning algorithms for both static and dynamic problems. There are four important concepts in Torch: DataSet, Machine, Trainer, Measurer. The DataSet produces one training example which is given to a Machine to compute an output by using the Measurer. The Trainer will use the output for tuning the Machine.

Weka is a collection of machine learning algorithms for data mining tasks written in Java. It contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. Weka is open source software issued under the GNU General Public License. Weka is organized in a hierarchy of packages. Each package contains a collection of related classes. There are packages for core components, associations, attribute selection, classifiers, clusterers, estimators, filters, experiments and graphical user interface.

Both Torch and Weka have several disadvantages which make them unsuitable for conducting QSPkR/qSPkR experiments. In Torch, there are no graphical user interface or pre-compiled programs, thus it is not easily usable without additional programming. In addition, Torch only contains a limited number of machines learning methods. Algorithms for statistical molecular design are not

available for both Torch and Weka. Both software also have a limited number of descriptor selection methods, especially wrapper methods, and have a limited number of methods to measure prediction capabilities of QSPkR/qSPkR models. Hence modifications to the two software are needed in order to use them for QSPkR/qSPkR experiments. However, there are two difficulties in modifying the two software. Firstly, the design and naming of Torch's C++ classes are different from the usual QSPkR terminologies, which create a steep learning curve in using the library. Secondly, both software are continuously being improved by their original authors and thus any modifications may become obsolete and become unusable in the newer versions. Hence a new software specifically for conducting QSPkR/qSPkR experiments is needed.

In this work, a machine learning library, YMLL, and a Microsoft Windows software, PHAKISO, were designed and created from scratch to enable QSPkR/qSPkR experiments to be conducted easily. Both YMLL and PHAKISO were coded in C++. Most of the algorithms in YMLL were implemented based on algorithms that were provided in the literatures. The remaining algorithms were implemented by either translation of existing freely available source codes to C++ or creating C++ wrappers around the existing freely available source codes. Table 3.1 lists the different types of machine learning algorithms that were implemented in YMLL, Torch and Weka

Table 3.1 Types of machine learning algorithms in YMLL, Torch and Weka

YMLL	Torch	Weka
1. Multiple linear regression for classification problems	1. Bayes classifier	1. AODE
2. Logistic regression	2. MLP	2. BayesNet
3. Partial least squares for classification problems	3. Speech MLP	3. ComplementNaiveBayes
4. Linear discriminant analysis	4. K-Means	4. NaiveBayes
5. C4.5 decision tree	5. MAP Diagonal GMM	5. NaiveBayesMultinomial
6. C4.5 decision rules	6. MAP HMM	6. NaiveBayesSimple
7. *k* nearest neighbour	7. Speech HMM	7. NaiveBayesUpdateable
8. AnnieNN for classification problems	8. Simple decoder speech HMM	8. LeastMedSq
9. TorchMLP for classification problems	9. KNN	9. LinearRegression
10. Feedforward backpropagation neural network for classification problems	10. Parzen machine	10. Logistic
11. Probabilistic neural network	11. SVM classification	11. MultilayerPerceptron
12. Master's probabilistic neural network	12. SVM regression	12. PaceRegression
13. SVMStar	13. Weighted sum machine	13. RBFNetwork
14. SVM^{light}		14. SMO
15. LibSVM		15. SMOreg
16. SVMTorch		16. SimpleLinearRegression
17. Multiple linear regression		17. SimpleLogistic
18. Principal component regression		18. VotedPerceptron
		19. Winnow
		20. IB1
		21. IBk
		22. KStar
		23. LBR
		24. LWL
		25. AdaBoostM1
		26. AdditiveRegression
		27. AttributeSelectedClassifier
		28. Bagging

19. Partial least squares
20. Continuum power regression
21. Continuum regression
22. AnnieNN
23. TorchMLP
24. Feedforward backpropagation neural network
25. General regression neural network
26. Master's general regression neural network
27. SVM*light* for regression problems
28. LibSVM for regression problems
29. SVMTorch for regression problems
29. ClassificationViaRegression
30. CostSensitiveClassifier
31. CVParameterSelection
32. Decorate
33. FilterClassifier
34. Grading
35. LogitBoost
36. MetaCost
37. MultiBoostAB
38. MultiClassClassifier
39. MultiScheme
40. OrdinalClassClassifier
41. RacedIncrementalLogitBoost
42. RandomCommittee
43. RegressionByDiscretization
44. Stacking
45. StackingC
46. ThresholdSelector
47. Vote
48. FLR
49. HyperPipes
50. VFI
51. ADTree
52. DecisionStump
53. Id3
54. J48
55. LMT
56. M5P
57. NBTree
58. RandomForest

	59. RandomTree
	60. REPTree
	61. UserClassifier
	62. ConjunctiveRule
	63. DecisionTable
	64. NNge
	65. OneR
	66. PART
	67. M5Rules
	68. Prism
	69 Ridor
	70. JRip
	71. ZeroR

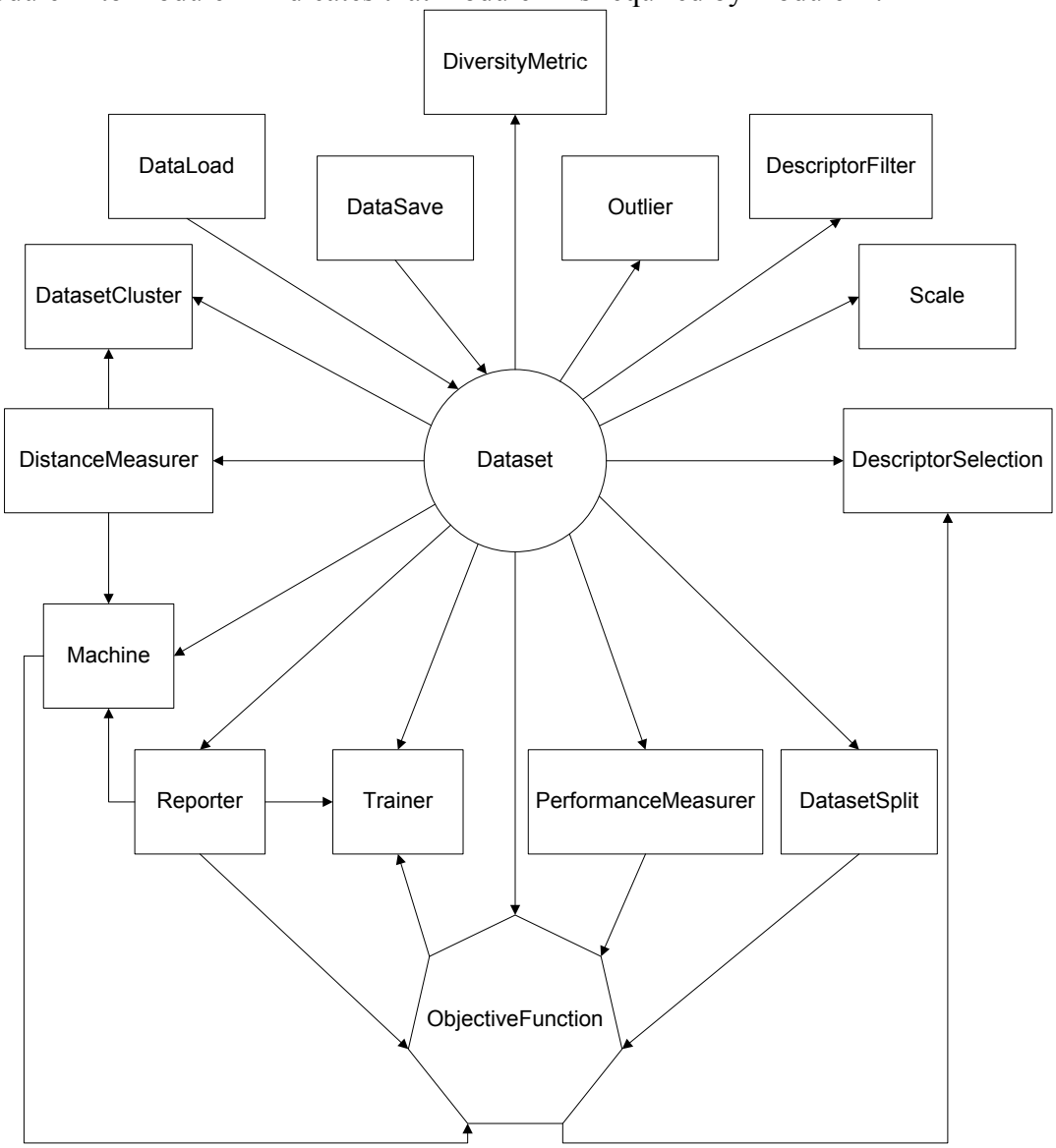
The source codes for both YMLL and PHAKISO are currently not available because of certain proprietary algorithms that were developed by the Bioinformatics and Drug Design (BIDD) group. However, the code and documentation of the header files, precompiled libraries of YMLL for various systems and the executable for PHAKISO are available freely on the PHAKISO website (http://www.phakiso.com) for non-commercial uses. In addition, certain parts of the source codes will be made available in the next release of YMLL and PHAKISO to aid in the development of additional algorithms by other programmers. The following sections describe the design and main features of YMLL and PHAKISO. A more detailed explanation of the usage of both YMLL and PHAKISO is provided on the PHAKISO website.

3.2 YMLL Organization

3.2.1 Overview

YMLL contains different modules which interact with one another to develop a QSPkR/qSPkR model. The modules in YMLL are Dataset, DataLoad, DataSave, DatasetSplit, DatasetCluster, DiversityMetric, Outlier, Machine, DescriptorFilter, DescriptorSelection, Scale, DistanceMeasurer, PerformanceMeasurer, Reporter, ObjectiveFunction and Trainer. Each module defines a standard interface to interact with other modules. The standardization of a module's interface enables different algorithms in the same module to work seamlessly with those in other modules and allow new algorithms to be easily added. The relationships between the different modules are shown in Figure 3.1. For example, to conduct a simple QSPkR/qSPkR experiment, we simply link the Dataset, DataLoad, Machine, and Reporter modules together. These modules will load a dataset into memory and pass to a machine learning algorithm to develop a QSPkR/qSPkR model. The prediction capability of the QSPkR/qSPkR model is then gauged and reported to the user. The programmer can choose different algorithms from the four different modules and the different algorithms are guaranteed to work with one another since they have to conform to the standard interface that is defined by their module.

Figure 3.1 Relationships between the different modules in YMLL. An arrow from module A to module B indicates that module A is required by module B.



3.2.2 Dataset, DataLoad, DataSave, DiversityMetric, DatasetSplit, DatasetCluster, and Outlier

The Dataset module is the most important in YMLL. Its main purpose is to store ADMET properties and descriptors of different compounds and to provide this information to other modules. The Dataset module also contains useful functions for merging of different datasets, removing of a portion of the dataset, changing of

descriptor set, removing of a portion of the descriptors and removing of compounds with the same descriptor values in the dataset.

DataLoad and DataSave modules contain multiple algorithms which enables information to be loaded from and saved to a variety of file formats, which includes comma-separated value (CSV) files, Microsoft Excel files, extensible markup language (XML) files, SVM^{light} (Joachims 1999) files, Torch (Collobert *et al.* 2002) files, and Weka (Witten *et al.* 2005) files.

A DiversityMetric module is available to compute the diversity of the dataset. Three popular diversity measures are provided: mean intermolecular dissimilarity (Perez 2005), average nearest neighbours (Agrafiotis *et al.* 1999) and cumulative property distribution (Agrafiotis 2001).

The aim of the DatasetSplit module is to divide a dataset into smaller portions. These smaller datasets can be used as training sets to train the machine learning method, or as testing sets to aid in the optimization of the descriptor subsets or machine learning parameters, or as validation sets to assess the prediction capability of the final QSPkR/qSPkR models. Currently, the dataset can be divided using simple methods like random selection and select every *N* compound, or using various statistical molecular design algorithms like Kennard and Stone (Kennard *et al.* 1969), sphere exclusion (Hudson *et al.* 1996), removal-until-done (Hobohm *et al.* 1992), and D-optimal design (Mitchell 1974), or using cross-validation methods like leave-one-out, *k* fold cross-validation, and bootstrap.

The DatasetCluster module can be used to separate the dataset into different clusters based on either hierarchical or non-hierarchical methods. Hierarchical methods include single linkage, complete linkage, group average, Wards, centroid and

median (Leach *et al.* 2003). Non-hierarchical methods include k-means (Forgy 1965) and a method proposed by Butina (Butina 1999).

The Outlier module is used to detect and remove outliers from a dataset. Presently, there is an algorithm proposed by Hadi (Hadi 1992), and three other algorithms proposed by Lu et al (Lu *et al.* 2003).

3.2.3 Machine

The Machine module contains various machine learning algorithms for both classification and regression problems. For classification problems, these include Bayes linear discriminant analysis (Tabachnick *et al.* 2000), logistic regression (Tabachnick *et al.* 2000), C4.5 decision tree (Quinlan 1993), C4.5 decision rules (Quinlan 1993), *k* nearest neighbours (Fix *et al.* 1951), probabilistic neural networks (Specht 1990), and support vector machine (Vapnik 1995). The C4.5 decision tree and C4.5 decision rules is a translation of the original Quinlan source codes from C to C++. Two different versions of probabilistic neural networks were provided. One is the implementation based on the algorithm that is provided in the literature and the other is a C++ wrapper for the source codes provided by Masters (Masters 1995). There are three four different versions of the support vector machine. The first is SVMStar, which is developed by the BIDD group and the rest are C++ wrappers for SVM^{light} (Joachims 1999), LibSVM (Chang *et al.* 2001) and SVMTorch (Collobert *et al.* 2002).

For regression problems, there are multiple linear regression (Tabachnick *et al.* 2000), principal component regression (Tabachnick *et al.* 2000), partial least squares (Geladi *et al.* 1986), continuum regression (de Jong *et al.* 2001), feedforward backpropagation neural network (Welstead 1994), general regression neural

network (Specht 1991) and support vector regression (Vapnik 1995). There are three different versions of feedforward backpropagation neural network. One is an implementation based on the algorithm that is provided in the literature and the other two are C++ wrappers for Annie (Shankar *et al.* 2004) and TorchMLP (Collobert *et al.* 2002).. There are two versions of the general regression neural network. The first is the implementation based on the algorithm that is provided in the literature and the second is a C++ wrapper for the source codes provided by Masters (Masters 1995). There are three different version of support vector regression. These are basically C++ wrappers for SVM^{light} (Joachims 1999), LibSVM (Chang *et al.* 2001) and SVMTorch (Collobert *et al.* 2002).

3.2.4 DescriptorFilter, DescriptorSelection, Scale

The DescriptorFilter and DescriptorSelection modules, which are used for descriptor selection, contain filter and wrapper algorithms respectively. For filter methods, there are CORCHOP (Livingstone *et al.* 1989), discrimination score (Guyon *et al.* 2002) and RELIEFF (Kononenko 1994). For wrapper methods, there are forward selection (Xu *et al.* 2001), backward elimination (Xu *et al.* 2001), stepwise regression (Xu *et al.* 2001), sequential floating forward selection (Pudil *et al.* 1994), generalized simulated annealing (Sutter *et al.* 1993), reverse elimination method of tabu search (Glover 1989), genetic algorithm (Siedlecki *et al.* 1989) and recursive feature elimination (Guyon *et al.* 2002). These wrapper methods are commonly used to select relevant descriptors (Sutter *et al.* 1993; Kohavi *et al.* 1997; Xu *et al.* 2001; Molina *et al.* 2002; Guyon *et al.* 2003). However, most of these algorithms were not present in Torch or Weka. Thus the implementation of these algorithms in YMLL is necessary to facilitate the development of relevant QSPkR/qSPkR models. All the

implementations of the wrapper methods are based on algorithms that were described in the literature. The genetic algorithm wrapper method is implemented with the help of GALib (Wall 2005).

A Scale module is also provided to enable ease of scaling of descriptors. The types of scaling methods that are available includes autoscaling, range scaling from 0 to 1, range scaling from -1 to 1, natural logarithm scaling, logarithm base 10 scaling, mean scaling and variance scaling.

3.2.5 DistanceMeasurer

The DistanceMeasurer module measures the distance or similarity (dissimilarity) between two compounds. Available distance metrics include Euclidean distance (Willett *et al.* 1998), Manhattan distance (Willett *et al.* 1998), Soergel distance (Willett *et al.* 1998), Gaussian distance (Zaknich 1999), Quadratic distance (Zaknich 1999), Tophat distance (Zaknich 1999) and Triangular distance (Zaknich 1999). Algorithms for similarity measures include Tanimoto coefficient (Willett *et al.* 1998), Dice coefficient (Willett *et al.* 1998), Cosine coefficient (Willett *et al.* 1998) and Pearson correlation coefficient (Weisstein).

3.2.6 PerformanceMeasurer and Reporter

The purpose of the PerformanceMeasurer module is to compute various statistics for assessing the prediction capability of QSPkR/qSPkR models. Statistics that can be computed for classification problems include sensitivity, specificity, concordance, absolute error rate, relative error rate, Matthews correlation coefficient (Matthews 1975) and Cohen Kappa coefficient (Chohan *et al.* 2005). For

regression problems, the following statistics can be calculated: correlation coefficient, coefficient of determination, adjusted coefficient of determination, mean absolute error, mean square error, root mean square error, Spearman rho coefficient, standard deviation, F statistics and average fold error (Obach *et al.* 1997).

The Reporter module is used to provide a report of the prediction capability of the QSPkR/qSPkR models to either the screen or to a file.

3.2.7 Trainer and ObjectiveFunction

The Trainer module is used for optimizing of the parameters for a machine learning method. Currently, the module can only optimize machine learning methods with a single parameter.

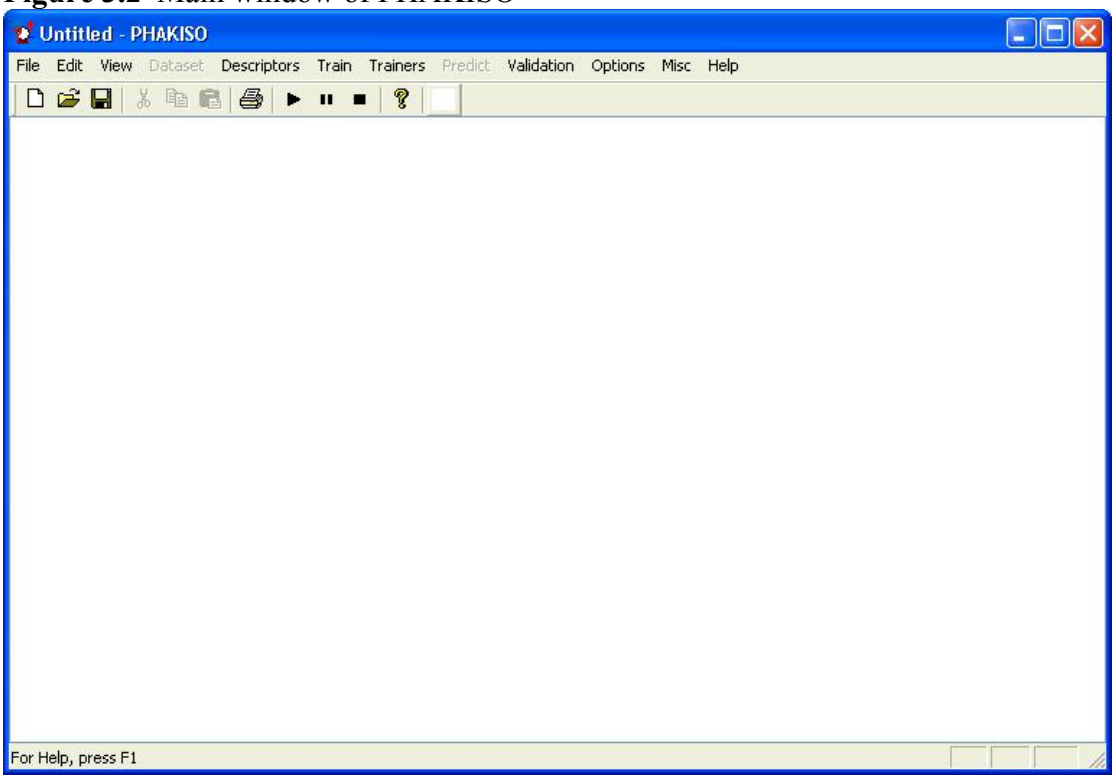
The ObjectiveFunction module is used to provide performance evaluation of a QSPkR/qSPkR model to the DescriptorSelection module or the Trainer module.

3.3 PHAKISO

3.3.1 Introduction

PHAKISO is a Microsoft Windows software, which uses the YMLL library, for performing QSPkR/qSPkR experiments. The aim of PHAKISO is to streamline the development of QSPkR/qSPkR models by offering a graphical user interface (GUI), which is shown in Figure 3.2, to the algorithms that are implemented in the YMLL library. This enables researchers to easily transform their data to a QSPkR/qSPkR model with just a few mouse clicks.

Figure 3.2 Main window of PHAKISO



3.3.2 Features

Table 3.2 lists the standard features of PHAKISO, which are the GUI versions of the algorithms in YMLL library. Table 3.3 lists some additional features of PHAKISO which are not found in the YMLL library.

Table 3.2 Standard features of PHAKISO

Measurement of dataset diversity
Determination of compound clusters in dataset
Determination of outliers in dataset
Statistical molecular design
Y-randomization of dataset
Scaling of descriptors
Objective descriptor selection
Subjective descriptor selection
Construction of a QSPkR/qSPkR model
Optimization of parameters for machine learning methods
Assess prediction capability of QSPkR/qSPkR models on other datasets
Validation of QSPkR/qSPkR models

Table 3.3 Additional features of PHAKISO

Display information on descriptors (mean, standard deviation, minimum and maximum values, etc)
Automatic filling in of values for descriptors with missing values
Principal component analysis

3.3.3 Organization

All the features of PHAKISO are organized into a few menu headings: ‘Dataset’, ‘Descriptors’, ‘Train’, ‘Trainers’, ‘Predict’, ‘Validation’, and ‘Options’. Some of the menu headings are initially disabled and will only be activated when the

features under the menu heading becomes available. For example, the ‘Predict’ menu will only be activated when a QSPkR/qSPkR model has been developed.

3.3.3.1 ‘Dataset’ menu

The ‘Dataset’ menu contains algorithms for diversity measurement, finding clusters of compounds in the dataset, removal of duplicate compounds, removal of outlier compounds, statistical molecular design, y-randomization and calculating basic statistics for the dataset.

3.3.3.2 ‘Descriptor’ menu

The ‘Descriptor’ menu contains algorithms for adding and removing descriptors, calculating correlation among the descriptors, calculating basic statistics for the descriptors, filling in of missing descriptor values, principal component analysis, scaling of the descriptors, objective descriptor selection and subjective descriptor selection.

3.3.3.3 ‘Train’ menu

The ‘Train’ menu contains all the machine learning methods which are available for developing a QSPkR/qSPkR model from a training set. Once a QSPkR/qSPkR models has been developed, the ‘Predict’ menu will be activated.

3.3.3.4 'Trainers' menu

The 'Trainers' menu contains algorithms for determining the optimum parameter values for the machine learning methods. Currently, the algorithms are only able to optimize a single parameter for the machine learning methods.

3.3.3.5 'Predict' menu

The 'Predict' menu contains algorithms for assessing the prediction capability of the developed QSPkR/qSPkR models. The QSPkR/qSPkR models can be used to predict the target property of compounds in the training set, testing set or a validation set.

3.3.3.6 'Validation' menu

The 'Validation' menu contains algorithms for validating the developed QSPkR/qSPkR models. The models can be validated by using cross-validation, bootstrapping, validation set or y-randomization.

3.3.3.7 'Options' menu

The 'Options' menu is used to adjust the parameters for all the machine learning methods. General settings such as the verbosity of the software can also be changed.

Chapter 4

Prediction of Drug Absorption

The prediction of absorption-related processes, in particular, human intestinal absorption (section 4.1), and p-glycoprotein substrates (section 4.2), is presented in this chapter. SVM was used to develop classification systems for identifying compounds that are absorbable by human intestine and compounds that are substrates of the p-glycoprotein transporter. The effect of recursive feature elimination (RFE), a method for identifying relevant descriptors, on the classification accuracies of the SVM classification systems is discussed (sections 4.1.3.1 and 4.2.3). Analysis of the RFE-selected descriptors and comparison with other classification studies are also presented (sections 4.1.3.2, 4.1.3.3 and 4.2.3).

4.1 Human intestinal absorption (HIA)

4.1.1 Introduction

Absorption is defined as the process by which unchanged drug proceeds from site of administration to site of measurement within the body. The oral route is the most convenient and widely used method of drug administration. Thus it is of interest during drug discovery to identify compounds that are suitable for this route of delivery. Drug absorption from the gastrointestinal (GI) tract is complex process. It primarily involves passive transport with a small portion of compounds being absorbed by active transport through various transporters (Pelkonen *et al.* 2001). A large number of factors, which can be classified into three categories, i.e.

physicochemical, physiological, and formulation related, affect GI absorption. Since formulation related factors are usually optimized experimentally while physiological factors cannot be controlled, prediction interests are centered on the extent of absorption as a function of physicochemical properties of the compounds (Boobis *et al.* 2002).

qSPkR models have been developed to determine compounds absorbable (*HIA*+) or nonabsorbable (*HIA*-) by human intestine. The overall accuracies of these ranged from 80.0% to 95.7% (Bergstrom *et al.* 2003; Niwa 2003; Zmuidinavicius *et al.* 2003; Pérez *et al.* 2004). These models employ a variety of molecular descriptors to characterize structural and physicochemical properties of molecules. Some of these descriptors were initially developed for the construction of quantitative structure activity relationship (QSAR) and quantitative structure property relationship (QSPR) of structurally related compounds. Thus these descriptors may not be universally applicable for other compounds or for the prediction of other properties. For instance, descriptors for the QSAR of relatively small sets of related compounds are not applicable for the analysis of chemical diversity (Bayada *et al.* 1999). The use of descriptors unrelated to a particular type of properties or biological activity will generate noise in a machine learning system, which may affect the prediction accuracy of that system (Bayada *et al.* 1999). In some cases, it is difficult to manually select descriptors useful for a particular property. Thus methods capable of automatic selection of molecular descriptors are desirable. The redundancy in molecular descriptors can be partially reduced by means of feature selection methods. It is thus of interest to examine whether feature selection methods can be explored for automatic selection of molecular descriptors and for improvement of the prediction accuracy of ADMET properties by machine learning method.

In this work, recursive feature elimination (RFE) is used as a feature selection method to automatically select molecular descriptors for support vector machine (SVM) prediction of HIA. The computed results are compared to those of earlier studies to examine whether our selected descriptors are capable of giving similar or better classification performance with respect to those derived from a preselected set of descriptors.

4.1.2 Methods

4.1.2.1 Selection of datasets

A "measured absorption rate" of 70% is used as the criterion for dividing compounds into *HIA+* and *HIA-* classes (Zhao *et al.* 2001; Abraham *et al.* 2002). A total of 131 *HIA+* and 65 *HIA-* compounds are collected. In general, a relatively smaller number of compounds with low intestinal absorption is specifically reported in the literature (Klopman *et al.* 2002). Thus, the number of known *HIA+* compounds is expected to be significantly larger than those of *HIA-* compounds.

4.1.2.2 Molecular descriptors

The molecular descriptors used in this work are selected from those commonly used in the literature (Todeschini *et al.* 2000). There are a total of 159 descriptors, given in Table 4.1, which can be divided into five classes based on their properties. These classes are simple molecular properties, molecular connectivity and shape, electrotopological state, quantum chemical properties, and geometrical properties.

Table 4.1 Molecular descriptors and their classes used for human intestinal absorption property prediction[a].

Descriptor class	Number of descriptors in class	Descriptors
Simple molecular properties	18	Molecular weight, Number of ring structures, number of rotatable bonds, number of H-bond donors, number of H-bond acceptors, Element counts
Molecular connectivity and shape	28	Molecular connectivity indices, Valence molecular connectivity indices, Molecular shape Kappa indices, Kappa alpha indices, Flexibility index
Electrotopological state	84	Electrotopological state indices and Atom type electrotopological state indices
Quantum chemical properties	13	Atomic charge on the most positively charged H atom, Largest negative charge on a non-H atom, Polarizability index, Hydrogen bond acceptor basicity (covalent HBAB), Hydrogen bond donor acidity (covalent HBDA), Molecular dipole moment, Absolute hardness, Softness, Ionization potential, Electron affinity, Chemical potential, Electronegativity index, Electrophilicity index
Geometrical properties	16	Molecular size vectors (distance of the longest separated atom pairs, combined distance of the longest separated three atoms, combined distance of the longest separated four atoms), Molecular van der Waals volume, Solvent accessible surface area, Molecular surface area, van der Waals surface area, Polar molecular surface area, Sum of solvent accessible surface areas of positively charged atoms, Sum of solvent accessible surface areas of negatively charged atoms, Sum of charge weighted solvent accessible surface areas of positively charged atoms, Sum of charge weighted solvent accessible surface areas of negatively charged atoms, Sum of van der Waals surface areas of positively charged atoms, Sum of van der Waals surface areas of negatively charged atoms, Sum of charge weighted van der Waals surface areas of positively charged atoms, Sum of charge weighted van der Waals surface areas of negatively charged atoms

[a] The total number of descriptors is 159.

There are 18 descriptors in the class of simple molecular properties, 28 descriptors in the class of molecular connectivity and shape, 84 descriptors in the class of electrotopological state, 13 descriptors in the class of quantum chemical properties, and 16 descriptors in the class of geometrical properties. These descriptors are computed using our own designed molecular descriptor computing program.

4.1.2.3 Computation procedure

The computation procedure used in this work is outlined as follows: The SVM classification system was trained by using a Gaussian kernel function. The training was conducted by sequential variation of the parameter σ in the special region against the whole training set. The prediction accuracy of this SVM system during the training process was evaluated by means of 5-fold cross-validation. In the first step, for a fixed σ, the SVM classifier is trained by using the complete set of descriptors. The second step is to compute the ranking criterion score *DJ(i)* for each descriptor in the current set by using equation (2.8). All of the computed *DJ(i)* values are subsequently ranked in descending order. The third step is to remove the *m* descriptors with smallest criterion scores. In this work, *m* was chosen to be 5, similar to that used in earlier studies (Yu *et al.* 2003). In the fourth step, the SVM classification system is retrained by using the remaining set of descriptors, and the corresponding prediction accuracy is computed by means of 5-fold cross-validation. The first to fourth steps are then repeated for other values of σ. After the completion of these procedures, the set of descriptors and parameter σ that give the best prediction accuracy are selected.

4.1.3 Results and discussion

4.1.3.1 Effect of feature selection on classification accuracy

The prediction accuracies of SVM classification systems using the RFE method (termed as SVM+RFE) and those without using RFE (termed as SVM) were evaluated by means of 5-fold cross-validation method. The computed sensitivity (SE) and specificity (SP) for each fold and the average accuracies of *HIA*+ and *HIA*- compounds as well as the overall prediction accuracy (Q) and Matthews correlation coefficient (MCC) are given in Table 4.2.

Table 4.2 SVM and SVM+RFE prediction accuracy of human intestinal absorption (*HIA*+) and nonabsorption (*HIA*-) of compounds by using 5-fold cross-validation.

Method	**Cross-**	***HIA*+**			***HIA*-**			**Q**	**MCC**
	validation	**TP**	**FN**	**SE (%)**	**TN**	**FP**	**SP (%)**	**(%)**	
SVM	1	22	5	81.5	7	5	58.3	74.4	0.40
	2	18	3	85.7	8	3	72.7	81.3	0.58
	3	37	3	92.5	7	5	58.3	84.6	0.54
	4	16	4	80.0	7	8	46.7	65.7	0.28
	5	18	5	78.3	12	3	80.0	79.0	0.57
	Average			**83.4**			**63.2**	**77.0**	**0.48**
RFE +	1	22	5	81.5	10	2	83.3	82.1	0.61
SVM	2	20	1	95.2	11	0	100.0	96.9	0.93
	3	35	5	87.5	8	4	66.7	82.7	0.53
	4	18	2	90.0	10	5	66.7	80.0	0.59
	5	22	1	95.7	13	2	86.7	92.1	0.83
	Average			**90.0**			**80.7**	**86.7**	**0.70**

The average accuracy for the SVM prediction of *HIA*+ and *HIA*- compounds is 83.4% and 63.2% respectively. By using RFE, the total number of descriptors is

significantly reduced from 159 to 27. The average accuracies for the prediction of HIA are substantially improved by using the reduced set of descriptors. These are 90.0% and 80.7% for *HIA*+ and *HIA*- compounds respectively. Our study seems to suggest that RFE is useful for removing redundant descriptors, which helps to increase the computational efficiency of statistical learning system. RFE is also capable of improving the accuracy of SVM classification of HIA behavior of compounds.

4.1.3.2 Comparison with other classification studies

The effect of feature selection on classification performance can be further evaluated by comparison with other classification studies of the same systems that use preselected descriptors. Direct comparison between our results and those from other studies may not be appropriate because of differences in the use of dataset, descriptors, evaluation, and classification methods. Nonetheless, a tentative comparison may provide some crude estimate regarding the approximate level of accuracy of our method with respect to those obtained by other studies that used more selective descriptors.

The reported accuracies of *HIA*+ predictions are 77%-87% by using partitioned total surface models (Bergstrom *et al.* 2003), 80% by using neural network methods together with 2D topological descriptors (Niwa 2003), and 97% by using structure activity relationship (SAR) models together with physicochemical and structural descriptors (Zmuidinavicius *et al.* 2003). The reported accuracy for *HIA*- prediction is 85% by using SAR models (Zmuidinavicius *et al.* 2003). Our prediction accuracy of 90.0% for *HIA*+ and 80.7% for *HIA*- by using SVM+RFE is thus comparable to the results from these methods that use selective sets of descriptors.

4.1.3.3 RFE selected molecular descriptors

Table 4.3 gives the descriptor classes of the RFE-method-selected descriptors. These descriptors along with their descriptor types are given in Table 4.4. It is found that hydrogen bonding and size are the dominant factors involved in the characterization of HIA property. This finding is consistent with the Lipinski's rule of five (Lipinski *et al.* 1997). In addition, hydrophobic and electrostatic interactions are also found to be important.

Table 4.3 Descriptor classes selected by the RFE method.

Descriptor class	Number of descriptors in descriptor class	Percentage in each class (%)
Electrostatic	4	14.8
Hydrogen bond acceptors	3	11.1
Hydrogen bond donors	6	22.2
Hydrophobic	6	22.2
Size	8	29.6

Table 4.4 Molecular descriptors in the reduced set selected by the RFE method

No	Descriptors	Description	Type
1	S(1)	Atom-type H Estate sum for -OH	Electrotopological state
2	S(5)	Atom-type H Estate sum for > NH	Electrotopological state
3	S(10)	Atom-type H Estate sum for :CH: (sp^2, aromatic)	Electrotopological state
4	S(13)	Atom-type H Estate sum for CH_n (unsaturated)	Electrotopological state
5	S(16)	Atom-type Estate sum for $-CH_3$	Electrotopological state
6	S(20)	Atom-type Estate sum for =CH-	Electrotopological state
7	S(25)	Atom-type Estate sum for =C<	Electrotopological state

8	S(26)	Atom-type Estate sum for : C:-	Electrotopological state
9	S(31)	Atom-type Estate sum for >NH	Electrotopological state
10	S(34)	Atom-type H Estate sum for =N-	Electrotopological state
11	S(35)	Atom-type Estate sum for :N:	Electrotopological state
12	S(39)	Atom-type H Estate sum for –OH	Electrotopological state
13	S(40)	Atom-type H Estate sum for =O	Electrotopological state
14	$^{2}\chi$	Simple molecular connectivity Chi indices for path order 02	Connectivity and shape
15	$^{3}\chi_{C}$	Simple molecular connectivity Chi indices for cluster	Connectivity and shape
16	$^{5}\chi_{CH}$	Simple molecular connectivity Chi indices for cycle of 5 atoms	Connectivity and shape
17	$^{6}\chi_{CH}$	Simple molecular connectivity Chi indices for cycle of 6 atoms	Connectivity and shape
18	$^{3}\chi^{v}_{C}$	Valence molecular connectivity Chi indices for cluster	Connectivity and shape
19	$^{5}\chi^{v}_{CH}$	valence molecular connectivity Chi indices for cycle of 5 atoms	Connectivity and shape
20	$^{6}\chi^{v}_{CH}$	valence molecular connectivity Chi indices for cycle of 6 atoms	Connectivity and shape
21	π_{i}	Polarizability index	Quantum chemical properties
22	ε_{a}	Hydrogen bond donor acidity (covalent HBDA)	Quantum chemical properties
23	A	Electron affinity	Quantum chemical properties
24	dis3	Length vectors (longest distance, longest third atom, 4th atom)	Geometrical properties
25	Sanc	Sum of solvent accessible surface areas of negatively charged atoms	Geometrical properties
26	Sancw	Sum of charge weighted solvent accessible surface areas of negatively	Geometrical properties

		charged atoms	
27	Ndonr	Number of H-bond donors	Simple molecular properties

The RFE selected descriptors describe polar properties, molecular size, cluster connectivity, and various +N–, –OH, and =O electrotopological properties, which are likely to be important for describing passive transport across membranes. These descriptors are primarily uncorrelated to each other. The majority of the descriptors removed by the RFE method, particularly those of electrotopological state, geometrical, and quantum chemical properties, were found to have at least a correlation coefficient of 0.7 to some of the descriptors selected. The rest of the RFE removed descriptors are mostly simple molecular properties (such as molecular weight, the number of specific types of atoms, and the number of rings), geometrical properties (such as molecular volume and surface areas), and connectivity properties (such as index for clusters and paths). These descriptors are not selected because they may not contain as much information as the current descriptor subset for describing the penetration of a compound through the intestinal membrane. For instance, Lipinski's rule of five (Lipinski *et al.* 1997) states that molecular weight is important for the prediction of drug absorption through the intestine. One reason why molecular weight was not selected by the RFE method in this study may be because it does not contain as much information as the current descriptor subset for describing the penetration of a compound through the intestinal membrane. Thus descriptors such as molecular connectivity and length vectors, which encode the shape and size of a molecule and have some degree of correlation with molecular weight, were included instead.

4.1.4 Conclusion

Statistical-learning methods have been developed for facilitating the prediction of pharmacokinetic and toxicological properties of compounds. These methods employ a variety of molecular descriptors to characterize structural and physicochemical properties of molecules. Some of these descriptors are specifically designed for the study of a particular type of properties or compounds, and their use for other properties or compounds might generate noise and affect the prediction accuracy of a statistical learning system. In this work, a feature selection method, RFE, is used to automatically select molecular descriptors for SVM prediction of HIA. RFE significantly reduces the number of descriptors need to develop a qSPkR model for HIA, thereby increasing the computational speed for their classification. The SVM prediction accuracies of HIA are substantially increased by RFE. These prediction accuracies are comparable to those of earlier studies derived from a selective set of descriptors. Our study suggests that molecular feature selection is useful for improving the speed and, in some cases, the accuracy of statistical learning methods for the prediction of pharmacokinetic and toxicological properties of chemical agents.

4.2 P-glycoprotein (P-gp) substrates

4.2.1 Introduction

P-gp, encoded by the highly conserved multidrug resistant (MDR) genes, is an ATP-dependent drug efflux pump which can transport a diverse range of structurally and functionally unrelated substrates across the plasma membrane (van Veen *et al.* 1998; Schmitt *et al.* 2002). Over expression of this protein may result in multidrug resistance and is a major cause of the failure of cancer chemotherapy (Gottesman *et al.* 1996; Ambudkar *et al.* 1999) and diminished efficacy of antibiotics and antiviral compounds (Kim *et al.* 1998; Delph 2000). Two approaches have been explored to circumvent MDR. One is the design of P-gp inhibitors (Klopman *et al.* 1997; Bakken *et al.* 2000) and another is to identify and eliminate drug candidates that are substrates of P-gp in early stage of drug discovery (Bain *et al.* 1997; Litman *et al.* 1997; Seelig 1998; Penzotti *et al.* 2002). Methods that facilitate the identification of P-gp substrates and inhibitors in a cost efficient and fast-speed manner are therefore useful for facilitating drug discovery.

Efforts have been directed at the development of computational methods for P-gp substrate prediction (Bain *et al.* 1997; Litman *et al.* 1997; Seelig 1998; Penzotti *et al.* 2002). Molecular mechanism of P-gp mediated transport is not well understood and the high-resolution structure of P-gp is unavailable (van Veen *et al.* 1998; Schmitt *et al.* 2002). Thus prediction methods are primarily based on statistical models derived from identification of structure-activity relationships (Bain *et al.* 1997; Litman *et al.* 1997), structural recognition elements (Seelig 1998), and multiple pharmacophores (Penzotti *et al.* 2002). In particular, the multiple-pharmacophore model showed promising capability of P-gp substrate prediction for a large variety of

compounds that conform to the known pharmacophores (Penzotti *et al.* 2002), achieving a prediction accuracy of 63 % for a set of 195 compounds. Not all of the pharmaceutically important substrates, agonists and antagonists have available pharmacophore models. Therefore methods that extend the prediction range beyond those compounds covered by known pharmacophore models are desired.

This work explored the use of SVM as a potential tool for the prediction of P-gp substrates. Known P-gp substrates and non-substrates were used for training and testing a SVM classification system for recognition of physicochemical features of P-gp substrates. Through this learning-by-examples process, the trained SVM system can then be used for classifying a chemical compound as either a substrate or a non-substrate of P-gp. The classification accuracy of this system was evaluated by using two methods, an independent set of compounds and 5-fold cross-validation, and it is compared to the 5-fold cross-validation prediction accuracies derived from three other machine learning methods using the same sets of data and molecular descriptors, so as to objectively examine whether SVM is useful for P-gp substrate prediction.

4.2.2 Methods

4.2.2.1 Selection of substrates and non-substrates of P-gp

P-gp substrates were collected from the literature (Seelig 1998; Penzotti *et al.* 2002). Non-substrates of P-gp are those specifically described as not transportable by P-gp. A total of 116 substrates and 85 non-substrates of P-gp were collected. These compounds were further separated into training and testing sets by two different methods. The first method is an independent validation set to evaluate the classification accuracy. The second method is 5-fold cross-validation.

In the first method, these compounds were separated into three sets: training, testing and independent validation set. The training set is used by SVM to develop a statistical model. The testing set is used by SVM to optimize the parameters of SVM classification algorithm and the independent validation set is used for assessing the classification accuracy of the model. These compounds were divided into the three sets by using the removal-until-done method (section 2.2.2.3).

4.2.2.2 Molecular descriptors

This study used the same set of 159 molecular descriptors as the HIA study (section 4.1.2.2). Redundant and un-related descriptors are further reduced by using RFE method with the same computation procedure as the HIA study (section 4.1.2.3).

4.2.2.3 Other statistical classification systems

To objectively examine whether SVM is useful for P-gp substrate prediction, prediction accuracies of the trained SVM system were compared with those derived from three other classification methods by using 5-fold cross-validation. These methods are *k* nearest neighbour (kNN), probabilistic neural network (PNN) and C4.5 decision tree (DT).

4.2.3 Results and discussion

SVM prediction of both substrates and non-substrates of P-gp was evaluated by means of independent validation set and 5-fold cross-validation. The results of these two methods are given in Table 4.5 and Table 4.6 respectively. The accuracy for the prediction of P-gp substrate using 5-fold cross-validation is 81.2% and that by

using independent validation set is 84.2% respective. Thus both methods appear to give consistent assessment about the prediction accuracy. This suggests that the trained SVM system is unlikely to overfit.

Table 4.5 SVM prediction accuracy for the substrates and non-substrates of P-gp by using independent validation sets.

Training set		Testing set				Independent validation set					
		Substrates		Nonsubstrates		Substrates			Nonsubstrates		
		TP	FN	TN	FP	TP	FN	SE (%)	TN	FP	SP (%)
74	68	22	0	12	0	16	3	84.2	4	2	66.7

Table 4.6 SVM prediction accuracy of the substrates and non-substrates of P-glycoprotein by using 5-fold cross-validation.

Cross-validation	Substrates			Non-substrates			Q (%)
	TP	FN	SE (%)	TN	FP	SP (%)	
1	17	7	70.8	12	4	75.0	72.5
2	15	2	88.2	11	5	68.8	78.8
3	30	8	78.9	13	1	92.9	82.7
4	15	4	78.9	15	3	83.3	81.1
5	16	2	88.9	16	5	76.2	82.1
Average			81.2			79.2	79.4
Standard error			7.5			9.2	4.2

A direct comparison with results from previous study is inappropriate because of differences in the use of dataset, molecular descriptors and classification methods. A tentative comparison suggests that our prediction accuracy for P-gp substrates is substantially improved with respect to the value of 63% derived from the ensemble pharmacophore model (Penzotti *et al.* 2002).

The prediction accuracy for non-substrates of P-gp is 79.2% using 5-fold cross-validation and 66.7% using independent validation set. The substantially lower accuracy derived from the independent validation set likely arises because of the small number of P-gp non-substrates in the set. Another factor is the inadequate sampling of the chemical space covered by non-substrates of P-gp. It is likely that the 85 non-substrates collected in this work only represent a portion of all possible classes of non-substrates of P-gp. Protein non-substrates are rarely described in the literature, thus additional efforts are needed to enable the collection of this information.

SVM classification results were further compared to those from other machine learning methods like kNN, PNN, and C4.5 DT to determine whether it is possible to use a simpler model for the prediction of P-gp substrates and non-substrates. The same sets of data and descriptors are used in these computations. The results are shown in Table 4.7 and it is found that the accuracy from SVM classification system is slightly better than those from other classification methods. This suggests that the SVM classification system developed in this study is not more flexible than is necessary and thus is unlikely to have overfitting problems.

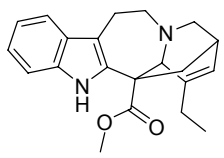
Table 4.7 Comparison of the prediction accuracy of the substrates and non-substrates of P-glycoprotein from different classification methods by using 5-fold cross-validation.

Method	SE (%)	SP (%)	Q (%)
kNN	79.2	61.6	70.8
PNN	77.3	71.4	74.4
C4.5 DT	74.6	69.9	71.5
SVM	**81.2**	**79.2**	**79.4**

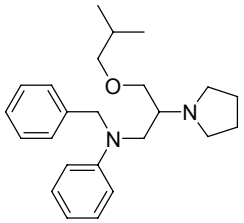
SVM typically uses a portion of the training set as support vectors for classification. In contrast, kNN and PNN use the whole training set for classification. Our own studies suggest that the number of support vectors of SVM is in the range of 40-70% of the training set. Thus the classification speed of SVM is usually 30-60% faster than that of kNN and PNN. On the other hand, the classification speed of SVM is slower than that of decision tree methods which conduct tests on descriptors to reach a decision leaf.

In the independent validation set, there are three and two incorrectly classified substrates and non-substrates of P-gp respectively, which are shown in Figure 4.1. The three P-gp substrates are catharantine, depredil and yohimbine, and the two non-substrates of P-gp are NSC364080 and NSC630357.

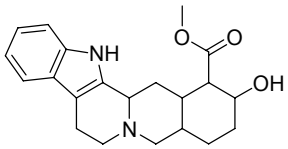
Figure 4.1 Structures of misclassified compounds in independent validation set.



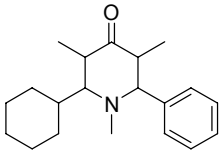
Catharantine (substrate)



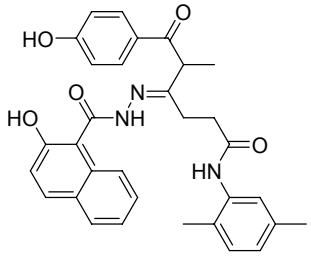
Depredil (substrate)



Yohimbine (substrate)



NSC364080 (nonsubstrate)



NSC630357 (nonsubstrate)

Table 4.8 gives the molecular descriptors selected from the feature selection method RFE. Those from the class of topological descriptor constitute the largest percentage of the descriptors selected. This is consistent with the findings from the classification of MDR compounds, many of which are P-gp substrates, by using structure-based descriptors and linear discriminant analysis, which showed that 60% of the molecular descriptors important for MDR are topological in nature (Bakken *et al.* 2000). A QSAR study of MDR compounds also identified several pharmacophores, e.g., a generic form of C-C-X-C-C with X=N, NH, or O (preferably a tertiary nitrogen), as a key structural element for MDR (Klopman *et al.* 1997). These

pharmacophores are primarily determined by electrotopological features and bond connectivity. In addition to the large percentage of electrotopological descriptors, RFE method also selected three molecular connectivity descriptors, which seems to correlate with the features of the pharmacophores identified from the QSAR study of MDR compounds.

Table 4.8 Molecular descriptors selected from the feature selection method for classification of P-gp substrates and non-substrates.

No	Descriptors	Description	Class
1	Ncocl	Count of Cl atoms	Simple molecular properties
2	Ndonr	Number of H-bond donors	Simple molecular properties
3	$^{5}\chi_{CH}$	Simple molecular connectivity Chi indices for cycle of 5 atoms	Connectivity and shape
4	$^{3}\chi^{v}_{P}$	Valence molecular connectivity Chi indices for path order 3	Connectivity and shape
5	$^{5}\chi^{v}_{CH}$	valence molecular connectivity Chi indices for cycle of 5 atoms	Connectivity and shape
6	Scar	Sum of Estate indices of carbon atoms	Geometrical properties
7	dis2	Length vector (longest third atom)	Geometrical properties
8	Sapcw	Sum of charge weighted solvent accessible surface areas of positively charged atoms	Geometrical properties
9	S(1)	Atom-type H Estate sum for -OH	Electrotopological state
10	S(9)	Atom-type H Estate sum for =CH- (sp^2)	Electrotopological state
11	S(12)	Atom-type H Estate sum for CH_n (Saturated)	Electrotopological state
12	S(13)	Atom-type H Estate sum for CH_n (unsaturated)	Electrotopological state
13	S(16)	Atom-type Estate sum for $-CH_3$	Electrotopological state

14	S(18)	Atom-type Estate sum for $>CH_2$	Electrotopological state
15	S(20)	Atom-type Estate sum for =CH-	Electrotopological state
16	S(21)	Atom-type Estate sum for : CH : (aromatic)	Electrotopological state
17	S(25)	Atom-type Estate sum for =C<	Electrotopological state
18	S(36)	Atom-type Estate sum for >N-	Electrotopological state
19	π_i	Polarizability index	Quantum chemical properties
20	q^+	Atomic charge on the most positively charged H atom	Quantum chemical properties
21	μ	Molecular dipole moment	Quantum chemical properties
22	ω	Electrophilicity index	Quantum chemical properties

The rest of the RFE selected descriptors are from the quantum chemical class and simple molecular property class. The selected quantum chemical descriptors determine polarizability, molecular dipole moment, electrophilicity, and the atomic charge of the positively charged hydrogen atoms in a molecule. The selected simple molecular property descriptors give the number of hydrogen bond donors and that of Cl atoms. With the exception of the last descriptor, the MolSurf counterparts of these quantum chemical and simple molecular property descriptors have used for the prediction of P-gp-interacting drugs by means of multivariate statistics method (Österberg *et al.* 2000). Based on structural comparison, it has been found that the number of electron donors and hydrogen bond acceptor groups are important elements for P-gp substrate recognition (Seelig 1998). An analysis of multiple pharmacophores of P-gp substrates has identified hydrophobe, hydrogen bond donor and acceptor as important elements for P-gp substrates (Penzotti *et al.* 2002). Thus these studies consistently suggested the importance of the selected quantum chemical

features and hydrogen-bond property for prediction of P-gp substrates and non-substrates.

The other RFE selected descriptor, the count of Cl atoms, has not been specifically used in other P-gp substrate studies. One possible reason is that the molecules used in those studies do not contain a Cl atom, thus it is unnecessary to introduce this descriptor in those studies. In this work, the descriptor for hydrogen bond acceptor was not selected by RFE, which has been found to be an important element for P-gp substrates in other studies (Seelig 1998; Penzotti *et al.* 2002). One likely reason for the exclusion of this descriptor is that it has a high level of redundancy with the relevant features covered by the quantum chemical descriptors such as electrophilicity, polarizability and molecular dipole moment when they are combined with the hydrogen bond donor descriptor.

4.2.4 Conclusion

SVM is a potentially useful computational method for facilitating the prediction of P-gp substrates. The SVM model developed in this work gave a prediction accuracy for P-gp substrates that is substantially improved against that obtained from the multiple-pharmacophore model. The prediction accuracy for nonsubstrates of P-gp is slightly better than those obtained from other statistical classification methods, including kNN, PNN, and C4.5 decision tree, that use the same sets of data and molecular descriptors. Prediction accuracy may be further improved by consideration of factors such as hydrogen bonding, active transport, and relationship with pharmacodynamic properties.

Chapter 5

Prediction of Drug Distribution

This chapter describes the prediction of a few important distribution processes, such as blood-brain barrier penetration, human serum albumin binding and milk-plasma ratio by using GRNN. The prediction accuracies of the GRNN-developed models were compared with those of QSPkR models developed by using MLR and MLFN. A new method for interpreting GRNN-developed QSPkR models, which enables relevant physicochemical and structural properties of a compound to be identified, is also introduced.

5.1 Introduction

Optimization of pharmacokinetic as well as the pharmacodynamic properties of a drug candidate is an important consideration in drug design process (Eddershaw *et al.* 2000; van de Waterbeemd *et al.* 2003). One important aspect of pharmacokinetic properties of a drug candidate is its distribution in the human body. A drug is required to achieve sufficient concentration at target site while possibly limiting its distribution elsewhere so as to produce desired therapeutic action with minimum side effects (Butina *et al.* 2002). Traditionally, the distribution properties of a drug candidate are obtained via *in vivo* and *in vitro* studies, which tend to be time-consuming and costly. Therefore, QSPkR modeling has recently been explored for predicting the distribution properties of drug candidates (Ekins *et al.* 2000c) in an effort to eliminate undesirable compounds in a fast and cost-effective manner.

The most common modeling methods for obtaining QSPkR models are linear methods such as multiple linear regression (MLR) (Geladi *et al.* 1986). These methods can be easily used and the derived models can be easily interpreted. However multiple mechanisms may be involved in determining a particular pharmacokinetic property. A variety of factors may interact in complex ways to affect the pharmacokinetic property of a compound. Therefore methods based only on linear relationships may not always be the most efficient approach for constructing a QSPkR model. Thus non-linear methods such as multi-layer feedforward neural networks (MLFN) (Wythoff 1993) and general regression neural network (GRNN) (Specht 1991) have increasingly been used for construction of QSPkR models.

GRNN has been explored for QSPkR modeling of human intestinal absorption (Niwa 2003) as well as for developing QSAR and QSPR of chemical compounds (Mosier *et al.* 2002). The prediction capability of GRNN has been found to be comparable to those of conventional non-linear methods such as MLFN but the former requires fewer descriptors (Mosier *et al.* 2002). Thus GRNN is expected to be equally useful for developing QSPkR models of other pharmacokinetic properties. This work is intended to test this feasibility by applying GRNN for developing QSPkR models of three distribution properties, blood-brain barrier (BBB) penetration, binding to human serum albumin (HSA) and milk-plasma (M/P) distribution. The performances of the GRNN-developed models were compared with those developed by using MLR and MLFN to determine whether GRNN produces more predictive QSPkR models.

The BBB exists at the choroids plexus and at the tissue capillary membranes between the blood and brain fluid, and BBB penetration is necessary for central nervous system (CNS) drugs (Hardman *et al.* 2002). Examples of these drugs are

antipsychotics, antiepileptics and antidepressants. For drugs not directed at targets in the brain, BBB penetration is undesirable because of potential CNS-related side effects. For example, the first generation antihistamines are known to penetrate the BBB leading to drowsiness (Meltzer 1990). The second generation antihistamines have a significantly reduced BBB penetration capability and are thus less likely to cause drowsiness (Kaliner 1992). One method for assessing the effects of a compound in the brain is to determine its concentration in the brain. This concentration can be calculated from the brain-blood (BB) ratio which is the concentration of this compound in the brain divided by that in the blood. Thus the BB ratio is an important pharmacokinetic property and a number of QSPkR models of BB ratio have been developed (Young *et al.* 1988; van de Waterbeemd *et al.* 1992; Abraham *et al.* 1994; Lombardo *et al.* 1996; Norinder *et al.* 1998; Clark 1999; Kelder *et al.* 1999; Luco 1999; Feher *et al.* 2000; Kaznessis *et al.* 2001; Keserü *et al.* 2001; Liu *et al.* 2001; Platts *et al.* 2001; Iyer *et al.* 2002; Hou *et al.* 2003), the majority of which were developed by using MLR and the computed r^2 values are in the range between 0.723-0.941.

Most drugs bind to serum proteins and such binding regulates drug distribution and subsequently its effect (Colmenarejo 2003). Albumin is the most abundant of all serum proteins and is the most common drug-binding protein in the circulatory system. Because of the important role of albumin-binding in regulation of drug distribution, QSPkR models for predicting the extent of albumin-binding have been developed (Gobburu *et al.* 1995; Colmenarejo *et al.* 2001; Kratochwil *et al.* 2002; Hall *et al.* 2003; Turner *et al.* 2003b), the majority of which were developed by using MLR and a congeneric series of compounds. In a study of a diverse set of 94 drugs and drug-like compounds, two QSPkR models developed by using MLR gave

computed r^2 values of 0.88 and 0.82 respectively on a separate testing set (Colmenarejo *et al.* 2001).

Breast milk is the best form of nutrition available to a newborn infant. Certain drugs administered to a nursing mother may be distributed into breast milk and thus transferred into the infant. The concentration of drug present in the breast milk can be used as an indicator of breast feed risk. The ratio of drug concentration in milk and plasma (M/P ratio) is the most widely used quantity for describing drug concentration in breast milk (Begg *et al.* 1993). However, the M/P ratio is seldom determined during clinical trials or after the drug has entered the market. In addition, M/P ratios were often obtained from studies involving a small number of women. This may lead to significant variations in the reported M/P ratio for a drug and makes it difficult for clinicians to advise women on the safety of breast-feeding. Methods for estimating the M/P ratios of drugs have been developed by using various modeling methods (Wilson 1981; Meskin *et al.* 1985; Fleishaker *et al.* 1987; Atkinson *et al.* 1990; Agatonovic-Kustrin *et al.* 2000; Agatonovic-Kustrin *et al.* 2002). In a recent study (Agatonovic-Kustrin *et al.* 2002), MLFN was used to train and test on 123 diverse compounds. The computed r^2 and mean square error (MSE) values from this model are 0.61 and 0.814 respectively.

5.2 Methods

5.2.1 MLFN algorithm

The algorithm of MLFN has been extensively described in literatures (Wythoff 1993; Erb 1995; Hudson *et al.* 1995). Thus only a brief description is given here. MLFN is composed of an input layer, a variable number of

hidden layers and an output layer. The input and output layers contain neurons representing the descriptors and response value respectively. In a fully connected MLFN, each neuron in the input layer sends its value to all neurons in the first hidden layer. Each neuron in the hidden layers receives inputs from all neurons in the previous layer and computes a weighted sum of the inputs. The neuron output is determined by passing the weighted sum through a transfer function, which is usually a linear or sigmoidal function. The single neuron in the output layer determines the predicted response value by computing a weighted sum of the outputs of all neurons in the last hidden layer. Weights for the connections between neurons in adjacent layers are initially randomly assigned. These weights are then refined via a backward propagation of error process during training of the MLFN. In this study, MLFN were performed using the YMLL library (section 3.2) and had a single hidden layer with ten neurons.

5.2.2 Molecular descriptors

A total of 1497 1D, 2D and 3D molecular descriptors were computed by using DRAGON (Todeschini *et al.* 2003). These descriptors, which can be divided into 18 classes, include 47 constitutional descriptors, 70 geometrical descriptors, 266 topological descriptors, 150 RDF descriptors (Hemmer *et al.* 1999), 21 molecular walk counts (Rücker *et al.* 1993), 160 3D-MoRSE descriptors (Schuur *et al.* 1996), 64 BCUT descriptors (Pearlman *et al.* 1999), 99 WHIM descriptors (Bravi *et al.* 1997), 21 Galvez topological charge indices (Galvez *et al.* 1994), 197 GETAWAY descriptors (Consonni *et al.* 2002), 96 2D autocorrelations, 121 functional groups, 14 charge descriptors, 120 atom-centred descriptors, 4 aromaticity indices (Randic 1975),

3 empirical descriptors, 41 Randic molecular profiles (Randic 1995) and 3 molecular properties.

5.2.3 Datasets

The BBB penetration dataset contains 175 compounds with experimental log BB values collected from various literature sources (Luco 1999; Kaznessis *et al.* 2001; Platts *et al.* 2001; Hou *et al.* 2003). Twelve compounds were identified as outliers by previous studies (Abraham *et al.* 1994; Lombardo *et al.* 1996; Clark 1999; Luco 1999; Kaznessis *et al.* 2001; Liu *et al.* 2001; Platts *et al.* 2001) and were removed from the dataset. The DRAGON software was unable to compute the descriptors of four compounds, argon, krypton, neon and xenon, and thus these compounds were also removed from the dataset. The final dataset of 159 compounds were divided into a training set of 129 compounds and a validation set of 30 compounds.

The HSA binding dataset was composed of 94 compounds with HSA binding constants, log Khsa, and was obtained from Colmenarejo et al (Colmenarejo *et al.* 2001). One compound, ebselen, was removed from the original dataset as its descriptors cannot be computed by the DRAGON software. These compounds were divided into a training set and a validation set of 75 and 18 compounds respectively.

The M/P distribution dataset consists of 123 compounds used in the Agatonovic-Kustrin's study (Agatonovic-Kustrin *et al.* 2002). An erroneous compound, norfluexetine, was identified and removed from the original dataset. The remaining compounds were split into a training set and validation set of 102 and 20 compounds respectively.

Kennard and Stone algorithm (section 2.2.2.2), which has been found to be useful for constructing representative training and validation sets from a dataset (Wu

et al. 1996; Zuegge *et al.* 2002; Rajer-Kanduc *et al.* 2003), was used in this work. In computing the Euclidean distance for the algorithm, principal component analysis (PCA) was used to select principal components (PC) whose eigenvalues were larger than one. The Euclidean distance was then calculated from the retained PCs. The selection process continues until approximately 80% ~ 85% of the compounds were selected for the training set. The remaining 15% ~ 20% of the compounds in the dataset were used as the validation set.

5.2.4 Descriptor selection

The first step involves the removal of all irrelevant descriptors such as constant descriptors and near-constant descriptors that have the same value for more than 80% of the compounds. All the remaining descriptors were autoscaled using equation (2.3). Genetic algorithm (section 2.3.3.2) was then used to further remove descriptors of low information content. In the mutation process of the genetic algorithm, descriptors may be randomly added to or deleted from a descriptor subset, subjected to an overall minimum and maximum of 3 and 10 descriptors respectively for each descriptor subset. At the end of the genetic algorithm-based descriptor selection process, the highest ranked subset was retained. As genetic algorithm is a heuristic method, the selection of relevant descriptor subset was repeated 10 times to improve the chances of finding the optimum descriptor subset. The best descriptor subset from these 10 runs was used to construct the QSPkR model.

In the descriptor selection process, the original training set was divided by using Kennard and Stone algorithm into a modeling training set and a modeling testing set by a 4:1 ratio. The modeling training set was used for constructing the QSPkR models in the genetic algorithm. The testing set was used to evaluate the

trained systems so that no overtrained systems are selected. The following cost function (Wessel *et al.* 1998; Mosier *et al.* 2002) was used as the fitness function during genetic algorithm optimization:

$$Cost = MSE_{train} + 0.4 \times \left| MSE_{train} - MSE_{test} \right| \quad (5.1)$$

where MSE_{train} and MSE_{test} are the mean square error of the modeling training set and testing set respectively and were calculated using the equation (2.35).

5.2.5 Model validation

Y-randomization was used to determine the probability of chance correlation during descriptor selection (Manly 1997; Leardia *et al.* 1998). The distribution properties of all the compounds in the modeling training set were first randomly rearranged. Descriptor selection using genetic algorithm was then used to find the optimum descriptor subset for the scrambled data and the cost of this descriptor subset was measured. The scrambling of the distribution properties and descriptor selection was repeated for 30 times. If the cost of all of the scrambled QSPkR models were significantly worse than the cost of the original QSPkR model, it can be concluded that the original QSPkR model was relevant and unlikely to arise as a result of chance correlation.

The validation set, not used in the derivation of the QSPkR models, was used to estimate the prediction capability of the final QSPkR models. Leave-one-out (LOO) and 10-fold cross-validation were not used for this purpose in this work because there are reports of the lack of correlation between cross-validation methods and the prediction capability of a QSAR model (Golbraikh *et al.* 2002; Kozak *et al.* 2003; Reunanen 2003; Olsson *et al.* 2004). In addition, cross-validation methods have a tendency to underestimate the prediction capability of a QSAR model, especially if

important molecular features are present in only a minority of the compounds in the training set (Mosier *et al.* 2002; Hawkins *et al.* 2004). Thus a model having low cross-validation results can still be quite predictive (Mosier *et al.* 2002).

5.2.6 Interpretation of GRNN-developed models

In multi-sigma GRNN-developed models, the contribution of each descriptor on the distribution property of a compound can be estimated from its σ value. Those descriptors with smaller σ values give higher contributions. From equation (2.27), it can be seen that the change in the distribution property is proportional to $1/\sigma^2$. A functional dependence study was also done using the procedures described in section 2.5.3.

5.3 Results and discussion

5.3.1 BBB penetration

A seven-descriptor subset was selected by the descriptor selection algorithm as the optimum set for GRNN model of BBB penetration, which is given in Table 5.1. Absolute pairwise correlation between the seven descriptors ranged from 0.032 to 0.561 with an absolute mean correlation of 0.287. For both MLR and MLFN, a nine-descriptor subset was obtained by the descriptor selection algorithm. The minimum costs of 30 scrambled QSPkR models developed using GRNN, MLR and MLFN were 0.375 (Mean: 0.430, SD: 0.033), 0.300 (Mean: 0.356, SD: 0.027) and 0.338 (Mean: 0.391, SD: 0.021) respectively. These were significantly larger than the cost of the original GRNN-, MLR- and MLFN-developed models, which were 0.042, 0.180 and

0.109 respectively. Thus y-randomization showed that the original QSPkR models were relevant and unlikely to arise from chance correlation. The prediction results of the QSPkR models, given in Table 5.2, show that the QSPkR model developed using GRNN was the best model.

Table 5.1 Descriptors selected for BBB GRNN model.

Descriptor	Type	Sigma	Range		Explanation
			Min	Max	
Ms	Constitutional	0.22	1.50	6.94	Mean electrotopological state
RBN	Constitutional	0.48	0	21	Number of rotatable bonds
piPC08	Topological	1.33	0.0	2682.2	Molecular multiple path count of order 08
GATS5e	2D autocorrelations	0.68	0.00	2.54	Geary autocorrelation - lag 5 / weighted by atomic Sanderson electronegativities
SPAM	Geometrical	0.32	0.30	0.72	Average span R
E1p	WHIM	0.68	0.20	0.89	1st component accessibility directional WHIM index / weighted by atomic polarizabilities
R2v	GETAWAY	0.66	0.10	1.14	R autocorrelation of lag 2 / weighted by atomic van der Waals volumes

Table 5.2 Predictive capabilities of BBB QSPkR models on independent validation set.

Method	r^2	R_s [a]	MSE
GRNN	**0.701**	**0.825**	**0.130**
MLR	0.649	0.782	0.154
MLFN	0.662	0.802	0.147

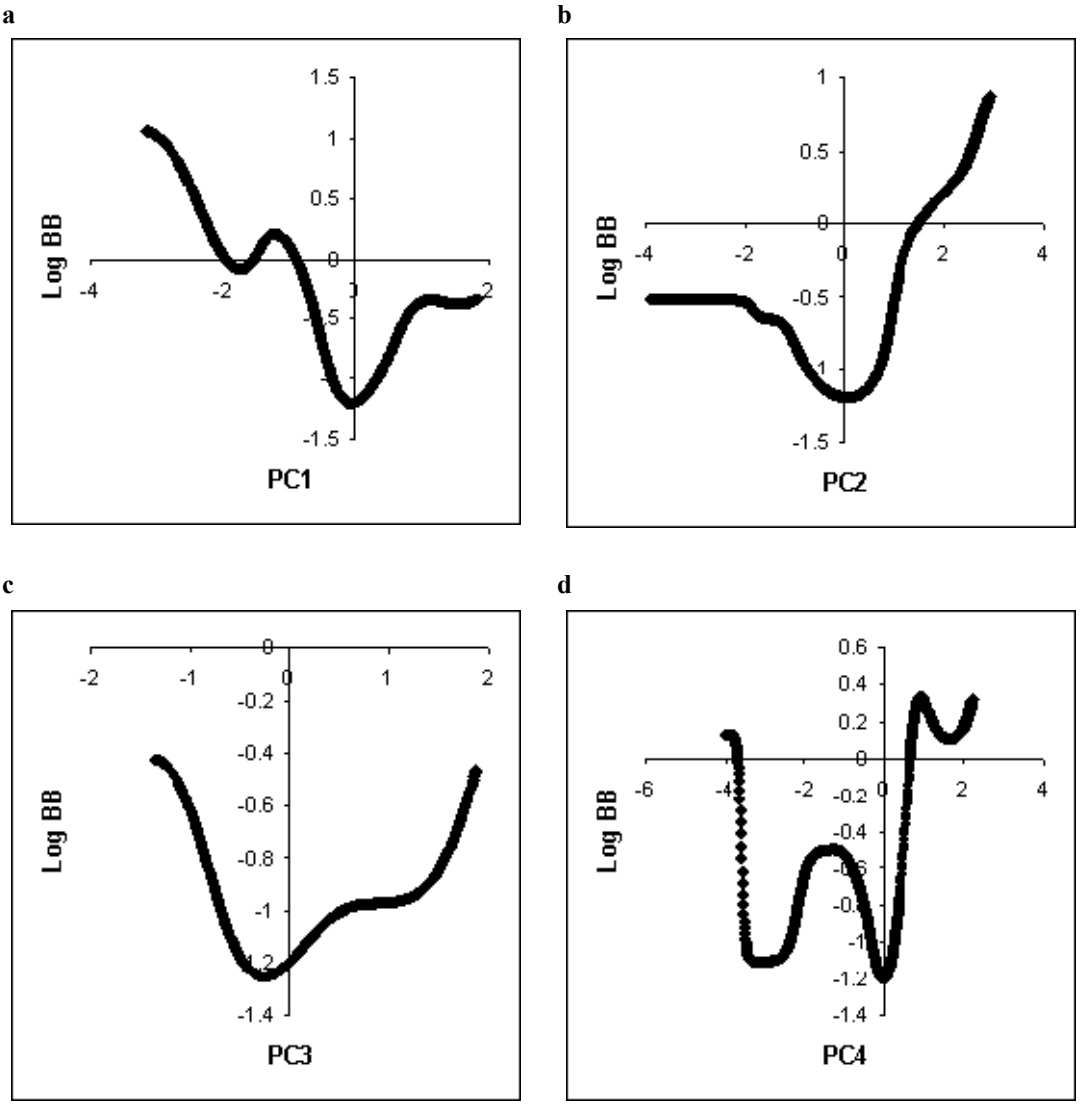
[a] Spearman rho coefficient

The ranking of the descriptors in the GRNN-developed model, which is determined by the individual sigma values, is in the following decreasing order: Ms, SPAM, RBN, R2v, GATS5e, E1p and piPC08. The frequency at which individual descriptors were selected during the ten genetic algorithm descriptor selection runs can be used to determine the relevance of the descriptors for the QSPkR model. During the ten genetic algorithm descriptor selections, Ms, RBN and SPAM were selected in 50%, 40% and 30% of the GRNN models respectively. Although R2v, GATS5e and E1p were not selected by the other nine GRNN models, other similar GETAWAY, 2D autocorrelations and WHIM descriptors which are correlated with R2v, GATS5e and E1p respectively were selected in five, three and two other GRNN models respectively. Only piPC08 had no similar descriptors in other models. Thus the majority of the descriptors in the GRNN model were selected more than once by the genetic algorithm descriptor selection method and hence these descriptors were likely to be important for the prediction of BBB penetration. The artificial testing sets prediction results for the first 4 principal components (PCs) of these seven descriptors are shown in Figure 5.1. Plots for the fifth to seventh PCs are not shown as they explained less than 17% of the total variance of the descriptors and thus likely to contain noise rather than useful information.

Explained variations of the descriptors showed that the first PC was primarily determined by SPAM, with some contributions from R2v, piPC08, Ms and GATS5e. SPAM is used to describe long chain molecules and is determined by the size and flexibility of a molecule. R2v encodes both molecular structure and van der Waals volume of a molecule. piPC08 belongs to the molecular path count type of descriptors, which are a useful measure of molecular size and complexity (Todeschini *et al.* 2000). Ms is an electrotopological state descriptor that encodes the electronic and topological

information of a molecule. GATS5e encodes both molecular structure and the group electronegativity of molecular substituents. The presence of these five descriptors in

Figure 5.1 Plots of log BB against the various PCs of BBB descriptor subset of GRNN.



the first PC suggests that the first PC is a measure of molecular size. The artificial testing sets show that BBB penetration generally increases with decreasing molecular size (Figure 5.1a). This is consistent with the findings that small molecular size is necessary for good BBB penetration (Pardridge 1998). On the other hand, this figure also suggests that large molecules have better BBB penetration than molecules of intermediate size. This finding is consistent with the results from other studies which

showed that increasing molecular volume seems to be correlated with increasing BBB penetration (Kaznessis *et al.* 2001; Platts *et al.* 2001).

E1p was the main contributor to the second PC and encodes information about the size, shape, symmetry, atom distribution and polarizability of a molecule (Bravi *et al.* 1997). As E1p encodes multiple characteristics of a molecule, it is not possible to clearly determine a relationship between a specific characteristic of a molecule and its BBB penetration. However, studies had consistently found the importance of size (Pardridge 1998; Kaznessis *et al.* 2001; Platts *et al.* 2001), shape (Ooms *et al.* 2002; Lobell *et al.* 2003a) and polarizability (Platts *et al.* 2001; Abraham 2004) of a molecule in determining the log BB of a molecule.

The third PC was formed mainly by RBN, and to a lesser extent, by piPC08. RBN is related to the flexibility of a molecule. The complex role of molecular flexibility in membrane permeation has been found by two studies. One found a positive correlation between flexibility and permeation (Iyer *et al.* 2002) while the other found a negative correlation (Veber *et al.* 2002). This seems to suggest that flexibility is an important factor in BBB penetration but its precise effects are dependent on the presence of other molecular characteristics. Using the artificial testing sets, it was found that compounds with 5 or 6 rotatable bonds had the lowest log BB values (Figure 5.1c).

The fourth PC was determined primarily by GATS5e and partially by Ms. As molecular size was described by the first PC, the fourth PC probably represented the electronegativity of a molecule. Electronegativity affects pKa of a compound and thus influences molecular charge at physiological pH. This is consistent with findings implicating molecular charge in the extent of BBB penetration (Lobell *et al.* 2003a).

5.3.2 HSA binding

Table 5.3 shows the optimum descriptor subset of the GRNN model of HSA binding. The descriptor subset contained only six descriptors compared to the 10-descriptor subset of MLR and eight-descriptor subset of MLFN. The absolute minimum, maximum and mean of the pairwise correlation between the six descriptors of the GRNN model are 0.012, 0.291 and 0.117 respectively. All of the scrambled QSPkR models produced during y-randomization had significantly larger costs than that of the corresponding original QSPkR models. The minimum cost of 30 scrambled QSPkR models for GRNN, MLR and MLFN were 0.234 (Mean: 0.344, SD: 0.048), 1.374 (Mean: 1.681, SD: 0.151) and 0.258 (Mean: 0.300, SD: 0.023) respectively and the cost for the corresponding original QSPkR models were 0.029, 0.053 and 0.050 respectively. Hence it is unlikely that the original QSPkR models were a result of chance correlation. Results of the validation set, given in Table 5.4, shows that the GRNN-developed model had a better prediction capability than that of the models developed by using MLR or MLFN.

Among the six descriptors, only Mor20p and GATS8e were selected in some of the other nine GRNN models. However, when similar and correlated descriptors were considered, Mor20p, GATS8e, C-040 and H-050 were present in 50%, 50%, 40% and 40% of the GRNN models respectively. Only RDF040m and SRW07 were not selected in other nine GRNN models. The plots of log Khsa against the first four PCs, obtained using the artificial testing sets, are shown in Figure 5.2. The last two PC accounted for less than 21% of the total variance of the descriptors and are not shown.

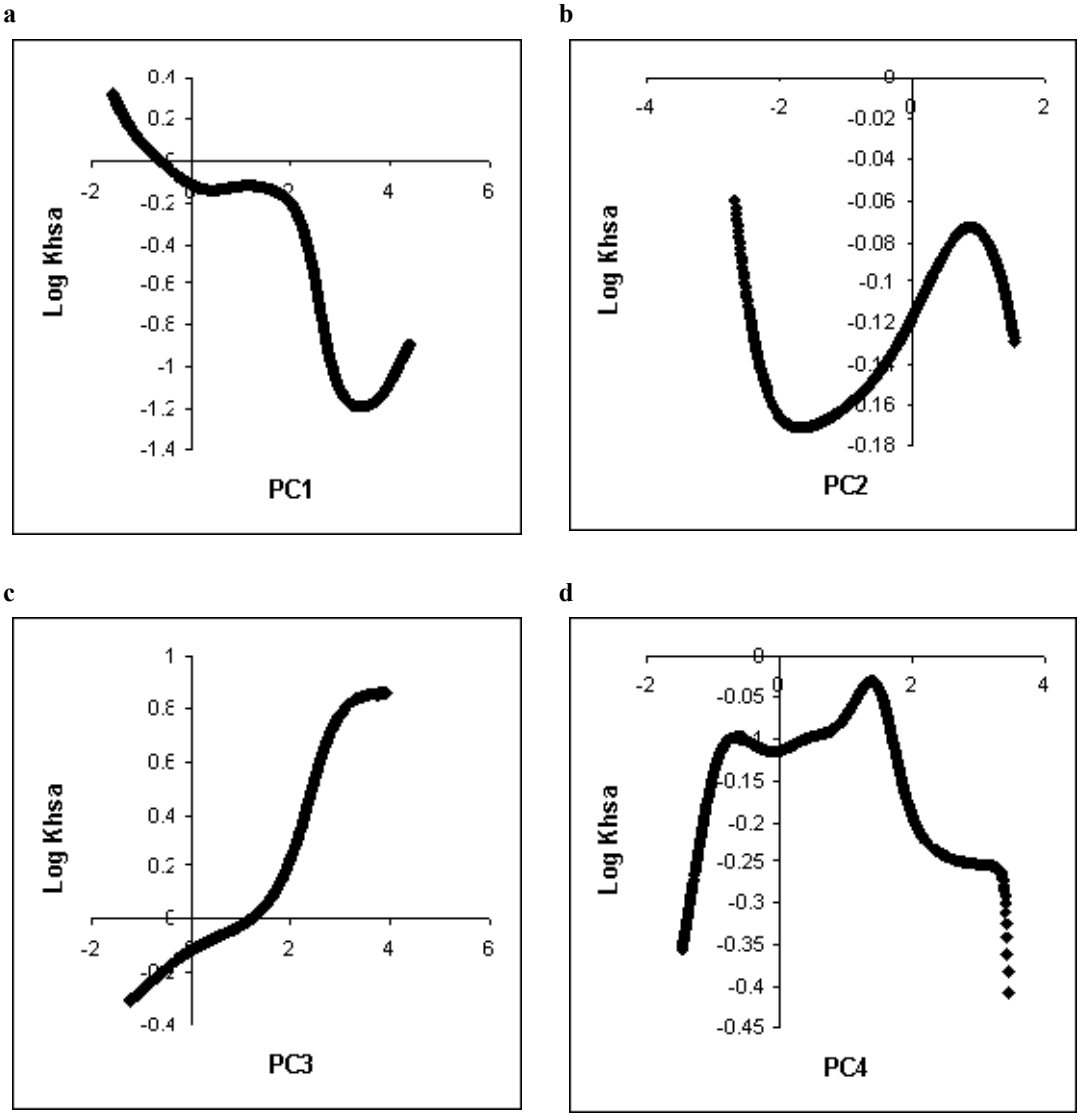
Table 5.3 Descriptors selected for HSA GRNN model.

Descriptor	Type	Sigma	Range		Explanation
			Min	Max	
SRW07	Molecular walk counts	1.389	0	518	Self-returning walk count of order 07
GATS8e	2D autocorrelations	0.500	0.00	4.26	Geary autocorrelation - lag 8 / weighted by atomic Sanderson electronegativities
RDF040m	RDF	1.297	0.24	23.51	Radial Distribution Function - 4.0 / weighted by atomic masses
Mor20p	3D-MoRSE	1.157	-0.45	2.23	3D-MoRSE - signal 20 / weighted by atomic polarizabilities
C-040	Atom-centred fragments	0.902	0	4	R-C(=X)-X / R-C#X / X-=C=X
H-050	Atom-centred fragments	0.568	0	7	H attached to heteroatom

Table 5.4 Predictive capabilities of HSA QSPkR models on independent validation set.

Method	r^2	R_s [a]	MSE
GRNN	**0.851**	0.825	**0.041**
MLR	0.770	0.822	0.079
MLFN	0.749	**0.851**	0.089

Figure 5.2 Plots of log Khsa against the various PCs of HSA descriptor subset of GRNN.



The first PC was determined mainly by H-050 and to a lesser extent by C-040. H-050 is related to hydrogen bond donating ability of a molecule while C-040 encodes information on hydrogen bond acceptors. Results from the artificial testing set suggest that binding affinity to HSA generally decreases with increasing hydrogen bonding ability of a molecule (Figure 5.2a). This is consistent with the findings of a HSA QSPkR model of beta-lactams (Hall *et al.* 2003).

Mor20p was the main contributor to the second PC. It is a representation of the 3D structure of a molecule and encodes information about the polarizability of a

molecule. Hence, in addition to hydrogen bonding, the binding affinity to HSA may also be affected by the polarizability of a molecule.

The third PC was primarily contributed by SRW07 and RDF040m. SRW07 is related to molecular branching and size and in general to the molecular complexity of the graph. RDF040m provides information about interatomic distances in the entire molecule and also other useful information such as bond distances, ring types, planar and non-planar systems, atom types and molecular weight. Thus, the third PC probably is a measure of the shape of a molecule. Shape of a molecule has also been identified as an important descriptor in other HSA binding studies (Colmenarejo *et al.* 2001; Kratochwil *et al.* 2002).

Most of the variances in GATS8e were explained by the fourth PC. GATS8e contains information about the group electronegativity of molecular substituents. Various QSAR models have identified charge distribution in a molecule (Kratochwil *et al.* 2002), electrostatic interactions (Colmenarejo 2003), and presence and electron accessibility of certain molecular substituents (Colmenarejo 2003) as important elements for HSA binding. Thus these studies consistently suggested the importance of electronic descriptors such as electronegativity in the prediction of HSA binding.

The ranking of the descriptors in the GRNN-developed model, in decreasing order, is GATS8e, H-050, C-040, Mor20p, RDF040m and SRW07. This suggests electronic properties are more important factors in determining the binding affinity to HSA than the shape of the molecule. This is consistent with findings that HSA can bind to a large variety of compounds with different shapes and sizes (Colmenarejo 2003). Lipophilic descriptors such as log P, which had been identified as an important factor for HSA binding in a number of studies (Colmenarejo *et al.* 2001; Kratochwil *et al.* 2002; Colmenarejo 2003), were absent from the current GRNN-developed

model. It is possible that descriptors such as log P do not contain as much information as the current descriptor subset for describing molecule-protein interactions (Agatonovic-Kustrin *et al.* 2002). Thus descriptors such as Mor20p and RDF040m which encode multiple characteristics of a molecule and have some degree of correlation with lipophilicity were included instead.

5.3.3 Milk-Plasma Distribution

Genetic algorithm descriptor selection found an optimum descriptor subset of seven descriptors for GRNN model of the M/P distribution, which is given in Table 5.5. These seven descriptors had an absolute minimum, maximum and mean pairwise correlation of 0.030, 0.476 and 0.169 respectively. A 10- and an eight-descriptor subset were found for MLR and MLFN respectively. The minimum cost of 30 scrambled QSPkR models for GRNN, MLR and MLFN were 1.582 (Mean: 2.080, SD: 0.407), 1.372 (Mean: 1.659, SD: 0.160) and 1.209 (Mean: 1.659, SD: 0.140) respectively. These were significantly larger than the cost of the corresponding original QSPkR models, which were 0.358, 0.985 and 0.412 respectively. Hence y-randomization showed that the original QSPkR models were relevant and unlikely to be a result of chance correlation. Table 5.6 shows the testing results of the QSPkR models by using the independent validation set. Among the three modeling methods, GRNN was the only one that produced a model with reasonable predictive ability. Both models developed by MLR and MLFN had computed r^2 values of less than 0.5. This suggests that GRNN is more suitable than either MLR or MLFN for developing QSPkR models of M/P distribution.

Table 5.5 Descriptors selected for M/P GRNN model.

Descriptor	Type	Sigma	Range		Explanation
			Min	Max	
TIE	Topological	0.750	4.54	495.50	E-state topological parameter
GATS3e	2D autocorrelations	0.650	0.46	1.79	Geary autocorrelation - lag 3 / weighted by atomic Sanderson electronegativities
HOMT	Aromaticity indices	0.220	-28.20	20.43	HOMA total
Mor23m	3D-MoRSE	1.661	-0.94	1.36	3D-MoRSE - signal 23 / weighted by atomic masses
Mor06p	3D-MoRSE	1.120	-2.39	2.80	3D-MoRSE - signal 06 / weighted by atomic polarizabilities
HATS5e	GETAWAY	0.138	0.00	1.12	Leverage-weighted autocorrelation of lag 5 / weighted by atomic Sanderson electronegativities
R4u	GETAWAY	0.650	0.33	2.99	R autocorrelation of lag 4 / unweighted

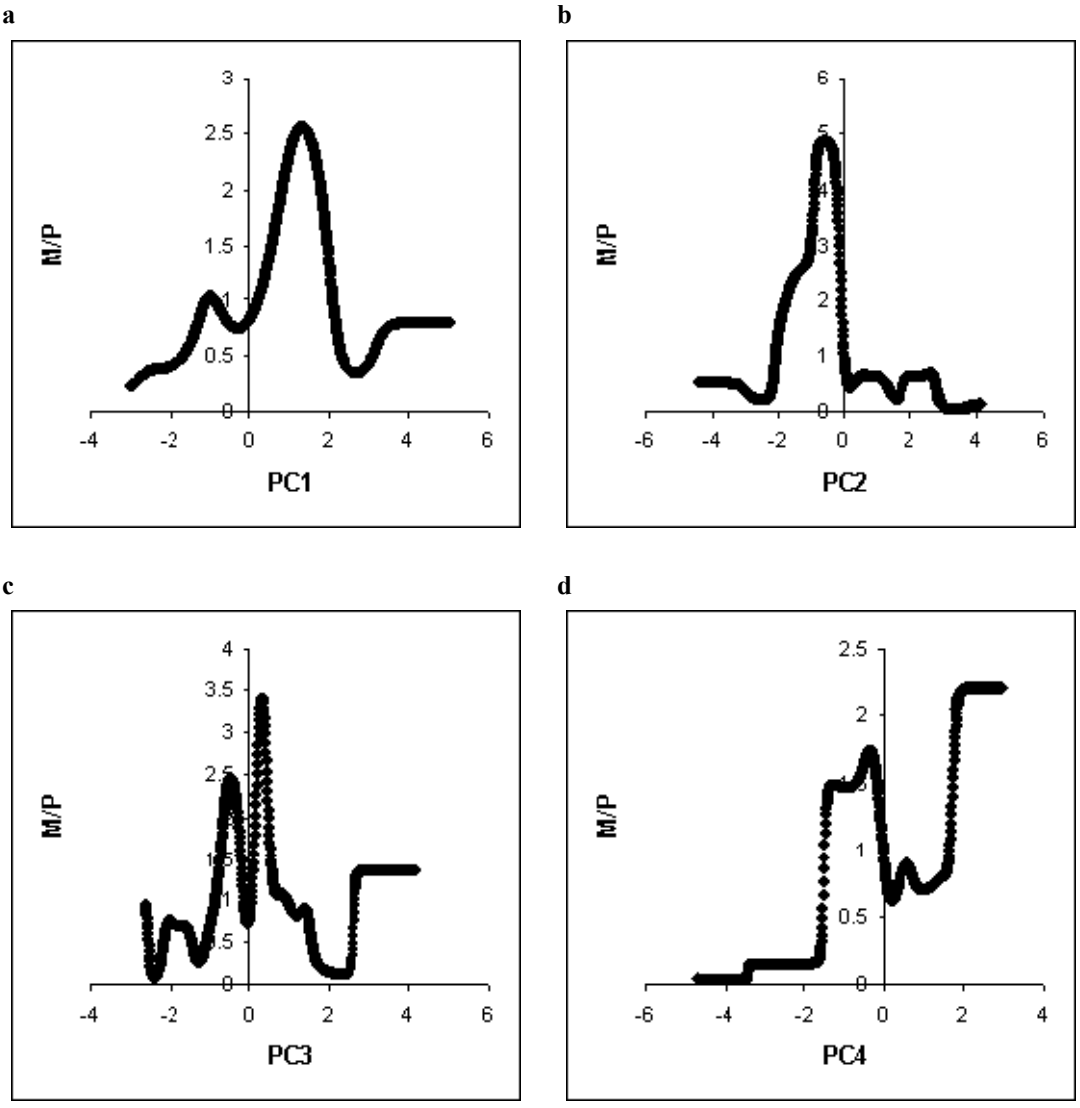
Table 5.6 Predictive capabilities of M/P QSPkR models on independent validation set.

Method	r^2	R_s [a]	MSE
GRNN	**0.677**	**0.769**	**0.206**
MLR	0.224	0.460	0.647
MLFN	0.201	0.408	0.587

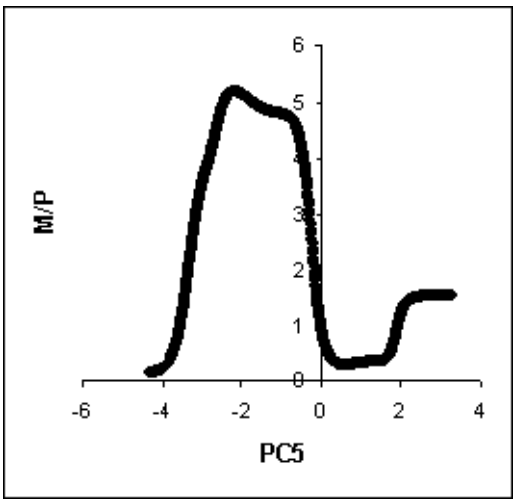
The descriptors TIE, GATS3e, HOMT, Mor23m, Mor06p, HATS5e, and R4u, or their similar and correlated counterparts were selected in 20%, 50%, 60%, 30%, 40%, 30% and 20% of the GRNN models respectively. Seven PCs are generated by the PCA of the descriptor subset of the GRNN-developed model. Only the first 5 PCs were significant. The plots of M/P ratios against these PCs, obtained from the artificial testing sets, are given in Figure 5.3. The sixth and seventh PCs were

responsible for less than 12% of the total variance and thus are not shown. The shapes of the plots are more complicated than those in Figure 5.1 and Figure 5.2, suggesting that the transfer of drugs from plasma into breast milk may involve multiple mechanisms.

Figure 5.3 Plots of M/P ratio against the various PCs of M/P descriptor subset of GRNN.



e



The first PC was contributed by Mor23m, R4u and TIE. Mor23m is a representation of the three-dimensional structure of a molecule and encodes information about atomic masses in a molecule. R4u and TIE encode the 3D and 2D structure of a molecule respectively. GATS3e, Mor06p and HATS5e were grouped together in the second PC. Both GATS3e and HATS5e encode molecular structure and the group electronegativity of molecular substituents. Mor06p is a representation of the three-dimensional structure of a molecule and encodes information about the polarizability of a molecule. The third PC was determined primarily by HATS5e and TIE. HOMT encodes information about the degree of electron delocalization in the molecule and was involved in the fourth PC with R4u. The fifth PC was contributed mainly by Mor06p.

The information encoded by the current descriptor subset can be broadly grouped into electronic factors such as π electrons distribution, charge-transfer properties and molecular charge and steric factors like molecular shape and size. These factors have also been found by other studies to be important for the prediction of M/P ratios (Atkinson *et al.* 1990; Begg *et al.* 1992; Agatonovic-Kustrin *et al.* 2000; Agatonovic-Kustrin *et al.* 2002). The sigma values of the descriptors suggest

electronic properties were more important than steric factors in determining M/P ratios.

5.3.4 General considerations

In the present study, the prediction capabilities of the GRNN-, MLR- MLFN-developed models were assessed by using independent validation sets. It is important that the results for the independent validation sets truly reflect the generalization ability of the QSPkR models. It has been suggested that both training and validation sets should be diverse and the validation sets should be representative of the training sets (Rajer-Kanduc *et al.* 2003; Schultz *et al.* 2003). The diversity index (DI) of the training and validation sets used in the present study are 0.321 and 0.405, 0.135 and 0.341, and 0.220 and 0.309 for BBB penetration, HSA binding and M/P distribution respectively. This suggests that the training and validation sets used in this work are sufficiently diverse. The representativity index (RI) between each of the training sets and its corresponding validation sets are 0.752, 0.590, and 0.645 for BBB penetration, HSA binding and M/P distribution respectively. These RI values suggest that the validation sets are representative of the training sets.

The GRNN models developed in this study may not be the optimum because of the correlations and overlapping among the chemical descriptors, while Parzen's nonparametric estimator (Equation (2.18)) normally requires that these descriptors are statistically independent. However, various studies have shown that descriptor correlation does not drastically affect the predictive ability of a GRNN model (Currit 2002; Mosier *et al.* 2002; Mosier *et al.* 2003; Niwa 2003). In one study, good predictive results were obtained even with pairwise correlation between descriptors of up to 0.94 (Currit 2002). The maximum pairwise correlation between descriptors for

the BBB penetration, HSA binding and milk-plasma distribution study is 0.561, 0.291 and 0.476 respectively. Thus intercorrelation among the descriptors is not expected to significantly affect the predictive ability of the GRNN models generated in this study. However, intercorrelation among the descriptors may increase the complexity of the GRNN models by obscuring models consisting of fewer or more interpretable descriptors (Mosier *et al.* 2003). This problem can be partially alleviated by using PCs instead of individual descriptors for the explanation of the QSPkR model. The grouping of similar descriptors in a single PC enables the explanation of the GRNN models to be made in terms of simple molecular characteristics instead of the more abstract molecular descriptors.

The CPU time needed for developing GRNN models is faster than that of MLFN but slower than that of MLR. During the prediction process, GRNN-developed models require substantially higher memory and CPU time than models developed by using MLR and MLFN, especially when large training sets are involved. This is because GRNN uses every compound in the training set to facilitate the prediction of the property of new compounds. Such a problem can be alleviated by the use of parallel computing algorithms.

5.4 Conclusion

Results from this work suggest that GRNN is a potentially useful method for developing QSPkR models from a diverse set of drug data. QSPkR models developed using GRNN for three drug distribution properties – BBB penetration, HSA binding, and M/P distribution – were tested and compared with those developed by using a linear method, MLR, and a non-linear method, MLFN. All the GRNN-developed models showed better prediction capability than the corresponding MLR- or MLFN-developed models. This suggests that the GRNN-developed models are not more flexible than is necessary and thus unlikely to overfit. Most of the non-linear methods including neural networks are incapable of providing explicit relationships between the predicted properties and the molecular features of the compounds. The use of multi-sigma GRNN models and PCA may be helpful for partially solving this problem. The individual values for each descriptor provide a useful hint about its contribution to the distribution properties. PCA, when coupled with specially designed artificial testing sets, may provide a rough guide for the influence of molecular characteristics on drug distribution properties. Future development of descriptors that can be easily translated back to the molecular structure will further enhance the interpretability of GRNN developed models.

Chapter 6

Prediction of Drug Metabolism and Elimination, Part I: Classification Methods

The use of consensus SVM model strategy to improve the prediction accuracies of substrates and inhibitors of three cytochrome P450 isoenzymes, 3A4, 2D6 and 2C9 is presented in this chapter. Physicochemical and structural properties of compounds that are important for the identification of substrates and inhibitors and factors that may affect the prediction accuracies are discussed.

6.1 Introduction

Drug metabolism is a process whereby a drug is modified by a metabolizing enzyme, and these processes play important roles in pharmacokinetics and therapeutic actions of drugs (van de Waterbeemd *et al.* 2003). For instance, lipophilic drugs need to be metabolized to hydrophilic metabolites so that they can be readily excreted (Smith *et al.* 1997a). Although the primary site of drug metabolism is in the liver, metabolism can also occur in the intestines, blood and other tissues.

Profiles of drug metabolism has increasingly become an important consideration in early stages of drug development because of the profound effect of metabolism on such important drug properties as metabolic stability, drug-drug interactions and drug toxicity (Li 2001; van de Waterbeemd *et al.* 2003). Lower metabolic stability of a drug generally reduces its efficacy as it becomes more difficult to reach an adequate therapeutic concentration at a target site. Whereas

higher metabolic stability of a drug may lead to harmful effect because of the prolonged half-life (Keseru 2001). A significant portion of adverse drug reactions has been attributed to drug-drug interactions that involve the interference of the normal metabolism of a drug due to the inhibition or induction of its metabolic enzyme by another drug (Ekins *et al.* 2001; Molnar *et al.* 2002). Drug metabolism is also known to produce metabolites more toxic than their parent compound (Li *et al.* 1995).

There are mainly two phases in drug metabolism processes. The first involves phase I enzymes responsible for drug oxidation, reduction or hydrolysis. The second involves phase II enzymes responsible for drug conjugation of the phase I metabolite with a water-solubilizing endogenous moiety (Long *et al.* 2003). The cytochrome P450 isoenzymes are responsible for most of the phase I metabolism processes (Smith *et al.* 1997a; de Groot *et al.* 2002), with CYP3A4, CYP2D6 and CYP2C9 mediating the metabolism of nearly 70% of all phase I metabolism (Lewis *et al.* 2002). CYP3A4 is responsible for the metabolism of over 50% of drugs (Smith *et al.* 1997a; Smith *et al.* 1997b; Zuegge *et al.* 2002) and its ability to metabolize a wide variety of drugs of varying molecular weight and physicochemical properties is attributed to its relatively large active site that facilitates weak hydrophobic interactions with its substrates (Smith *et al.* 1997a; Smith *et al.* 1997b; Long *et al.* 2003). CYP2D6 is a polymorphic enzyme primarily responsible for the metabolism of substrates containing a basic nitrogen (Langowski *et al.* 2002), which includes antiarrhythmics, antidepressants and beta-blockers (Susnow *et al.* 2003). Its metabolism activity is in many cases facilitated by an ion pair interaction between an aspartic acid residue at the active site and a protonated nitrogen atom of the substrate (Langowski *et al.* 2002). CYP2C9 is primarily involved in the metabolism of many polar drugs that are ionized at physiological pH, such as ibuprofen, naproxen, diclofenac and

sulphaphenazole (Smith *et al.* 1997b; Ekins *et al.* 2000a). Most of the substrates of CYP2C9 contain an aromatic group, and drug-enzyme interaction has been attributed to the π-π interactions between the aromatic groups of the substrate and specific residue at the binding site (Langowski *et al.* 2002) and hydrogen bonding (de Groot *et al.* 2002). Therefore, prediction of inhibitors, substrates and inducers of these P450 isoenzymes is important for analysis of drug metabolism and for developing efficient tools for screening drugs of appropriate metabolism profiles.

Several computer prediction systems have been developed by using statistical learning methods for identification of inhibitors of specific P450 isoenzymes. Zuegge et al (Zuegge *et al.* 2002) developed a filter for predicting CYP3A4 inhibition by using a linear partial least square-based approach, which gives an accuracy of 93% for 29 inhibitors and 86% for 21 non-inhibitors. Another filter for prediction of CYP3A4 inhibition was developed by Molnar and Keseru (Molnar *et al.* 2002) by using neural networks, which gives an accuracy of 91.7% for 36 inhibitors and 88.9% for 36 non-inhibitors respectively. A consensus filter for predicting CYP2D6 inhibitors was developed by Susnow and Dixon (Susnow *et al.* 2003) using recursive partitioning, which gives an accuracy of 100% for 10 inhibitors and 76% for 41 non-inhibitors. Ekin et al (Ekins *et al.* 2003) also used recursive partitioning to develop filters for predicting CYP3A4 and CYP2D6 inhibitors, which gives a Spearman's ρ value of 0.48 and 0.61 for a test set of 98 compounds respectively. The success of these methods raises an interest in the exploration of other statistical learning methods that have been used in a variety of drug studies (Trotter *et al.* 2001; Doniger *et al.* 2002; Cai *et al.* 2003).

The aim of this work is to explore the use of support vector machine (SVM) for facilitating the prediction of substrates and non-substrates, and inhibitors and non-

inhibitors of P450 isoenzymes. A genetic algorithm-based descriptor selection method (Gao *et al.* 2002; Frohlich *et al.* 2003) is used to select relevant molecular descriptors for SVM classification of the substrates and inhibitors of P450 isoenzymes. Because of the high number of redundant and overlapping descriptors, many sets of descriptors, which describe similar overall physicochemical properties but are derived from slightly different algorithms and parameters, can be selected by this genetic algorithm (GA) with different random seed. Consensus modeling strategy has been introduced for developing prediction systems based on multiple descriptor sets (Gramatica *et al.* 2004). In this work, this strategy was applied to the development of consensus SVM (CSVM) classification systems for the prediction of inhibitors and substrates of P450 isoenzymes by using multiple descriptor sets generated from GA of different seeds.

Our method was first applied to the prediction of the inhibitors of CYP3A4 and CYP2D6 by using a substantially higher number of inhibitors and non-inhibitors than those in earlier studies (Molnar *et al.* 2002; Zuegge *et al.* 2002; Susnow *et al.* 2003), which serves as a test of the capability of our method. It was then used for the prediction of the inhibitors of CYP2C9 and substrates of CYP3A4, CYP2D6 and CYP2C9. The relevance of the selected descriptors by the CSVM methods to drug interactions with P450 isoenzymes is discussed.

6.2 Methods

6.2.1 Datasets

Inhibitors and substrates of CYP3A4, CYP2D6 and CYP2C9 P450 isoenzymes were collected from various sources (Lacy *et al.* 2002; Rendic 2002;

Flockhart 2003; MICROMEDEX 2003a). In order to ensure that interlaboratory variations in experimental protocols do not significantly affect the quality of the data sets, the most common range of Ki values for the compounds investigated in more than one sources was used to select compounds as inhibitors or substrates (Susnow *et al.* 2003). The generated datasets are composed of 241 inhibitors and 368 substrates for CYP3A4, 180 inhibitors and 198 substrates for CYP2D6, and 167 inhibitors and 144 substrates for CYP2C9. Non-inhibitors and non-substrates are seldom described in the literature and few of these compounds are specified in a known chemical database. For instance, a comprehensive search of the literature sources (Lacy *et al.* 2002; Rendic 2002; Flockhart 2003; MICROMEDEX 2003a) identified only seven non-inhibitors and six non-substrates for CYP3A4, nine non-inhibitors and eight non-substrates for CYP2D6, and eight non-inhibitors and seven non-substrates for CYP2C9. In an earlier study of the prediction of CYP3A4 inhibitors (Molnar *et al.* 2002), non-inhibitors of the enzyme were selected from those well-studied compounds that are known inhibitors/substrates/agonists of proteins other than that enzyme and there is no report that any of these is an inhibitor of that enzyme. Such a method is based on the assumption that, as they have been well studied, if these compounds have not been reported to be inhibitors or substrates of a specific enzyme, it is highly likely that they are not. In this work, this method was used to generate non-inhibitors or non-substrates of the P450 isoenzymes. From this procedure, 461 non-inhibitors and 334 non-substrates for CYP3A4, 522 non-inhibitors and 504 non-substrates for CYP2D6, and 535 non-inhibitors and 558 non-substrates for CYP2C9 were generated. Substrates and inhibitors of an isoenzyme were denoted as belonging to the positive class ($D+$) and non-substrates and non-inhibitors of the isoenzyme were denoted as belonging to the negative class ($D-$) of the isoenzyme.

Representative training and validation sets were constructed from the datasets by using the removal-until-done method (section 2.2.2.3). The number of compounds in the training and validation sets for the inhibitors or substrates of each of these enzymes are given in Table 6.1.

Table 6.1 Number of compounds in the training, independent validation, modeling training and modeling testing sets for the inhibitors/substrates of different cytochrome P450 isoenzymes.

Dataset	**CYP**	**Training set**		**Validation set**		**Modeling training set**		**Modeling testing set**	
		D+[a]	***D-***[b]	***D+***[a]	***D-***[b]	***D+***[a]	***D-***[b]	***D+***[a]	***D-***[b]
Inhibitors /	3A4	216	386	25	75	196	306	20	80
non-inhibitors	2D6	160	442	20	80	143	359	17	83
	2C9	149	453	18	82	134	368	15	85
Substrates /	3A4	312	290	56	44	256	246	56	44
non-substrates	2D6	169	433	29	71	149	353	20	80
	2C9	130	472	14	86	121	381	9	91

[a] Inhibitors or substrates
[b] Non-inhibitors or non-substrates

Prediction accuracy of statistical learning systems is known to be strongly affected by the diversity of samples used in the training set (Rajer-Kanduc *et al.* 2003; Schultz *et al.* 2003). Independent validation sets have frequently been used for evaluating the predictive performance of these classification systems, and these need also to be diverse and representative of the samples studied in order to accurately assess the capabilities of the prediction systems (Rajer-Kanduc *et al.* 2003; Schultz *et al.* 2003). The diversity index (DI) of the six training sets and six validation sets are in the range between 0.001 and 0.005 and between 0.002 and 0.020 respectively. The low DI value of the *D*+ compounds and *D*- compounds for all of the training and

validation sets suggest that these datasets are sufficiently diverse. The representativity index (RI) value between each of the training sets and its corresponding validation set is in the range between 0.446 and 0.511, which suggests that these validation sets are representative of their corresponding training sets and these validation sets are suitable for assessing the systems developed in this work.

6.2.2 Molecular structures and descriptors

This study used the same set of 1497 DRAGON molecular descriptors as the distribution study (section 5.2.2). Moreover, an additional set of 105 electrotopological state descriptors (Kier *et al.* 1999) and 5 linear solvation energy relationship descriptors (Platts *et al.* 1999) were computed by using our own developed code. Our code has been tested on a number of compounds used in earlier studies to ensure the accuracy of the computed descriptors.

6.2.3 Descriptor selection

A GA (section 2.3.3.2) was used to remove descriptors irrelevant to the prediction of CYP450 inhibitors and substrates. The retained descriptors from this process were used for representing the compounds studied in this work. All of the descriptors in the training set were first normalized in the range of -1 to 1 by using equation (2.4) before applying the GA-based descriptor selection method. At the end of the GA-based descriptor selection process, the highest ranked descriptor subset was used to construct the final SVM classification system.

In the descriptor selection process, ranking of the different descriptor subsets can be determined by using either 10-fold cross-validation, 5-fold cross-validation or

a modeling testing set. Our analysis of the 30 P450 isoenzyme SVM classification systems derived from each of these cross-validation methods showed that the modeling testing method gives the best performance, and thus this validation method was used in all of the descriptor selection processes in this study. The modeling testing set was derived by dividing the original training set into a modeling training set and modeling testing set of 502 and 100 compounds respectively by using the removal-until-done method (section 2.2.2.3). The modeling training and modeling testing sets for the inhibitors or substrate of each of these enzymes are given in Table 6.1 above. The modeling training set was used for constructing the SVM classification systems in the GA. Matthews correlation coefficient (MCC) (equation (2.33)) was used as the fitness function for GA optimization.

6.2.4 CSVM methods

Two types of CSVM methods were used. The first is a 'positive majority' consensus SVM classification system (PM-CSVM), which classifies a compound as *D*+ if the majority of its SVM classification systems classify the compound as *D*+ (Eriksson *et al.* 2003). A PM-CSVM requires an odd number of SVM classification systems to prevent ambiguity in its prediction. The second is a 'positive probability' consensus SVM classification system (PP-CSVM), which explicitly computes the probability for a compound to be *D*+ using the following formulas (McDowell *et al.* 2002):

$$\Pr(S_i^+ \mid P_i) = \frac{\Pr(S_{i-1}^+ \mid P_{i-1})\alpha_i^+}{(1-\alpha_i^-) + (\alpha_i^+ + \alpha_i^- - 1) \times \Pr(S_{i-1}^+ \mid P_{i-1})} \tag{6.1}$$

$$\Pr(S_i^+ \mid P_i) = \frac{\Pr(S_{i-1}^+ \mid P_{i-1}) \times (1-\alpha_i^+)}{\alpha_i^- - (\alpha_i^+ + \alpha_i^- - 1) \times \Pr(S_{i-1}^+ \mid P_{i-1})} \tag{6.2}$$

where $\Pr(S_i^+ \mid P_i)$ is the posterior probability that a compound is *D*+ given the classification result from SVM classification system *i* and α_i^+ and α_i^- is the sensitivity (SE) and specificity (SP) of SVM classification system *i* respectively. Equation (6.1) or (6.2) was used when SVM classification system *i* classifies the compound as *D*+ or *D*- respectively. In the absence of the knowledge about the ratio of *D*+ to *D*- compounds in the population, the prior probability of a compound to be *D*+ is tentatively set at 0.5. Sensitivity and specificity of SVM classification system *i* were estimated by using the validation method of the descriptor selection process.

There are two methods for using GA-based descriptor selection process to find all optimized SVM classification systems for consensus modeling. The first method is to perform a single run of GA-based descriptor selection and record all the SVM classification systems in the final population that have a certain level of accuracy. The second method is to perform multiple runs of GA-based descriptor selection using different random seeds and select the best SVM classification system from each run for consensus modeling. The current study uses the second method to obtain SVM classification systems for consensus modeling because our analysis of the two methods showed that the top few SVM classification systems from the first method tends to be similar to one another whereas SVM classification systems from the second method tends to be more diverse.

To determine an appropriate number of SVM classification systems for the CSVM methods, the descriptor selection process was repeated for 101 times, producing a pool of SVM classification systems. SVM classification systems were randomly selected, with replacement, from the pool of SVM classification systems to form nine classes of CSVMs, each containing 11, 21, 31, 41, 51, 61, 71, 81 or 91 SVM classification systems. This random selection of SVM classification systems

from the pool of SVM classification systems and construction of CSVMs were repeated 1000 times. Our analysis of these nine CSVMs classes showed that the best accuracies for the two types of CSVM methods were obtained when at least 81 SVM classification systems were used to develop CSVMs, and the accuracies roughly level off at higher number of SVM classification systems. Thus, 81 SVM classification systems appear to be the optimum number of systems for constructing CSVMs, which are used for developing CSVMs for all the datasets in this work.

6.3 Results

The SVM classification system with the best cross-validation accuracies was selected from the 81 SVM classification systems as the "best-trained" single SVM classification system. This selection method has been used by other studies that used GA as the descriptor selection method (Sutherland *et al.* 2003b). A PM-CSVM and a PP-CSVM were constructed by using the 81 SVM classification systems. The prediction accuracies of these three systems were determined by using the independent validation set, which are given in Table 6.2.

Table 6.2 Accuracies of the "best-trained" single SVM classification systems, PM-CSVM and PP-CSVM for the prediction of CYP3A4 and CYP2D6 inhibitors/non-inhibitors by using the independent validation sets.

CYP	Classification system	TP	FN	TN	FP	SE	SP	Q	MCC
3A4	"Best-trained" single SVM classification system	20	5	72	3	80.0	96.0	92.0	0.782
	PM-CSVM	21	4	75	0	84.0	**100.0**	**96.0**	**0.893**
	PP-CSVM	23	2	73	2	**92.0**	97.3	**96.0**	**0.893**
2D6	"Best-trained" single SVM classification system	15	5	77	3	75	96.3	92.0	0.742
	PM-CSVM	16	4	78	2	80.0	**97.5**	**94.0**	0.807
	PP-CSVM	18	2	76	4	**90.0**	95.0	**94.0**	**0.821**

It is found that both CSVM methods give better accuracies than that of the "best-trained" single SVM classification system. Moreover, PP-CSVM gives similar sensitivities and slightly better specificities, while PM-CSVM gives slightly lower sensitivities and slightly better specificities than those of earlier classification systems for prediction of inhibitors of CYP3A4 (Molnar *et al.* 2002; Zuegge *et al.* 2002) and CYP2D6 (Susnow *et al.* 2003). Thus PP-CSVM appears to be more useful than PM-CSVM for predicting inhibitors and substrates of P450 isoenzymes.

The accuracies of PP-CSVM for the prediction of inhibitors of CYP2C9 and substrates of CYP3A4, CYP2D6 and CYP2C9 are given in Table 6.3. The prediction accuracies of these CSVMs are at a similar level as those of the inhibitors of CYP3A4 and CYP2D6, which suggest that these CSVM methods, particularly PP-CSVM, are generally useful for predicting both the inhibitors and substrates of different P450 isoenzymes.

Table 6.3 Accuracies of PP-CSVM for the prediction of CYP2C9 inhibitors/non-inhibitors and CYP3A4, CYP2D6, and CYP2C9 substrates/non-substrates by using the independent validation sets.

Dataset	CYP	TP	FN	TN	FP	SE (%)	SP (%)	Q (%)	MCC
Inhibitors / non-inhibitors	2C9	16	2	79	3	88.9	96.3	95.0	0.835
Substrates / non-substrates	3A4	55	1	40	4	98.2	90.9	95.0	0.899
	2D6	28	1	67	4	96.6	94.4	95.0	0.884
	2C9	12	2	85	1	85.7	98.8	97.0	0.872

6.4 Discussion

6.4.1 Overall prediction accuracies

The difference between the specificities of the current CSVMs and those of classification systems from earlier studies may be due to the difference in the number and diversity of *D*- compounds used for training the classification systems. In our work, the number of *D*- compounds in the training set ranges from 290 to 472, whereas earlier classification systems were developed by using 41 to 145 *D*-compounds. Statistical learning methods require a large number of compounds for development of classification systems. In addition, diversity of the training sets has been shown to affect the applicability domain of qSPkR models (Dimitrov *et al.* 2005). Therefore it is not surprising that the methods of the current work, which uses a more diverse and larger number of *D*- compounds, give higher specificities than those of earlier studies. Another possible reason for the improved specificities is the use of SVM, which has been found to be consistently superior to other classification methods in most classification problems (Burbidge *et al.* 2001; Czerminski *et al.* 2001; Meyer *et al.* 2003).

For all of the datasets, with the exception of the CYP3A4 substrates/non-substrates dataset, the number of *D*- compounds is always higher than the number of *D*+ compounds. This may create a bias of the SVM classification systems to predict unknown compounds as *D*-, resulting in higher number of false negatives. However, previous studies suggest that SVM are not significantly affected by unbalanced datasets (Cai *et al.* 2003; Lessmann 2004), especially if there are more than 80-100 compounds of each class in the training set (Han *et al.* 2004). All of the datasets used in this work contains at least 130 compounds of each class in the training set and thus the unbalanced dataset is not expected to significantly affect the predictive ability of the SVM classification systems.

6.4.2 Evaluation of prediction performance

The results of our SVM systems were compared with those of several statistical learning methods including multiple linear regression (MLR), partial least squares (PLS), logistic regression (LR), C4.5 decision tree (DT) and *k* nearest neighbour (kNN). GA was used to determine the optimum descriptor subsets for each of these classification methods by using 30 different random seeds, from which 30 separate classification models were generated for each method. The prediction accuracies of these classification models were determined by using the independent validation set. Table 6.4 gives the results for CYP3A4 substrates/non-substrates. The accuracies for the other P450 isoenzymes datasets are similar and thus are not given here. It was found that the SVM classification systems give the highest prediction accuracies than those of other methods.

Table 6.4 Average accuracies of different statistical learning classification systems for the prediction of CYP3A4 substrates/non-substrates by using independent validation sets.

Classification method	SE (%)[a]	SP (%)[a]	Q (%)[a]	MCC[a]
MLR	86.1 (3.9)	71.4 (4.4)	79.6 (2.9)	0.586 (0.060)
LR	83.8 (3.9)	71.0 (5.1)	78.1 (3.0)	0.555 (0.063)
PLS	79.9 (5.8)	72.5 (5.2)	76.7 (3.7)	0.528 (0.073)
C4.5 DT	75.5 (6.8)	66.4 (6.7)	71.5 (4.3)	0.423 (0.087)
kNN	92.4 (2.0)	82.6 (3.4)	88.1 (1.7)	0.759 (0.034)
SVM	**98.0 (1.4)**	**85.3 (3.1)**	**92.4 (1.2)**	**0.849 (0.024)**

[a] Numbers in parenthesis are the standard deviations.

To determine whether the selected descriptors of the SVM classification systems include those irrelevant for the prediction of the inhibitors or substrates of the respective enzymes, 10 groups of classification systems were generated by using the GA-based descriptor selection method. These groups are SVM_{100}, SVM_{200}, SVM_{300}, SVM_{400}, SVM_{500}, SVM_{600}, SVM_{700}, SVM_{800}, SVM_{900}, and SVM_{1000}, in which the subscript denotes the number of descriptors used. Each group contains 30 SVM classification systems. The prediction accuracies of these SVM classification systems were determined by using the independent validation sets. Table 6.5 gives the results for the CYP3A4 substrates/non-substrates, which shows that prediction accuracies begin to decrease when more than 400 descriptors are used in a SVM classification system. This suggests that the maximum number of relevant descriptors for the CYP3A4 substrates/non-substrates dataset is around 400. Because the original 81 SVM classification systems for the CYP3A4 substrates/non-substrates dataset contain 214 to 402 descriptors, our results seem to suggest that the original 81 SVM classification systems are unlikely to contain irrelevant descriptors. Similar

conclusions are also made for the rest of the P450 isoenzymes datasets based on our computational studies.

Table 6.5 Average accuracies of 10 groups of SVM classification systems for the prediction of CYP3A4 substrates/non-substrates by using independent validation sets.

Number of descriptors	SE (%)[a]	SP (%)[a]	Q (%)[a]	MCC[a]
100	93.0 (3.1)	80.4 (4.4)	87.5 (2.7)	0.747 (0.054)
200	96.7 (2.0)	83.0 (3.3)	90.7 (1.9)	0.814 (0.039)
300	98.0 (1.6)	85.6 (3.6)	92.6 (1.9)	0.853 (0.037)
400	98.0 (1.3)	82.4 (3.4)	91.1 (1.6)	0.825 (0.032)
500	98.2 (1.0)	80.9 (3.1)	90.6 (1.4)	0.815 (0.028)
600	98.6 (0.8)	74.5 (3.3)	88.0 (1.5)	0.769 (0.028)
700	99.3 (0.9)	66.4 (5.4)	84.8 (2.3)	0.715 (0.040)
800	100.0 (0.0)	51.5 (3.1)	78.7 (1.4)	0.611 (0.024)
900	99.9 (0.3)	45.7 (2.4)	76.1 (1.0)	0.565 (0.017)
1000	100.0 (0.0)	37.3 (3.2)	72.4 (1.4)	0.500 (0.026)

[a] Numbers in parenthesis are the standard deviations.

It has been shown that chance correlations may occur during descriptor selection especially if the number of descriptors available for selection is large (Topliss *et al.* 1979; Jouan-Rimbaud *et al.* 1996). Y-randomization (section 2.5.2) has been frequently used to determine the probability of chance correlation during descriptor selection processes (Manly 1997; Leardia *et al.* 1998). In this work, y-randomization was repeated for 81 times. The average Matthews correlation coefficient of these scrambled SVM classification systems derived by using the independent validation sets were found to be in the range between 0.189 and 0.288, which are significantly lower than those of the original SVM classification systems, which are in the range between 0.783 and 0.852. This suggests that the original SVM

classification systems are relevant and unlikely to arise as a result of chance correlation.

A frequently used method for checking whether a prediction system is overfitted is to compare the prediction accuracies determined by using cross-validation methods with those determined by using independent validation sets (Hawkins 2004). Because descriptor selection was performed by using the modeling testing sets as the cross-validation method, an overfitted classification system is expected to have much higher prediction accuracy for the modeling testing sets than for the independent validation sets. As shown in Table 6.6, the prediction accuracies of the SVM systems based on the modeling testing sets and those based on independent validation sets are similar. This suggests that the SVM classification systems in this work are unlikely to overfit.

Table 6.6 Comparison of the average accuracies of SVM classification systems for the prediction of inhibitors/substrates of different P450 isoenzymes by using modeling testing sets and independent validation sets.

Dataset	CYP	Modeling testing set[a]				Independent validation set[a]			
		SE (%)	SP (%)	Q (%)	MCC	SE (%)	SP (%)	Q (%)	MCC
Inhibitors / non-inhibitors	3A4	76.5 (6.2)	98.8 (1.3)	94.3 (0.8)	0.817 (0.026)	82.1 (4.5)	97.9 (1.5)	93.9 (1.3)	0.835 (0.036)
	2D6	79.1 (7.3)	98.5 (1.4)	95.2 (0.8)	0.828 (0.028)	79.3 (5.4)	96.7 (1.6)	93.2 (1.7)	0.783 (0.054)
	2C9	81.9 (4.7)	98.8 (1.0)	96.3 (0.6)	0.851 (0.025)	86.4 (5.0)	97.3 (1.3)	95.3 (1.1)	0.842 (0.039)
Substrates / non-substrates	3A4	96.3 (1.5)	86.7 (2.7)	92.1 (0.8)	0.841 (0.015)	98.0 (1.3)	85.2 (3.0)	92.4 (1.3)	0.849 (0.026)
	2D6	84.6 (5.0)	98.9 (1.3)	96.0 (0.6)	0.874 (0.018)	86.9 (4.7)	96.9 (1.5)	94.0 (1.7)	0.852 (0.043)
	2C9	77.0 (8.2)	98.9 (1.0)	97.0 (0.8)	0.810 (0.047)	72.3 (7.9)	99.2 (0.9)	95.4 (1.1)	0.801 (0.051)

[a] Numbers in parenthesis are the standard deviations.

6.4.3 The selected descriptors

The majority of the selected descriptors in our SVM classification systems are composite descriptors, which can be divided into three groups: 3D-MoRSE, RDF and Randic molecular profiles. 3D-MoRSE descriptors, which are representations of the 3D structure of a molecule and encode features such as molecular weight, van der Waals volume, electronegativities and polarizabilities, have been used for the classification of dopamine D1 and D2 agonists and modeling the binding of steroids to corticosteroid binding globulin (Schuur *et al.* 1996). RDF descriptors provide information about bond lengths, ring types, planar and nonplanar systems, atom types,

and molecular weight and have been used for pharmacokinetic studies (Wegner *et al.* 2004). Randic molecular profiles measure interactions between atoms in a molecule and encode information on molecular shape, which is an important factor in ligand-enzyme interactions. Because shape and chemical complementarily between a ligand and an enzyme are important for ligand-enzyme binding, it is not surprising that these three classes of 3D descriptors, which provide information on hydrophobicity, electronegativities, polarizabilities and shape of a molecule, are frequently selected by the descriptor selection process.

Because composite descriptors encode multiple physicochemical and structural aspects of the molecule, it is difficult to extract from these descriptors information about which specific molecular characteristics are important for the inhibitors and substrates of these P450 isoenzymes. Nonetheless, it is possible to infer some information from non-composite descriptors. As many descriptors are overlapping and some of them are redundant, it is more appropriate to group them into classes of descriptors of similar properties and discuss their contribution to the inhibitor/substrates predictions at the class level. Table 6.7 gives the classes of non-composite descriptors selected by our computations. It is found that shape is the dominant factor involved in ligand-P450 isoenzyme interaction. This is not surprising because shape complementarity is important for ligand-protein interactions. In addition to the shape descriptors, electrostatic and hydrophobic interactions are found to be the dominant forces involved in ligand-P450 isoenzyme interaction. Descriptors that describe hydrogen bonding, also appear to be important for the ligand-P450 isoenzyme interactions, which is consistent with the findings that hydrogen bonds are involved in the ligand-P450 isoenzyme interactions (de Groot *et al.* 2002).

Table 6.7 Important descriptor classes selected for the prediction of inhibitors/substrates of different P450 isoenzymes.

Dataset	CYP	Electrostatic (%)	Hydrogen bond acceptors (%)	Hydrogen bond donors (%)	Hydrophobic (%)	Shape (%)	Size (%)
Inhibitors / non-inhibitors	3A4	20.4	3.6	3.3	8.8	56.8	7.1
	2D6	20.5	2.4	2.5	10.0	57.1	7.5
	2C9	20.1	2.0	2.9	8.8	59.0	7.2
Substrates / non-substrates	3A4	21.0	2.8	1.9	9.5	57.2	7.5
	2D6	18.9	3.1	3.5	8.5	59.7	6.3
	2C9	19.1	3.5	3.0	9.4	58.2	6.8

It is also possible to roughly distinguish between *D*+ and *D*- compounds and to roughly distinguish between inhibitors and substrates from the values of six selected descriptors, S, nHAcc, nHDon, MLOGP, MW, and SPH. These descriptors are representative of the four dominant interaction forces, electrostatic, hydrogen bond acceptor, hydrogen bond donor and hydrophobicity, and size and shape of the compounds respectively. S is the combined dipolarity/polarizability, nHAcc and nHDon, are the number of acceptor and donor atoms for hydrogen bonds respectively, MLOGP is the Moriguchi Log P (Moriguchi *et al.* 1992), MW is the molecular weight and SPH is the spherosity. The average values of these four descriptors for *D*+ and *D*- compounds of all the various datasets are given in Table 6.8. Substrates of CYP3A4 are generally larger in size, less spherical in shape, more hydrophobic and have more hydrogen bonding sites than non-substrates. Inhibitors of CYP3A4 are generally less hydrophobic than substrates but are larger in size and contained more hydrogen bond donors and acceptors. Substrates of CYP2D6 are generally smaller in size, more hydrophobic than non-substrates and contain one hydrogen bond donor. There are

only minor differences between inhibitors and substrates of CYP2D6, which suggest that there is considerable overlap between the inhibitors and substrates of CYP2D6. Substrates of CYP2C9 generally are more hydrophobic than inhibitors of CYP2C9 but are smaller in size and have lesser hydrogen bonding capacity.

Table 6.8 Differences in the values of descriptors important for distinguish between *D*+ and *D*- compounds.

Dataset	CYP	Descriptor	Average value[a]	
			***D*+**	***D*-**
Inhibitors / non-inhibitors	3A4	S	2.56 (1.24)	2.36 (1.12)
		nHAcc	6.47 (4.05)	4.59 (2.64)
		nHDon	2.27 (2.44)	1.23 (1.40)
		MLogP	1.83 (2.02)	1.96 (2.06)
		MW	417 (185)	313 (116)
		SPH	0.77 (0.13)	0.77 (0.13)
	2D6	S	2.17 (1.00)	2.52 (1.20)
		nHAcc	4.57 (2.70)	5.47 (3.48)
		nHDon	1.57 (1.81)	1.59 (1.92)
		MLogP	2.54 (1.76)	1.70 (2.09)
		MW	355 (125)	346 (159)
		SPH	0.78 (0.13)	0.77 (0.13)
	2C9	S	2.56 (1.21)	2.39 (1.15)
		nHAcc	5.31 (2.65)	5.21 (3.50)
		nHDon	1.49 (1.52)	1.62 (1.99)
		MLogP	1.78 (2.11)	1.96 (2.02)
		MW	351 (123)	348 (159)
		SPH	0.76 (0.13)	0.78 (0.13)
Substrates / non-substrates	3A4	S	2.56 (1.15)	2.29 (1.17)
		nHAcc	5.53 (3.45)	4.91 (3.14)
		nHDon	1.72 (1.99)	1.44 (1.75)

	MLogP	2.20 (1.99)	1.60 (2.06)
	MW	379 (157)	315 (137)
	SPH	0.76 (0.13)	0.78 (0.13)
2D6	S	2.19 (1.08)	2.53 (1.18)
	nHAcc	4.10 (2.13)	5.68 (3.58)
	nHDon	1.15 (1.22)	1.76 (2.07)
	MLogP	2.51 (1.74)	1.68 (2.11)
	MW	319.6 (99.8)	360 (166)
	SPH	0.78 (0.14)	0.77 (0.13)
2C9	S	2.52 (1.26)	2.41 (1.14)
	nHAcc	4.69 (2.52)	5.38 (3.48)
	nHDon	1.03 (1.14)	1.73 (2.01)
	MLogP	2.05 (2.04)	1.88 (2.05)
	MW	326 (112)	354 (160)
	SPH	0.75 (0.14)	0.78 (0.13)

[a] Numbers in parenthesis are the standard deviations.

CYP3A4 has a relatively large active site that facilitates weak hydrophobic interactions with its substrates (Smith *et al.* 1997a; Smith *et al.* 1997b; Long *et al.* 2003). A pharmacophoric model of the substrates suggests that there are four important features: two hydrogen bond acceptor, one hydrogen bond donor and one hydrophobic region (Ekins *et al.* 1999b). Some of the descriptor classes frequently selected by the SVM classification systems for the prediction of substrates and non-substrates of CYP3A4 are related to the hydrophobicity and hydrogen bonding ability of the molecule. Examples of descriptors in these classes include ARR, which is the aromatic ratio, aaCH and aasC, which are electrotopological descriptors for carbons in aromatic rings, nHAcc and nHDon. The differences in the distribution of intermolecular forces between inhibitors and substrates of CYP3A4 suggest that the

inhibitors have less electrostatic and hydrophobic interactions and more hydrogen bonding at the binding site than the substrates.

The pharmacophoric model for substrates of CYP2D6 consists of a basic nitrogen atom and a flat hydrophobic region (Ekins *et al.* 2001; Langowski *et al.* 2002). Some of the frequently selected descriptor classes by SVM classification systems for predicting substrates and non-substrates of CYP2D6 match this model. Examples of descriptors in these classes include MAXDP, which is the maximal electrotopological positive variation topological descriptor and is related to the electrophilicity of the molecule, nN, which is the number of nitrogen atoms, and BLI, which is the Kier benzene-likeness index. These descriptor classes are also selected by the SVM classification systems for predicting inhibitors and non-inhibitors of CYP2D6. However, differences in the distribution of intermolecular forces between inhibitors of CYP2D6 suggest that the inhibitors may have increased electrostatic and hydrophobic interactions at the active site. This is consistent with the findings from pharmacophoric studies of inhibitors of CYP2D6 which suggests that the inhibitors have an additional region in which functional groups with lone pairs enhance inhibitory potency and a region for hydrophobic groups (Ekins *et al.* 2001).

Descriptors encoding aromaticity, polarity and hydrogen bond donors are frequently selected by SVM classification systems for predicting substrates and non-substrates of CYP2C9. These include aasC, which is the electrotopological state atom index for aromatic carbons, MAXDN, which is the maximal electrotopological negative variation topological descriptor and is related to the nucleophilicity of the molecule, and nHDon. These selected descriptors are consistent with the findings that the substrates of CYP2C9 are primarily polar compounds that contains an aromatic group and that drug-CYP2C9 interaction is mediated by both hydrogen bonding (de

Groot *et al.* 2002) and π-π interactions at the binding site (Langowski *et al.* 2002). The differences in the distribution of intermolecular forces between inhibitors and substrates of CYP2C9 suggest that the inhibitors have fewer hydrogen bonds but increased electrostatic interactions at the active site than the substrates.

6.4.4 Potential training errors and misclassified compounds

In this work, non-inhibitors and non-substrates were selected from those without a report identifying them as an inhibitor or a substrate. There is also a certain level of overlapping between non-inhibitors of different CYP subtypes, between non-inhibitors and non-substrates of a specific CYP subtype, and between non-inhibitors and substrates of a particular CYP subtype. A potential problem with this method is that a small number of true inhibitors or substrates may be selected as non-inhibitors or non-substrates (false negatives). The extent of training errors caused by false negatives can be roughly estimated by using experimentally confirmed non-inhibitors/non-substrates. However, there is only a limited number of experimentally confirmed non-inhibitors/non-substrates. In the CYP3A4 substrate/non-substrate validation set, only irbesartan is a known non-substrate (MICROMEDEX 2003a). In the CYP2C9 inhibitor/non-inhibitor validation set, only reboxetine is experimentally determined to be a non-inhibitor (MICROMEDEX 2003a). In the CYP2D6 substrate/non-substrate validation set, only nilvadipine is a known non-substrate (MICROMEDEX 2003a). In the CYP2D6 inhibitor/non-inhibitor validation set, only gatifloxacin is a known non-inhibitor (MICROMEDEX 2003a). All of these compounds, except irbesartan, were correctly predicted by the CSVMs to be non-inhibitors/non-substrates. These results, together with the reported high accuracies of the SVM classification systems for other systems (Sorich *et al.* 2003; Xue *et al.*

2004c), suggest that by using SVM (Vapnik 1995), the training errors caused by false negatives can be kept at a minimum.

Table 6.9 gives the list of compounds misclassified by more than 50% of the SVM classification systems for each dataset. A possible reason for the misclassification of some of these compounds is that some descriptor subsets may be inadequate to properly describe these compounds. Examples of these compounds are carbamazepine, chlorphenamine, cinnarizine, doxepin, methadone, olanzapine and zuclopenthixol, which contain two aromatic rings separated by an atom and irbesartan and lorsartan, which contain a highly polar tetrazole ring. Among the misclassified non-inhibitors or non-substrates, only irbesartan is a known non-substrate (MICROMEDEX 2003a). Oxomemazine is a known inducer and flurithromycin is a known inhibitor of CYP3A4 (Rendic 2002). Thus it may be possible that both oxomemazine and flurithromycin are actually false negatives as more than 60% of the CYP3A4 inhibitors in the dataset are both CYP3A4 inhibitors and substrates. Similarly, doxepin, which is a known CYP2D6 substrate (Rendic 2002), may also be a false negative as nearly 50% of the CYP2D6 substrates are both CYP2D6 substrates and inhibitors.

6.4.5 Comparison of the two CSVM systems

The results from our studies show that PP-CSVM gives slightly better accuracies than PM-CSVM. This is because individual SVM classification systems in PP-CSVM are ranked according to their accuracies and SVM classification systems with better accuracies have more influence on the final classification of a compound.

Table 6.9 List of misclassified compounds in this work[a].

Dataset	CYP	Misclassified compounds
Inhibitors / non-inhibitors	3A4	Pilocarpine (*D*+)
		Stiripentol (*D*+)
		Olanzapine (*D*+)
		Cyclophosphamide (*D*+)
	2D6	Lobeline (*D*+)
		Propafenone (*D*+)
		Reboxetine (*D*+)
		Sulconazole (*D*+)
		Doxepin (*D*-)
		Isoconazole (*D*-)
	2C9	Stiripentol (*D*+)
		Sulconazole (*D*+)
		Isoconazole (*D*-)
Substrates / non-substrates	3A4	Chlorphenamine (*D*+)
		Flurithromycin (*D*-)
		Irbesartan (*D*-)
		Oxomemazine (*D*-)
		Pargyline (*D*-)
		Pentazocine (*D*-)
		Sulindac (*D*-)
	2D6	Carbamazepine (*D*+)
		Cinnarizine (*D*+)
		Zuclopenthixol (*D*+)
		Domperidone (*D*-)
		Emedastine (*D*-)
	2C9	Cinnarizine (*D*+)
		Losartan (*D*+)
		Methadone (*D*+)

[a] All of the compounds misclassified by more than 50% of the 81 classification systems are included

This is different from PM-CSVM where all individual SVM classification systems, regardless of their accuracies, contribute equally to the final classification of a compound. Thus it is expected that PP-CSVM, by reducing the contribution from SVM classification systems with lower accuracies, gives better or at least equal accuracies as PM-CSVM.

There are two potential problems with PP-CSVM. The first is that the prior probability, which was tentatively set at 0.5, may not always be the most appropriate value for representing the ratio of *D*+ to *D*- compounds in the population. This problem can be partially solved by using a large number of individual SVM classification systems to construct a CSVM so that the influence of prior probability on the final classification result is reduced. In this study, we have found that the same classification results were obtained even when the prior probability was varied from 0.05 to 0.95 when 81 SVM classification systems used to construct the CSVM. The second problem is the difficulty in determining the true sensitivities and specificities of the individual SVM classification systems, which are required by equations (6.1) and (6.2). In the present study, sensitivities and specificities of the SVM classification systems were estimated by using the modeling testing set and have a mean absolute difference of 2.0% and 3.4% respectively from those derived by using the independent validation set. If sensitivities and specificities of the individual SVM classification systems derived from the independent validation set are used in PP-CSVM, the resultant CSVMs are found to give slightly higher accuracies, suggesting a possible need for a more accurate estimate of the performance of some SVM classification systems.

6.5 Conclusion

Results from this work are consistent with earlier studies which suggest that consensus classification systems give better predictive performance than single classification systems. All of the PP-CSVMs for predicting inhibitors/substrates of the three P450 isoenzymes, CYP3A4, CYP2D6 and CYP2C9, show high prediction accuracies, with improved specificities compared to earlier studies. A potential problem of this work is that the selection criteria for non-inhibitors and non-substrates may result in a small number of false negatives. However, the use of SVM in this work can help to achieve a balance between training errors and prediction accuracies. The accuracies of the SVM classification systems may also be improved by the addition of a correction factor to the SVM decision function. The present CSVMs are only suitable for distinguishing between inhibitors and non-inhibitors or substrates and non-substrates. With the availability of more detailed experimental data, it is possible to use multi-class SVM (Angulo *et al.* 2003) for classification of non-inhibitors, weak inhibitors and strong inhibitors or SVM regression (Smola *et al.*) for quantitative prediction of the Ki values of inhibitors. Our computational results suggest PP-CSVM is better than PM-CSVM for constructing CSVMs for classifying inhibitors and substrates of various P450 isoenzymes. Thus CSVMs, particularly PP-CSVM, are potentially useful for developing filters for prediction of inhibitors and substrates of P450 isoenzymes.

Chapter 7

Prediction of Drug Metabolism and Elimination, Part II: Regression Methods

This chapter describes three machine learning approaches for the prediction of total clearance. Several different sets of descriptors are compared for their usefulness in modeling total clearance. Important physicochemical and structural properties of a compound are also identified by using the new method that is introduced in Chapter 5.

7.1 Introduction

Drug clearance is measured by a quantity, total clearance (CL_{tot}), which is a proportionality constant describing the relationship between a substance's rate of transfer, in amount per unit time, and its concentration, in an appropriate reference fluid (Wilkinson 1981). Drug clearance occurs by perfusion of blood to the organs of extraction, which are generally the liver and the kidney (Smith *et al.* 2001b). The CL_{tot} value of a drug is an important pharmacokinetic parameter because it is directly related to bioavailability and drug elimination and can be used to determine the dosing rate and steady-state concentration of a drug (Toutain *et al.* 2004). Thus it is important to predict the CL_{tot} value of drug leads during drug discovery so that compounds with acceptable metabolic stability can be identified and those with poor bioavailability can be eliminated.

Traditionally, the CL_{tot} value of a drug candidate is obtained via *in vivo* and *in vitro* studies (Naritomi *et al.* 2001; Zuegge *et al.* 2001; Wajima *et al.* 2003a; Wajima *et al.* 2003b), which tends to be time-consuming and costly. Therefore, QSPkR modeling has recently been explored for predicting the CL_{tot} value of drug candidates (Karalis *et al.* 2002; Karalis *et al.* 2003; Turner *et al.* 2003b; Ng *et al.* 2004; Turner *et al.* 2004b) in an effort to eliminate undesirable compounds in a fast and cost-effective manner. An initial partial least squares (PLS) study conducted by Karalis et al (Karalis *et al.* 2002) using 272 structurally unrelated compounds failed to find any correlation between CL_{tot} and a large variety of molecular descriptors used in that study. Karalis et al (Karalis *et al.* 2003) then developed a partial least square (PLS) model and non-linear regression model for CL_{tot} by using 23 cephalosporins. The r^2 and q^2 values of the PLS-developed model are 0.775 and 0.731, while the r^2 value of the non-linear regression model is 0.804. These two studies suggest that multiple mechanisms may be involved in CL_{tot} and thus linear methods may not always be suitable for constructing QSPkR models for CL_{tot}. Another study for the prediction of CL_{tot} was done by Turner et al (Turner *et al.* 2003b) who used artificial neural network (ANN), which gives a r^2 value of 0.982 for a training set of 16 cephalosporins and a r^2 value of 0.998 for a validation set of 4 cephalosporins. Subsequently, Turner et al (Turner *et al.* 2004b) used a larger training set of 56 compounds to develop an ANN-based QSPkR model, which gives a r^2 value of 0.731 for a validation set of 6 compounds. These results suggest that non-linear methods may be useful for developing models for CL_{tot} prediction of structurally unrelated compounds. Two QSPkR models for CL_{tot} were developed by Ng et al (Ng *et al.* 2004) by using *k* nearest neighbour (kNN) and PLS. The kNN-developed QSPkR model gives a q^2 value of 0.77 for a training set of 38 antimicrobial compounds and a r^2

value of 0.94 for a validation set of 6 antimicrobial compounds. There are 68% of the 44 compounds having predicted CL_{tot} within twofold of actual values. For the PLS-developed QSPkR model, there are only 50% of the 44 compounds having predicted CL_{tot} within twofold of actual values and the q^2 value of this model is 0.09 for the training set and its r^2 value is 0.35 for the validation set. These results are consistent with the study of Turner (Turner *et al.* 2004b) and further confirm the usefulness of non-linear methods for developing QSPkR models for predicting CL_{tot}. All of the previous QSPkR models for predicting CL_{tot} have primarily been developed and tested by using a relatively small number of compounds (<70), which is significantly smaller in number and diversity than the number of compounds with known CL_{tot} data. Thus it is of interest to evaluate the prediction capabilities of QSPkR models that are developed by using much larger and more diverse datasets.

This work is intended to evaluate the capability of several statistical learning methods for predicting CL_{tot} by using 503 compounds found from a comprehensive literature search, which is substantially larger in number and more diverse in structure than those used in earlier studies. The methods used include general regression neural network (GRNN), support vector regression (SVR) and kNN. Different descriptor sets, which encode different combination of the structural and physiochemical properties of a compound, were also compared for their usefulness for constructing QSPkR models to predict CL_{tot}. Consensus modeling strategy has been introduced for developing prediction systems based on multiple models (Mosier *et al.* 2003; Asikainen *et al.* 2004). In this work, this strategy was also applied to the development of consensus QSPkR (cQSPkR) models for the prediction of CL_{tot} by using QSPkR models generated from different statistical learning methods.

7.2 Method

7.2.1 Dataset

Compounds with known human CL_{tot} values were selected from several sources including Micromedex (MICROMEDEX 2003b), a classic pharmacology textbook (Hardman *et al.* 2002) and a number of publications (Ito *et al.* 1998; Obach 1999; Naritomi *et al.* 2001; Turner *et al.* 2003b; Wajima *et al.* 2003a; Ng *et al.* 2004; Turner *et al.* 2004b). In order to ensure that experimental variations in determining CL_{tot} do not significantly affect the quality of our data sets, only CL_{tot} values obtained from healthy adult males and from intravenous administration were used for constructing the dataset. In addition, a number of compounds were excluded because they are known to possess certain molecular characteristics which do not permit reliable calculations of the molecular descriptors used in this study (Karalis *et al.* 2002). Examples of these compounds are quaternary ammonium compounds, molecules with complex chemical structures like amphotericin-B, aminoglycosides, vancomycin, and compounds containing one or more metal atoms. A total of 503 compounds were selected from this process and these were used as the dataset for this work. The CL_{tot} value for each of these compounds was log-transformed ($logCL_{tot}$) to normalize the data and to reduce unequal error variances (Neter *et al.* 1996). Representative training set and validation set were constructed from our dataset by using the removal-until-done method (section 2.2.2.3).

7.2.2 Molecular structures and descriptors

Six different sets of descriptors were used to describe the structural and physico-chemical properties of the compounds. The first set (DS-MIXED) contains a

number of commonly used descriptors, including 21 constitutional descriptors, six geometrical descriptors, 72 topological descriptors and 108 electrotopological state descriptors (Kier *et al.* 1999). The second set (DS-3DMoRSE) includes 224 3D-MoRSE descriptors (Schuur *et al.* 1996), which are representations of the 3D structure of a molecule and encode features such as molecular weight, van der Waals volume, electronegativities and polarizabilities. The third set (DS-ATS) is composed of 209 Moreau-Broto topological autocorrelation (ATS) descriptors (Moreau *et al.* 1980), which describes how molecular properties such as polarizability, charge, electronegativity, are distributed along the topological structure. The fourth set (DS-GETAWAY) consists of 340 GETAWAY descriptors (Consonni *et al.* 2002), which encodes both molecular structure and chemical information such as atomic mass, polarizability, van der Waals volume and electronegativity. The fifth set (DS-RDF) contains 203 RDF descriptors (Hemmer *et al.* 1999), which provides information about interatomic distances in the entire molecule and also other useful information such as bond distances, ring types, planar and non-planar systems, atom types and molecular weight. The last set (DS-WHIM) includes 126 WHIM descriptors (Bravi *et al.* 1997), which encodes information about the size, shape, symmetry, atom distribution and polarizability of a molecule. All of the descriptors were computed from the 3D structure of each compound using MODEL (Li *et al.* 2005b).

Objective feature selection is applied to all of the six sets of descriptors to remove descriptors irrelevant or redundant to the CL_{tot} of the compounds, so as to improve computation speed, performance and interpretability of predictive models. The first step involves the removal of all irrelevant descriptors such as constant descriptors. Redundant descriptors were then eliminated by removing one of the two descriptors with pairwise correlation coefficient of greater than 0.90 (Wessel *et al.*

1998; Mosier *et al.* 2002). The final number of descriptors for each descriptor set is 84, 109, 142, 155, 111 and 44 for DS-MIXED, DS-3DMoRSE, DS-ATS, DS-GETAWAY, DS-RDF, and DS-WHIM, respectively. All of the remaining descriptors in each descriptor set were autoscaled to a mean value of zero and a variance of one (equation (2.3)) to ensure that all descriptors have equal potential to affect the QSPkR model (Livingstone 1995b).

7.2.3 Optimization of the parameters of GRNN, SVR and kNN

Optimization of the parameters for GRNN, SVR and kNN was conducted by scanning the parameter value through a range from 1 to 30. The predictive capability of the QSPkR model developed from a particular parameter value can be determined by using cross-validation methods, such as 5-fold cross-validation, 10-fold cross-validation and modeling testing set. Our cytochrome P450 study has shown that the use of a modeling testing set gives the best performance for assessing the predictive capability of a model (section 6.2.3). Thus this validation method was used to select the optimum parameter for each statistical learning method in this study. The following function was used to measure the predictive capability of a QSPkR model (Wessel *et al.* 1998; Mosier *et al.* 2002):

$$F = MAE_{train} + \left| MAE_{train} - MAE_{test} \right| \qquad (7.1)$$

where MAE_{train} and MAE_{test} are the mean absolute error of the modeling training set and modeling testing set respectively. The modeling testing set was derived by dividing the original training set into a modeling training set and modeling testing set of 303 and 95 compounds respectively by using the removal-until-done method (section 2.2.2.3).

7.2.4 cQSPkR method

In this work, consensus QSPkR (cQSPkR) models were developed by combining QSPkR models generated from different statistical learning methods. cQSPkR models compute the predicted CL_{tot} of a compound by averaging the predicted CL_{tot} of that compound from the different QSPkR models (Sutherland *et al.* 2003b).

7.2.5 Evaluation of QSPkR models

The validation set, not used in the derivation of the QSPkR models, was used to estimate the prediction capability of the QSPkR models. The fold-error for each compound and the percentage of compounds in the validation set where the fold-error is less than two or three were calculated. The predictive capability of the QSPkR models can be measured by the Spearman rank correlation coefficient (R_s) and average-fold error (Obach *et al.* 1997). R_s is used to assess the ability of the QSPkR models to rank compounds based on their CL_{tot}. The average-fold error is the geometric mean of the ratio of predicted and actual values, and QSPkR models that predicts CL_{tot} perfectly gives a value of 1 and a model with an average-fold error of less than 2 is considered to be a successful one (Obach *et al.* 1997). The predicted $\log CL_{tot}$ values of the compounds were converted back to CL_{tot} prior to the calculation of fold-errors and average-fold errors. A functional dependence study was also done using the procedures described in section 2.5.3.

7.3 Results and discussion

7.3.1 Dataset analysis

The diversity index (DI) of the training set and validation set used in this study and those of several reference datasets are given in Table 7.1.

Table 7.1 Diversity indices of the datasets used in this and other studies.

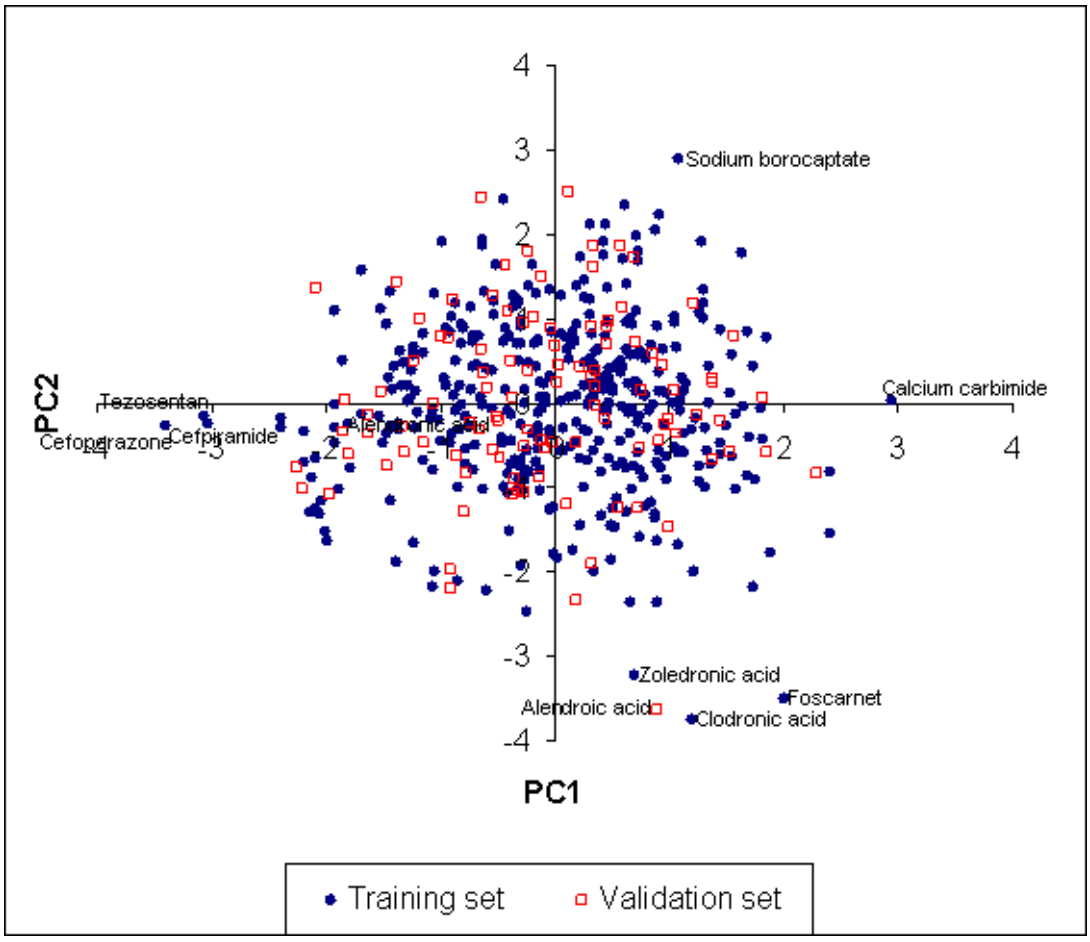
	Dataset	Number of compounds	Diversity index
Datasets used in this work	Training set	398	0.067
	Validation set	105	0.068
Highly diverse datasets	Satellite structures (Oprea *et al.* 2001)	8	0.076
	FDA approved drugs	1121	0.069
	NCI Diversity set (NCI/NIH)	1804	0.124
Congeneric datasets	Penicillins	59	0.452
	Cephalosporins	73	0.568
	Fluoroquinolones	39	0.579
QSAR, QSPR datasets	Estrogen receptor ligands (Sutherland *et al.* 2003a)	1009	0.274
	Benzodiazepine receptor ligands (Sutherland *et al.* 2003a)	405	0.314
	Dihydrofolate reductase (DHFR) inhibitors (Sutherland *et al.* 2003a)	756	0.384
	Cyclooxygenase 2 (COX2) inhibitors (Sutherland *et al.* 2003a)	467	0.584

It is found that the DI values of the training set and validation set is very small, as low as 0.067, which is at the level of those of highly diverse datasets. For comparison, the DI values of datasets containing congeneric compounds are typically

greater than 0.452, and those of the compounds used in QSAR and QSPR studies are typically in the range of 0.274 to 0.584. This suggests that the training set and validation set are sufficiently diverse. The representativity index (RI) value between the training set and validation set is 0.881, which suggests that the validation set is representative of the training set and thus is suitable for assessing the predictive capability of the QSPkR models developed in this work.

Principal component analysis (PCA) (Wold *et al.* 1987) was performed by using the dataset of 503 compounds to identify outliers and clusters. Two principal components (PCs) were derived which is able to explain 73.2% of the total variance in the descriptors. Components one and two are able to explain 60.9% and 12.3% of the variance respectively. Figure 7.1 shows a score plot of the compounds in the training set and validation set by using the first two PCs. Score plots are useful for comparing the distribution of compounds in the chemical space between two datasets and to identify clusters of compounds and single compounds that may be outliers (Wold *et al.* 1987; Doddareddy *et al.* 2006). There are no distinct clusters in the training set and validation set. The validation set is evenly distributed throughout the score space of the training set, confirming the representativeness of the validation set. Four compounds, alendronic acid, clodronic acid, foscarnet and zoledronic acid, were found to be farther away from the majority of compounds and are located at the bottom right of the score space. Other compounds that are farther away from the majority of compounds are cefoperazone, cefpiramide and tezosentan, which are located at the left of the score space, and carbimide and borocaptate, which are located at the right and top right of the score space respectively. There seems to be no evidence to suggest that these compounds are outliers. Thus they are retained in the training set and validation set.

Figure 7.1 Score plot of the first two principal components for training set and validation set.



7.3.2 Analysis of descriptor sets

The computed R_s values and average-fold errors of the QSPkR models developed by using different descriptor sets are shown in Table 7.2. Comparison of the QSPkR models based on the six descriptor sets shows that models based on the DS-MIXED descriptor set generally give higher R_s values and lower average-fold errors than those based on other descriptors sets. This suggests that models based on the DS-MIXED descriptor set are more useful and it may be advantageous to use a variety of descriptors for prediction of pharmacokinetic properties than to use a specialized descriptor set which may partially neglect some important features.

The descriptors in the six descriptor sets were combined to form a new descriptor set (DS-ALL). The G-ALL, S-ALL and K-ALL models developed by using DS-ALL have higher predictive capabilities compared to models developed by using individual descriptor sets. This suggests that all of the three statistical learning methods are able to extract useful information from the different descriptor sets and to effectively combine them to develop more predictive QSPkR models.

Table 7.2 Average-fold errors of QSPkR models developed by using different statistical learning methods and different descriptors sets[a].

Statistical learning methods	Model	Descriptor set	Optimum parameter	R_s	Average-fold error
GRNN	G-MIXED	DS-MIXED	2	0.636	1.73
	G-3DMoRSE	DS-3DMoRSE	3	0.540	1.75
	G-ATS	DS-ATS	3	0.448	1.86
	G-GETAWAY	DS-GETAWAY	3	0.520	1.80
	G-RDF	DS-RDF	3	0.558	1.80
	G-WHIM	DS-WHIM	2	0.302	1.96
	G-ALL	DS-ALL	7	0.633	1.63
SVR	S-MIXED	DS-MIXED	3	0.558	1.73
	S-3DMoRSE	DS-3DMoRSE	4	0.518	1.81
	S-ATS	DS-ATS	7	0.548	1.74
	S-GETAWAY	DS-GETAWAY	8	0.564	1.78
	S-RDF	DS-RDF	4	0.607	1.76
	S-WHIM	DS-WHIM	5	0.346	1.95
	S-ALL	DS-ALL	13	0.643	1.66
kNN	K-MIXED	DS-MIXED	2	0.523	2.00
	K-3DMoRSE	DS-3DMoRSE	2	0.360	2.23
	K-ATS	DS-ATS	3	0.406	2.03

	K-GETAWAY	DS-GETAWAY	2	0.522	2.00
	K-RDF	DS-RDF	3	0.447	1.98
	K-WHIM	DS-WHIM	3	0.392	2.01
	K-ALL	DS-ALL	2	0.513	1.90
PLS	P-MIXED	DS-MIXED	17	0.528	1.89
	P-3DMoRSE	DS-3DMoRSE	8	0.377	2.26
	P-ATS	DS-ATS	7	0.562	2.09
	P-GETAWAY	DS-GETAWAY	10	0.474	1.92
	P-RDF	DS-RDF	6	0.468	1.99
	P-WHIM	DS-WHIM	28	0.282	2.10
	P-ALL	DS-ALL	5	0.559	1.96

[a] The average-fold errors were assessed by using the validation set.

7.3.3 Predictive capability of QSPkR and cQSPkR models

Table 7.2 above shows the predictive capabilities of the QSPkR models developed by using GRNN, SVR and kNN. PLS was used as a reference QSPkR method for comparison of the predictive capabilities of the different models. The results for the corresponding PLS-developed QSPkR models are also given in Table 7.2. All of the GRNN- and SVR-developed QSPkR models have average-fold errors less than 2 while kNN-developed models have average-fold errors near 2, which are similar to those of PLS-developed models. GRNN- and SVR-developed QSPkR models were also found to generally give higher R_s values than the corresponding kNN- and PLS-developed models. This suggests that both GRNN and SVR are more useful than either kNN or PLS for developing QSPkR models of drug clearance and both GRNN- and SVR-developed models are not more flexible than is necessary and thus unlikely to overfit.

To assess the performance of the three statistical learning methods for CL_{tot} prediction of a more diverse set of compounds, it is useful to examine whether the predictive capability of these methods is at a similar level as those derived from the use of a significantly smaller set of compounds. It is noted that, a direct comparison with results from previous studies is inappropriate because of the differences in the dataset, molecular descriptors, and computing algorithms used. Although desirable, it is impossible to conduct a separate comparison using results directly from other studies without full information about the algorithms of molecular descriptors and modeling methods used in each study. Nonetheless, a tentative comparison may provide some crude estimate regarding the approximate level of predictive capability of the QSPkR models studied in this work.

Table 7.3 gives the prediction results of the G-ALL, S-ALL, K-ALL and P-ALL models from this work along with those derived from previous studies. The percentage of compounds in the validation set with predicted CL_{tot} within two-fold error of actual values of G-ALL and S-ALL models are comparable and in some cases slightly better than those of earlier studies that were tested by using a much smaller number of compounds. Diversity of the training sets has been shown to affect the applicability domain of QSPkR models (Dimitrov *et al.* 2005). Thus the results suggest that using a more diverse and larger number of compounds and applying statistical learning methods, particularly GRNN and SVR, are useful for prediction of CL_{tot}. A possible reason for the better performance of GRNN and SVR is that multiple mechanisms are involved in determining CL_{tot}. A variety of factors may interact in complex ways to affect the CL_{tot} of a compound. Therefore methods based only on linear relationships, such as PLS, may not be the most efficient approach for constructing a QSPkR model for predicting CL_{tot}. Thus non-linear methods, such as

GRNN and SVR, which do not require prior knowledge about the molecular mechanism or structure-activity relationship of a particular drug property may be more suitable.

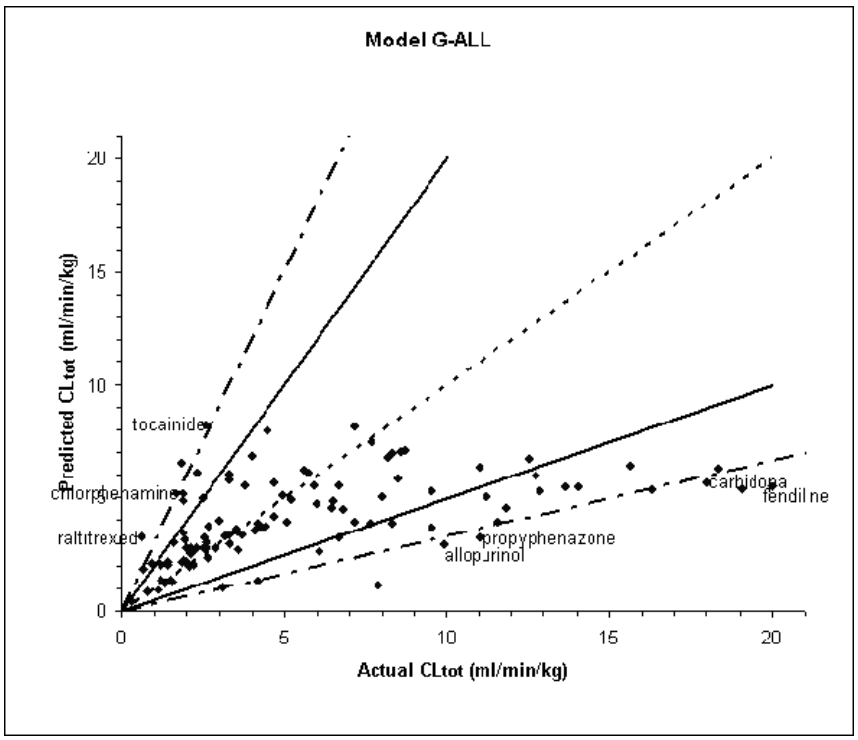
Table 7.3 Number of compounds with the predicted CL_{tot} within two-fold error of the actual CL_{tot} from this work and other studies.

Model	Number of compounds	Number (percentage) of compounds with fold-errors < 2
G-ALL (this work)	105	73 (69.5%)
S-ALL (this work)	105	78 (74.3%)
K-ALL (this work)	105	65 (61.9%)
P-ALL (this work)	105	63 (60.0%)
Multiple linear regression (Wajima *et al.* 2003a)	68	44 (64.7%)
kNN (Ng *et al.* 2004)	44	30 (68.2%)
PLS (Ng *et al.* 2004)	44	22 (50.0%)
Parallel tube (Obach 1999)	29	16 (55.2%)

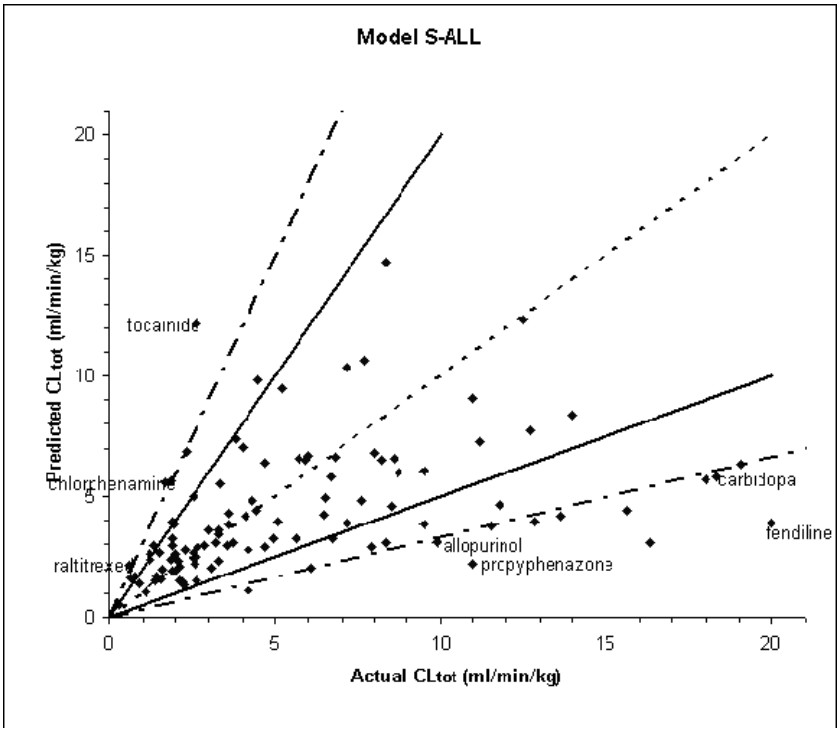
Plots of the predicted CL_{tot} against the actual values for the G-ALL and S-ALL models are shown in Figure 7.2. These plots show that both models tend to under-predict the CL_{tot} value of compounds rather than over-predicting the CL_{tot}. Under-prediction of CL_{tot} is more desirable than over-prediction of CL_{tot} during drug development because over-prediction results in more frequent dosing of a drug candidate during clinical trials which may lead to higher rates of adverse drug reactions. For compounds with fold-errors greater than 2, the G-ALL model underpredicted 22 and overpredicted 10 of these compounds respectively. The corresponding values for the S-ALL model are 18 and 11 respectively.

Figure 7.2 (a) Plot of predicted CL_{tot} vs actual CL_{tot} for the G-ALL model. (b) Plot of predicted CL_{tot} vs actual CL_{tot} for the S-ALL model.

a

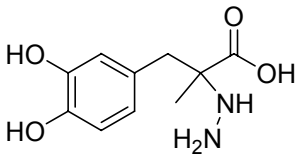


b

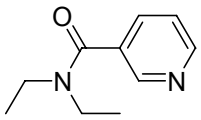


The dotted line represents line of unity. The area between the two solid lines and between the two dotted-dash lines represents an area between two-fold and three-fold error respectively. Compounds in validation set with fold-error greater than 3 for both G-ALL and S-ALL models are identified.

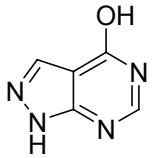
Figure 7.3 Chemical structures of compounds in validation set with fold-errors greater than three for both G-ALL and S-ALL models[a].



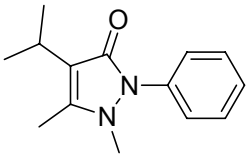
Carbidopa
G-ALL: 3.2
S-ALL: 3.2



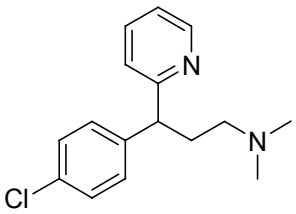
Tocainide
G-ALL: 3.2
S-ALL: 4.7



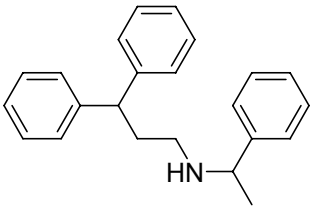
Allopurinol
G-ALL: 3.3
S-ALL: 3.2



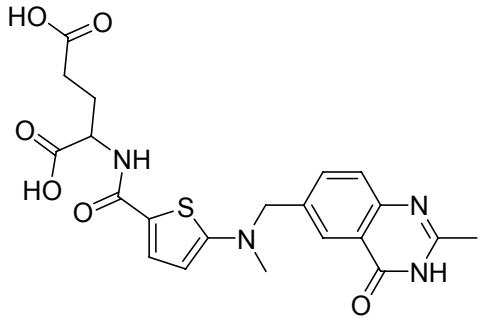
Propyphenazone
G-ALL: 3.4
S-ALL: 5.0



Chlorphenamine
G-ALL: 3.1
S-ALL: 3.3



Fendiline
G-ALL: 3.6
S-ALL: 5.2



Raltitrexed
G-ALL: 5.5
S-ALL: 3.7

[a] The numbers represent the fold-errors of each compound for both G-ALL and S-ALL models.

There are 7 compounds having fold-errors greater than 3 for both models and their chemical structures are shown in Figure 7.3. A possible reason for the high fold-errors of some of these compounds is that the descriptors used in this study may be inadequate to properly describe these compounds. Examples of these compounds are chlorphenamine and fendiline, which contain two aromatic rings separated by an atom, allopurinol, which contains a complex two ring system with multiple heteroatoms, carbidopa, which has a hydrazine group that is a highly reactive reducing agent, and raltitrexed, which has two carboxylic acid groups that makes it highly charged at physiological pH. Our previous studies (sections 4.2.3, 6.4.4, and 8.1.3.3) have suggested that compounds containing these structural features may not be adequately represented by currently available descriptors. Thus by using the currently available algorithm, these compounds are misrepresented and incorrectly positioned in the chemical space, leading to inaccurate prediction of their CL_{tot} values.

A cQSPkR model was developed by using G-ALL and S-ALL models. The K-ALL model was not used because its prediction capability is significantly lower than those of the G-ALL and S-ALL models and hence may reduce the prediction capability of the cQSPkR model. The cQSPkR model has an average-fold error of 1.61. Thus the cQSPkR model had slightly better prediction capability than either the G-ALL or S-ALL model. The relatively small average-fold error suggests that the model is useful for the prediction of CL_{tot}. The cQSPkR model correctly predicted 77 (73.3%) compounds in validation set to be within two-fold error of actual CL_{tot}. For compounds with fold-errors greater than 2, the cQSPkR model under-predicted 19 and over-predicted 9 of these compounds respectively. None of the under-predicted or over-predicted compounds have fold-errors greater than 4.5. This is significantly improved over that of the G-ALL and S-ALL models which have two and four

compounds with fold-errors greater than 4.5 respectively. The cQSPkR model gives an R_s value of 0.652 which suggests that the model may be useful for ranking compounds according to their CL_{tot} in large chemical libraries.

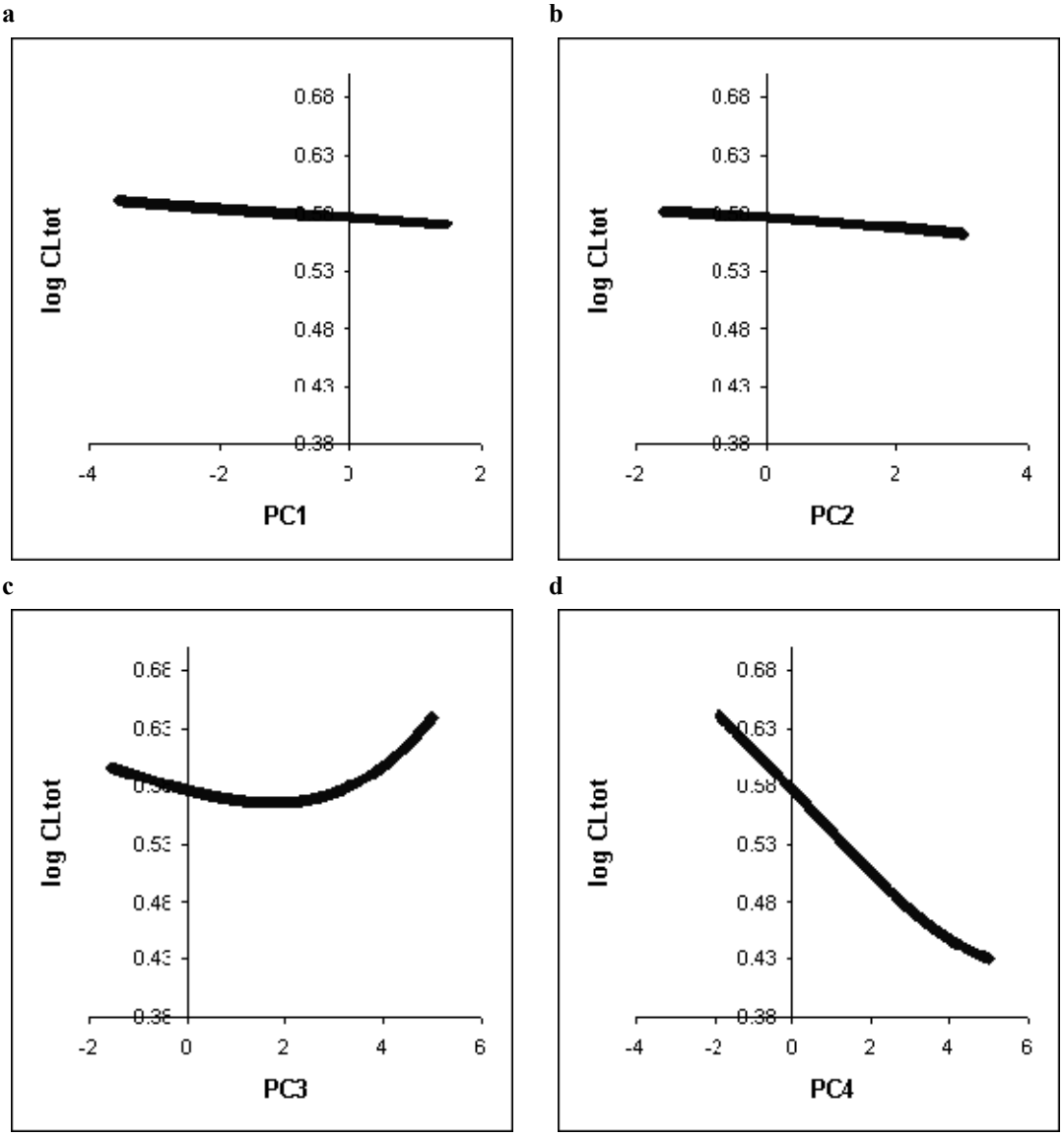
7.3.4 Functional dependence analysis

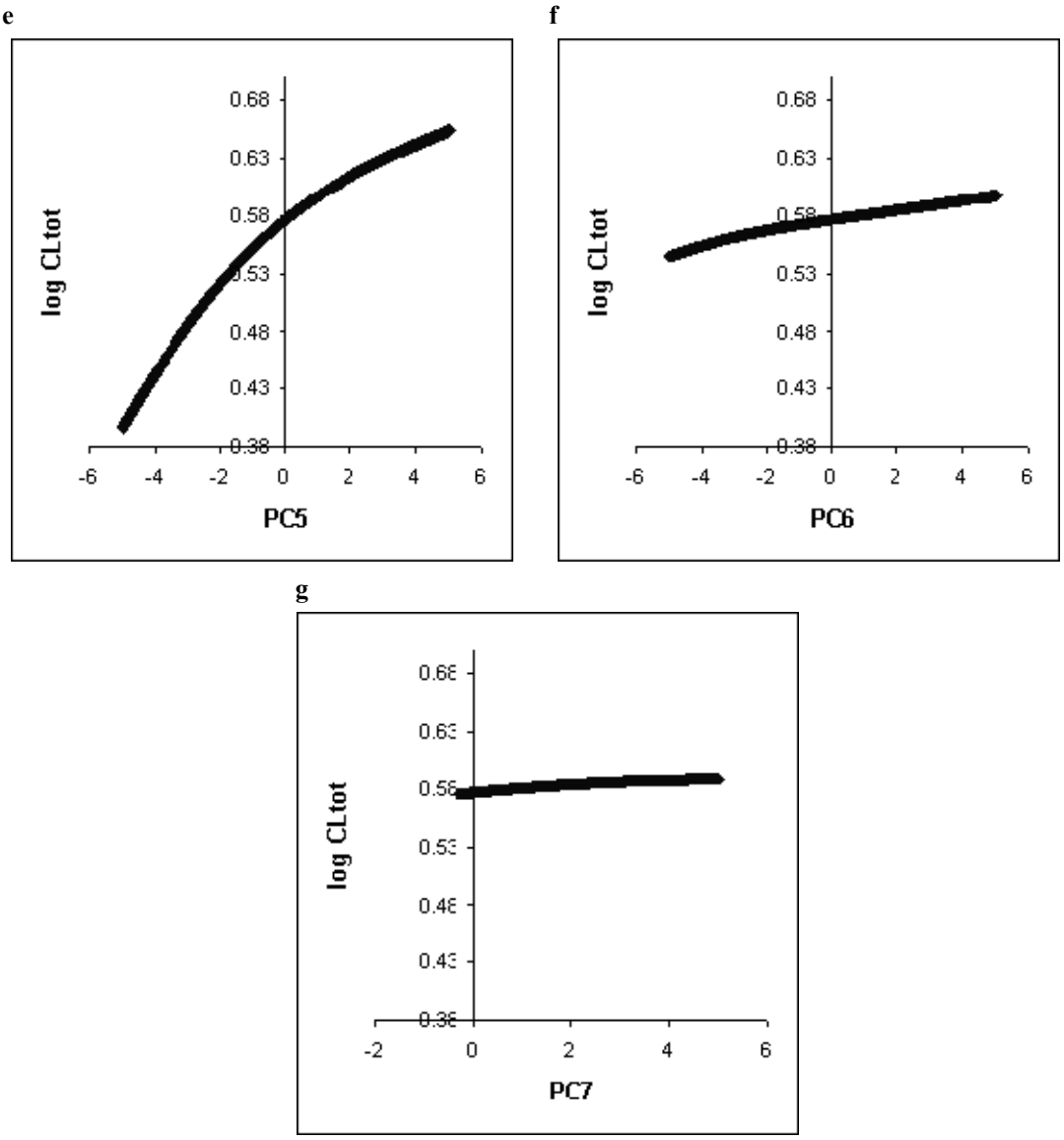
Multiple elimination processes are involved in drug clearances. Thus it is difficult to determine which molecular characteristics are important in affecting CL_{tot}. Nonetheless, it is possible to infer some information from a functional dependence study of the QSPkR models. It is noted that the results of a functional dependence study may vary with respect to different QSPkR models. Thus the following interpretation of the descriptors must be taken in light of the absolute predictive ability of the QSPkR models. Figure 7.4 shows the prediction results of the first seven PCs of the G-MIXED model by using artificial testing sets. The first seven PCs are able to explain approximately 60% of the total variance of the descriptors. Plots of $logCL_{tot}$ against the PCs for the S-MIXED model are similar and thus are not given here. The DS-MIXED descriptor set was used to determine the relationship between a specific molecular characteristic and CL_{tot} because models developed by using this descriptor set have higher predictive capabilities than those developed by using other descriptor sets. In addition, it is relatively easier to assign the descriptors in the DS-MIXED descriptor set to specific molecular characteristics. Table 7.4 gives the list of the dominant descriptors and the corresponding molecular characteristic in different PCs.

Table 7.4 The dominant descriptors and the corresponding molecular characteristic in different principal components.

PC	Dominant Descriptors	Corresponding Molecular Characteristic
First	3D-Wiener index (Basak *et al.* 1999)	Molecular shape
	Valence molecular connectivity Chi index for path order 2 (Kier *et al.* 1999)	
Second	Atom-type Estate sum for :CH: (sp^2, aromatic) (Kier *et al.* 1999)	Lipophilicity
	Atom-type Estate sum for :C:- (Kier *et al.* 1999)	
	AlogP- (Viswanadhan *et al.* 1993)	
Third	Kier flexibility index (Kier 1990)	Flexibility
Fourth	Average molecular weight	Molecular size
	Gravitational 3D index (Wessel *et al.* 1998)	
Fifth	Atom-type Estate sum for =S=< (Kier *et al.* 1999)	Charge and molecular solvation
	Solvation molecular connectivity Chi index for path order 2 (Todeschini *et al.* 2000)	
	Mean topological charge index for path order 1 (Galvez *et al.* 1994)	
Sixth	Number of H-bond acceptors	Hydrogen bond accepting capability
Seventh	Number of H-bond donors	Hydrogen bond donating capability

Figure 7.4 Plots of log CL_{tot} against the various PCs for G-ALL model. Increasing values of PC1 denotes increasing sphericity of a compound. Increasing values of PC2 denotes decreasing lipophilicity of a compound. Increasing values of PC3 denotes decreasing flexibility of a compound. Increasing values of PC4 denotes increasing molecular size of a compound. Increasing values of PC6 denotes increasing hydrogen bond accepting ability of a compound. Increasing values of PC7 denotes increasing hydrogen bond donating ability of a compound.





Analysis of the variations of the descriptors shows that the first PC is primarily determined by topological descriptors. These include 3D-Wiener index, which decreases with increasing sphericity of a structure, and $^{2}\chi^{v}$, which is the valence molecular connectivity Chi index for path order 2 and encodes the relative degree of branching in a compound (Kier *et al.* 1986). Prediction results by using the artificial testing set show that CL_{tot} generally decreases with increasing value of the 3D-Wiender index and $^{2}\chi^{v}$ (Figure 7.4a). This suggests that spherically-shaped

molecules with fewer side chain branching tend to have higher CL_{tot} than that of aspherical molecules with multiple branches.

Electrotopological state descriptors like Estate_aaCH and Estate_aasC, which describe electrotopological properties of carbons in aromatic rings, and AlogP (Viswanadhan *et al.* 1993), which measures the partition coefficient of a compound, are the main contributor to the second PC. Thus it is likely that the second PC is a measure of the lipophilicity of a compound. Results of the artificial testing set suggest CL_{tot} increases with increasing lipophilicity of a compound.

The third PC is determined primarily by KierFlexibilityIndex, which is related to the flexibility of a molecule. The complex role of molecular flexibility in membrane permeation has been found by two studies. One found a positive correlation between flexibility and permeation (Iyer *et al.* 2002) while the other found a negative correlation (Veber *et al.* 2002). Using the artificial testing set, it was found that compounds with low or high flexibility have higher CL_{tot} than those with moderate flexibility. This may partially explain the apparent contradiction between the two earlier studies.

The fourth PC is formed mainly by AMW, which is the average molecular weight, and Gravitational3DIndex. These are related to the volume of a molecule and the distribution of atomic masses within the molecular space. The contribution of these two descriptors to the fourth PC suggests that the fourth PC is a measure of molecular size. The artificial testing sets show that CL_{tot} generally increases with decreasing molecular size. This is consistent with the findings that small molecular size is necessary for good membrane penetration (Pardridge 1998).

The main contributors to the fifth PC are Estate_ddssS, which is the electrotopological descriptor for sulfur atoms, ${}^{2}\chi^{s}$, which is the solvation molecular

connectivity Chi index for path order 2, and $Mean^{1}G^{c}$, which is the mean topological charge index for path order 1. It is difficult to attribute these descriptors to a single molecular characteristic. Nonetheless, studies have consistently shown that charge and molecular solvation are important in determining the metabolism (Smith *et al.* 1997b; de Groot *et al.* 2002) and renal clearance (Turner *et al.* 2004b; Venturoli *et al.* 2005) of a molecule.

The sixth and seventh PCs are determined primarily by descriptors encoding the hydrogen bond acceptor and donor properties of a compound respectively. Figure 7.4f and Figure 7.4g shows that CL_{tot} increases with increasing hydrogen bonding capability of a compound. Studies have found that binding affinity to human serum albumin generally decreases with increasing hydrogen bonding capability of these compounds (Hall *et al.* 2003; Yap *et al.* 2005b). Many compounds bind to serum albumin and the albumin-bound fraction is not available for hepatic metabolism or renal clearance (Colmenarejo 2003). Thus factors which decrease serum albumin binding are expected to increase the CL_{tot} of a compound.

The rate of change in CL_{tot} per unit change in the PC values can provide a useful hint about the contribution of a molecular characteristic to the clearance of a compound. The plots in Figure 7.4 show the effect of changing the value of each PC on the clearance of a compound in the following order: PC5 > PC4 > PC3 > PC6 > PC1 ≈ PC2 > PC7. Thus charge, molecular solvation, molecular size and flexibility are the most important molecular properties which influence clearance of a compound.

7.4 Conclusion

Our study suggests that both GRNN and SVR are potentially useful for developing QSPkR models to predict drug clearance from a large diverse set of compound data. QSPkR models developed by using GRNN, SVR and KNN were tested and compared with those developed by using a linear method, PLS. All of the GRNN- and SVR-developed models show better prediction capability than the corresponding KNN- or PLS-developed models. The predictive capabilities of the QSPkR models developed in this study are comparable to those of previous studies and can be further improved by using consensus modeling methods.

A collection of constitutional, geometrical, topological and electrotopological descriptors seems to be more useful for modeling drug clearance than specialized descriptor sets such as 3DMoRSE, ATS, GETAWAY, RDF and WHIM. An individual descriptor set tends to partially neglect some important features and thus the use of all the available descriptors may help to alleviate such type of feature bias. The three statistical learning methods used in this work appears to be capable of combining the information encoded in the different descriptor sets effectively to develop more predictive QSPkR models.

Chapter 8

Toxicity Prediction

This chapter describes two important drug toxicities: genotoxicity (section 8.1) and torsade de pointes (section 8.2). The classification accuracies of the qSPkR models for prediction of genotoxic potential and torsade-causing potential of compounds developed by using SVM and other classification methods are presented. The possible reasons for misclassification of some compounds are also discussed.

8.1 Genotoxicity

8.1.1 Introduction

Adverse drug reactions (ADRs) are responsible for the failure of a substantial percentage of investigational drugs and the withdrawal of marketed drugs (Johnson *et al.* 2000; van de Waterbeemd *et al.* 2003). Up to one-third of all drug failures are due to ADRs (Kennedy 1997). A variety of toxicological tests and clinical safety evaluations need to be conducted and evaluated by the drug regulatory authorities for drug safety assessment. Because of the high cost of conducting toxicity tests and clinical trials, effort has been directed at developing low-cost and efficient tools for predicting ADRs aimed at eliminating unsafe drug candidates in the early stages of drug development (Kennedy 1997; van de Waterbeemd *et al.* 2003).

Genotoxicity is one of the ADRs closely evaluated in drug discovery and approval processes. The molecular mechanisms of genotoxicity include DNA intercalation by aromatic ring of a drug, DNA methylation, DNA adduct formation

and strand break, and unscheduled DNA synthesis (Bolzan *et al.* 2002). Some genotoxic (*GT*+) compounds require metabolic activation and their *GT*+ effect are mediated via N-dialkylation (Snyder *et al.* 2004). These events subsequently result in chromosomal aberrations, micronuclei, sister chromatid exchanges, and cell death which contribute to drug ADR (Bolzan *et al.* 2002).

Tools for fast and efficient prediction of drug *GT*+ potential, particularly those based on computational methods, are being developed (Kramer 1998; Schwetz *et al.* 1998). For instance, expert systems that use structural alerts for predicting *GT*+ as well as other toxicological profiles are now commercially available. These include Deductive Estimation on Risk from Existing Knowledge (DEREK), Multiple Computer Automated Structure Evaluation (MCASE) and Toxicity Prediction by Komputer Assisted Technology (TOPKAT). Specific details about these computational databases can be found in the review by Greene (Greene 2002). qSPkR models have been developed for predicting *GT*+ potential of several groups of related chemicals (Marchant 1996; Cash 2001). However the qSPkR models of a majority of chemical groups are yet to be determined which hinders the practical application of this method.

Statistical learning methods have recently been explored as a new approach for genotoxicity prediction without the restriction on the features of structures or types of molecules (He *et al.* 2003; Mattioni *et al.* 2003; Mosier *et al.* 2003). Instead of focusing on specific structural feature or a particular group of related molecules, these methods classify molecules into *GT*+ and non-genotoxic (*GT*-) compounds based on their general structural and physicochemical properties regardless of their structural and chemical types. Therefore, in principle, these methods are expected to be applicable to a diverse set of molecules. However, the performance of these methods

can be practically limited by the quality of molecular descriptors, diversity of training and testing data, and the efficiency of statistical learning algorithm.

So far, three statistical learning methods, linear discriminant analysis (LDA), *k* nearest neighbour (kNN), and probabilistic neural network (PNN), have been used and achieved a prediction accuracy of up to 73.8% for *GT*+ and 92.8% for *GT*- compounds respectively (He *et al.* 2003; Mattioni *et al.* 2003; Mosier *et al.* 2003). However, these methods have been developed and tested by using no more than 394 *GT*+ and *GT*- compounds (Snyder *et al.* 2004), which is significantly smaller in number and diversity than the 860 known *GT*+ and *GT*- compounds found from our recent literature search. Therefore, there is a need to examine if a similar level of accuracy can be achieved for the more diverse set of molecules. It is also of interest to determine if the *GT*+ accuracy can be further improved by a training set composed of a more diverse set of *GT*+ compounds. Moreover, other statistical learning methods such as support vector machine (SVM) and C4.5 decision tree (DT) have shown promising potential, and it is useful to evaluate these methods.

This work is intended to evaluate several statistical learning methods by using 860 *GT*+ and *GT*- compounds. These methods include SVM, PNN, kNN and C4.5 DT. Recursive feature elimination (RFE) is used in this work for selecting the molecular descriptors relevant to the classification of *GT*+ and *GT*- compounds. To adequately assess the prediction accuracy of the methods used in this work, two different evaluation methods are used. One is 5-fold cross-validation, and the other is the use of an external independent validation set.

8.1.2 Methods

8.1.2.1 Selection of GT+ and GT- compounds

A total of 860 *GT*+ and *GT*- compounds with known genotoxicity test results are selected from several sources including the 1999-2002 Physician's Desk Reference, National Toxicology Program, and a number of publications (Snyder *et al.* 2001; He *et al.* 2003; Mattioni *et al.* 2003; Mosier *et al.* 2003; Snyder *et al.* 2004). Genotoxicity tests for generating these data include the pre-ICH four standard batteries (Ames test, *in vitro* cytogenetics, *in vivo* cytogenetics, mouse lymphoma assay) and the Salt-Overly-Sensitive (SOS) chromotest (which is a rapid alternative genotoxicity test based on the detection of the DNA damage through the SOS pathway) (Quillardet *et al.* 1993; Vasilieva 2002). Compounds with genotoxicity test results are divided into *GT*+ and *GT*- groups according to whether these genotoxicity test results showed at least one positive finding. Under this definition, there are a total of 229 *GT*+ compounds and 631 *GT*- compounds.

These compounds are further separated into training and testing sets by either 5-fold cross-validation or removal-until-done method (section 2.2.2.3) depending on the evaluation method used. For evaluation by an independent validation set, these compounds are divided into training, testing, and independent validation set. The generated training, testing and independent evaluation set contains 577 (166 *GT*+, 411 *GT*-), 160 (36 *GT*+, 124 *GT*-) and 123 (27 *GT*+, 96 *GT*-) compounds respectively.

8.1.2.2 Molecular descriptors

In this work, a set of 199 molecular descriptors, which include 143 topological, 31 quantum chemical, and 25 geometrical descriptors, are computed using our own

designed molecular descriptor computing program. The remaining redundant and un-related descriptors are further reduced by using RFE method. The computation procedure for RFE is the same as that in the human intestinal absorption study (section 4.1.2.3).

8.1.3 Results and discussion

8.1.3.1 Overall prediction accuracies

Prediction results of SVM without RFE and SVM with RFE (SVM+RFE) by using 5-fold cross-validation are given in Table 8.1. The accuracies of SVM+RFE are 75.5% for *GT*+ compounds and 90.6% for *GT*- compounds, which are slightly better than the values of 69.4% for *GT*+ compounds and 88.2% for *GT*- compounds derived from SVM without RFE. The *GT*+ prediction accuracy is noticeably improved, which indicates the usefulness of RFE in selecting the proper set of descriptors for the prediction of *GT*+ and *GT*- compounds. The use of these RFE-selected descriptors also slightly improves the prediction accuracy of the other three statistical methods. The *GT*+ accuracies are improved from 70.4% to 74.1% for PNN and from 44.4% to 55.6% for DT respectively, and that of kNN remains roughly unchanged. The *GT*- accuracy of kNN is improved from 82.2% to 86.5%, and those of PNN and C4.5 DT are roughly unchanged. These results showed that descriptor selection by using RFE plays the important role in improving the prediction capability for the above methods in general. Similar prediction accuracies are also found from two additional 5-fold cross-validation studies conducted by using training-testing sets separately generated from different random number seed parameters.

Table 8.1 SVM and SVM+RFE prediction accuracy of the *GT*+ and *GT*- compounds by using 5-fold cross-validation.

Method	Cross - validation	Genotoxicity			Non-genotoxicity			Q (%)	MCC
		TP	FN	SE (%)	TN	FP	SP (%)		
SVM	1	32	17	65.3	109	11	90.8	83.4	0.59
	2	30	10	75.0	115	14	89.1	85.8	0.62
	3	32	13	71.1	119	21	85.0	81.6	0.53
	4	32	19	62.7	106	11	90.6	82.1	0.56
	5	32	12	72.7	107	18	85.6	82.2	0.56
	average			**69.4**			**88.2**	**83.0**	**0.57**
	SD[a]			4.6			2.5	1.5	0.03
	SE[b]			1.9			1.0	0.6	0.01
SVM +	1	35	14	71.4	111	9	92.5	86.4	0.66
RFE	2	32	8	80.0	118	11	91.5	88.8	0.69
	3	35	10	77.8	123	17	87.9	85.4	0.62
	4	35	16	68.6	109	8	93.2	85.7	0.65
	5	35	9	79.5	110	15	88.0	85.8	0.65
	average			**75.5**			**90.6**	**86.4**	**0.66**
	SD[a]			4.6			2.3	1.2	0.02
	SE[b]			1.9			0.9	0.5	0.01

[a] standard deviation
[b] standard error

Table 8.2 gives the *GT*+ and *GT*- prediction accuracies derived from the four methods SVM, PNN, kNN and C4.5 DT by using the independent validation set and the RFE-selected molecular descriptors. The *GT*+ accuracies are in the range of 55.6%-77.8% and the *GT*- accuracies are in the range of 75.0%-92.7%. Similar level of accuracies are obtained for SVM, PNN and kNN, with SVM giving the highest value of 77.8% and 92.7% for *GT*+ and *GT*- compounds respectively. C4.5 DT appears to give substantially lower accuracies, which is concordant with other

experimental comparison results (Brown *et al.* 2000; Huang *et al.* 2002). A possible reason for this lower accuracy is that C4.5 DT uses information gain to find the optimum set of descriptors, which may not be the most effective approach for every problem. It has been pointed out that filter methods, such as information gain, may not be as efficient as wrapper methods, such as RFE, for determining the subset of descriptors relevant to a particular problem (Saeys *et al.* 2004).

Table 8.2 Comparison of the prediction accuracies of *GT+* and *GT-* compounds derived from different machine learning methods by using the independent validation set in this work.

Method	Parameter	TP	FN	TN	FP	SE (%)	SP (%)	Q (%)
C4.5 DT	-	15	12	72	24	55.6	75.0	70.7
PNN	σ=0.2	20	7	77	19	74.1	80.2	78.9
k-NN	k=3	19	8	83	13	70.4	86.5	82.9
SVM	σ=3	21	6	89	7	77.8	92.7	89.4

8.1.3.2 Relevance of selected features to genotoxicity study

Apart from the quality of datasets used, selection of descriptors relevant to genotoxicity study is important for optimizing the prediction system by reducing the noise in a statistical learning process. A total of 39 molecular descriptors are selected by the RFE method, as given in Table 8.3. Most of these are found to be relevant to the assessment of genotoxicity potential of molecules. For instance, an important characteristics of some *GT+* compounds is their ability to intercalate DNA (He *et al.* 2003). The selected electrotopological state descriptors S(10) and S(14) describe atom-type H estate sum for :CH: sp^2 aromatic structures and atom-type H estate sum for CH_n aromatic structures respectively.

Table 8.3 Molecular descriptors selected from the RFE method for SVM classification of *GT+* and *GT-* compounds.

Descriptors	Description	Class
Nrot	Number of rotatable bonds	Simple molecular properties
ndonr	Number of H-bond donors	Simple molecular properties
$^{3}\chi_{C}$	Simple molecular connectivity Chi indices for cluster	Connectivity and shape
$^{4}\chi_{PC}$	Simple molecular connectivity Chi indices for path/cluster	Connectivity and shape
$^{3}\chi^{v}_{C}$	Valence molecular connectivity Chi indices for cluster	Connectivity and shape
$^{4}\chi^{v}_{PC}$	Valence molecular connectivity Chi indices for path/cluster	Connectivity and shape
S(2)	Atom-type H Estate sum for =NH	Electrotopological state
S(4)	Atom-type H Estate sum for $-NH_2$	Electrotopological state
S(10)	Atom-type H Estate sum for :CH: (sp^2, aromatic)	Electrotopological state
S(13)	Atom-type H Estate sum for CH_n (unsaturated)	Electrotopological state
S(14)	Atom-type H Estate sum for CH_n (aromatic)	Electrotopological state
S(16)	Atom-type Estate sum for $-CH_3$	Electrotopological state
S(25)	Atom-type Estate sum for =C<	Electrotopological state
S(26)	Atom-type Estate sum for : C:-	Electrotopological state
S(27)	Atom-type Estate sum for : C ::	Electrotopological state
S(30)	Atom-type Estate sum for =NH	Electrotopological state
S(34)	Atom-type Estate sum for =N-	Electrotopological state
S(35)	Atom-type Estate sum for :N:	Electrotopological state
S(41)	Atom-type Estate sum for -O-	Electrotopological state
Tradi	PetitJohn R2 Index	Electrotopological state
Tpeti	PetitJohn I2 Index	Electrotopological state
M	Molecular dipole moment	Quantum chemical properties
μ_{cp}	Chemical potential	Quantum chemical properties

χ_{en}	Electronegativity index	Quantum chemical properties
ω	Electrophilicity index	Quantum chemical properties
$Q_{H, Max}$	Most positive charge on H atoms	Quantum chemical properties
$Q_{N, Max}$	Most positive charge on N atoms	Quantum chemical properties
$Q_{O, Max}$	Most positive charge on O atoms	Quantum chemical properties
$Q_{H, Min}$,	Most negative charge on H atoms	Quantum chemical properties
Rpc	Relative positive charge	Quantum chemical properties
Rnc	Relative negative charge	Quantum chemical properties
Rugty	Molecular rugosity	Geometrical properties
Gloty	Molecular globularity	Geometrical properties
Shpl	Hydrophilic region	Geometrical properties
Shpb	Hydrophobic region	Geometrical properties
Capty	Capacity factor	Geometrical properties
Hiwpl	Hydrophilic integy moment	Geometrical properties
Hiwpb	Hydrophobic integy moment	Geometrical properties
Hiwpa	Amphiphilic moment	Geometrical properties

Many *GT+* compounds are known to structurally modify or form a covalent bond to DNA via chemical reactions. A substantial portion of the RFE selected descriptors are from the class of electrotopological state that describe characteristics of specific types of functional groups involved in DNA modification. There are also a substantial number of descriptors from the quantum chemical class that determine molecular dipole moment, chemical potential, electronegativity, electrophilicity, relative positive and negative charge, and the atomic charge on H, N and O atoms in a molecule. These properties are important for describing features of chemical reactions involved in the modification of DNA.

The size, shape, and polar property of a molecule have also been found to play a role in genetic damages caused by *GT+* compounds (He *et al.* 2003). Eight of the

selected descriptors are VolSurf descriptors (Cruciani *et al.* 2000b). These are molecular rugosity, molecular globularity, capacity factor, hydrophilic and hydrophobic region, hydrophilic integy moment, hydrophobic moment and amphiphilic moment. These descriptors primarily describe the size, shape, and polar property of a molecule. In general VolSurf descriptors, which are one-dimensional descriptors extracted from the computed 3D molecular field maps, were developed specifically for pharmacokinetics and pharmacodynamics applications (Crivori *et al.* 2000; Cruciani *et al.* 2000b). It is thus not surprising that the VolSurf descriptors related to the molecular size, shape, and polar property are selected.

Molecular connectivity is another feature known to be important for discriminating between some *GT*+ compounds from their *GT*- analogs. For instance, 4-amino-3-nitro-2,5-dimethylaniline is a *GT*+ compound, while its analog 4-amino-3-nitro-2,6-dimethylaniline is *GT*- (Chung *et al.* 1997). Four molecular connectivity descriptors, ${}^{3}\chi_{C}$, ${}^{4}\chi_{PC}$, ${}^{3}\chi^{v}{}_{C}$, and ${}^{4}\chi^{v}{}_{PC}$, are selected by RFE in this work. These descriptors are simple molecular connectivity chi indices for cluster, simple molecular connectivity chi indices for path/cluster, valence molecular connectivity chi indices for cluster, and valence molecular connectivity Chi indices for path/cluster respectively.

8.1.3.3 Performance evaluation

To assess the performance of the statistical learning methods for genotoxicity prediction of the more diverse set of molecules, it is useful to examine whether the accuracy from these methods is at a similar level as those derived by the use of a significantly smaller set of molecules. It is noted that, a direct comparison with results from previous studies is inappropriate because of the differences in the dataset and

molecular descriptors used. Table 8.4 gives the prediction results of the four statistical methods from this work along with those derived from previous studies.

Table 8.4 Overview of the prediction accuracies of *GT+* and *GT-* compounds from this work as with those from other studies[a].

Study	Method	Number of compounds	SE (%)	SP (%)	Q (%)
Snyder RD	MCASE	394	48.1	95.1	89.6
(Snyder *et al.* 2004)[b]	DEREK	394	51.9	75.1	73.6
	TOPKAT	394	43.4	88.1	81.7
Philip D. Mosier	k-NN	140	66.7	92.9	85.0
(Mosier *et al.* 2003)					
Linnan He	Consensus model	227	73.8	84.3	81.2
(He *et al.* 2003)	developed with k-NN,				
	LDA, and PNN				
	classifiers				
Brian E. Mattioni	k-NN	334	69.3	74.1	72.2
(Mattioni *et al.* 2003)					
This work	C4.5	860	55.6	75.0	70.7
	PNN	860	74.1	80.2	78.9
	k-NN	860	70.4	86.5	82.9
	SVM	860	77.8	92.7	89.4

[a] Prediction accuracies of this work listed here are based on independent evaluation sets, which are similar to those based on 5-fold cross-validation. Since different groups used different sets of descriptors, the accuracies given in this table only reflect the relative efficiency of each method.
[b] Best performance characteristics of the three programs were selected.

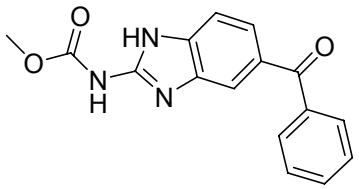
The *GT+* accuracies of these four methods are comparable and in some cases slightly better than those of earlier studies derived from kNN (Mattioni *et al.* 2003; Mosier *et al.* 2003) and the consensus model developed with kNN, LDA, and PNN (He *et al.* 2003). The *GT-* accuracies of these four methods are comparable to

those of earlier studies (He *et al.* 2003; Mattioni *et al.* 2003; Mosier *et al.* 2003; Snyder *et al.* 2004). The results from all of these statistical learning methods are substantially better than those obtained by DEREK, TOPKAT, MCASE programs (Snyder *et al.* 2004). Diversity of the training sets has been shown to affect the applicability domain of QSPkR models (Dimitrov *et al.* 2005). Thus the results suggest that the better prediction performance of the qSPkR models developed in this work is likely due to the use of a more diverse and larger number of compounds and the capability of statistical learning methods for classification of a more diverse range of molecules than that of structural alert-based approaches.

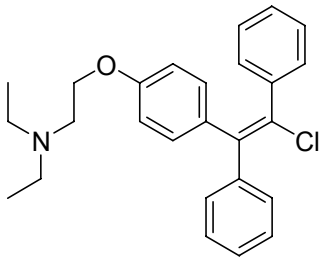
Overall, our study suggests that statistical learning methods, particularly SVM, kNN and PNN, are useful for genotoxicity assessment of a broad range of compounds. The prediction accuracy of these methods is at a similar level as those of earlier studies that were tested by using a much smaller number of molecules. Another advantage of these methods is that they do not require knowledge about the molecular mechanism or structure activity relationship (SAR) of a particular drug property. Moreover, the classification speed of these methods is generally fast. For instance, the number of compounds which can be classified per second by using SVM, kNN, PNN and C4.5 DT method is approximately 4000, 3000, 2000 and 62000 respectively on a P4 3.6Ghz machine. SVM typically uses a portion of the training set as support vectors for classification. In contrast, kNN and PNN use the whole training set for classification. The number of support vectors of SVM is in the range of 45-75% of the training set. Thus the classification speed of SVM is usually 25-55% faster than that of kNN and PNN. On the other hand, the classification speed of SVM is slower than that of C4.5 DT which uses a set of rules to reach a decision leaf.

There are six *GT+* and seven *GT-* compounds in the independent validation set that were misclassified by SVM, which are shown in Figure 8.1 and Figure 8.2 respectively.

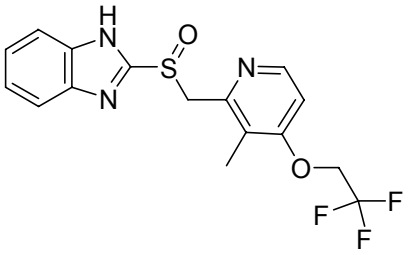
Figure 8.1 Six structures of misclassified *GT+* compounds in the independent validation set. Chemical name and relevant Chemical Abstracts Service (CAS) number of these compounds are shown in the figure.



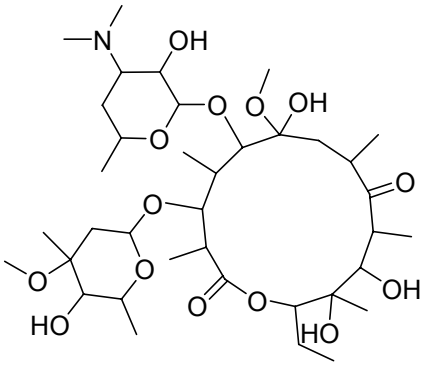
Mebendazole (31431-39-7)



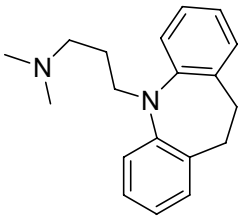
Clomiphene (911-45-5)



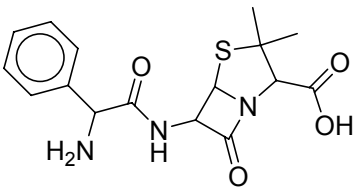
Lansoprazole (103577-45-3)



Clarithromycin (81103-11-9)

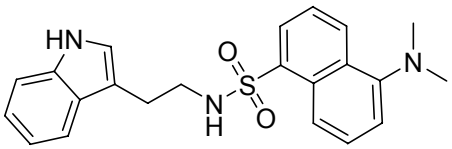


Imipramine (50-49-7)

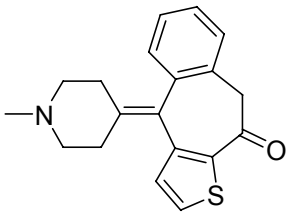


Ampicillin (69-53-4)

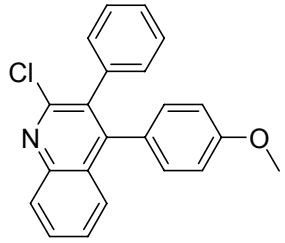
Figure 8.2 Seven structures of misclassified *GT*- compounds in the independent validation set. Chemical name and relevant Chemical Abstracts Service (CAS) number of these compounds are shown in the figure.



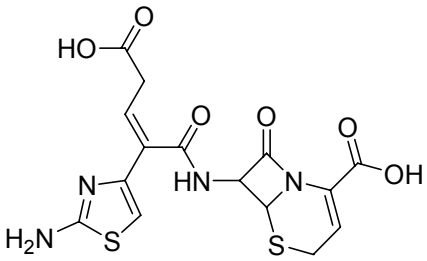
Dansyltryptamine (13285-17-1)



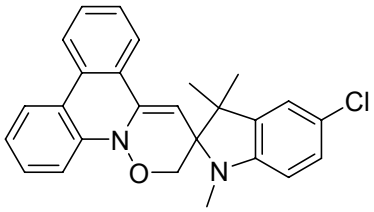
Ketotifen (34580-13-7



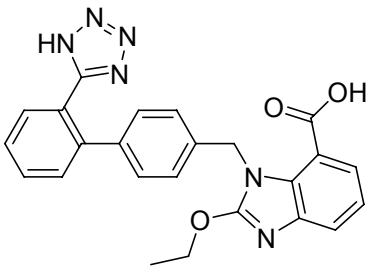
2-chloro-4-(4-methoxyphenyl)-3-phenylquinoline (37118-70-0)



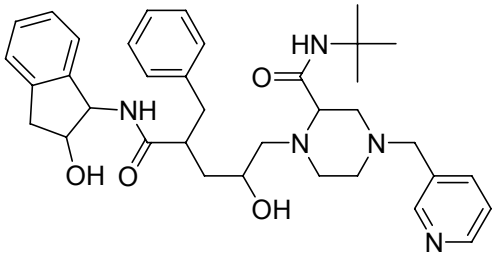
Ceftibuten (97519-39-6)



5-chloro-1,3-dihydro-1,3,3-trimethylspiro(indole-2,3'-phenanthr(9,10-B)oxazine) (119980-37-9)



Candesartan (139481-59-7)



Indinavir (150378-17-9)

The six misclassified *GT*+ compounds are mebendazole, clomiphene, lansoprazole, clarithromycin, imipramine and ampicillin. From the study of Snyder et. al (Snyder *et al.* 2004), ampicillin, imipramine and lansoprazole were also misclassified by MCASE, DEREK and TOPKAT. Clomiphene was misclassified by MCASE and TOPKAT, but correctly classified by DEREK which alerts the halogenated alkene structure (Snyder *et al.* 2004). Mebendazole was misclassified by DEREK but predicted as equivocal genotoxicity by TOPKAT and by MCASE as *GT*+ with 57% probability (Snyder *et al.* 2004). To the best of our knowledge, there is no computational study on clarithromycin, which has been found to be *GT*+ in the *in vitro* cytogenetics tests (Snyder *et al.* 2001) but *GT*- in other assays such as bacterial mutation (Ames), mouse lymphoma assay (MLA), and *in vivo* cytogenetics.

DEREK is a knowledge-based expert system of qualitative estimation model (Greene 2002). MCASE performs a quantitative prediction by generating each test molecule into 2-10 atoms fragments by consideration of their physicochemical properties (Greene 2002). TOPKAT uses electrotopological states as well as shape, symmetry, molecular weight and logP as descriptors in a QSAR model for prediction (Greene 2002). Although each of these methods is able to correctly predict one of the six *GT*+ compounds misclassified by our method, there are also *GT*+ compounds, such as naloxone and pentobarbital (Snyder *et al.* 2004), correctly predicted by our method but misclassified by each of these methods. While all of the methods misclassified some of the *GT*+ compounds due to the general inadequacy for fully representing all of the properties of these molecules, each method appears to be more useful to specific types of compounds than other methods. For instance, clomiphene is correctly predicted by DEREK because of the use of knowledge-based alert for halogenated alkene structure, while it is misclassified by our method because

of the lack of a descriptor to properly represent halogen atoms. Thus the use of multiple methods may be useful to cover a more diverse set of compounds.

The seven misclassified *GT-* compounds are dansyltryptamine, ketotifen, 2-chloro-4-(4-methoxyphenyl)-3-phenylquinoline, ceftibuten, 5-chloro-1,3-dihydro-1,3,3 trimethylspiro, candesartan, and indinavir. Both candesartan and indinavir were correctly classified by MCASE, DEREK and TOPKAT (Snyder *et al.* 2004). Ketotifen was correctly classified by MCASE and DEREK, but misclassified by TOPKAT (Snyder *et al.* 2004). Ceftibuten was correctly classified by MCASE and TOPKAT, but misclassified by DEREK (Snyder *et al.* 2004). The first two compounds contain aromatic amines, the third contains an α,β-unsaturated ketone group, the fourth is composed of an α,β-unsaturated amide group, These chemical groups can be easily distinguished from the structural alerts of genotoxicity (Ashby 1985) used in MCASE, DEREK and TOPKAT, but they are not properly described by the commonly used molecular descriptors. This is perhaps the reason why our method failed to correctly classify these four compounds. Dansyltryptamine, 2-chloro-4-(4-methoxyphenyl)-3-phenylquinoline and 5-chloro-1,3-dihydro-1,3,3 trimethylspiro were correctly predicted by using LDA, kNN, PNN and their consensus model in an earlier study (He *et al.* 2003). These are polycyclic aromatic compounds that contain either chlorine atom or aromatic amine and a N-dimethyl group. One possible reason for the correct prediction of these compounds in that study (He *et al.* 2003) is that it focused on polycyclic aromatic compounds only and thus was easier to select all of the relevant features without the concern of introducing noise for other types of chemical groups. In contrast, our study includes a diverse set of compounds, and our feature selection method can only pick up those descriptors that are both relevant to the polycyclic aromatic compounds and without significant noise to other types of

compounds. It is also noted that there are polycyclic aromatic compounds, such as 9-aminophenanthrene and ethyl 5-hydroxy-2-methylindole-3-carb-oxylate that were correctly predicted by our method and misclassified in the earlier study (He *et al.* 2003). This seems to suggest that the currently available descriptors may not be fully representative of the polycyclic aromatic compounds.

In general, the main reason for the SVM misclassification of these *GT+* and *GT-* compounds is that none of the current descriptors adequately represents the compounds containing multi-rings with various heteroatoms such as nitrogen, oxygen, sulphur, fluorine and chlorine. Current topological descriptors are capable of representing molecular shape, connectivity, and some level of molecular flexibility (Basak *et al.* 1999; Luco 1999; Wegner *et al.* 2004). However, because of the limited coverage of the number of bond links in a heteroatom loop, these descriptors are not yet capable of describing the special features of a complex multi-ring structure that contains multiple heteroatoms. Another reason for the misclassification of some of these compounds is that none of the current descriptors can be used to fully represent molecules containing a long flexible chain. Therefore, there is a need to explore different combination of descriptors and to select more optimum set of descriptors by using more refined feature selection algorithms and parameters. However, indiscriminate use of many existing topological descriptors, which are overlapping and redundant to each others, may introduce noise as well as extending the coverage of some aspects of these special features. Thus, it may be necessary to introduce more appropriate descriptors for representing these and other special features.

8.1.4 Conclusion

This study shows that statistical learning methods, particularly SVM, kNN, and PNN, are useful for facilitating the prediction of *GT*+ potential of a diverse set of molecules without requiring the intrinsic mechanism knowledge of chemical compounds. The prediction accuracy of these methods may be further improved by introducing molecular descriptors that can better represent complex ring structures and flexible long chains and by selection of descriptors most relevant to genotoxicity prediction by means of more refined feature selection methods and parameters. Current efforts are directed at the improvement of the efficiency and speed of feature selection methods (Furlanello *et al.* 2003), which can further help to optimally select molecular descriptors and enable the development of more accurate and efficient computational tools for genotoxicity prediction. Moreover, recent works on the introduction of weighting function into SVM descriptors (Chapelle *et al.* 2002) may also be helpful in developing SVM into a practical tool for the prediction of toxicological properties of compounds.

8.2 Torsade de Pointes

8.2.1 Introduction

In an effort to improve the efficiency of drug discovery, computational tools for ADR prediction have been developed, aimed at facilitating the elimination of ADR causing compounds in early stages of drug development (Kennedy 1997; van de Waterbeemd *et al.* 2003). Mechanism-based knowledge systems (Sanderson *et al.* 1991; Smithing *et al.* 1992) and statistical models describing the correlation between specific ADR and structure-derived physicochemical features (Klopman 1992; Prival 2001) have been developed. Moreover, ligand-protein docking methods have also been explored for the prediction of ADR by screening ADR-inducing drug-protein interactions (Chen *et al.* 2001; Rockey *et al.* 2002). These methods have shown promising potential in the prediction of such ADRs as carcinogenicity, mutagenicity, teratogenicity, irritation, sensitization, immunotoxicity and neurotoxicity (Cronin *et al.* 1994; Kulkarni *et al.* 1999; Benigni *et al.* 2000; Devillers 2000).

So far, attention has not been sufficiently paid to the development of methods for prediction of serious ADRs that occur less frequently. While these ADRs are tolerated to a certain extent for the approval of drugs used in serious diseases urgently needing effective or more treatment options such as AIDS and cancer (Somers *et al.* 1990), they are nonetheless important safety issues for the approval of drugs intended for minor illnesses with availability of alternative treatment options. Examples of these illnesses are rhinitis, cough, pain, inflammation and hypertension. Therefore, there is a need to develop computational methods for facilitating the prediction of the ADRs of these drugs.

One such ADR is torsade de pointes (TdP), which is an atypical rapid ventricular tachycardia with periodic waxing and waning of amplitude of the QRS complexes on the electrocardiogram as well as rotation of the complexes about the isoelectric line (Saunders 2000). TdP may be self-limited or may progress to ventricular fibrillation (Saunders 2000). This ADR is uncommon (Darpo 2001) and thus difficult to detect during clinical trials. There are cases of TdP-causing drugs which were initially approved and later withdrawn after post-marketing surveillance revealed their TdP-causing potential (De Ponti *et al.* 2002; Layton *et al.* 2003).

Not all mechanisms of TdP are completely understood (Moss 1999). TdP is frequently associated with QT prolongation, which is the lengthening of the time between the start of ventricular depolarization and the end of ventricular repolarization. This arises from the disruption of the balance between inward and outward currents during the cardiac action potential repolarization phase (Malik *et al.* 2001). Drugs that induce QT prolongation usually cause disruption of the outward potassium currents by blocking potassium ion channels, particularly human ether-a-gogo related gene (HERG) K^+ channel (Vandenberg *et al.* 2001). This correlation between QT prolongation and blockade of relevant channels had been exploited in the development of computational methods for the prediction of the QT prolongation risk of drugs using artificial neural network (Roche *et al.* 2002) and pharmacophore models (Cavalli *et al.* 2002).

There is no definitive correlation between QT prolongation and TdP (Malik *et al.* 2001; Muzikant *et al.* 2002). For instance, verapamil causes QT prolongation but does not induce TdP, whereas procainamide and disopyramide cause TdP but are not potent inhibitors of the HERG K^+ channel (Muzikant *et al.* 2002). Thus, it is desirable

to develop a method capable of prediction of TdP of multiple mechanisms without complete knowledge of these mechanisms.

A useful method for classification of systems with multiple mechanisms without requiring their knowledge is SVM. This work explores the use of SVM as a potential tool for TdP prediction.

8.2.2 Methods

8.2.2.1 Selection of TdP- and non-TdP-causing compounds

TdP-causing (*TdP+*) compounds were collected from ArizonaCERT (ArizonaCERT). These compounds were identified from human studies and can be divided into 4 classes: Class 1 contains compounds with risk of TdP, class 2 includes compounds with possible risk of TdP, class 3 is composed of compounds to be avoided by congenital long QT patients and class 4 contains compounds which have been weakly associated with TdP. Only compounds from class 1, 2 and 3 were used for training the SVM system. Compounds in class 4 were not considered because it is unclear which of the compounds definitely induces TdP. Thus 67 *TdP+* compounds were selected and used as the training set.

To objectively assess the prediction accuracy of our SVM system, an additional set of *TdP+* compounds, also identified from human studies, were collected from Micromedex (MICROMEDEX 2003b), Drug Information Handbook (Lacy *et al.* 2002), Meyler's side effects of drugs (Dukes 1996) and a list of compounds compiled by De Ponti et al (De Ponti *et al.* 2001), The selection criteria for the compounds are: (1) compounds with known TdP side effects and (2) compounds from De Ponti's list satisfying either criterion Ia or IIIa. Criterion Ia is the existence of clinical studies

and/or case reports associating the compound with the occurrence of TdP/ventricular tachyarrhythmias. Criterion IIIa is the presence of official warnings in the labeling on QT prolongation or occurrence of TdP. The exclusion criteria are: (1) compounds known to be involved in QT prolongation without information about their effect on TdP, (2) compounds in class 1, 2, 3 or 4 of the ArizonaCERT list. This gives an independent validation set of 39 *TdP+* compounds.

Like in the case of other classification systems, training of a SVM system requires information about non-TdP-causing (*TdP-*) compounds. In this work, 243 *TdP-* compounds were obtained from the search of Micromedex, Drug Information Handbook and American Hospital Formulary Service (AHFS) (Bethesda 2001) for compounds with no reported case of TdP in humans. 39 of these compounds were randomly selected and used as part of the independent validation set to assess the prediction accuracy of the SVM system on *TdP-* compounds, while the rest were used in the training set.

8.2.2.2 Chemical descriptors

In this work, linear solvation energy relationships (LSER) descriptors (Kamlet *et al.* 1981; Kamlet *et al.* 1987; Abraham 1993) were used for the modeling of *TdP-*causing potential of compounds. LSER descriptors describe solvent-solute interactions and contain three main terms: a cavity term, a polar term, and hydrogen-bond term. The cavity term is a measure of the endoergic cavity-forming process, which is the free energy necessary to separate the solvent molecules, overcoming solvent-solvent cohesive interactions, and provides a suitably size cavity for the solute. The polar term measures the exoergic balance of solute-solvent and solute-solute

dipolarity/polarizability interactions and the hydrogen-bond term measures the exoergic effects of the complexation between solutes and solvents.

LSER was initially developed for the estimation of the effects of different solvents on properties of specific solutes or the solubilities, lipophilicities, or other properties of a set of different solutes in a specific solvent. It has since been extended for analysis of biological properties including toxicological properties of compounds (Wilson *et al.* 1991; He *et al.* 1995; Sixt *et al.* 1995; Dai *et al.* 2001; Yu *et al.* 2002; Liu *et al.* 2003), cell permeation (Platts *et al.* 2000), intestinal absorption (Zhao *et al.* 2001) and blood-brain barrier penetration (Platts *et al.* 2001). LSER descriptors encode the size, polarity and hydrogen bonding capability of a chemical which have been found to be important for the passive transport of a chemical through biological membranes (Gratton *et al.* 1997; Kramer *et al.* 2001). In addition, it has been shown that complex systems, such as receptor sites, can be approximately described as a solvent system and LSER methods provide useful insights into important binding features (Cramer *et al.* 1992). Thus, the polar term may represent the binding action via dispersion forces of a chemical in the polar regions of a receptor molecule and the hydrogen bond term represents the hydrogen-bonding effect between the chemical and the receptor molecule (Lowrey *et al.* 1997; Liu *et al.* 2003). Since toxicity of a compound involves the transport of the compound to a site and its interaction with a molecular target, LSER descriptors are thus likely to be useful for TdP modeling.

The LSER descriptors used in this study was calculated using our own developed software based on the method developed by Platts (Platts *et al.* 1999). The accuracy of these calculated descriptors for some of the compounds has been verified using the demo version of the software Absolv (Sirius 2000). These descriptors are

excess molar refraction, combined dipolarity/polarizability, overall solute hydrogen bond acidity, overall solute hydrogen bond basicity and McGowan's characteristic volume.

8.2.2.3 Validation of SVM classification system

In this work, the SVM classification system was optimized and validated using leave-one-out (LOO) cross-validation. Y-randomization was also used to validate the trained SVM classification system. The randomization is repeated 10 times and LOO accuracies of the new classification system from each run are compared to that of the original classification system. If the scrambled training set gives significantly lower LOO accuracies than the original training set, the original classification system is unlikely to arise as a result of chance correlation.

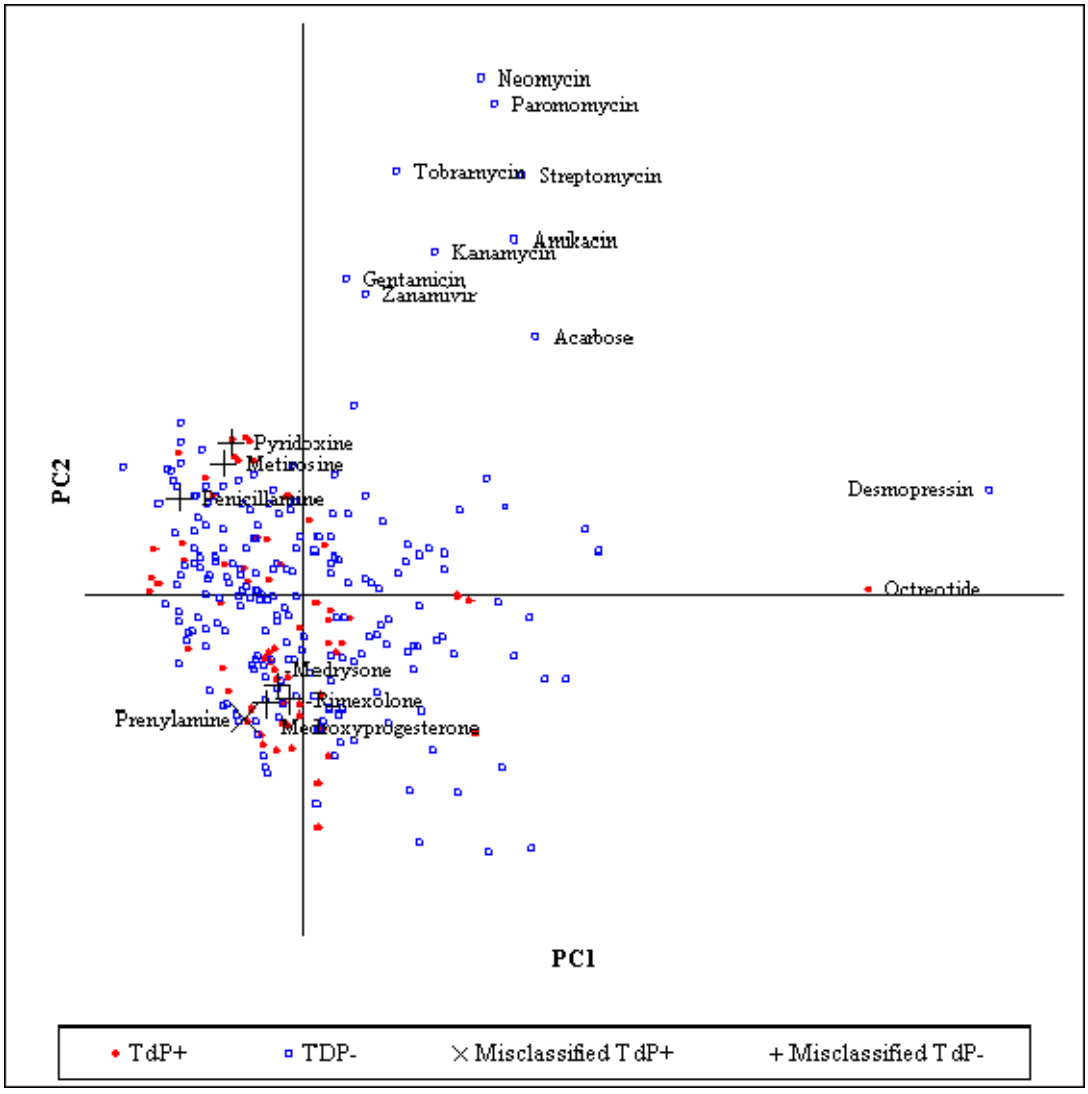
The final SVM classification system was then tested by using the independent validation set to objectively assess its predictive capability. Prediction accuracy of the final SVM classification system using this independent validation was compared with those derived from three other classification methods useful for the prediction of multiple mechanisms. These methods are PNN, kNN and C4.5 DT. The three classification systems were trained using the same training set, descriptors and procedure as those used in SVM. They were tested using the same independent validation set.

8.2.3 Results

A principal component analysis (PCA) (Wold *et al.* 1987) on all of the five LSER descriptors was performed using the training set. PCA resulted in two principal

components (PCs) which explained 84.6% of the total variance in the five LSER descriptors. Component one and two explained 70.2% and 14.4% of the variance respectively. Figure 8.3 shows a score plot of the compounds in the training set using the first two PCs. Score plots are useful for comparing the distribution of compounds in the chemical space between two datasets and to identify clusters of compounds and single compounds that may be outliers (Wold *et al.* 1987; Doddareddy *et al.* 2006).

Figure 8.3 Score plot of first two principal components for training set.



Octreotide, a *TdP*+ compound, and desmopressin, a *TdP*- compound, was found to be far out to the right of the score space. Both of these compounds are large in size, with molecular weight of approximately 1019 and 1069 respectively. There is also a cluster of *TdP*- compounds at the top of the score plot. This cluster mainly contains the aminoglycoside antibiotics like amikacin and gentamicin together with two other compounds, acarbose and zanamivir. Other than the aminoglycosides' cluster, the score plot showed that *TdP*+ and *TdP*- compounds cannot be easily separated using their PCs.

LOO cross-validation was used to derive the optimum sigma parameter for the Gaussian kernel used by SVM and the optimum SVM classification system was found to have a LOO *TdP*+ accuracy of 71.6% and LOO *TdP*- accuracy of 86.3%. Both of these accuracies are significantly greater than 50%, indicating that the trained SVM classification system is significantly better than a random classifier.

To determine whether it results from chance correlation, the SVM classification system was further tested by repeating y randomization for 10 times. The average LOO *TdP*+ accuracy from these ten scrambled classification systems is 21.2% and the average LOO *TdP*- accuracy is 77.3%. Both of these accuracies are worse than that of the original SVM classification system, indicating that the SVM classification system is produced as a result of actual correlation between LSER descriptors and TdP-causing potential of the chemicals and not due to chance.

There has been no reported computational study of the TdP-causing potential of a compound. Thus to objectively assess the usefulness of SVM for TdP prediction, its prediction accuracy is compared with those obtained from three other classification methods, C4.5 DT, kNN and PNN, using the same independent validation set. The optimum parameters, k for kNN and σ for PNN, were found by using LOO cross-

validation. The optimum parameters for SVM, PNN and kNN and the accuracy results are given in Table 8.5. SVM has the highest overall accuracy among the four classification methods. Its *TdP*+ accuracy of 97.4% is substantially higher than the other three classification methods which have *TdP*+ accuracies of 38.5-89.7%. Its *TdP*- accuracy of 84.6% is comparable to the other three methods which have *TdP*- accuracies of 84.6-92.3%. These results suggest that SVM is potentially useful for facilitating the prediction of TdP causing risk of investigative compounds and likely other ADRs with multiple mechanisms. In addition, the SVM classification system is not more flexible than is necessary and thus unlikely to overfit.

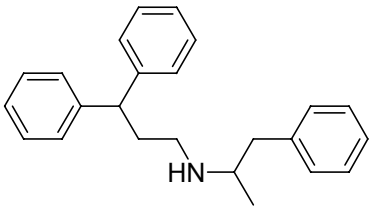
Table 8.5 Results of various classification methods on independent validation set.

Method	Optimum parameter	*TdP*+			*TdP*-			Q (%)
		TP	FN	SE (%)	TN	FP	SP (%)	
C4.5 DT	-	15	24	38.5	36	3	92.3	65.4
kNN	3	35	4	89.7	34	5	87.2	88.5
PNN	0.1	28	11	71.8	33	6	84.6	78.2
SVM	0.3	38	1	97.4	33	6	84.6	91.0

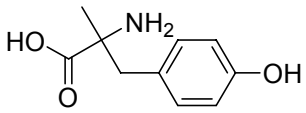
In the training set, there are several aminoglycoside antibiotics grouped together in a cluster which does not overlap significantly with the main cluster of compounds. To examine whether this cluster of aminoglycoside antibiotics contribute in some way to the high *TdP*+ accuracy, a new SVM classification system was trained with all of the aminoglycoside antibiotics removed from the training set. The new SVM classification system gives the same *TdP*+ and *TdP*- accuracies as the original system. This suggests that the aminoglycoside antibiotics are not responsible for the high *TdP*+ accuracy of the SVM classification system.

There are seven compounds incorrectly classified by our SVM system, which are shown in Figure 8.4. These include one *TdP+* compound (prenylamine) and six *TdP-* compounds (medroxyprogesterone, medrysone, metirosine, penicillamine, pyridoxine, rimexolone). Their location on the score plot of the training set is shown in Figure 8.3 above. Prenylamine is incorrectly classified by SVM, PNN and C4.5 DT. Metirosine and pyridoxine are incorrectly classified by SVM, kNN and PNN while penicillamine is incorrectly classified by both SVM and PNN. Medroxyprogesterone, medrysone and rimexolone have a common steroidal structure and are consistently misclassified by all the four classification methods. This may indicate that the LSER descriptors are unable to fully describe the properties of steroidal compounds thus resulting in their misclassifications by all the four classification methods.

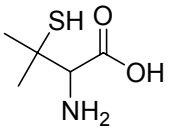
Figure 8.4 Incorrectly classified compounds in the independent validation set.



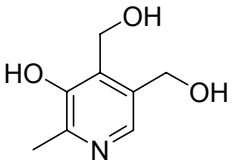
Prenylamine



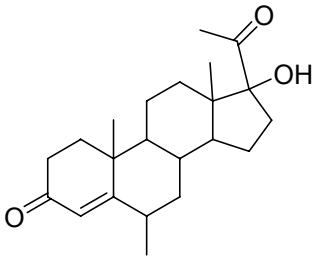
Metirosine



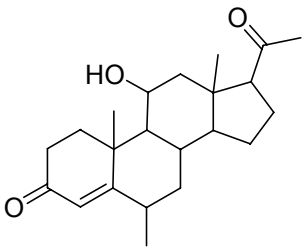
Penicillamine



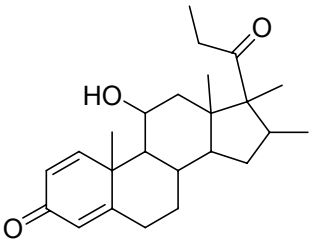
Pyridoxine



Medroxyprogesterone



Medrysone



Rimexolone

To determine whether the LSER descriptors are sufficient for TdP prediction, we analyzed 490 commonly used descriptors for their relevance in TdP classification and used those essential descriptors to construct a separate SVM classification system. Results using that system are compared with the results using LSER descriptors. These descriptors can be broadly classified into four classes. The first class includes

descriptors for global properties of a molecule such as molecular weight, count of atoms, rings and rotatable bonds. The second class contains topological descriptors such as molecular connectivity indices (Kier *et al.* 1986), electrotopological indices (Kier *et al.* 1999), shape indices (Kier 1985) and flexibility indices (Kier 1990). The third class is composed of geometric descriptors including molecular volume, surface area and polar surface area. The fourth class contains chemical descriptors such as dipole moment, polarizability and some of the VolSurf descriptors (Cruciani *et al.* 2000a). A preliminary screening was done to reduce the pool of descriptors by eliminating those descriptors that contained little information. Descriptors that have the same value for more than 50% of the compounds were also removed. Backward elimination was then used to produce an optimum subset of descriptors. During backward elimination, LOO cross-validation was used to assess the performance of each subset of descriptors. In the end, the best subset of descriptors consists of 108 descriptors that are not highly correlated with one another. These 108 descriptors were used to train the SVM classification system and the resultant system has *TdP*+ and *TdP*- accuracies of 92.3% and 84.6% on the independent validation set. These results are comparable to that of the current study. This suggests that LSER descriptors are equally useful for prediction of TdP as those using a more diverse set of descriptors.

8.2.4 Discussion

In this study, SVM classification system is compared with three other classification methods and the results suggest that SVM classification system has the best predictive ability among the four methods. All of these classification methods were developed primarily in the machine learning literature and use different

algorithms than standard statistical methods. Thus to fully evaluate the performance of SVM classification system, a standard statistical method, logistic regression, was applied to the classification of the same *TdP*+ and *TdP*- datasets. The *TdP*+ prediction accuracy using the independent validation set using logistic regression is only 20.5%. In addition, y randomization validation tests showed that the LOO *TdP*- accuracy of the logistic regression model is less than the mean LOO *TdP*- accuracies of the scrambled models. Thus the logistic regression model, as a method for systems with unique mechanism, is not suitable for TdP classification which is intrinsically a multi-mechanism problem.

The possible reason for the usefulness of LSER descriptors for TdP prediction is that they roughly encode most of the essential characteristics related to the TdP causing capability of a compound. Excess molar refraction represents the tendency of a compound to interact with a receptor through n- and π-electron pairs and thus is a measure of the hydrophobic interaction between the compound and receptor. The combined dipolarity/polarizability, on the other hand, represents the ability of electrons to move and be delocalized in the chemical and is a measure of the polar interaction between the compound and receptor.

The overall solute hydrogen bond acidity, overall solute hydrogen bond basicity represents the ability of the compound to form hydrogen bonds with the receptor. This, together with the hydrophobic and polar interactions encoded by the excess molar refraction and combined dipolarity/polarizability, determines the binding affinity of the chemical for the receptor.

The McGowan's characteristic volume influences the passage of a chemical through biological membranes. A compound with a large volume may have difficulty passing through biological membranes and thus may not exhibit toxicity as it is

unable to reach its toxicity receptor. In addition, the binding site of a receptor is usually a cavity that can accommodate compounds of a specific range of sizes and shapes.

Currently, with the exception of C4.5 DT which is able to generate decision rules, the other three classification methods are unable to determine the relative importance of individual LSER descriptor. This limits the scope of the application of SVM classification systems in drug design to tasks such as high-throughput screening. With further improvement of SVM algorithm such as the introduction of weighting function to the descriptors (Chapelle *et al.* 2002), specific rules of the descriptors may be derived which in turn extend the application range of SVM classification systems.

As with all other *in silico* predictions of toxicological properties of chemical compounds, prediction of TdP-causing potential by SVM should be assessed together with pharmacokinetic and pharmacodynamic properties of the chemical compounds in order to determine their clinical significance. This is because a potential TdP-causing drug is not the sole factor in precipitating TdP in a patient. Variability in drug concentrations, drug/drug interactions and individual patient's susceptibility are some of the numerous factors that affect the occurrence of TdP in patients. Thus a positive TdP-causing risk of a drug-like molecule may not preclude its use in the clinical setting (Malik *et al.* 2001). For example, both halofantrine and terfenadine can potentially cause TdP. However, halofantrine is still in use whereas terfenadine has been withdrawn from the US market as halofantrine is useful for resistant malaria treatment but for terfenadine, there are other safer alternatives, like fexofenadine available (Malik *et al.* 2001). Despite the limitations of *in silico* prediction of TdP, it may be used as part of the overall risk-benefit analysis of investigative drugs to evaluate their usefulness in the clinical setting.

8.2.5 Conclusion

As a statistical learning method for the prediction of systems with multiple mechanisms, SVM is potentially useful for facilitating the prediction of TdP causing risk of investigative drugs. The availability of more extensive information about various ADR-causing compounds and associated mechanisms and more comprehensive descriptors for toxicity prediction will enable the development of SVM and other computational methods into useful tools for facilitating the prediction of different types ADRs in early stage of drug development.

Chapter 9

Conclusions

This last chapter summarizes the major findings (section 9.1) and contributions (section 9.2) of this work to the progress of using machine learning approaches for pharmacokinetics and toxicity predictions. Limitations of the present work (section 9.3) and possible areas for future studies (section 9.4) are also discussed.

9.1 Major Findings

In chapters 4, 6 and 8, support vector machine (SVM) was shown to be a useful computational method for facilitating the prediction of ADMET properties like human intestinal absorption (HIA), p-glycoprotein (P-gp) substrates, cytochrome (CYP) P450 isoenzymes inhibitors and substrates, genotoxicity and torsade de pointes (TdP), without requiring the intrinsic mechanism knowledge of chemical compounds. Thus it is likely that SVM will be an efficient computational tool for the prediction of ADMET properties of chemical compounds.

In chapters 4 and 8, recursive feature elimination (RFE) was found to be capable of automatic selection of molecular descriptors and reduction of the noise generated by the use of overlapping and redundant molecular descriptors. This reduction appears to be helpful in enhancement of the performance of SVM for the prediction of ADMET properties of chemical compounds.

In chapter 6, our study suggests that consensus classification systems give better predictive performance than single classification systems. This result is consistent with earlier studies. All of the 'positive probability' consensus SVM classification systems (PP-CSVMs) for predicting inhibitors/substrates of the three P450 isoenzymes, CYP3A4, CYP2D6 and CYP2C9, show high prediction accuracies, with improved specificities compared to those of earlier studies. Our computational results suggest PP-CSVM is better than 'positive majority' consensus SVM classification system (PM-CSVM) for constructing consensus SVMs (CSVMs) for classifying inhibitors and substrates of various P450 isoenzymes. Thus CSVMs, particularly PP-CSVM, are potentially useful for developing filters for prediction of inhibitors and substrates of P450 isoenzymes and other ADMET properties.

In chapter 5, our results suggest that general regression neural network (GRNN) is a potentially useful method for developing QSPkR models from a diverse set of drug data. QSPkR models developed by using GRNN for three drug distribution properties, blood-brain barrier (BBB) penetration, human serum albumin (HSA) binding, and milk-plasma (M/P) distribution, were tested and compared with those developed by using a linear method, multiple linear regression (MLR), and a non-linear method, multilayer feedforward neural network (MLFN). All the GRNN-developed models showed better prediction capability than the corresponding MLR- or MLFN-developed models.

In chapter 7, our study suggests that both GRNN and support vector regression (SVR) are potentially useful for developing QSPkR models to predict drug clearance from a large diverse set of compound data. QSPkR models developed by using GRNN, SVR and *k* nearest neighbour (kNN) were tested and compared with those developed by using a linear method, partial least squares (PLS). All of the GRNN- and SVR-

developed models show better prediction capability than the corresponding kNN- or PLS-developed models. The predictive capabilities of the QSPkR models developed in this study are comparable to those of previous studies and are further improved by using consensus modeling methods.

In chapter 7, we also found that a collection of constitutional, geometrical, topological and electrotopological descriptors seems to be more useful for modeling drug clearance than specialized descriptor sets such as 3DMoRSE, ATS, GETAWAY, RDF and WHIM. A possible reason is that an individual descriptor set tends to partially neglect some important features and thus the use of different types of descriptors may help to alleviate such type of feature bias. The three statistical learning methods, GRNN, SVR and kNN, appears to be capable of combining the information encoded in the different descriptor sets effectively to develop more predictive QSPkR models.

Non-linear methods, such as SVM, GRNN, and SVR, are useful for developing QSPkR/qSPkR models involving multiple mechanisms because they belong to the class of distance-based methods. In a diverse dataset, compounds having the same mechanism of actions will be close to one another in the chemical space and compounds having different mechanism of actions will be far apart. In distance-based methods, multiple localized models were developed for each mechanism and these were then combined into a single model. The ADMET property of a compound is predicted by measuring the distance between the compound and the various localized models and then using the localized model which is closest to the compound.

9.2 Contributions

This work has improved the quality of previous QSPkR/qSPkR models for ADMET prediction. All the QSPkR/qSPkR models developed in this work have higher prediction capability than the corresponding models developed by other workers. The use of known, relative new machine learning methods, such as SVM, SVR and GRNN, and consensus modeling was found to be useful for improving prediction capability of QSPkR/qSPkR models for HIA, P-gp substrates, BBB penetration, HSA binding, M/P, CYP isoenzymes substrates and inhibitors, total body clearance, and genotoxicity. A qSPkR model was also constructed for TdP, a rare but serious adverse drug reaction, which have not received sufficient attention. These models were also developed by using a larger number and more diverse groups of compounds, as well as compounds with known human ADMET data. Thus the models are expected to have better generalization ability than the previous models and are directly applicable for the prediction of human ADMET property without the need for allometric scaling to convert predicted animal ADMET property to human ADMET property like in the previous models. Hence, the QSPkR/qSPkR models developed in this work are potentially useful to be incorporated as part of the strategy for reducing the cost and improving the speed of drug development.

This work introduces a novel principal component analysis (PCA) based method to improve the interpretability of QSPkR models. Most of the non-linear methods including neural networks are incapable of providing explicit relationships between the predicted properties and the molecular features of the compounds. Results from this work suggest that the use of multi-sigma GRNN models and PCA can partially solve this problem. The individual σ values for each descriptor provide a useful hint about its contribution to the ADMET properties. PCA, when coupled with

specially designed artificial testing sets, may provide a rough guide for the influence of molecular characteristics on ADMET properties. Hence the development of the novel PCA-based functional dependence study approach in this work has helped to improve the interpretability of non-linear models, which were previously difficult to interpret.

A new machine learning library, YMLL, and a new Microsoft Windows software, PHAKISO, were designed and developed in this work to enable QSPkR/qSPkR models to be developed and validated easily. The library and software were better than existing software, Torch and Weka, for developing QSPkR/qSPkR models because Torch and Weka were developed for general machine learning problems whereas YMLL and PHAKISO were developed specifically for QSPkR/qSPkR problems. Thus YMLL and PHAKISO contain algorithms which were more relevant for QSPkR/qSPkR problems. In addition, PHAKISO presents a user-friendly graphical user interface to enable QSPkR/qSPkR models to be built with a few mouse clicks. Torch, on the other hand, does not have a graphical user interface and thus is difficult to be used by scientists who are not familiar with programming. Both Torch and Weka also have a limited number of descriptor selection methods, especially wrapper methods, and have a limited number of methods to measure prediction capabilities of QSPkR/qSPkR models. Hence the development of YMLL and PHAKISO in this work and the availability of both software for non-commercial uses are expected to aid scientists in creating QSPkR/qSPkR models rapidly and thus speed up the drug development process.

9.3 Limitations

The performance of machine learning methods critically depends on the diversity of compounds in a training dataset and the appropriate representation of these compounds. The datasets used in this work are not expected to be fully representative of all of the compounds possessing and not possessing a specific ADMET property. This is particularly true for compounds not possessing a specific property given the vast chemical space of millions of known compounds in the currently available chemical database. Various degrees of inadequate compound representation in these studies likely affect, to a certain extent, the prediction accuracy of the developed QSPkR/qSPkR models.

In chapter 6, a potential problem is that the selection criteria for non-inhibitors and non-substrates of CYP isoenzymes may result in a small number of false negatives. However, the use of SVM was found to help in achieving a balance between training errors and prediction accuracies. The CSVMs are presently only suitable for distinguishing between inhibitors and non-inhibitors or substrates and non-substrates. The availability of more detailed experimental data will enable the use of multi-class SVM (Angulo *et al.* 2003) for the classification of non-inhibitors, weak inhibitors and strong inhibitors, or SVM regression (Smola *et al.*) for quantitative prediction of the Ki values of inhibitors.

Most of the descriptors used in this study are 1-D and 2-D in information content about the molecule. Some of the descriptors have some 3-D content such as molecular dipole moment, but properties related specifically to conformational entropy, are not present. In addition, molecular dipole moment is ill-defined since the associated “active conformation” for the dipole is unknown. The insufficient use of 3-D descriptors may affect the prediction capability and interpretation of the

QSPkR/qSPkR models. For example, in BBB penetration, transport across a series of tightly packed biological membranes is involved. There are no meaningful 1-D and 2-D descriptors that can distinguish between a long, thin and flexible conformation of a molecule from a spherical, balled-up and rigid confirmation of a molecule and this difference in conformational preference is a controlling factor for BBB penetration.

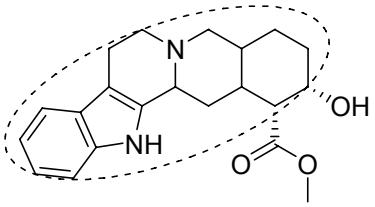
Some of the descriptors used in this study may be highly correlated and thus are very likely redundant in information content. Descriptors which are essentially the same do not necessarily equate to a large information content regarding distinct molecular properties. Hence the interpretation of the descriptors should be more appropriately conducted at the descriptor class level where redundant and overlapping descriptors are grouped into one class. PCA can also be used to group descriptors sharing the same information content into one principal component (PC) and performing functional dependence study on the individual PCs rather than on the individual descriptors.

Currently, three-levels of characterizing a mechanism of action from a QSPkR/qSPkR model are usually reported in the literature. The first level is stating the specific groups and their interactions of a molecule responsible for activity. The second level is providing a pharmacophore needed for expressing the activity, and the last level is stating general molecular features that often seem to present in molecules exhibiting a given type of activity. All the QSPkR/qSPkR models that were developed in this study fall into the third, least specific category.

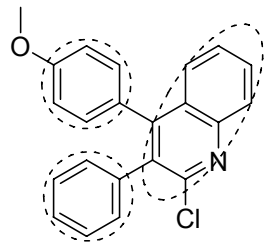
Examinations of incorrectly classified compounds in this work have consistently suggested that the current molecular descriptors are not sufficient to adequately represent some of the compounds that contain complex structural or chemical configurations (Figure 9.1). These include compounds containing long

flexible chains, highly polar tetrazole rings, multiple ionisable groups, polycyclic aromatic structures, complex two ring system with multiple heteroatoms, aromatic rings separated by a specific atom, compounds with multiple heteroatoms and compounds with complicated ring structure. Due to the limited coverage of the number of bond links in a heteroatom loop, topological descriptors are not yet capable of describing the special features of a complex multi-ring structure that contains multiple heteroatoms. It appears that none of the currently available descriptors can be used to fully represent molecules containing a long flexible chain.

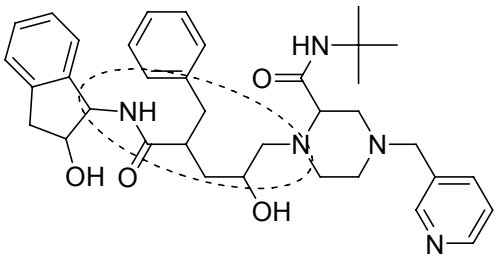
Figure 9.1 Examples of compounds not-well-represented by the currently available molecular descriptors. The not-well-represented part of the structure is indicated by a dashed line.



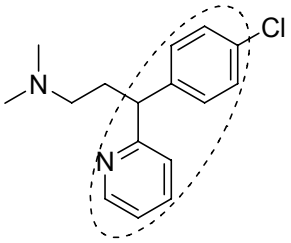
Inflexible multi-ring
(Yohimbine)



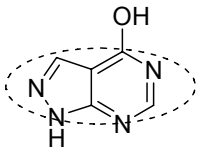
Polycyclic aromatic structure
(2-chloro-4-(4-methoxyphenyl)-3-phenylquinoline)



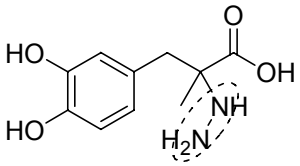
Long flexible chain
(Indinavir)



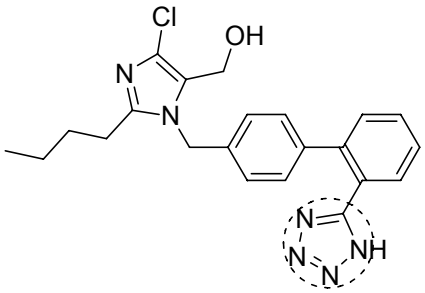
Two aromatic rings separated by a specific atom
(Chlorphenamine)



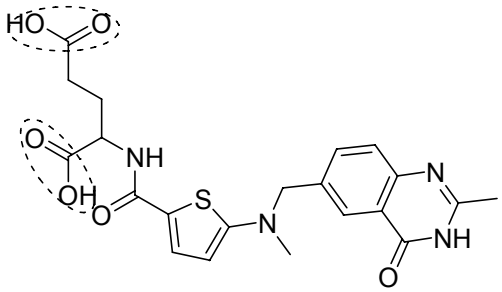
Complex two ring system with multiple heteroatoms
(Allopurinol)



Hydrazine group
(Carbidopa)



Highly polar tetrazole ring
(Losartan)



Multiple ionisable groups
(Raltitrexed)

9.4 Suggestions for Future Studies

Some of the ADMET properties are known to correlate to specific pharmacodynamic properties which lead to clinical significance. For instance, a drug with high BB ratio may not have effects in the brain either because of the absence of target receptors or insufficient potencies towards the target receptors in the brain. Conversely, a drug with a relatively low BB ratio may still have effects in the brain because of its high potency towards specific receptors (Hallstrom *et al.* 1980). Such correlations, which have not been adequately considered in machine learning methods so far, may need to be incorporated in developing QSPkR/qSPkR models for predicting those ADMET properties that are known to correlate to certain pharmacodynamic property.

In this work, it is assumed that sensitivity and specificity of the SVM classification system are equally important. However, in a drug discovery project, these accuracies may have different importance at different stages of the design cycle. For example, in the initial target and hit identification phase, it may be more important not to miss potential leads. Thus, it is more important to have a classification system which has very high sensitivity (small number of false negatives) and reasonably good specificity. At later stages, it becomes increasingly important to focus on a manageable number of candidates. Thus a classification system with very high specificity (small number of false positives) and reasonably good sensitivity may become more important. It is possible to alter the SVM classification systems to suit these different needs. There are two possible approaches for modifying the SVM classification systems. The first approach uses different training error penalties (equation (2.14)) for $D+$ and $D-$. For example, a higher training error penalty for $D+$ and lower training error penalty for $D-$ can be used to increase the sensitivity of the

SVM classification systems. The second approach adds a correction factor to the SVM decision function (equation (2.12)). A positive or negative correction factor will improve the sensitivity or specificity of the SVM classification system respectively.

There is a need to explore different combination of descriptors and to select more optimum set of descriptors by using more refined feature selection algorithms and parameters. However, indiscriminate use of many existing topological descriptors, which are overlapping and redundant to each other, may introduce noise as well as extending the coverage of some of the aspects of these special features. Thus, it may be necessary to introduce new descriptors for more appropriately representing these and other special features. The new descriptors should ideally be able to be translated back to the molecular structures. This will improve the interpretability of the QSPkR/qSPkR models.

In this work, RFE is incorporated into SVM classification systems for dividing molecules into two classes according to specific ADMET property. This method can also be applied to the prediction of ADMET properties in a continuous fashion. Future studies can combine RFE with SVR for providing non-linear QSPkR of specific ADMET properties.

Genetic programming (GP), an evolutionary programming approach, has been found to be useful for the development of qSPkR models for oral bioavailability prediction of a diverse group of compounds (Bains *et al.* 2002). This is because GP implements the IF logic to capture multiple mechanism of action within a single model. Thus evolutionary programming approaches, which have the potential to identify and optimize all independent QSPkR models consistent with the training set data, may be potentially useful for the prediction of the ADMET properties of chemical compounds.

The lack of structural diversity in the training sets may limit the applicability of the models developed by machine learning. However, it may be possible to use analog compound training sets to provide benchmarks as to what upper-level models are possible from a given method for a given endpoint. Future studies can model high analog datasets as a way to evaluate how much accuracy and reliability is lost in modeling structurally diverse data sets for a given machine learning approach.

Bibliography

Abraham MH (1993). Scales of solute hydrogen-bonding: Their construction and appplication to physicochemical and biochemical processes. *Chemical Society Reviews* **22**: 73-83.

Abraham MH (2004). The factors that influence permeation across the blood-brain barrier. *European Journal of Medicinal Chemistry* **39**(3): 235-240.

Abraham MH, Chadha HS, Martins F, Mitchell RC, Bradbury MW and Gratton JA (1999). Hydrogen bonding part 46: A review of the correlation and prediction of transport properties by an LFER method: Physicochemical properties, brain penetration and skin permeability. *Pesticide Science* **55**(1): 78-88.

Abraham MH, Chadha HS and Mitchell R (1994). Hydrogen bonding. 33. Factors that influence the distribution of solutes between blood and brain. *Journal of Pharmaceutical Sciences* **83**(9): 1257-1268.

Abraham MH, Zhao YH, Le J, Hersey A, Luscombe CN, Reynolds DP, Beck G, Sherborne B and Cooper I (2002). On the mechanism of human intestinal absorption. *European Journal of Medicinal Chemistry* **37**(7): 595-605.

Accelrys (2005). DS ViewerPro. Accelrys.

Adenot M and Lahana R (2004). Blood-brain barrier permeation models: Discriminating between potential CNS and non-CNS drugs including p-glycoprotein substrates. *Journal of Chemical Information and Computer Sciences* **44**(1): 239-248.

Agatonovic-Kustrin S, Beresfordb R, Pauzi A and Yusof M (2001). Theoretically-derived molecular descriptors important in human intestinal absorption. *Journal of Pharmaceutical and Biomedical Analysis* **25**(2): 227-237.

Agatonovic-Kustrin S, Ling LH, Tham SY and Alany RG (2002). Molecular descriptors that influence the amount of drugs transfer into human breast milk. *Journal of Pharmaceutical and Biomedical Analysis* **29**(1-2): 103-119.

Agatonovic-Kustrin S, Tucker IG, Zecevic M and Zivanovic LJ (2000). Prediction of drug transfer into human milk from theoretically derived descriptors. *Analytica Chimica Acta* **418**(2): 181-195.

Agrafiotis DK (1996). *Stochastic algorithms for maximizing molecular diversity*. 3rd Electronic Computational Chemistry Conference.

Agrafiotis DK (2001). A constant time algorithm for estimating the diversity of large chemical libraries. *Journal of Chemical Information and Computer Sciences* **41**(4): 159-167.

Agrafiotis DK and Lobanov VS (1999). An efficient implementation of distance-based diversity measures based on k-d trees. *Journal of Chemical Information and Computer Sciences* **39**(1): 51-58.

Ajay, Bemis GW and Murcko MA (1999). Designing libraries with CNS activity. *Journal of Medicinal Chemistry* **42**(24): 4942-4951.

Almuallim H and Dietterich TG (1994). Learning Boolean concepts in the presence of many irrelevant features. *Artificial Intelligence* **69**: 279-306.

Ambudkar SV, Dey S, Hrycyna CA, Ramachandra M, Pastan I and Gottesman MM (1999). Biochemical, cellular, and pharmacological aspects of the multidrug transporter. *Annual Review of Pharmacology and Toxicology* **39**: 361-398.

Andrea TA and Kalayeh H (1991). Applications of neural networks in quantitative structure-activity relationships of dihydrofolate reductase inhibitors. *Journal of Medicinal Chemistry* **34**: 2824-2836.

Andrews CW, Bennett L and Yu LX (2000). Predicting human oral bioavailability of a compound: development of a novel quantitative structure-bioavailability relationship. *Pharmaceutical Research* **17**(6): 639-644.

Angulo C, Parra X and Catala A (2003). K-SVCR. A support vector machine for multi-class classification. *Neurocomputing* **55**(1-2): 57-77.

Ankrest M, Breunig M, Kriegel H and Sander J (1999). OPTICS: Ordering points to identify the clustering structure. *Proceedings of the ACM SIGMOD International Conference on Management of Data***:** 49-60.

ArizonaCERT. (2003). Drugs that prolong the QT interval and/or induce torsades de pointes ventricular arrhythmia. University of Arizona CERT. Retrieved 18 November 2003, from http://www.arizonacert.org/medical-pros/drug-lists/drug-lists.htm.

Ashby J (1985). Fundamental structural alerts to potential carcinogenicity or noncarcinogenicity. *Environmental Mutagenesis* **7**(6): 919-921.

Asikainen AH, Ruuskanen J and Tuppurainen KA (2004). Performance of (consensus) kNN QSAR for predicting estrogenic activity in a large diverse set of organic compounds. *SAR and QSAR in Environmental Research* **15**(1): 19-32.

Atkinson HC and Begg EJ (1990). Prediction of drug distribution into human milk from physiochemical characteristics. *Clinical Pharmacokinetics* **18**(2): 151-167.

Bai JPF, Utis A, Crippen G, He H-D, Fischer V, Tullman R, Yin H-Q, Hsu C-P, Jiang L and Hwang K-K (2004). Use of classification regression tree in predicting oral absorption in humans. *Journal of Chemical Information and Computer Sciences* **44**(6): 2061-2069.

Bain LJ, McLachlan JB and LeBlanc GA (1997). Structure-activity relationships for xenobiotic transport substrates and inhibitory ligands of P-glycoprotein. *Environmental Health Perspectives* **105**(8): 812-818.

Bains W, Gilbert R, Sviridenko L, Gascon JM, Scoffin R, Birchall K, Harvey I and Caldwell J (2002). Evolutionary computational methods to predict oral bioavailability QSPRs. *Current Opinion in Drug Discovery and Development* **5**(1): 44-51.

Bakken GA and Jurs PC (2000). Classification of multidrug-resistance reversal agents using structure-based descriptors and linear discriminant analysis. *Journal of Medicinal Chemistry* **43**(23): 4534-4541.

Balaban AT (1986). Chemical graphs. 48. Topological index J for heteroatom-containing molecules taking into account periodicities of element properties. *MATCH* **21**: 115-122.

Balakin KV, Ekins S, Bugrim A, Ivanenkov YA, Korolev D, Nikolsky YV, Skorenko AV, Ivashchenko AA, Savchuk NP and Nikolskaya T (2004). Kohonen maps for prediction of binding to human cytochrome P450 3A4. *Drug Metabolism and Disposition: The Biological Fate of Chemicals* **32**(10): 1183-1189.

Basak SC, Gute BD and Drewes LR (1996). Predicting blood-brain transport of drugs: a computational approach. *Pharmaceutical Research* **13**(5): 775-778.

Basak SC, Gute BD and Ghatak S (1999). Prediction of complement-inhibitory activity of benzamidines using topological and geometric parameters. *Journal of Chemical Information and Computer Sciences* **39**: 255-260.

Basak SC and Magnuson VR (1983). Molecular topology and narcosis. A quantitative structure-activity relationship (QSAR) study of alcohols using complimentary information content (CIC). *Arzneimittel-Forschung/Drug Research* **33**: 501-503.

Bayada DM, Hamersma H and van Geerestein VJ (1999). Molecular diversity and representativity in chemical databases. *Journal of Chemical Information and Computer Sciences* **39**(1): 1-10.

Begg EJ and Atkinson HC (1993). Modelling of the passage of drugs into milk. *Pharmacology and Therapeutics* **59**(3): 301-310.

Begg EJ, Atkinson HC and Duffull SB (1992). Prospective evaluation of a model for the prediction of milk:plasma drug concentrations from physicochemical characteristics. *British Journal of Clinical Pharmacology* **33**(5): 501-505.

Benigni R, Guiliani A, Franke R and Gruska A (2000). Quantitative structure-activity relationships of mutagenic and carcinogenic aromatic amines. *Chemical Reviews* **100**(10): 3697-3714.

Bergstrom CA, Strafford M, Lazorova L, Avdeef A, Luthman K and Artursson P (2003). Absorption classification of oral drugs based on molecular surface properties. *Journal of Medicinal Chemistry* **46**(4): 558-570.

Bethesda (2001). *AHFS drug information*, American Society of Health-System Pharmacists, Inc.

Bolzan AD and Bianchi MS (2002). Genotoxicity of streptozotocin. *Mutation Research* **512**(2-3): 121-134.

Boobis A, Gundert-Remy U, Kremers P, Macheras P and Pelkonen O (2002). In silico prediction of ADME and pharmacokinetics. Report of an expert meeting organised by COST B15. *European Journal of Pharmaceutical Sciences* **17**(4-5): 183-193.

Box GEP, Hunter WG and Hunter JS (1978). *Statistics for experimenters: An introduction to design, data analysis, and model building*. New York, Wiley.

Brassard G and al. e (1996). *Fundamentals of algorithms*. New Jersey, Prentice Hall.

Bravi G, Gancia E, Mascagni P, Pegna M, Todeschini R and Zaliani A (1997). MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: A comparative 3D QSAR study in a series of steroids. *Journal of Computer-Aided Molecular Design* **11**(1): 79-92.

Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet C, Ares JM and Haussler D (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences of the United States of America* **97**(1): 262-267.

Burbidge R, Trotter M, Buxton B and Holden S (2001). Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Computers and Chemistry* **26**(1): 5-14.

Burges CJC (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* **2**(2): 127-167.

Butina D (1999). Unsupervised data base clustering based on Daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets. *Journal of Chemical Information and Computer Sciences* **39**(4): 747-750.

Butina D, Segall MD and Frankcombe K (2002). Predicting ADME properties in silico: Methods and models. *Drug Discovery Today* **7**(11 (Suppl)): S83-S88.

Cacoullos T (1966). Estimation of a multivariate density. *Annals of the Institute of Statistical Mathematics* **18**: 179-189.

Cai CZ, Han LY, Ji ZL, Chen X and Chen YZ (2003). SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Research* **31**(13): 3692-3697.

Caldwell J, Gardner I and Swales N (1995). An introduction to drug disposition: the basic principles of absorption, distribution, metabolism, and excretion. *Toxicologic Pathology* **23**(2): 102-114.

Cardie C (1993). Using decision trees to improve case-based learning. *Proceedings 10th International Conference on Machine Learning*. Los Altos, Morgan Kaufmann**:** 25-32.

Carhart RE, Smith DH and Venkataraghavan R (1985). Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences* **25**(2): 64-73.

Carnahan B, Meyer G and Kuntz L-A (2003). Comparing statistical and machine learning classifiers: Alternatives for predictive modeling in human factors research. *Human Factors* **45**(3): 408-423.

Cash GG (2001). Prediction of the genotoxicity of aromatic and heteroaromatic amines using electrotopological state indices. *Mutation Research* **491**(1-2): 31-37.

Cavalli A and Poluzzi E (2002). Toward a pharmacophore for drugs inducing the long QT syndrome: Insights from a CoMFA study of HERG K+ channel blockers. *Journal of Medicinal Chemistry* **45**(18): 3844-3853.

Chang CC and Lin CJ (2001). LIBSVM: A library for support vector machines. http://csie.ntu.edu.tw/~cjlin/libsvm.

Chapelle O, Vapnik V, Bousquet O and Mukherjee S (2002). Choosing multiple parameters for support vector machines. *Machine Learning* **46**(1-3): 131-159.

Charton M (1975). Steric effects. I. Esterification and acid catalysed hydrolysis of esters. *Journal of the American Chemical Society* **97**: 1552-1556.

Chen YZ and Ung CY (2001). Prediction of potential toxicity and side effect protein targets of a small molecule by a ligand-protein inverse docking approach. *Journal of Molecular Graphics and Modelling* **20**(3): 199-218.

Cheng A, Diller DJ, Dixon SL, Egan WJ, Lauri G and Merz KMJ (2002). Computation of the physio-chemical properties and data mining of large molecular collections. *Journal of Computational Chemistry* **23**(1): 172-183.

Chohan KK, Paine SW, Mistry J, Barton P and Davis AM (2005). A rapid computational filter for cytochrome P450 1A2 inhibition potential of compound libraries. *Journal of Medicinal Chemistry* **48**(16): 5154-5161.

Chung KT, Kirkovsky L, Kirkovsky A and Purcell WP (1997). Review of mutagenicity of monocyclic aromatic amines: Quantitative structure-activity relationships. *Mutation Research* **387**(1): 1-16.

Clark DE (1999). Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 2. Prediction of blood-brain barrier penetration. *Journal of Pharmaceutical Sciences* **88**(8): 815-821.

Clark RD (1997). OptiSim: An extended dissimilarity selection method for finding diverse representative subsets. *Journal of Chemical Information and Computer Sciences* **37**(6): 1181-1188.

Clark RD and Wolohan PR (2003). Molecular design and bioavailability. *Current Topics in Medicinal Chemistry* **3**(11): 1269-1288.

Collobert R, Bengio S and Mariéthoz J (2002). Torch: A modular machine learning software library. *Technical Report IDIAP-RR 02-46, IDIAP*.

Colmenarejo G (2003). In silico prediction of drug-binding strengths to human serum albumin. *Medicinal Research Reviews* **23**(3): 275-301.

Colmenarejo G, Alvarez-Pedraglio A and Lavandera JL (2001). Cheminformatic models to predict binding affinities to human serum albumin. *Journal of Medicinal Chemistry* **44**(25): 4370-4378.

Consonni V, Todeschini R and Pavan M (2002). Structure/Response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. *Journal of Chemical Information and Computer Sciences* **42**(3): 682-692.

Craig AJ, Weininger D and Delany J (2005). Fingerprints - Screening and Similarity. *Daylight Theory Manual*, Daylight Chemical Information Systems, Inc.

Cramer CJ and Truhlar DG (1992). An SCF solvation model for the hydrophobic effect and absolute free energies of aqueous solvation. *Science* **256**(5054): 213-217.

Crivori P, Cruciani G, Carrupt PA and Testa B (2000). Predicting blood-brain barrier permeation from three-dimensional molecular structure. *Journal of Medicinal Chemistry* **43**(11): 2204-2216.

Cronin MTD and Basketter DA (1994). Multivariate QSAR analysis of a skin sensitization database. *SAR and QSAR in Environmental Research* **2**(3): 159-179.

Cronin MTD and Schultz TW (2003). Pitfalls in QSAR. *Journal of Molecular Structure: THEOCHEM* **622**(1-2): 39-51.

Cruciani G, Crivori P, Carrupt PA and Testa B (2000a). Molecular fields in quantitative structure-permeation relationships: the VolSurf approach. *Journal of Molecular Structure: THEOCHEM* **503**(1-2): 17-30.

Cruciani G, Pastor M and Guba W (2000b). Volsurf: a new tool for the pharmacokinetic optimization of lead compounds. *European Journal of Pharmaceutical Sciences* **11**(Suppl. 2): S29-S39.

Cummins DJ, Andrews CW, Bentley JA and Cory M (1996). Molecular diversity in chemical databases: Comparison of medicinal chemistry knowledge bases and databases of commerically available compounds. *Journal of Chemical Information and Computer Sciences* **36**(4): 750-763.

Currit N (2002). Inductive regression: overcoming OLS limitations with the general regression neural network. *Computers, Environment and Urban Systems* **26**(4): 335-353.

Czerminski R, Yasri A and Hartsough D (2001). Use of support vector machine in pattern classification: Application to QSAR studies. *Quantitative Structure-Activity Relationships* **20**(3): 227-240.

Dai J, Jin L, Yao S and Wang L (2001). Prediction of partition coefficient and toxicity for benzaldehyde compounds by their capacity factors and various molecular descriptors. *Chemosphere* **42**(8): 899-907.

Darpo B (2001). Spectrum of drugs prolonging QT interval and the incidence of torsade de pointes. *European Heart Journal* **2001**(3 Suppl): K70-80.

Daszykowski M, Walczak B and Massart DL (2002). Representative subset selection. *Analytica Chimica Acta* **468**(1): 91-103.

de Groot MJ and Ekins S (2002). Pharmacophore modeling of cytochromes P450. *Advanced Drug Delivery Reviews* **54**(3): 367-383.

de Jong S, Wise BM and Ricker NL (2001). Canonical partial least squares and continuum power regression. *Journal of Chemometrics* **15**(2): 85-100.

De Ponti F, Poluzzi E, Cavalli A, Recanatini M and Montanaro N (2002). Safety of non-antiarrhythmic drugs that prolong the QT interval or induce torsade de pointes: An overview. *Drug Safety* **25**(4): 263-286.

De Ponti F, Poluzzi E and Montanaro N (2001). Organising evidence on QT prolongation and occurrence of torsades de pointes with non-antiarrhythmic drugs: a call for consensus. *European Journal of Clinical Pharmacology* **57**(3): 185-209.

Delph Y. (2000). P-glycoprotein: a tangled web waiting to be unraveled. from http://www.aidsinfonyc.org/tag/science/pgp.html.

Devillers J (2000). A neural network SAR model for allergic contact dermatitis. *Toxicology Methods* **10**(3): 181-193.

Dewar MJS and Stewart JJP (1984). A new procedure for calculating molecular polarizabilities; Applications using MNDO. *Chemical Physics Letters* **111**(4,5): 416-420.

DiMasi JA, Hansen RW and Grabowski HG (2003). The price of innovation: new estimates of drug development costs. *Journal of Health Economics* **22**(2): 151-185.

Dimitrov S, Dimitrova G, Pavlov T, Dimitrova N, Patlewicz G, Niemela J and Mekenyan O (2005). A stepwise approach for defining the applicability domain of SAR and QSAR models. *Journal of Chemical Information and Modeling* **45**(4): 839-849.

Doddareddy MR, Cho YS, Koh HY, Kim DH and Pae AN (2006). In silico renal clearance model using classical Volsurf approach. *Journal of Chemical Information and Modeling* **46**(3): 1312-1320.

Doniger S, Hofmann T and Yeh J (2002). Predicting CNS permeability of drug molecules: comparison of neural network and support vector machine algorithms. *Journal of Computational Biology* **9**(6): 849-864.

Dorronsoro I, Chana A, Abasolo MI, Castro A, Gil C, Stud M and Martinez A (2004). CODES/Neural network model: A useful tool for in silico prediction of oral absorption and blood-brain barrier permeability of structurally diverse drugs. *Quantitative Structure-Activity Relationships* **23**(2-3): 89-98.

Drews J (2000). Drug discovery: A historical perspective. *Science* **287**(5460): 1960-1964.

Drucker H, Wu DH and Vapnik VN (1999). Support vector machine for spam categorization. *IEEE Transactions on Neural Networks* **10**(5): 1048-1054.

Dukes MNG (1996). *Meyler's side effects of drugs*. Amsterdam, Excerpta Medica.

Durant JL, Leland BA, Henry DR and Nourse JG (2002). Reoptimization of MDL keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences* **42**(6): 1273-1280.

Ecker GF and Noe CR (2004). In silico prediction models for blood-brain barrier permeation. *Current Medicinal Chemistry* **11**(12): 1617-1628.

Eddershaw PJ, Beresford AP and Bayliss MK (2000). ADME/PK as part of a rational approach to drug discovery. *Drug Discovery Today* **5**(9): 409-414.

Ekins S, Berbaum J and Harrison RK (2003). Generation and validation of rapid computational filters for CYP2D6 and CYP3A4. *Drug Metabolism and Disposition* **31**(9): 1077-1080.

Ekins S, Bravi G, Binkley S, Gillespie JS, Ring BJ, Wikel JH and Wrighton SA (1999a). Three and four dimensional-quantitative structure activity relationship (3D/4D-QSAR) analyses of CYP2D6 inhibitors. *Pharmacogenetics* **9**: 477-489.

Ekins S, Bravi G, Binkley S, Gillespie JS, Ring BJ, Wikel JH and Wrighton SA (2000a). Three- and four-dimensional-quantitative structure activity relationship (3D/4D-QSAR) analyses of CYP2C9 inhibitors. *Drug Metabolism and Disposition: The Biological Fate of Chemicals* **28**(8): 994-1002.

Ekins S, Bravi G, Wikel JH and Wrighton SA (1999b). Three-dimensional-quantitative structure activity relationship analysis of cytochrome P-450 3A4 substrates. *Journal of Pharmacology and Experimental Therapeutics* **291**(1): 424-433.

Ekins S, de Groot MJ and Jones JP (2001). Pharmacophore and three-dimensional quantitative structure-activity relationship methods for modeling cytochrome P450 active sites. *Drug Metabolism and Disposition: The Biological Fate of Chemicals* **29**(7): 936-944.

Ekins S, Ring BJ, Grace J, McRobie-Belle DJ and Wrighton SA (2000b). Present and future in vitro approaches for drug metabolism. *Journal of Pharmacological and Toxicological Methods* **44**(1): 313-324.

Ekins S, Waller CL, Swaan PW, Cruciani G, Wrighton SA and Wikel JH (2000c). Progress in predicting human ADME parameters in silico. *Journal of Pharmacological and Toxicological Methods* **44**(1): 251-272.

Engkvist O, Wrede P and Rester U (2003). Prediction of CNS activity of compound libraries using substructure analysis. *Journal of Chemical Information and Computer Sciences* **43**(1): 155-160.

Erb RJ (1995). The backpropagation neural network--a Bayesian classifier. Introduction and applicability to pharmacokinetics. *Clinical Pharmacokinetics* **29**(2): 69-79.

Eriksson L, Jaworska J, Cronin M, Worth A, Gramatica P and McDowell R (2003). Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environmental Health Perspectives* **111**(10): 1361-1375.

Eriksson L, Johansson E, Kettaneh-Wold N and Wade KM (2001a). *Multi- and megavariate data analysis - Principles and applications*. Umea, Sweden, Umetrics AB.

Eriksson L, Johansson E, Kettaneh-Wold N and Wade KM (2001b). PCA. *Multi- and megavariate data analysis - Principles and applications*. Umea, Sweden, Umetrics AB**:** 43-70.

Ertl P, Rohde B and Selzer P (2000). Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *Journal of Medicinal Chemistry* **43**(20): 3714-3717.

Ester M, Kriegel HP, Sander J and Xu X (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining***:** 226-231.

Evgeniou T and Pontil M (2001). Support vector machines: theory and applications. *Machine learning and its applications. Advanced lectures*. Paliouras G, Karkaletsis V and Spyropoulos CD. New York, Springer**:** 249-257.

Farnum M, DesJarlais R and Agrafiotis DK (2003). Molecular diversity. *Handbook of chemoinformatics : From data to knowledge*. Gasteiger J. Chichester, Wiley-VCH. **4:** 1640-1686.

Feher M, Sourial E and Schmidt JM (2000). A simple model for the prediction of blood-brain partitioning. *International Journal of Pharmaceutics* **201**(2): 239-247.

Fix E and Hodges JL (1951). Discriminatory analysis: Non-parametric discrimination: Consistency properties. Texas, USAF School of Aviation Medicine, Randolph Field**:** 261-279.

Fleishaker JC, Desai N and McNamara PJ (1987). Factors affecting the milk-to-plasma drug concentration ratio in lactating women: physical interactions with protein and fat. *Journal of Pharmaceutical Sciences* **76**(3): 189-193.

Flockhart DA. (2003). Cytochrome P450 drug-interaction table. Retrieved November 2003, from http://medicine.iupui.edu/flockhart/table.htm.

Forgy E (1965). Cluster analysis of multivariate data: Efficiency vs interpretability of classifications. *Biometrics* **21**: 768-780.

Frohlich H, Chapelle O and Scholkopf B (2003). Feature selection for support vector machines by means of genetic algorithm. *Proceedings. 15th IEEE International Conference on Tools with Artificial Intelligence***:** 142-148.

Fujita T, Iwasa J and Hansch C (1964). A new substituent constants, p, derived from partition coefficients. *Journal of the American Chemical Society* **86**: 5175-5180.

Furlanello C, Serafini M, Merler S and Jurman G (2003). An accelerated procedure for recursive feature ranking on microarray data. *Neural Networks* **16**: 641-648.

Galvez J, Garcia R, Salabert MT and Soler R (1994). Charge indexes. New topological descriptors. *Journal of Chemical Information and Computer Sciences* **34**(3): 520-525.

Gao H, Lajiness MS and Van Drie J (2002). Enhancement of binary QSAR analysis by a GA-based variable selection method. *Journal of Molecular Graphics and Modelling* **20**(4): 259-268.

Geladi P and Kowalski BR (1986). Partial least squares regression: A tutorial. *Analytica Chimica Acta* **185**: 1-17.

Gillet VJ, Willett P, Fleming PJ and Green DVS (2002). Designing focused libraries using MoSELECT. *Journal of Molecular Graphics and Modelling* **20**(6): 491-498.

Glover F (1989). Tabu search - Part I. *ORSA Journal on Computing* **1**: 190-206.

Gobburu JV and Shelver WH (1995). Quantitative structure-pharmacokinetic relationships (QSPR) of beta blockers derived using neural networks. *Journal of Pharmaceutical Sciences* **84**(7): 862-865.

Golbraikh A and Tropsha A (2002). Beware of q2! *Journal of Molecular Graphics and Modelling* **20**(4): 269-276.

Gottesman MM, Pastan I and Ambudkar SV (1996). P-glycoprotein and multidrug resistance. *Current Opinion in Genetics and Development* **6**(5): 610-617.

Gramatica P, Pilutti P and Papa E (2004). Validated QSAR prediction of OH tropospheric degradation of VOCs: Splitting into training-test sets and consensus modeling. *Journal of Chemical Information and Computer Sciences* **44**(5): 1794-1802.

Gratton JA, Abraham MH, Bradbury MW and Chadha HS (1997). Molecular factors influencing drug transfer across the blood-brain barrier. *Journal of Pharmacy and Pharmacology* **49**(12): 1211-1216.

Greene N (2002). Computer systems for the prediction of toxicity: An update. *Advanced Drug Delivery Reviews* **54**(3): 417-431.

Guyon I and Elisseeff A (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research* **3**: 1157-1182.

Guyon I, Weston J, Barnhill S and Vapnik V (2002). Gene selection for cancer classification using support vector machines. *Machine Learning* **46**(1-3): 389-422.

Hadi AS (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society, Series B* **54**(3): 761-771.

Hall LH, Kellogg GE and Haney DN (2002). Molconn-Z. eduSoft, LC.

Hall LM, Hall LH and Kier LB (2003). QSAR modeling of beta-lactam binding to human serum proteins. *Journal of Computer-Aided Molecular Design* **17**(2-4): 103-118.

Hall LM, Hall LH and Kier LB (2004). Modeling drug albumin binding affinity with E-state topological structure representation. *Journal of Chemical Information and Computer Sciences* **43**(6): 2120-2128.

Hallstrom C and Lader MH (1980). Diazepam and N-desmethyldiazepam concentrations in saliva, plasma and CSF. *British Journal of Clinical Pharmacology* **9**(4): 333-339.

Han JW and Kamber M (2001). *Data mining : concepts and techniques*. San Francisco, Morgan Kaufmann Publishers.

Han LY, Cai CZ, Lo SL, Chung MCM and Chen YZ (2004). Prediction of RNA-binding proteins from primary sequence by support vector machine approach. *RNA* **10**(3): 355-368.

Hansch C (1972). Quantitative relationships between lipophilic character and drug metabolism. *Drug Metabolism Reviews* **1**: 1-14.

Hansch C, Leo A, Mekapati SB and Kurup A (2004). QSAR and ADME. *Bioorganic and Medicinal Chemistry* **12**(12): 3391-3400.

Hardman JG, Limbird LE and Goodman Gilman A (2002). *Goodman and Gilman's the pharmacological basis of therapeutics*. New York, McGraw-Hill.

Hassan M, Bielawski JP, Hempel JC and Waldman M (1996). Optimization and visualization of molecular diversity of combinatorial libraries. *Molecular Diversity* **2**(1-2): 64-74.

Hawkins DM (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences* **44**(1): 1-12.

Hawkins DM, Basak SC and Mills D (2004). Assessing model fit by cross-validation. *Journal of Chemical Information and Computer Sciences* **43**(2): 579-586.

He L, Jurs PC, Custer LL, Durham SK and Pearl GM (2003). Predicting the genotoxicity of polycyclic aromatic compounds from molecular structure with different classifiers. *Chemical Research in Toxicology* **16**(12): 1567-1580.

He YB, Wang LS, Liu ZT and Zhang Z (1995). Acute toxicity of alkyl (1-phenylsulfonyl)cycloalkane-carboxylates to Daphnia magna and quantitative structure--activity relationships. *Chemosphere* **31**(2): 2739-2746.

Hemmer MC, Steinhauer V and Gasteiger J (1999). Deriving the 3D structure of organic molecules from their infrared spectra. *Vibrational Spectroscopy* **19**(1): 151-164.

Herman RA and Veng-Pedersen P (1994). Quantitative structure-pharmacokinetic relationships for systemic drug distribution kinetics not confined to a congeneric series. *Journal of Pharmaceutical Sciences* **83**(3): 423-428.

Higo J and Go N (1989). Algorithm for rapid calculation of excluded volume of large molecules. *Journal of Computational Chemistry* **10**(3): 376-379.

Hobohm U, Scharf M, Schneider R and Sander C (1992). Selection of representative protein data sets. *Protein Science* **1**(3): 409-417.

Hou TJ and Xu XJ (2002). ADME evaluation in drug discovery. 1. Applications of genetic algorithms to the prediction of blood-brain partitioning of a large set of drugs. *Journal of Molecular Modeling* **8**(12): 337-349.

Hou TJ and Xu XJ (2003). ADME evaluation in drug discovery. 3. Modeling blood-brain barrier partitioning using simple molecular descriptors. *Journal of Chemical Information and Computer Sciences* **43**(6): 2137-2152.

Huang C, Davis LS and Townshend JRG (2002). An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing* **23**: 725-749.

Huang L, Lu HM and Dai Y (2003). Feature selection of support vector regression for quantitative structure-activity relationships (QSAR). *Proceedings of the International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences***:** 88-93.

Hudson BD, Hyde RM, Rahr E, Wood J and Osman J (1996). Parameter based methods for compound selection from chemical databases. *Quantitative Structure-Activity Relationships* **15**: 285-289.

Hudson PTW and Postma EO (1995). Choosing and using a neural net. *Artificial neural networks : an introduction to ANN theory and practice*. Thuijsman F, Weijters AJMM and Braspenning PJ. Berlin, Springer-Verlag**:** 273-287.

Ito K, Iwatsubo T, Kanamitsu S, Nakajima Y and Sugiyama Y (1998). Quantitative prediction of in vivo drug clearance and drug interactions from in vitro data on metabolism, together with binding and transport. *Annual Review of Pharmacology and Toxicology* **38**: 461-499.

Iyer M, Mishru R, Han Y and Hopfinger AJ (2002). Predicting blood-brain barrier partitioning of organic molecules using membrane-interaction QSAR analysis. *Pharmaceutical Research* **19**(11): 1611-1621.

Izrailev S and Agrafiotis DK (2004). A method for quantifying and visualizing the diversity of QSAR models. *Journal of Molecular Graphics and Modelling* **22**(4): 275-284.

Jarvis RA and Patrick EA (1973). Clustering using a similarity measure based on shared near neighbours. *IEEE Transactions in Computers* **C-22**: 1025-1034.

Joachims T (1999). Making large-Scale SVM Learning Practical. *Advances in Kernel Methods: Support Vector Learning*. Schölkopf B, Burges CJC and Smola AJ. Cambridge, MIT-Press**:** 169-184.

Johnson DE and Wolfgang GH (2000). Predicting human safety: Screening and computational approaches. *Drug Discovery Today* **5**(10): 445-454.

Johnson MA and Maggiora GM (1990). *Concepts and applications of molecular similarity*. New York, Wiley.

Johnson RA and Wichern DW (1982). *Applied multivariate statistical analysis*. Englewood Cliffs, NJ, Prentice Hall.

Jouan-Rimbaud D, Massart DL and de Noord OE (1996). Random correlation in variable selection for multivariate calibration with a genetic algorithm. *Chemometrics and Intelligent Laboratory Systems* **35**(2): 213-220.

Kaliner MA (1992). Nonsedating antihistamines: pharmacology, clinical efficacy and adverse effects. *American Family Physician* **45**(3): 1337-1342.

Kaliszan R and Markuszewski M (1996). Brain/blood distribution described by a combination of partition coefficient and molecular mass. *International Journal of Pharmaceutics* **145**(1-2): 9-16.

Kamlet MJ, Abbound J-LM and Taft RW (1981). An examination of linear solvation energy relationships. *Progress in Physical Organic Chemistry*. Taft RW. New York, Wiley. **13:** 485-630.

Kamlet MJ, Doherty PJ, Taft RW, Abraham MH, Veith GD and Abraham DJ (1987). Solubility properties in polymers and biological media. 8. An analysis of the factors that influence toxicities of organic nonelectrolytes to the golden orfe fish (*Leuciscus idus melanotus*). *Environmental Science and Technology* **21**: 149-155.

Karalis V, Tsantili-Kakoulidou A and Macheras P (2002). Multivariate statistics of disposition pharmacokinetic parameters for structurally unrelated drugs used in therapeutics. *Pharmaceutical Research* **19**(12): 1827-1834.

Karalis V, Tsantili-Kakoulidou A and Macheras P (2003). Quantitative structure-pharmacokinetic relationships for disposition parameters of cephalosporins. *European Journal of Pharmaceutical Sciences* **20**(1): 115-123.

Karelson M and et al. (1996). Quantum-chemical descriptors in QSAR/QSPR studies. *Chemical Reviews* **96**(3): 1027-1043.

Katritzky AR, Karelson M and Lobanov V (1997). QSPR as a means of predicting and understanding chemical and physical properties in terms of structure. *Pure and Applied Chemistry* **69**(2): 245-248.

Katritzky AR, Mu L, Lobanov VS and Karelson M (1996). Correlation of boiling points with molecular structure. 1. A training set of 298 diverse organics and a test set of 9 simple inorganics. *Journal of Physical Chemistry* **100**: 10400-10407.

Kaznessis YN, Snow ME and Blankley CJ (2001). Prediction of blood-brain partitioning using Monte Carlo simulations of molecules in water. *Journal of Computer-Aided Molecular Design* **15**(8): 697-708.

Kelder J, Grootenhuis PDJ, Bayada DM, Delbressine LPC and Ploemen JP (1999). Polar molecular surface area as a dominating determinant for oral absorption and brain penetration of drugs. *Pharmaceutical Research* **16**(10): 1514-1519.

Kennard RW and Stone L (1969). Computer aided design of experiments. *Technometrics* **11**: 137-148.

Kennedy T (1997). Managing the drug discovery/development interface. *Drug Discovery Today* **2**(10): 436-444.

Keseru GM (2001). A virtual high throughput screen for high affinity cytochrome P450cam substrates. Implications for in silico prediction of drug metabolism. *Journal of Computer-Aided Molecular Design* **15**(7): 649-657.

Keserü GM and Molnár L (2001). High-throughput prediction of blood-brain partitioning: a thermodynamic approach. *Journal of Chemical Information and Computer Sciences* **41**(1): 120-128.

Kier LB (1985). A shape index from molecular graphs. *Quantitative Structure-Activity Relationships* **4**: 109-116.

Kier LB (1990). Indexes of molecular shape from chemical graphs. *Computational chemical graph theory*. Rouvray DH. New York, Nova Science Publishers**:** 151-174.

Kier LB (1997). Kappa shape indices for similarity analysis. *Medicinal Chemistry Research* **7**: 394-406.

Kier LB and Hall LH (1986). *Molecular connectivity in structure-activity analysis*. Letchworth, Hertfordshire, England; New York, Research Studies Press; Wiley.

Kier LB and Hall LH (1999). *Molecular structure description: The electrotopological state*. San Diego, Academic Press.

Kim RB, Fromm MF, Wandel C, Leake B, Wood AJ, Roden DM and Wilkinson GR (1998). The drug transporter P-glycoprotein limits oral absorption and brain

entry of HIV-1 protease inhibitors. *Journal of Clinical Investigation* **101**(2): 289-294.

Kirpichenok MA and Zefirov NS (1987). Electronegativity and molecular geometry. I. General basis of the developed approach and determination of the effect of closer electrostatic interactions on bond lengths in organic molecules. *Zhurnal Organicheskoi Khimii* **23**: 673-691.

Klein C, Kaiser D, Kopp S, Chiba P and Ecker GF (2002). Similarity based SAR (SIBAR) as tool for early ADME profiling. *Journal of Computer-Aided Molecular Design* **16**(11): 785-793.

Klopman G (1992). MULTI-CASE: 1. A hierarchical computer automated structure evaluation program. *Quantitative Structure-Activity Relationships* **11**: 176-184.

Klopman G, Shi LM and Ramu A (1997). Quantitative structure-activity relationship of multidrug resistance reversal agents. *Molecular Pharmacology* **52**(2): 323-334.

Klopman G, Stefan LR and Saiakhov RD (2002). ADME evaluation. 2. A computer model for the prediction of intestinal absorption in humans. *European Journal of Pharmaceutical Sciences* **17**(4-5): 253-263.

Kohavi R and John GH (1997). Wrappers for feature subset selection. *Artificial Intelligence* **97**(1-2): 273-324.

Kononenko I (1994). *Estimating attributes: analysis and extensions of RELIEF*. Machine Learning: ECML-94. European Conference on Machine Learning. Proceedings.

Koymans LA, Vermeulen NPE, Van Acker SABE, Tekoppele JM, Heykants JJP, Lavrijsen K, Meuldermans W and Donne-Op Den Kelder GM (1992). A predictive model for substrates of cytochrome P450-debrisoquine 2D6. *Chemical Research in Toxicology* **5**(2): 211-219.

Kozak A and Kozak R (2003). Does cross validation provide additional information in the evaluation of regression models? *Canadian Journal of Forest Research* **33**(6): 976-987.

Kramer PJ (1998). Genetic toxicology. *Journal of Pharmacy and Pharmacology* **50**(4): 395-405.

Kramer SD and Wunderli-Allenspach H (2001). Physicochemical properties in pharmacokinetic lead optimization. *Farmaco* **56**(1-2): 145-148.

Kratochwil NA, Huber W, Muller F, Kansy M and Gerber PR (2002). Predicting plasma protein binding of drugs: A new approach. *Biochemical Pharmacology* **64**(9): 1355-1374.

Kubat M, Holte RC and Matwin S (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning* **30**(2-3): 195-215.

Kubinyi H (2003). Drug research: Myths, hype and reality. *Nature Reviews Drug Discovery* **2**(8): 665-668.

Kulkarni AS and Hopfinger AJ (1999). Membrane-interaction QSAR analysis: Application to the estimation of eye irritation by organic compounds. *Pharmaceutical Research* **16**(8): 1244-1252.

Labute P (2000). A widely applicable set of descriptors. *Journal of Molecular Graphics and Modelling* **18**: 464-477.

Lacy CF and et al. (2002). *Drug information handbook*. Hudson, Ohio, Lexi-Comp, Inc.

Lam RLH, Welch WJ and Young SS (2002). Uniform coverage designs for molecule selection. *Technometrics* **44**(2): 99-109.

Langowski J and Long A (2002). Computer systems for the prediction of xenobiotic metabolism. *Advanced Drug Delivery Reviews* **54**(3): 407-415.

Layton D, Key C and Shakir SA (2003). Prolongation of the QT interval and cardiac arrhythmias associated with cisapride: Limitations of the pharmacoepidemiological studies conducted and proposals for the future. *Pharmacoepidemiology and Drug Safety* **12**(1): 31-40.

Leach AR and Gillet VJ (2003). Selecting diverse sets of compounds. *An introduction to chemoinformatics*. Boston, Kluwer Academic Publisher**:** 123-145.

Leardia R and González AL (1998). Genetic algorithms applied to feature selection in PLS regression: How and when to use them. *Chemometrics and Intelligent Laboratory Systems* **41**(2): 195-207.

Lessmann S (2004). Solving unbalanced classification problems with support vector machines. *Proceedings of the International Conference on Artificial Intelligence, IC-AI'04*. **1:** 214-220.

Lewis DF, Modi S and Dickins M (2002). Structure-activity relationship for human cytochrome P450 substrates and inhibitors. *Drug Metabolism Reviews* **34**(1-2): 69-82.

Li AP (2001). Screening for human ADME/Tox drug properties in drug discovery. *Drug Discovery Today* **6**(7): 357-366.

Li AP, Kaminski DL and Rasmussen A (1995). Substrates of human hepatic cytochrome P450 3A4. *Toxicology* **104**(1-3): 1-8.

Li H, Ung CY, Yap CW, Xue Y, Li ZR, Cao ZW and Chen YZ (2005a). Prediction of genotoxicity of chemical compounds by statistical learning methods. *Chemical Research in Toxicology* **18**(6): 1071-1080.

Li ZR, Han LY and Chen YZ (2005b). MODEL - Molecular Descriptor Lab (http://jing.cz3.nus.edu.sg/cgi-bin/model/model.cgi). http://jing.cz3.nus.edu.sg/cgi-bin/model/model.cgi. Bioinformatics & Drug Design group: Singapore.

Lipinski CA, Lombardo F, Dominy BW and Feeney PJ (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews* **23**: 3-25.

Litman T, Zeuthen T, Skovsgaard T and Stein WD (1997). Structure-activity relationships of P-glycoprotein interacting drugs: kinetic characterization of their effects on ATPase activity. *Biochimica et Biophysica Acta* **1361**(2): 159-168.

Liu R, Sun H and So SS (2001). Development of quantitative structure-property relationship models for early ADME evaluation in drug discovery. 2. Blood-brain barrier penetration. *Journal of Chemical Information and Computer Sciences* **41**(6): 1623-1632.

Liu XH, Wang B, Huang Z, Han SK and Wang LS (2003). Acute toxicity and quantitative structure-activity relationships of alpha-branched phenylsulfonyl acetates to Daphnia magna. *Chemosphere* **50**(3): 403-408.

Liu Y (2004). A comparative study on feature selection methods for drug discovery. *Journal of Chemical Information and Computer Sciences* **44**(5): 1823-1828.

Livingstone DJ (1995a). *Data analysis for chemists: Applications to QSAR and chemical product design*. Oxford, Oxford University Press.

Livingstone DJ (1995b). Data pre-treatment. *Data analysis for chemists: Applications to QSAR and chemical product design*. Oxford, Oxford University Press**:** 48-64.

Livingstone DJ and Manallack DT (2003). Neural networks in 3D QSAR. *QSAR & Combinatorial Science* **22**(5): 510-518.

Livingstone DJ and Rahr E (1989). Corchop - An interactive routine for the dimension reduction of large QSAR data sets. *Quantitative Structure-Activity Relationships* **8**: 103-108.

Lobell M, Molnár L and Keserü GM (2003a). Recent advances in the prediction of blood-brain partitioning from molecular structure. *Journal of Pharmaceutical Sciences* **92**(2): 360-370.

Lobell M and Sivarajah V (2003b). In silico prediction of aqueous solubility, human plasma protein binding and volume of distribution of compounds from calculated pKa and AlogP98 values. *Molecular Diversity* **7**(1): 69-87.

Lombardo F, Blake JF and Curatolo WJ (1996). Computation of brain-blood partitioning of organic solutes via free energy calculations. *Journal of Medicinal Chemistry* **39**(24): 4750-4755.

Long A and Walker JD (2003). Quantitative structure-activity relationships for predicting metabolism and modeling cytochrome P450 enzyme activities. *Environmental Toxicology and Chemistry* **22**(8): 1894-1899.

Lowrey AH, Famini GR, Loumbev V, Wilson LY and Tosk JM (1997). Modeling drug-melanin interaction with theoretical linear solvation energy relationships. *Pigment Cell Research* **10**(5): 251-256.

Lu CT, Chen DC and Kou YF (2003). *Algorithms for spatial outlier detection*. Proceedings of the Third IEEE International Conference on Data Mining.

Luco JM (1999). Prediction of the brain-blood distribution of a large set of drugs from structurally derived descriptors using partial least-squares (PLS) modeling. *Journal of Chemical Information and Computer Sciences* **39**(2): 396-404.

Malik M and Camm AJ (2001). Evaluation of drug-induced QT interval prolongation: implications for drug approval and labelling. *Drug Safety* **24**(5): 323-351.

Manallack DT and Livingstone DJ (1999). Neural networks in drug discovery: have they lived up to their promise? *European Journal of Medicinal Chemistry* **34**(3): 195-208.

Mandagere AK and Jones B (2003). Prediction of bioavailability. *Drug bioavailability: Estimation of solubility, permeability, absorption and bioavailability*. van de Waterbeemd H, Lennernas H and Artursson P. Weinheim, Wiley-VCH. **18:** 444-460.

Manly BFJ (1997). *Randomization bootstrap and Monte Carlo methods in biology*. London, Chapman and Hall.

Maran U and Sild S (2003). QSAR modeling of genotoxicity on non-congeneric sets of organic compounds. *Artificial Intelligence Review* **20**(1-2): 13-38.

Marchant CA (1996). Prediction of rodent carcinogenicity using the DEREK system for 30 chemicals currently being tested by the National Toxicology Program. *Environmental Health Perspectives* **104S**(5): 1065-1073.

Markin RS, Murray WJ and Boxenbaum H (1988). Quantitative structure-activity study on human pharmacokinetic parameters of benzodiazepines using the graph theoretical approach. *Pharmaceutical Research* **5**(4): 201-208.

Masters T (1995). *Advanced algorithms for neural networks : a C++ sourcebook*. New York, Wiley.

Matthews BW (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta* **405**(2): 442-451.

Mattioni BE, Kauffman GW, Jurs PC, Custer LL, Durham SK and Pearl GM (2003). Predicting the genotoxicity of secondary and aromatic amines using data subsetting to generate a model ensemble. *Journal of Chemical Information and Computer Sciences* **43**(3): 949-963.

McDowell R and Jaworska J (2002). Bayesian analysis and inference from QSAR predictive model results. *SAR and QSAR in Environmental Research* **13**: 111-125.

McGowan JC (1963). Partition coefficients and biological activities. *Nature* **200**: 1317-1317.

Meltzer EO (1990). Performance effects of antihistamines. *Journal of Allergy and Clinical Immunology* **86**(4 Part 2): 613-619.

Menard PR, Mason JS, Morize I and Bauerschmidt S (1998). Chemical space metrics in diversity analysis, library design, and compound selection. *Journal of Chemical Information and Computer Sciences* **38**(6): 1204-1213.

Meskin MS and Lien EJ (1985). QSAR analysis of drug excretion into human breast milk. *Journal of Clinical and Hospital Pharmacy* **10**(3): 269-278.

Meyer D, Leischa F and Hornik K (2003). The support vector machine under test. *Neurocomputing* **55**(1-2): 169-186.

MICROMEDEX (2003a). DRUGDEX® System. MICROMEDEX, Inc.: Greenwood Village, Colorado. Edition expires 12/2003

MICROMEDEX (2003b). MICROMEDEX. MICROMEDEX, Inc.: Greenwood Village, Colorado. Edition expires 12/2003

Miller JL, Bradley EK and Teig SL (2002). Luddite: An information-theoretic library design tool. *Journal of Chemical Information and Computer Sciences* **43**(1): 47-54.

Mitchell TJ (1974). An algorithm for the construction of "D-optimal" experimental designs. *Technometrics* **16**: 203-210.

Mitchell TM (1997). *Machine learning*. New York, McGraw-Hill.

Molina LC, Belanche L and Nebot A (2002). Evaluating feature selection algorithms. *Topics in Artificial Intelligence: 5th Catalonian Conference on AI, CCIA 2002*. Escrig MT, Toledo F and Golobardes E, Springer-Verlag Heidelberg. **2504:** 216-227.

Molnar L and Keseru GM (2002). A neural network based virtual screening of cytochrome P450 3A4 inhibitors. *Bioorganic and Medicinal Chemistry Letters* **12**(3): 419-421.

Moreau G and Broto P (1980). The autocorrelation of a topological structure: A new molecular descriptor. *Nouveau Journal de Chimie* **4**: 359-360.

Moriguchi I, Hirono S, Liu Q, Nakagome I and Matsushita Y (1992). Simple method of calculating octanol/water partition coefficient. *Chemical and Pharmaceutical Bulletin* **40**(1): 127-130.

Mosier PD and Jurs PC (2002). QSAR/QSPR studies using probabilistic neural networks and generalized regression neural networks. *Journal of Chemical Information and Computer Sciences* **42**(6): 1460-1470.

Mosier PD, Jurs PC, Custer LL, Durham SK and Pearl GM (2003). Predicting the genotoxicity of thiophene derivatives from molecular structure. *Chemical Research in Toxicology* **16**(6): 721-732.

Moss AJ (1999). The QT interval and torsade de pointes. *Drug Safety* **21**(Suppl 1): 5-10.

Mount J, Ruppert J, Welch W and Jain AN (1999). IcePick: A flexible surface-based system for molecular diversity. *Journal of Medicinal Chemistry* **42**(1): 60-66.

Muzikant AL and Penland RC (2002). Models for profiling the potential QT prolongation risk of drugs. *Current Opinion in Drug Discovery and Development* **5**(1): 127-135.

Narayanan R and Gunturi SB (2005). In silico ADME modelling: prediction models for blood-brain barrier permeation using a systematic variable selection method. *Bioorganic and Medicinal Chemistry* **13**(8): 3017-3028.

Narendra PM and Fukunaga K (1977). A branch and bound algorithm for feature subset selection. *IEEE Transactions on Computers* **26**: 917-922.

Naritomi Y, Terashita S, Kimura S, Suzuki A, Kagayama A and Sugiyama Y (2001). Prediction of human hepatic clearance from in vivo animal experiments and in vitro metabolic studies with liver microsomes from animals and humans. *Drug Metabolism and Disposition* **29**(10): 1316-1324.

NCI/NIH. (2005). Developmental therapeutics program. Retrieved 5 July 2005, from http://dtp.nci.nih.gov/index.html.

Neter J, Kutner MH, Nachtsheim CJ and Wasserman W (1996). Diagnostics and remedial measures. *Applied linear statistical models*. Chicago, Irwin**:** 95-151.

Ng C, Xiao YD, Putnam W, Lum B and Tropsha A (2004). Quantitative structure-pharmacokinetic parameters relationships (QSPKR) analysis of antimicrobial

agents in humans using simulated annealing k-nearest-neighbor and partial least-square analysis methods. *Journal of Pharmaceutical Sciences* **93**(10): 2535-2544.

Nikolic S, Trinajstic N and Mihalic Z (1995). The Wiener index: Development and applications. *Croatica Chemica Acta* **68**(1): 105-129.

Niwa T (2003). Using general regression and probabilistic neural networks to predict human intestinal absorption with topological descriptors derived from two-dimensional chemical structures. *Journal of Chemical Information and Computer Sciences* **43**(1): 113-119.

Norinder U (2003). Support vector machine models in drug design: applications to drug transport processes and QSAR using simplex optimisations and variable selection. *Neurocomputing* **55**(1-2): 337-346.

Norinder U and Österberg T (2001). Theoretical calculation and prediction of drug transport processes using simple parameters and partial least squares projections to latent structures (PLS) statistics. The use of electrotopological state indices. *Journal of Pharmaceutical Sciences* **90**(8): 1076-1085.

Norinder U, Österberg T and Artursson P (1999). Theoretical calculation and prediction of intestinal absorption of drugs in humans using MolSurf parametrization and PLS statistics. *European Journal of Pharmaceutical Sciences* **8**(1): 49-56.

Norinder U, Sjöberg P and Österberg T (1998). Theoretical calculation and prediction of brain-blood partitioning of organic solutes using MolSurf parametrization and PLS statistics. *Journal of Pharmaceutical Sciences* **87**(8): 952-959.

Nuttakorn T and Boonserm K (2001). Support vector machine for Thai phoneme recognition. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **9**(6): 803-813.

Obach RS (1999). Prediction of human clearance of twenty-nine drugs from hepatic microsomal intrinsic clearance data: an examination of in vitro half-life approach and nonspecific binding to microsomes. *Drug Metabolism and Disposition* **27**(11): 1350-1359.

Obach RS, Baxter JG, Liston TE, Silber BM, Jones BC, Macintyre F, Rance DJ and Wastall P (1997). The prediction of human pharmacokinetic parameters from preclinical and in vitro metabolism data. *Journal of Pharmacology and Experimental Therapeutics* **283**(1): 46-58.

Olsson I-M, Gottfries J and Wold S (2004). D-optimal onion designs in statistical molecular design. *Chemometrics and Intelligent Laboratory Systems* **73**(1): 37-46.

Ooms F, Weber P, Carrupt PA and Testa B (2002). A simple model to predict blood-brain barrier permeation from 3D molecular fields. *Biochimica et Biophysica Acta* **1587**(2-3): 118-125.

Oprea TI and Gottfries J (1999). Toward minimalistic modeling of oral drug absorption. *Journal of Molecular Graphics and Modelling* **17**(5-6): 261-274.

Oprea TI and Gottfries J (2001). Chemography: the art of navigating in chemical space. *Journal of Combinatorial Chemistry* **3**(2): 157-166.

Osterberg T and Norinder U (2001). Prediction of drug transport processes using simple parameters and PLS statistics. The use of ACD/logP and ACD/ChemSketch descriptors. *European Journal of Pharmaceutical Sciences* **12**(3): 327-337.

Österberg T and Norinder U (2000). Theoretical calculation and prediction of P-glycoprotein-interacting drugs using MolSurf parametrization and PLS statistics. *European Journal of Pharmaceutical Sciences* **10**(4): 295-303.

Palm K, Stenberg P, Luthman K and Artursson P (1997). Polar molecular surface properties predict the intestinal absorption of drugs in humans. *Pharmaceutical Research* **14**(5): 568-571.

Pan D, Iyer M, Liu J, Li Y and Hopfinger AJ (2004). Constructing optimum blood brain barrier QSAR models using a combination of 4D-molecular similarity measures and cluster analysis. *Journal of Chemical Information and Computer Sciences* **44**(6): 2083-2098.

Pardridge WM (1998). CNS drug design based on principles of blood-brain barrier transport. *Journal of Neurochemistry* **70**: 1781-1792.

Park BK, Kitteringham NR, Powell H and Pirmohamed M (2000). Advances in molecular toxicology - Towards understanding idiosyncratic drug toxicity. *Toxicology* **153**(1-3): 39-60.

Parzen E (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics* **33**(3): 1065-1076.

Patterson DE, Cramer RD, Ferguson AM, Clark RD and Weinberger LE (1996). Neighborhood behavior: A useful concept for validation of "Molecular Diversity" descriptors. *Journal of Medicinal Chemistry* **39**(16): 3049-3059.

Pauling L and Pressman D (1945). The serological properties of simple substances. IX. Hapten inhibition of precipitation of antisera homologous to the o-, m-, and p-Azophenylarsonic acid groups. *Journal of the American Chemical Society* **67**: 1003-1012.

Pearlman RS *CONCORD User's Manual*. St. Louis, MO, Tripos Inc.

Pearlman RS and Smith KM (1999). Metric validation and the receptor-relevant subspace concept. *Journal of Chemical Information and Computer Sciences* **39**(1): 28-35.

Pelkonen O, Boobis AR, Gundert-Remy U and 1 ACBWG (2001). In vitro prediction of gastrointestinal absorption and bioavailability: an experts' meeting report. *European Journal of Clinical Pharmacology* **57**(9): 621-629.

Penzotti JE, Lamb ML, Evensen E and Grootenhuis PDJ (2002). A computational ensemble pharmacophore model for identifying substrates of P-glycoprotein. *Journal of Medicinal Chemistry* **45**(9): 1737-1740.

Perez JJ (2005). Managing molecular diversity. *Chemical Society Reviews* **34**(2): 143-152.

Pérez MAC, Sanz MB, Torres LR, Avalos RG, González MP and Díaz HG (2004). A topological sub-structural approach for predicting human intestinal absorption of drugs. *European Journal of Medicinal Chemistry* **39**(11): 905-916.

Pintore M, van de Waterbeemd H, Piclin N and Chrétien JR (2003). Prediction of oral bioavailability by adaptive fuzzy partitioning. *European Journal of Clinical Pharmacology* **38**(4): 427-431.

Platts JA, Abraham MH, Hersey A and Butina D (2000). Estimation of molecular linear free energy relationship descriptors. 4. Correlation and prediction of cell permeation. *Pharmaceutical Research* **17**(8): 1013-1018.

Platts JA, Abraham MH, Zhao YH, Hersey A, Ijaz L and Butina D (2001). Correlation and prediction of a large blood-brain distribution data set - an LFER study. *European Journal of Medicinal Chemistry* **36**(9): 719-730.

Platts JA, Butina D, Abraham MH and Hersey A (1999). Estimation of molecular free energy relation descriptors using a group contribution approach. *Journal of Chemical Information and Computer Sciences* **39**(5): 835-845.

Potter T and Matter H (1998). Random or rational design? Evaluation of diverse compound subsets from chemical structure databases. *Journal of Medicinal Chemistry* **41**(4): 478-488.

Prival MJ (2001). Evaluation of the TOPKAT system for predicting the carcinogenicity of chemicals. *Environmental and Molecular Mutagenesis* **37**(1): 55-69.

Pudil P, Novoviová J and Kittler J (1994). Floating search methods in feature selection. *Pattern Recognition Letters* **15**(11): 1119-1125.

Quillardet P and Hofnung M (1993). The SOS chromotest: A review. *Mutation Research* **297**(3): 235-279.

Quinlan JR (1993). *C4.5 : programs for machine learning*. San Mateo, Calif, Morgan Kaufmann.

Rajer-Kanduc K and Zupan JM, N. (2003). Separation of data on the training and test set for modelling: a case study for modelling of five colour properties of a white pigment. *Chemometrics and Intelligent Laboratory Systems* **65**(2): 221-229.

Randic M (1975). Graph theoretical approach to local and overall aromaticity of benzenoid hydrocarbons. *Tetrahedron* **31**(11-12): 1477-1481.

Randic M (1991). Novel graph theoretical approach to heteroatom in quantitative structure-activity relationship. *Chemometrics and Intelligent Laboratory Systems* **10**: 213-227.

Randic M (1995). Molecular profiles. Novel geometry-dependent molecular descriptors. *New Journal of Chemistry* **19**: 781-791.

Rendic S (2002). Summary of information on human CYP enzymes: human P450 metabolism data. *Drug Metabolism Reviews* **34**(1-2): 83-448.

Reunanen J (2003). Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research* **3**: 1371-1382.

Roche O, Trube G, Zuegge J, Pflimlin P, Alanine A and Schneider G (2002). A virtual screening method for prediction of the hERG potassium channel liability of compound libraries. *Chembiochem* **3**(5): 455-459.

Rockey WM and Elcock AH (2002). Progress toward virtual screening for drug side effects. *Proteins* **48**(4): 664-671.

Rohrbaugh R and Jurs PC (1987). Descriptions of molecular shape applied in studies of structure/activity and structure/property relationships. *Analytica Chimica Acta* **199**: 99-109.

Rose K, Hall LH and Kier LB (2002). Modeling blood-brain barrier partitioning using the electrotopological state. *Journal of Chemical Information and Computer Sciences* **42**(3): 651-666.

Rücker G and Rücker C (1993). Counts of all walks as atomic and molecular descriptors. *Journal of Chemical Information and Computer Sciences* **33**(5): 683-695.

Saeys Y, Degroeve S, Aeyels D, Rouze P and Van de Peer Y (2004). Feature selection for splice site prediction: a new method using EDA-based feature ranking. *BMC Bioinformatics* **5**: 64-.

Sanderson DM and Earnshaw CG (1991). Computer prediction of possible toxic action from chemical structure: The DEREK system. *Human and Experimental Toxicology* **10**(4): 261-273.

Saunders WB (2000). *Dorland's illustrated medical dictionary*. London.

Schmitt L and Tampe R (2002). Structure and mechanism of ABC transporters. *Current Opinion in Structural Biology* **12**(6): 754-760.

Schultz TW, Netzeva TI and Cronin MTD (2003). Selection of data sets for QSARs: analyses of Tetrahymena toxicity from aromatic compounds. *SAR and QSAR in Environmental Research* **14**(1): 59-81.

Schuur JH, Setzer P and Gasteiger J (1996). The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. *Journal of Chemical Information and Computer Sciences* **36**(2): 334-344.

Schwetz BA and Casciano DA (1998). Genetic toxicology: Impact on the next generation of toxicology. *Environmental and Molecular Mutagenesis* **31**(1): 1-3.

Seelig A (1998). A general pattern for substrate recognition by P-glycoprotein. *European Journal of Biochemistry* **251**: 252-261.

Segarra V, Lopez M, Ryder H and Palacios JM (1999). Prediction of drug permeability based on grid calculations. *Quantitative Structure-Activity Relationships* **18**(5): 474-481.

Seydel JK and Schaper KJ (1981). Quantitative structure-pharmacokinetic relationships and drug design. *Pharmacology and Therapeutics* **15**: 131-182.

Shankar A and Pacovský O (2004). Annie - Artificial neural network library. www.sourceforge.net/projects/annie.

Sheridan RP, SanFeliciano SG and Kearsley SK (2000). Designing targeted libraries with genetic algorithms. *Journal of Molecular Graphics and Modelling* **18**(4-5): 320-334.

Siedlecki W and Sklansky J (1989). A note on genetic algorithms for large-scale feature selection. *Pattern Recognition Letters* **10**: 335-347.

Sirius (2000). Absolv. Sirius Analytical Instruments Ltd.

Sixt S, Altschuh J and Brueggemann R (1995). Quantitative structure-toxicity relationships for 80 chlorinated compounds using quantum chemical descriptors. *Chemosphere* **30**(12): 2397-2414.

Sjoberg P (1997). MolSurf - a generator of chemical descriptors for QSAR. *Computer-assisted lead finding and optimization: current tools for medicinal chemistry*. van de Waterbeemd H, Testa B and Folkers G. Basel; Weinheim, VHCA; Wiley-VCH**:** 83-92.

Smith DA, Ackland MJ and Jones BC (1997a). Properties of cytochrome P450 isoenzymes and their substrates Part 1: active site characteristics. *Drug Discovery Today* **2**(10): 406-414.

Smith DA, Acklanda MJ and Jones BC (1997b). Properties of cytochrome P450 isoenzymes and their substrates Part 2: properties of cytochrome P450 substrates. *Drug Discovery Today* **2**(11): 479-486.

Smith DA, van de Waterbeemd H and Walker DK (2001a). High(er) throughput ADME studies. *Pharmacokinetics and metabolism in drug design*. Weinheim ; Chichester, Wiley-VCH. **13:** 133-141.

Smith DA, van de Waterbeemd H and Walker DK (2001b). *Pharmacokinetics and metabolism in drug design*. Weinheim ; Chichester, Wiley-VCH.

Smithing MP and Darvas F (1992). HazardExpert: an expert system for predicting chemical toxicity. *Food Safety Assessment*. Finlay JW, Robinson SF and Armstrong DJ. Washington, DC, American Chemical Society**:** 191-200.

Smola AJ and Scholkopf B *A tutorial on support vector regression*. NeuroCOLT2 Technical Report Series.

Snarey M, Terrett NK, Willett P and Wilton DJ (1997). Comparison of algorithms for dissimilarity-based compound selection. *Journal of Molecular Graphics and Modelling* **15**(6): 372-385.

Snyder RD and Green JW (2001). A review of the genotoxicity of marketed pharmaceuticals. *Mutation Research* **488**(2): 151-169.

Snyder RD, Pearl GS, Mandakas G, Choy WN, Goodsaid F and Rosenblum IY (2004). Assessment of the sensitivity of the computational programs DEREK, TOPKAT, and MCASE in the prediction of the genotoxicity of pharmaceutical molecules. *Environmental and Molecular Mutagenesis* **43**(3): 143-158.

Somers E, Kasparek MC and Pound J (1990). Drug regulation--the Canadian approach. *Regulatory Toxicology and Pharmacology* **12**(3): 214-223.

Somol P and Pudil P (2000). Oscillating search algorithms for feature selection. *Proceedings of the 15th International Conference on Pattern Recognition*. Barcelona. **2:** 406-409.

Somol P, Pudila P, Novoviová J and Paclí P (1999). Adaptive floating search methods in feature selection. *Pattern Recognition Letters* **20**(11-13): 1157-1163.

Sorich MJ, Miners JO, McKinnon RA, Winkler DA, Burden FR and Smith PA (2003). Comparison of linear and nonlinear classification algorithms for the prediction of drug and chemical metabolism by human UDP-glucuronosyltransferase isoforms. *Journal of Chemical Information and Computer Sciences* **43**(6): 2019-2024.

Specht DF (1990). Probabilistic neural networks. *Neural Networks* **3**(1): 109-118.

Specht DF (1991). A general regression neural network. *IEEE Transactions on Neural Networks* **2**(6): 568-576.

Stanton DT and Jurs PC (1990). Development and use of charged partial surface area structural descriptors in computer assisted quantitative structure-property relationship studies. *Analytical Chemistry* **62**: 2323-2329.

Stenberg P, Luthman K and Artursson P (2000). Virtual screening of intestinal permeability. *Journal of Controlled Release* **65**: 231-243.

Subramanian G and Kitchen DB (2003). Computational models to predict blood-brain barrier permeation and CNS activity. *Journal of Computer-Aided Molecular Design* **17**(10): 643-664.

Sun HM (2004). A universal molecular descriptor system for prediction of LogP, LogS, LogBB, and absorption. *Journal of Chemical Information and Computer Sciences* **44**(2): 748-757.

Susnow RG and Dixon SL (2003). Use of robust classification techniques for the prediction of human cytochrome P450 2D6 inhibition. *Journal of Chemical Information and Computer Sciences* **43**(4): 1308-1315.

Sutherland JJ, O'Brien LA and Weaver DF (2003a). Spline-fitting with a genetic algorithm: A method for developing classification structure-activity relationships. *Journal of Chemical Information and Computer Sciences* **43**(6): 1906-1915.

Sutherland JJ and Weaver DF (2003b). Development of quantitative structure-activity relationships and classification models for anticonvulsant activity of hydantoin analogues. *Journal of Chemical Information and Computer Sciences* **43**(3): 1028-1036.

Sutter JM and H. KJ (1993). Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection. *Microchemical Journal* **47**(1-2): 60-66.

Svetnik V, Wang T, Tong C, Liaw A, Sheridan RP and Song QH (2005). Boosting: An ensemble learning tool for compound classification and QSAR modeling. *Journal of Chemical Information and Modeling* **45**(3): 786-799.

Tabachnick BG and Fidell LS (2000). *Using multivariate statistics*. Boston, MA, Allyn and Bacon.

Taft RW (1952). Polar and steric substituent constants for aliphatic and o-benzoate groups from rates of esterification and hydrolysis of esters. *Journal of the American Chemical Society* **74**: 3120-3128.

Todeschini R and Consonni V (2000). *Handbook of molecular descriptors*. Weinheim, Wiley-VCH.

Todeschini R, Consonni V, Mauri A and Pavan M (2003). DRAGON Web version. Talete SRL: Milan.

Todeschini R, Consonni V, Mauri A and Pavan M (2005). DRAGON. Talete SRL: Milan.

Toon S and Rowland M (1983). Structure-pharmacokinetic relationships among the barbiturates in the rat. *Journal of Pharmacology and Experimental Therapeutics* **225**(3): 752-763.

Topliss JG and Edwards RP (1979). Chance factors in studies of quantitative structure-activity relationships. *Journal of Medicinal Chemistry* **22**(10): 1238-1244.

Toutain PL and Bousquet-Melou A (2004). Plasma clearance. *Journal of Veterinary Pharmacology and Therapeutics* **27**(6): 415-425.

Tropsha A, Gramatica P and Gombar VK (2003). The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR & Combinatorial Science* **22**(1): 69-77.

Trotter MWB, Buxton BF and Holden SB (2001). Support vector machines in combinatorial chemistry. *Measurement and Control* **34**(8): 235-239.

Trotter MWB and Holden SB (2003). Support vector machines for ADME property classification. *QSAR & Combinatorial Science* **22**(5): 533-548.

Turner JV, Glass BD and Agatonovic-Kustrin S (2003a). Prediction of drug bioavailability based on molecular structure. *Analytica Chimica Acta* **485**(1): 89-102.

Turner JV, Maddalena DJ and Agatonovic-Kustrin S (2004a). Bioavailability prediction based on molecular structure for a diverse series of drugs. *Pharmaceutical Research* **21**(1): 68-82.

Turner JV, Maddalena DJ and Cutler DJ (2004b). Pharmacokinetic parameter prediction from drug structure using artificial neural networks. *International Journal of Pharmaceutics* **270**(1-2): 209-219.

Turner JV, Maddalena DJ, Cutler DJ and Agatonovic-Kustrin S (2003b). Multiple pharmacokinetic parameter prediction for a series of cephalosporins. *Journal of Pharmaceutical Sciences* **92**(3): 552-559.

van de Waterbeembd H, Camenisch G, Folkers G and Raevsky OA (1996). Estimation of Caco-2 cell permeability using calculated molecular descriptors. *Quantitative Structure-Activity Relationships* **15**: 480-490.

van de Waterbeemd H and Gifford E (2003). ADMET in silico modelling: towards prediction paradise? *Nature Reviews. Drug Discovery* **2**(3): 192-204.

van de Waterbeemd H and Kansy M (1992). Hydrogen-bonding capacity and brain penetration. *Chimia* **46**: 299-303.

van Veen HW and Konings WN (1998). Structure and function of multidrug transporters. *Advances in Experimental Medicine and Biology* **456**: 145-158.

Vandenberg JI, Walker BD and Campbell TJ (2001). HERG K+ channels: friend and foe. *Trends in Pharmacological Sciences* **22**(5): 240-246.

Vapnik VN (1995). *The nature of statistical learning theory*. New York, Springer.

Vasilieva S (2002). Chromotest methodology for fundamental genetic research. *Research in Microbiology* **153**(7): 435-440.

Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW and Kopple K (2002). Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry* **45**(12): 2615-2623.

Venturoli D and Rippe B (2005). Ficoll and dextran vs. globular proteins as probes for testing glomerular permselectivity: effects of molecular size, shape, charge, and deformability. *American Journal of Physiology. Renal Physiology* **288**(4): F605-F613.

Veropoulos K (2001). Machine learning approaches to medical decision making, University of Bristol.

Viswanadhan VN, Reddy MR, Bacquet RJ and Erion MD (1993). Assessment of methods used for predicting lipophilicity: application to nucleosides and nucleoside bases. *Journal of Computational Chemistry* **14**(9): 1019-1026.

Votano JR, Parham M, Hall LH and Kier LB (2004). New predictors for several ADME/Tox properties: Aqueous solubility, human oral absorption, and Ames genotoxicity using topological descriptors. *Molecular Diversity* **8**(4): 379-391.

Wajima T, Fukumura K, Yano Y and Oguma T (2003a). Prediction of human clearance from animal data and molecular structural parameters using multivariate regression analysis. *Journal of Pharmaceutical Sciences* **91**(12): 2489-2499.

Wajima T, Fukumura K, Yano Y and Oguma T (2003b). Prediction of human pharmacokinetics from animal data and molecular structural parameters using multivariate regression analysis: Oral clearance. *Journal of Pharmaceutical Sciences* **92**(12): 2427-2440.

Waldman M, Li H and Hassan M (2000). Novel algorithms for the optimization of molecular diversity of combinatorial libraries. *Journal of Molecular Graphics and Modelling* **18**(4-5): 412-426.

Wall M (2005). GAlib. http://lancet.mit.edu/ga/. Massachusetts Institute of Technology. A C++ Library of Genetic Algorithm Components

Watari N and et al. (1988). Prediction of hepatic first-pass metabolism and plasma levels following intravenous and oral administration of barbiturates in the rabbit based on quantitative structure-pharmacokinetic relationships. *Journal of Pharmacokinetics and Biopharmaceutics* **16**(3): 279-301.

Wegner JK (2005). JOELib/JOELib2. http://www-ra.informatik.uni-tuebingen.de/software/joelib/index.html.

Wegner JK, Fröhlich H and Zell A (2004). Feature selection for descriptor based classification models. 2. Human intestinal absorption (HIA). *Journal of Chemical Information and Computer Sciences* **44**(3): 931-939.

Weisstein EW. (1999). Correlation coefficient. MathWorld - A Wolfram Web Resource. from http://mathworld.wolfram.com/CorrelationCoefficient.html.

Welstead ST (1994). *Neural network and fuzzy logic applications in C/C++*. New York, Wiley.

Wessel MD, Jurs PC, Tolan JW and Muskal SM (1998). Prediction of human intestinal absorption of drug compounds from molecular structure. *Journal of Chemical Information and Computer Sciences* **38**(4): 726-735.

White RE (2000). High-throughput screening in drug metabolism and pharmacokinetic support of drug discovery. *Annual Review of Pharmacology and Toxicology* **40**: 133-157.

Wilkinson GR (1981). Clearance approaches in pharmacology. *Pharmacological Reviews* **39**(1): 1-47.

Willett P, Barnard JM and Downs GM (1998). Chemical similarity searching. *Journal of Chemical Information and Computer Sciences* **38**(6): 983-996.

Wilson JT (1981). *Drugs in breast milk*. Sydney, ADIS Press.

Wilson LY and Famini GR (1991). Using theoretical descriptors in quantitative structure-activity relationships: some toxicological indices. *Journal of Medicinal Chemistry* **34**(5): 1668-1674.

Winkler DA and Burden FR (2004). Modelling blood-brain barrier partitioning using Bayesian neural nets. *Journal of Molecular Graphics and Modelling* **22**(6): 499-505.

Witten IH and Frank E (2005). *Data mining: Practical machine learning tools and techniques*. San Francisco, Morgan Kaufmann.

Wold S and Eriksson L (1995). Statistical validation of QSAR results. *Chemometric methods in molecular design*. van de Waterbeemd H. Weinheim; New York; Basel; Cambridge; Tokyo, VCH**:** 309-318.

Wold S, Esbensen K and Geladi P (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* **2**: 37-52.

Wu W, Walczaka B, Massarta DL, Heuerdingb S, Ernib F, Lastc IR and Prebble KA (1996). Artificial neural networks in classification of NIR spectral data: Design of the training set. *Chemometrics and Intelligent Laboratory Systems* **33**(1): 35-46.

Wythoff BJ (1993). Backpropagation neural networks. A tutorial. *Chemometrics and Intelligent Laboratory Systems* **18**(2): 115-155.

Xu L and Zhang WJ (2001). Comparison of different methods for variable selection. *Analytica Chimica Acta* **446**: 475-481.

Xue CX, Zhang RS, Liu HX, Yao XJ, Liu MC, Hu ZD and Fan BT (2004a). QSAR models for the prediction of binding affinities to human serum albumin using the heuristic method and a support vector machine. *Journal of Chemical Information and Computer Sciences*.

Xue Y, Li ZR, Yap CW, Sun LZ, Chen X and Chen YZ (2004b). Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. *Journal of Chemical Information and Computer Sciences* **44**(5): 1630-1638.

Xue Y, Yap CW, Sun LZ, Cao ZW, Wang JF and Chen YZ (2004c). Prediction of p-glycoprotein substrates by support vector machine approach. *Journal of Chemical Information and Computer Sciences* **44**(4): 1497-1505.

Yamazaki K and Kanaoka M (2004). Computational prediction of the plasma protein-binding percent of diverse pharmaceutical compounds. *Journal of Pharmaceutical Sciences* **93**(6): 1480-1494.

Yap CW, Cai CZ, Xue Y and Chen YZ (2004). Prediction of torsade-causing potential of drugs by support vector machine approach. *Toxicological Sciences* **79**(1): 170-177.

Yap CW and Chen YZ (2005a). Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. *Journal of Chemical Information and Modeling* **45**(4): 982-992.

Yap CW and Chen YZ (2005b). Quantitative structure-pharmacokinetic relationships for drug distribution properties by using general regression neural network. *Journal of Pharmaceutical Sciences* **94**(1): 153-168.

Yoshida F and Topliss JG (2000). QSAR model for drug human oral bioavailability. *Journal of Medicinal Chemistry* **43**(13): 2575-2585.

Young RC, Mitchell RC, Brown TH, Ganellin CR, Griffith R, Jones M, Rana KK, Saundesr D, Smith IR, Sore NE and Wilks TJ (1988). Development of a new physicochemical model for brain penetration and its application to the design of centrally acting H2 receptor histamine antagonists. *Journal of Medicinal Chemistry* **31**: 656-671.

Yu H, Yang J, Wang W and Han J (2003). Discovering compact and highly discriminative features or feature combinations of drug activities using support vector machines. *Proceeding of the IEEE computer society bioinformatics conference (CSB)*: 220-228.

Yu RL, Hu GR and Zhao YH (2002). Comparative study of four QSAR models of aromatic compounds to aquatic organisms. *Journal of Environmental Sciences (China)* **14**(4): 552-557.

Yuan Z and Huang BX (2004). Prediction of protein accessible surface areas by support vector regression. *Proteins* **57**(3): 558-564.

Zaknich A (1999). Efficient kernel functions for the general regression and modified pobabilistic neural networks. *Proceedings of the International Joint Conference on Neural Networks*. **2:** 1446-1449.

Zhao YH, Le J, Abraham MH, Hersey A, Eddershaw PJ, Luscombe CN, Boutina D, Beck G, Sherborne B, Cooper I and Platts JA (2001). Evaluation of human intestinal absorption data and subsequent derivation of a quantitative structure-activity relationship (QSAR) with the Abraham descriptors. *Journal of Pharmaceutical Sciences* **90**(6): 749-784.

Zmuidinavicius D, Didziapetris R, Japertas P, Avdeef A and Petrauskas A (2003). Classification structure-activity relations (C-SAR) in prediction of human intestinal absorption. *Journal of Pharmaceutical Sciences* **92**(3): 621-633.

Zuegge J, Fechner U, Roche O, Parrott NJ, Engkvist O and Schneider G (2002). A fast virtual screening filter for cytochrome P450 3A4 inhibition liability of compound libraries. *Quantitative Structure-Activity Relationships* **21**(3): 249-256.

Zuegge J, Schneider G, Coassolo P and Lave T (2001). Prediction of hepatic metabolic clearance: comparison and assessment of prediction models. *Clinical Pharmacokinetics* **40**(7): 553-563.

Appendix

Table 1: HIA+ compounds.

Acebutolol	Diazepam	Lornoxicam	Praziquantel
Acetaminophen	Diclofenac	Meloxicam	Prazosin
Acetylsalicylic acid	Dihydrocodeine	Mercaptoethane sulfonic acid	Progesterone
Acrivastine	Disulfiram	Methadone	Propiverine
Alprazolam	Ethambutol	Methotrexate	Propranolol
Alprenolol	Ethinylestradiol	Methylprednisolone	Propylthiouracil
Aminopyrine	Famciclovir	Metoprolol	Quinidine
Amoxicillin	Felodipine	Mexiletine	Recainam
Amrinone	Fenclofenac	Morphine	Saccharin
Antipyrine	Flecainide	Moxonidine	Salicylic acid
Atropine	Fluconazole	Naloxone	Sorivudine
Betaxolol	Flumazenil	Naproxen	Sotalol
Bromazepam	Fluvastatin	Nefazodone	Spironolactone
Bumetanide	Gallopamil	Nicotine	Stavudine
Bupropion	Glyburide	Nicotinic acid	Sudoxicam
Caffeine	Granisetron	Nisoldipine	Sulindac
Camazepam	Guanabenz	Nitrendipine	Sultopride
Captopril	Hydrocortisone	Nizatidine	Telmisartan
Cefadroxil	Ibuprofen	Nordiazepam	Tenidap
Cefatrizine	Imipramine	Norfloxacin	Tenoxicam
Ceftizoxime	Isoniazid	Ofloxacin	Terazosin
Cephalexin	Isoxicam	Omeprazole	Testosterone

Chloramphenicol	Isradipine	Ondansetron	Theophylline
Cicaprost	Ketoprofen	Oxatomide	Tiagabine
Cisapride	Ketorolac	Oxazepam	Timolol
Clofibrate	Labetalol	Oxprenolol	Tolbutamide
Clonidine	Lamivudine	Oxyfedrine	Tolmesoxide
Codeine	Lamotrigine	Phenglutarimide	Topiramate
Corticosterone	Lansoprazole	Phenytoin	Torasemide
Cycloserine	Levodopa	Pindolol	Toremifene
Cyproterone-acetate	Levonorgestrel	Piroxicam	Tramadol
Desipramine	Loracarbef	Piroximone	Trapidil
Dexamethasone	Lormetazepam	Practolol	

Table 2: HIA- compounds.

Acarbose	Chlorothiazide	Lactulose	Pirbuterol
Acyclovir	Cimetidine	Lincomycin	Pravastatin
Amiloride	Ciprofloxacin	Lisinopril	Raffinose
Ampicillin	Doxorubicin	Lovastatin	Ranitidine
Ascorbic-acid	Eflornithine	Mannitol	Reproterol
Atenolol	Enalapril	Metaproterenol	Ribavirin
Azithromycin	Erythromycin	Metformin	Rimiterol
Azosemide	Etoposide	Methyldopa	Streptomycin
Aztreonam	Famotidine	Metolazone	Sulfasalazine
Benazepril	Fenoterol	Mibefradil	Sulpiride
Benzylpenicillin	Furosemide	Nadolol	Sumatriptan
Bretyliumtosylate	Gabapentin	Neomycin	Terbutaline
Bromocriptine	Ganciclovir	Netivudine	Thiacetazone
Capreomycin	Gliclazide	Olsalazine	Tranexamicacid
Cefetamet-pivoxil	Guanoxan	Ouabain	
Ceftriaxone	Hydrochlorothiazide	Pafenolol	
Cefuroxime	Kanamycin	Phenoxymethylpenicillin	

Table 3: P-glycoprotein substrates.

Compound	Set	Compound	Set
Corticosterone	Training	Prazosin	Training
Doxorubicin	Training	Promazine	Training
Quinidine	Training	Ritonavir	Training
Vinblastine	Training	Tetraphenylphosphonium	Training
Acetamido-deoxypodophyllotoxin	Training	Bisantrene	Training
Fluphenazine	Training	Endosulfan	Training
Hydrocortisone	Training	Estriol	Training
Digoxin	Training	Ivermectin	Training
Dexamethasone	Training	Leupeptin	Training
Daunomycin	Training	Mithramycin	Training
HOE33342	Training	Pararosaniline	Training
GF120918-1	Training	Rapamycin	Training
Diltiazem	Training	S9788	Training
Colchicine	Training	Safingol	Training
Cyclosporin-A	Training	Phenoxazine	Training
Dibucaine	Training	Vindoline	Training
Phodamine123	Training	Epirubicin	Testing
Digitoxigenin	Training	Quinine	Testing
Staurosporine	Training	Vincristine	Testing
Isosafrole	Training	Cis-flupenthixol	Testing
Lovastatin	Training	Digitoxin	Testing
Fexofenadine	Training	Methylprednisolone	Testing
Nimodipine	Training	Idarubicin	Testing
Nelfinavir	Training	Verapamil	Testing
Methadone	Training	Pafenolol	Testing
Trifluoperazine	Training	Digoxigenin	Testing
Monensin	Training	Terfenadine	Testing

Ondansetron	Training	Spiperone	Testing
Indinavir	Training	Cinchonidine	Testing
Dexniguldipine	Training	Methylreserpate	Testing
Saquinavir	Training	Celiprolol	Testing
S-farnesylcysteine-methylester	Training	Cepharanthine	Testing
Reserpine	Training	Puromycin	Testing
LY335979	Training	Docetaxel	Testing
Mitoxantrone	Training	Mitomycin-C	Testing
Topotecan	Training	Morphine	Testing
Dipyridamole	Training	Valinomycin	Testing
Haloperidol	Training	Teniposide	Testing
Estradiol	Training	Epothilone_A	Testing
Azidopine	Training	Acebutolol	Validation
Toremifene	Training	Adriamycin	Validation
Paclitaxel	Training	Aldosterone	Validation
Thioridazine	Training	Calphostin_C	Validation
Morphine-6-glucuronide	Training	Catharantine	Validation
Nifedipine	Training	Chlorpromazine	Validation
Actinomycin_D	Training	CP100356	Validation
Cefoperazone	Training	Depredil	Validation
Triflupromazine	Training	Domperidone	Validation
Amiodarone	Training	Emetine	Validation
Cefazolin	Training	Etoposide	Validation
Cefotetan	Training	Gallopamil	Validation
Clotrimazole	Training	Hydroxyrubicin	Validation
Erythromycin	Training	k02	Validation
Flunitrazepam	Training	Losartan	Validation
Loperamide	Training	Nicardipine	Validation
Methotrexate	Training	Perphenazine	Validation

Phenobarbital	Training	Rifampicin	Validation
Phenytoin	Training	Yohimbine	Validation

Table 4: P-glycoprotein non-substrates.

Compound	Set	Compound	Set
4 (Penzotti *et al.* 2002)	Training	NSC268251	Training
NSC667558	Training	NSC606532	Training
NSC676602	Training	NSC617286	Training
NSC667532	Training	NSC639677	Training
Prednisolone	Training	NSC648403	Training
Aminodeoxy	Training	NSC666331	Training
Cortexolone	Training	NSC671400	Training
Methoxychlor	Training	NSC686028	Training
Chlorambucil	Training	S_farnesyl_cysteine	Training
NSC674570	Training	Aminocarb	Training
NSC49899	Training	Atrazine	Training
Deoxypodophyllotoxin	Training	Chaps	Training
PSC833	Training	Dialifos	Training
NSC630148	Training	Dieldrin	Training
NSC630721	Training	Leptophos	Training
3 (Penzotti *et al.* 2002)	Training	Mirex	Training
Progesterone	Training	Phosmet	Training
Aldoxycarb	Training	Systeine_methylester	Training
L767679	Training	Triforine	Training
BIBW22	Training	Trypan_blue	Training
NSC633528	Training	Vinclozolin	Training
Nigericin	Training	NSC667551	Training
NSC653278	Training	NSC676615	Training
NSC623083	Training	Epipodophyllotoxin	Training
NSC668354	Training	Deoxycorticosterone	Training
Reserpic_acid	Training	1 (Penzotti *et al.* 2002)	Testing
Fluazifop-butyl	Training	2 (Penzotti *et al.* 2002)	Testing

NSC664565	Training	Farnesol	Testing
Tamoxifen	Training	Melphalan	Testing
NSC667560	Training	Mevinphos	Testing
Cytarabine	Training	Paraquat	Testing
NSC615985	Training	Propiconazole	Testing
NSC678047	Training	NSC676593	Testing
NSC676610	Training	NSC676618	Testing
Carbaryl	Training	NSC674508	Testing
Aldicarb	Training	NSC309132	Testing
Carmustine	Training	NSC364080	Validation
Cyclophosphamide	Training	NSC630357	Validation
Epinephrine	Training	NSC667533	Validation
Fluorouracil	Training	NSC676617	Validation
Lindane	Training	NSC676616	Validation
NSC314622	Training	Podophyllotoxin	Validation
Midazolam	Training		

Table 5: Blood brain barrier penetration dataset.

Compound	Log BB	Set	Remarks
2	-0.04	Training	
4	-1.3	Outlier	(Kaznessis *et al.* 2001; Platts *et al.* 2001)
11	-1.17	Training	
12	-2.15	Outlier	(Luco 1999; Kaznessis *et al.* 2001)
13	-0.67	Training	
14	-0.66	Validation	
15	-0.12	Training	
16	-0.18	Training	
17	-1.15	Validation	
18	-1.57	Training	
19	-1.54	Training	
20	-1.12	Outlier	(Abraham *et al.* 1994; Platts *et al.* 2001)
21	-0.73	Outlier	(Abraham *et al.* 1994; Platts *et al.* 2001)
22	-0.27	Training	
23	-0.28	Training	
24	-0.46	Validation	
25	-0.24	Training	
26	-0.02	Training	
27	0.69	Validation	
28	0.44	Training	
30	0.22	Training	
33	-0.3	Validation	
34	-1.34	Validation	
35	-1.82	Training	
69	-0.16	Training	
111-trichloroethane	0.4	Training	
111-trifluoro-2-chloroethane	0.08	Training	
1-hydroxymidazolam	-0.07	Training	

22-dimethylbutane	1.04	Training	
2-methylpentane	0.97	Validation	
2-methylpropanol	-0.17	Training	
2-propanol	-0.15	Training	
3-methylhexane	0.9	Training	
3-methylpentane	1.01	Training	
4-hydroxymidazolam	-0.3	Validation	
9-OH-risperidone	-0.67	Training	
Acetone	-0.15	Training	
Acetylsalicylic acid	-0.5	Training	
Alprazolam	0.04	Validation	
Aminopyrine	0	Training	
Amitriptyline	0.89	Training	
Amobarbital	0.04	Training	
Argon	0.03	Not used	Descriptors cannot be computed
Atenolol	-1.42	Training	
Benzene	0.37	Training	
Bretazenil	-0.09	Training	
Bromperidol	1.38	Training	
Butanone	-0.08	Training	
C15	0.39	Training	
C17	1.2	Training	
C7	0.11	Training	
Caffeine	-0.05	Validation	
Carbamazepine	0	Training	
Carbamazepine epoxide	-0.34	Training	
Carmustine	-0.52	Training	
Chlorambucil	-1.7	Training	
Chlorpromazine	1.06	Validation	
Cimetidine	-1.42	Training	

Clobazam	0.35	Training	
Clonidine	0.11	Training	
Codeine	0.55	Validation	
CS2	0.6	Training	
Cyclohexane	0.92	Training	
Cyclopropane	0	Training	
Desipramine	1.2	Validation	
Desmethylclobazam	0.36	Validation	
Desmethyldesipramine	1.06	Validation	
Desmonomethylpromazine	0.59	Validation	
Di-(2-fluoroethene) ether	0.13	Training	
Diazepam	0.52	Validation	
Dichloromethane	-0.11	Training	
Didanosine	-1.3	Training	
Diethyl ether	0	Training	
Divinyl ether	0.11	Training	
Domperidone	-0.78	Training	
Enflurane	0.24	Training	
Ethanol	-0.16	Training	
Ether	0	Validation	
Ethylbenzene	0.2	Training	
Flumazenil	-0.29	Training	
Flunitrazepam	0.06	Training	
Fluphenazine	1.51	Outlier	(Platts *et al.* 2001)
Fluroxene	0.13	Training	
Haloperidol	1.34	Training	
Halothane	0.35	Training	
Heptane	0.81	Training	
Hexane	0.8	Training	
Hexobarbital	0.1	Training	

Hydroxyzine	0.39	Training	
Ibuprofen	-0.18	Training	
Icotidine	-2	Outlier	(Abraham *et al.* 1994; Lombardo *et al.* 1996)
Imipramine	1.06	Training	
Indinavir	-0.74	Training	
Indometacin	-1.26	Training	
Isoflurane	0.42	Training	
Krypton	-0.16	Not used	Descriptors cannot be computed
Lupitidine	-1.06	Training	
Mepyramine	0.49	Training	
Mesoridazine	-0.36	Training	
Methane	0.04	Training	
Methohexital	-0.06	Training	
Methoxyflurane	0.25	Training	
Methylcyclopentane	0.93	Training	
Mianserin	0.99	Training	
Midazolam	0.36	Validation	
Mirtazapine	0.53	Validation	
Morphine	-0.16	Training	
m-Xylene	0.29	Training	
Neon	0.2	Not used	Descriptors cannot be computed
Nevirapine	0	Training	
Nitrogen	0.03	Outlier	(Clark 1999; Liu *et al.* 2001)
Nitrous oxide	0.03	Training	
Nor-1-chlorpromazine	1.37	Validation	
Nor-2-chlorpromazine	0.97	Training	
Nordazepam	0.5	Training	
Northioridazine	0.75	Training	
Org 12962	1.64	Outlier	(Kaznessis *et al.* 2001; Platts *et al.* 2001)
Org 13011	0.16	Training	

Org 30526	0.39	Training	
Org 32104	0.52	Validation	
Org 34167	0	Training	
Org 4428	0.82	Training	
Org 5222	1.03	Validation	
Oxazepam	0.61	Validation	
o-Xylene	0.37	Training	
Paracetamol	-0.31	Training	
Paraxanthine	0.06	Training	
Pentane	0.76	Training	
Pentobarbital	0.12	Training	
Phenazone	-0.1	Training	
Phenserine	1	Training	
Phenylbutazone	-0.52	Training	
Phenytoin	-0.04	Training	
Physostigmine	0.079	Training	
Promazine	1.23	Training	
Propanol	-0.16	Training	
Propanone	-0.15	Validation	
Propranolol	0.64	Training	
p-Xylene	0.31	Training	
Quinidine	-0.46	Training	
Ranitidine	-1.23	Outlier	(Abraham *et al.* 1994; Platts *et al.* 2001)
Risperidone	-0.02	Validation	
RO19-4603	-0.25	Training	
Salicylic acid	-1.1	Training	
Salicyluric acid	-0.44	Training	
SB 222200	0.3	Training	
SF6	0.36	Training	
SKF101468	0.25	Validation	

SKF89124	-0.43	Training	
SKF93319	-1.3	Training	
Sulforidazine	0.18	Validation	
Teflurane	0.27	Training	
Temelastine	-1.88	Training	
Terbutylchlorambucil	1	Training	
Theobromine	-0.28	Validation	
Theophylline	-0.29	Training	
Thiopental sodium	-0.14	Training	
Thioridazine	0.24	Outlier	(Platts *et al.* 2001)
Tibolone	0.4	Training	
Tiotidine	-0.82	Training	
Toluene	0.37	Training	
Triazolam	0.74	Training	
Trichloroethene	0.34	Training	
Trichloromethane	0.29	Training	
Trifluoperazine	1.44	Training	
Valproic acid	-0.22	Training	
Verapamil	-0.7	Training	
Xenon	0.03	Not used	Descriptors cannot be computed
Y-G14	-0.3	Validation	
Y-G15	-0.06	Training	
Y-G16	-0.42	Training	
Y-G19	-1.3	Outlier	(Kaznessis *et al.* 2001; Platts *et al.* 2001)
Y-G20	-1.4	Outlier	(Kaznessis *et al.* 2001; Platts *et al.* 2001)
Zidovudine	-0.72	Training	
Zolantidine	0.14	Training	

Table 6: Human serum albumin binding dataset.

Compound	Log Khsa	Set	Remarks
Acebutolol	-0.21	Training	
Acetylsalicylic acid	-1.39	Training	
Acrivastine	-0.02	Training	
Alprenolol	0.04	Validation	
Amoxicillin	-1.21	Training	
Atenolol	-0.48	Validation	
Bumetanide	-0.03	Training	
Bupropion	-0.05	Training	
Caffeine	-0.92	Training	
Camptothecin	-0.08	Training	
Carbamazepine	-0.1	Training	
Cefalexin	-1.11	Validation	
Cefuroxime	-1.33	Training	
Cefuroxime axetil	-0.56	Training	
Chloramphenicol	-0.46	Training	
Chlorpromazine	1.1	Validation	
Chlorpropamide	-0.44	Training	
Cimetidine	-0.44	Training	
Ciprofloxacin	0.14	Training	
Clofibrate	0.27	Training	
Clonidine	-0.13	Training	
Clotrimazole	1.34	Training	
Cromoglicic acid	-1.07	Training	
Dansylglycine	-0.26	Training	
Desipramine	0.61	Training	
Digitoxin	0.13	Training	
Doxycycline	0.01	Validation	
Droperidol	0.43	Training	

Ebselen	-1.04	Not used	Descriptors cannot be computed
Estradiol	0.68	Training	
Etoposide	-0.49	Training	
Flucytosine	-1.11	Training	
Furosemide	-0.13	Training	
Fusidic acid	0.33	Training	
Glibenclamide	0.68	Training	
Hydrochlorothiazide	-0.42	Training	
Hydrocortisone	-0.4	Validation	
Imipramine	0.75	Validation	
Indometacin	0.47	Training	
Itraconazole	1.04	Training	
Ketoconazole	0.84	Training	
Ketoprofen	0.03	Training	
Labetalol	0.14	Training	
Lamotrigine	-0.13	Training	
Levofloxacin	0.14	Training	
Lidocaine	-0.23	Training	
Methotrexate	-0.77	Training	
Methylprednisolone	-0.22	Validation	
Metoprolol	-0.29	Training	
Minocycline	0.21	Training	
Nadolol	-0.4	Training	
Naproxen	0.25	Training	
Norfloxacin	0.14	Validation	
Novobiocin	0.35	Training	
Ondansetron	0.37	Training	
Oxprenolol	-0.15	Validation	
Oxyphenbutazone	-0.02	Validation	
Paracetamol	-0.81	Training	

Phenazone	-0.69	Training
Phenoxymethylpenicillin	-0.69	Validation
Phenylbutazone	0.19	Training
Phenytoin	0	Training
Pindolol	-0.13	Validation
Prazosin	0.06	Validation
Prednisolone	-0.4	Training
Procaine	-0.19	Training
Progesterone	0.59	Training
Promazine	0.92	Training
Propranolol	0.28	Training
Propylthiouracil	-0.75	Training
Quinidine	0.44	Training
Quinine	0.49	Validation
Ranitidine	-0.1	Training
Salicylic acid	-0.66	Training
Sancycline	0.21	Validation
Scopolamine	-0.34	Training
Sotalol	-0.44	Training
Sulfaphenazole	-0.21	Training
Sulfasalazine	0.56	Training
Sumatriptan	-0.05	Training
Terazosin	-0.16	Training
Terbinafine	1.17	Training
Testosterone	0.74	Validation
Tetracaine	0.32	Training
Tetracycline	-0.08	Validation
Timolol	-0.33	Training
Tolazamide	-0.42	Training
Tolbutamide	-0.22	Training

Triflupromazine	1.05	Training
Trimethoprim	-0.26	Training
Tryptophan	-0.78	Training
Verapamil	0.52	Training
Warfarin	-0.04	Training
Zidovudine	-1.02	Training

Table 7: Milk-plasma distribution dataset.

Compound	M/P	Set	Remarks
Acyclovir	2.35	Training	
Amitriptyline	1.53	Validation	
Amoxycillin	0.028	Validation	
Amphetamine	5.15	Training	
Ampicillin	0.295	Validation	
Aspirin	1.63	Training	
Astemizole	4.4	Training	
Atenolol	2.1	Validation	
Bupivacaine	0.34	Training	
Bupropion	5.545	Training	
Caffeine	0.711	Validation	
Cannabis	4.24	Training	
Carbamazepine	0.465	Training	
Carbamazepine 10,11-epoxide	0.79	Training	
Carbenicillin	0.02	Training	
Cefotaxime	0.16	Training	
Cefoxitin	0	Training	
Ceftriaxone	0.045	Training	
Cephalexin	0.012	Training	
Chloramphenicol	0.655	Training	
Chlorprothixene	1.48	Training	
Cimetidine	1.7	Training	
Ciprofloxacin	1.495	Training	
Citalopram	2.1	Training	
Clemastine	0.375	Training	
Clofazimine	1.35	Training	
Clomipramine	1.03	Training	

Clonazepam	0.33	Training
Clozapine	3.555	Training
Codeine	2.16	Validation
Cotinine	0.78	Training
Decarboetoxyloratadine	0.8	Training
Demethylcitalopram	1.75	Validation
Desipramine	0.915	Validation
Desmethyldoxepin	1.275	Validation
Diazepam	0.7	Training
Diltiazem	0.98	Training
Disopyramide	0.9	Training
Dothiepin	1.59	Validation
Dothiepsulfoxide	1.18	Training
Doxepin	1.37	Training
Doxycycline	0.34	Training
Erythromycin	0.455	Training
Ethanol	0.9	Training
Ethosuximide	0.8	Training
Flunitrazepam	0.54	Training
Fluoxetine	0.78	Validation
Gentamicin	0.44	Training
Haloperidol	0.64	Training
Ibuprofen	0	Training
Imipramine	0.76	Validation
Indomethacin	0.19	Training
Labetalol	1.7	Training
Lamotrigine	0.425	Training
Lidocaine	1.07	Training
Loratadine	1.2	Training
Lorazepam	0.205	Training

Medroxyprogesterone	0.72	Training	
Mefloquine	0.145	Training	
Mepindolol	2.6	Validation	
Methadone	0.44	Training	
Methotrexate	0.04	Training	
Methyldopa	0.265	Training	
Metoprolol	2.55	Training	
Metronidazole	0.95	Training	
Mexiletine	1.34	Training	
Mianserin	2.2	Training	
Minoxidil	0.76	Training	
Moclobemide	0.72	Training	
Morphine	2.46	Training	
Nadolol	4.6	Training	
N-desmethylsertraline	1.64	Training	
Nefopam	1.2	Training	
Nicotine	2.25	Training	
Nitrazepam	0.27	Validation	
Nitrendipine	0.35	Training	
Nitrofurantoin	2.25	Training	
Nordothiepin	0.85	Training	
Nordothiepin sulfoxide	1.86	Validation	
Norethindrone	0.19	Validation	
Norfluexetine	0.56	Not used	Erroneous compound
Norfluoxetine	0.56	Training	
Nortriptyline	1.18	Training	
Noscapine	0.29	Training	
O-desmethylvenlafaxine	3.3	Training	
Oxazepam	0.1	Training	
Oxprenolol	0.37	Training	

Paracetamol	0.88	Training
Paroxetine	0.75	Training
Penicillin G	0.315	Validation
Penicillin V	0.37	Training
Perfenazine	0.9	Training
Phenacetin	0.67	Training
Phenobarbitone	0.5	Training
Phenytoin	0.363	Training
Prednisolone	0.13	Training
Procainamide	3.2	Training
Propranolol	0.403	Training
Quazepam	4.13	Training
Quinapril	0.12	Training
Rosaramicin	0.12	Training
Roxithromycin	0.035	Training
Sertraline	1.275	Validation
Sotalol	5.4	Training
Sulfamethoxazole	0.1	Training
Sumatriptan	4.9	Training
Suprofen	0.014	Training
Temazepam	0.14	Validation
Tetracycline	0.95	Validation
Theobromine	0.82	Training
Theophylline	0.7	Training
Tiapamil	0.44	Training
Timolol	0.8	Training
Tinidazole	1.005	Training
Tolmetin	0.005	Training
Triprolidine	0.53	Training
Valproic acid	0.053	Training

Venlafaxine	3.8	Training
Verapamil	0.6	Training
Vigabatrin	1	Training
Zolpidem	0.13	Training
Zonisamide	0.93	Training
Zopiclone	0.555	Training

Table 8: CYP P450 datasets. I – inhibitor/non-inhibitor datasets. S – substrates/non-substrates datasets.

Compound	3A4-I	3A4-S	2D9-I	2D9-S	2C9-I	2C9-S
Abacavir	P-	P-	P-	P-	P-	P-
Abecarnil	P-	P-	P-	P-	P-	P-
Abiraterone	P-	P-	P-	P-	P-	P-
Acebutolol	P-	P-	P+	P-	P-	P-
Aceclofenac	P-	P-	P-	P-	P-	P+
Acenocoumarol	P-	P+	P-	P-	P+	P+
Acetanilide	P-	P-	P-	P-	P-	P-
Acetazolamide	P+	P-	P-	P-	P-	P-
Acetone	P-	P-	P-	P-	P-	P-
Acetylsalicylic acid	P-	P-	P-	P-	P-	P+
Adinazolam	P-	P+	P-	P-	P-	P-
Ajmaline	P-	P-	P+	P-	P-	P-
Albendazole	P-	P+	P-	P+	P-	P+
Alfentanil	P-	P+	P-	P-	P-	P-
Almotriptan	P+	P+	P+	P+	P-	P-
Alosetron	P-	P+	P-	P-	P-	P+
Alpidem	P-	P+	P-	P-	P-	P-
Alprazolam	P-	P+	P-	P-	P-	P-
Alprenolol	P-	P-	P+	P+	P-	P-
Ambroxol	P+	P+	P-	P-	P-	P-
Amfetamine	P-	P-	P-	P+	P-	P-
Amiflamine	P-	P-	P-	P+	P-	P-
Amifloxacin	P-	P-	P-	P-	P-	P-
Aminoglutethimide	P-	P-	P-	P-	P-	P-
Aminopyrine	P-	P+	P-	P+	P-	P+
Amiodarone	P+	P+	P+	P+	P+	P+

Amitriptyline	P+	P+	P+	P+	P+	P+
Amlodipine	P+	P+	P+	P-	P+	P-
Amodiaquine	P-	P-	P-	P+	P-	P-
Amoxapine	P-	P+	P+	P+	P-	P+
Amprenavir	P+	P+	P-	P+	P-	P+
Anastrozole	P+	P+	P-	P-	P+	P-
Aniline	P-	P-	P-	P-	P-	P-
Anthraquinone	P-	P-	P-	P-	P-	P-
Apomorphine	P-	P+	P-	P-	P-	P-
Aprepitant	P+	P+	P-	P-	P-	P-
Aprindine	P-	P-	P-	P+	P-	P-
Aranidipine	P+	P-	P-	P-	P-	P-
Argatroban	P-	P+	P-	P-	P-	P-
Aripiprazole	P-	P+	P-	P+	P-	P-
Artemisinin	P-	P+	P-	P+	P-	P-
Artesunate	P-	P-	P-	P-	P-	P-
Astemizole	P+	P+	P+	P+	P+	P-
Atamestane	P-	P-	P-	P-	P-	P-
Atazanavir	P+	P-	P-	P-	P+	P-
Atomoxetine	P-	P+	P-	P+	P-	P+
Atorvastatin	P+	P+	P+	P-	P+	P-
Atovaquone	P-	P-	P-	P-	P+	P-
Avasimibe	P+	P-	P-	P-	P-	P-
Avitriptan	P-	P+	P-	P+	P-	P-
Azacyclonol	P-	P-	P+	P-	P-	P-
Azamulin	P+	P-	P-	P-	P-	P-
Azapropazone	P-	P-	P-	P-	P+	P-
Azatadine	P-	P-	P-	P-	P-	P-
Azelastine	P+	P+	P+	P+	P+	P-

Azimilide	P-	P+	P-	P-	P-	P-
Azithromycin	P+	P-	P-	P-	P-	P-
Barnidipine	P+	P+	P+	P-	P+	P-
Beclometasone	P-	P-	P-	P-	P-	P-
Benidipine	P+	P-	P+	P-	P+	P-
Benzbromarone	P-	P-	P-	P-	P+	P-
Benzene	P-	P-	P-	P-	P-	P-
Benzfetamine	P-	P+	P-	P-	P-	P-
Benzydamine	P-	P+	P-	P+	P-	P-
Bepridil	P-	P+	P+	P-	P-	P-
Betamethasone	P+	P-	P-	P-	P-	P-
Betaxolol	P-	P-	P+	P+	P-	P-
Bexarotene	P-	P+	P-	P-	P-	P-
Bezafibrate	P-	P+	P-	P-	P-	P-
Bifluranol	P-	P-	P-	P-	P-	P-
Bifonazole	P+	P-	P-	P-	P-	P-
Biperiden	P-	P-	P+	P-	P-	P-
Bisoprolol	P-	P+	P-	P+	P-	P-
Boldenone	P-	P+	P-	P-	P-	P-
Bortezomib	P-	P+	P-	P+	P-	P+
Bosentan	P-	P+	P-	P-	P-	P+
Brinzolamide	P-	P+	P-	P-	P-	P-
Brofaromine	P-	P-	P-	P+	P-	P-
Bromazepam	P-	P+	P-	P-	P-	P-
Bromocriptine	P+	P+	P-	P-	P-	P-
Bromperidol	P-	P+	P-	P+	P-	P-
Bropirimine	P-	P-	P-	P-	P-	P-
Brotizolam	P-	P+	P-	P-	P-	P-
Budesonide	P-	P+	P-	P-	P-	P-

Budipine	P-	P-	P+	P-	P-	P-
Buflomedil	P-	P-	P-	P+	P-	P-
Bufuralol	P-	P+	P+	P+	P-	P+
Bunitrolol	P-	P-	P-	P+	P-	P-
Bupivacaine	P-	P+	P-	P+	P-	P-
Bupranolol	P-	P-	P+	P+	P-	P-
Buprenorphine	P+	P+	P+	P+	P+	P-
Bupropion	P-	P+	P+	P-	P-	P-
Buspirone	P-	P+	P-	P+	P-	P-
Busulfan	P-	P+	P-	P-	P-	P-
Caffeine	P-	P+	P-	P+	P-	P+
Calcium folinate	P+	P-	P-	P-	P-	P-
Candesartan	P-	P-	P-	P-	P+	P+
Capravirine	P-	P-	P-	P-	P-	P-
Captopril	P-	P-	P-	P+	P-	P-
Carbamazepine	P+	P+	P-	P+	P+	P+
Carbaril	P-	P+	P-	P+	P-	P+
Carbimazole	P-	P-	P-	P-	P-	P-
Carisoprodol	P-	P-	P-	P-	P-	P-
Carteolol	P-	P-	P-	P+	P-	P-
Carvedilol	P+	P+	P-	P+	P-	P+
Cathinone	P-	P-	P+	P-	P-	P-
Celecoxib	P-	P+	P+	P+	P-	P+
Cerivastatin	P+	P+	P+	P-	P+	P-
Cetirizine	P+	P-	P-	P-	P-	P-
Cevimeline	P-	P+	P-	P+	P-	P-
Chenodeoxycholic acid	P-	P+	P-	P-	P-	P-
Chloral hydrate	P-	P-	P-	P-	P-	P-
Chloramphenicol	P+	P-	P+	P-	P+	P-

Chlordiazepoxide	P-	P+	P-	P-	P-	P-
Chlormadinone	P+	P-	P-	P-	P-	P-
Chloroquine	P+	P+	P+	P+	P-	P-
Chlorphenamine	P-	P+	P+	P+	P-	P-
Chlorproguanil	P-	P-	P-	P-	P-	P-
Chlorpromazine	P+	P+	P+	P+	P+	P-
Chlorpyrifos	P+	P+	P-	P+	P-	P+
Chlorzoxazone	P+	P+	P-	P+	P-	P-
Cibenzoline	P-	P+	P-	P+	P-	P-
Ciclosporin	P+	P+	P-	P-	P-	P-
Cilnidipine	P-	P+	P-	P-	P+	P-
Cilostazol	P-	P+	P-	P-	P-	P-
Cimetidine	P+	P+	P+	P-	P+	P-
Cinnarizine	P-	P-	P-	P+	P-	P+
Ciprofibrate	P-	P+	P-	P-	P-	P-
Ciprofloxacin	P+	P-	P-	P-	P-	P-
Cisapride	P+	P+	P+	P+	P-	P+
Citalopram	P-	P+	P+	P+	P+	P-
Clarithromycin	P+	P+	P-	P-	P-	P-
Clemastine	P+	P-	P+	P-	P-	P-
Clindamycin	P+	P+	P-	P-	P-	P-
Clofazimine	P+	P-	P-	P-	P-	P-
Clofibrate	P-	P+	P-	P-	P-	P-
Clofibric acid	P-	P+	P-	P-	P-	P-
Clomethiazole	P-	P+	P-	P-	P-	P-
Clomipramine	P-	P+	P+	P+	P-	P+
Clonazepam	P-	P+	P-	P-	P-	P-
Clopidogrel	P-	P+	P-	P-	P+	P-
Clotrimazole	P+	P+	P+	P-	P+	P-

Clozapine	P+	P+	P+	P+	P+	P+
Cocaine	P+	P+	P+	P-	P-	P-
Codeine	P-	P+	P+	P+	P-	P-
Colchicine	P+	P+	P-	P-	P+	P-
Colecalciferol	P-	P-	P+	P-	P+	P-
Colestyramine	P-	P-	P-	P-	P-	P-
Corticosterone	P+	P+	P-	P-	P-	P-
Cortisol	P+	P+	P-	P-	P-	P-
Cortisone	P-	P+	P-	P-	P-	P-
Cotinine	P-	P-	P-	P-	P-	P-
Coumarin	P+	P-	P-	P-	P-	P-
Cyclobenzaprine	P-	P+	P-	P+	P-	P-
Cyclophosphamide	P+	P+	P-	P+	P-	P+
Cyproterone	P-	P+	P-	P-	P-	P-
Dacarbazine	P-	P-	P-	P-	P-	P-
Dalfopristin	P+	P-	P-	P-	P-	P-
Danazol	P+	P-	P-	P-	P-	P-
Dantrolene	P-	P+	P-	P-	P-	P-
Dapsone	P-	P+	P-	P+	P-	P+
Daunorubicin	P+	P-	P-	P-	P-	P-
Dazoxiben	P-	P-	P-	P-	P-	P-
Debrisoquine	P-	P-	P+	P+	P-	P-
Delapril	P+	P-	P-	P-	P-	P-
Delavirdine	P+	P+	P+	P+	P+	P+
Demethylcitalopram	P-	P-	P-	P+	P-	P-
Desipramine	P-	P-	P+	P+	P-	P-
Desloratadine	P+	P-	P+	P-	P+	P-
Desogestrel	P-	P-	P-	P-	P-	P+
Dexamethasone	P+	P+	P-	P-	P-	P-

Dexloxiglumide	P-	P-	P-	P-	P+	P-
Dexmedetomidine	P+	P-	P+	P-	P+	P-
Dextromethorphan	P-	P+	P+	P+	P-	P+
Dextropropoxyphene	P+	P-	P+	P-	P+	P-
Dextrorphan	P-	P+	P-	P+	P-	P-
Diazepam	P+	P+	P-	P-	P-	P+
Diclofenac	P+	P+	P-	P-	P+	P+
Dicoumarol	P-	P-	P-	P-	P+	P+
Dieldrin	P-	P-	P-	P-	P-	P-
Diethylcarbamazine	P-	P-	P-	P-	P-	P-
Diethylstilbestrol	P-	P+	P-	P-	P-	P-
Difloxacin	P-	P-	P-	P-	P-	P-
Digitoxin	P-	P+	P-	P-	P-	P-
Digoxin	P+	P-	P-	P-	P-	P-
Dihydralazine	P+	P+	P-	P-	P-	P-
Dihydrocodeine	P-	P+	P-	P+	P-	P-
Dihydroergotamine	P+	P+	P-	P-	P-	P-
Diltiazem	P+	P+	P+	P+	P+	P+
Dimethyl sulfoxide	P+	P-	P+	P-	P+	P-
Diosmin	P-	P-	P-	P-	P-	P-
Diphenhydramine	P-	P-	P+	P+	P-	P-
Dipotassium clorazepate	P-	P+	P-	P-	P-	P-
Diprafenone	P-	P-	P-	P+	P-	P-
Dirithromycin	P+	P-	P-	P-	P-	P-
Disopyramide	P-	P+	P-	P-	P-	P+
Disulfamide	P+	P-	P-	P-	P-	P-
Disulfiram	P-	P-	P-	P-	P+	P-
Ditiocarb sodium	P+	P-	P-	P-	P-	P-
Docetaxel	P+	P+	P-	P-	P-	P-

Dofetilide	P-	P+	P-	P-	P-	P-
Dolasetron	P-	P+	P-	P+	P-	P+
Domperidone	P+	P-	P-	P-	P-	P-
Donepezil	P-	P+	P-	P+	P-	P-
Dorzolamide	P-	P+	P-	P-	P-	P+
Doxepin	P-	P+	P-	P+	P-	P+
Doxorubicin	P+	P+	P+	P-	P-	P-
Doxycycline	P+	P+	P-	P-	P-	P-
Dronabinol	P-	P+	P-	P-	P-	P-
Drospirenone	P+	P+	P-	P-	P+	P-
Dutasteride	P-	P+	P-	P-	P-	P-
Ebastine	P-	P+	P-	P-	P-	P-
Ebrotidine	P+	P-	P-	P-	P-	P-
Ecabapide	P-	P+	P-	P-	P-	P-
Econazole	P+	P-	P-	P-	P-	P-
Efavirenz	P+	P+	P+	P-	P+	P-
Efonidipine	P+	P-	P-	P-	P-	P-
Eletriptan	P-	P+	P-	P-	P-	P+
Emedastine	P-	P+	P-	P-	P-	P-
Emivirine	P-	P+	P-	P-	P-	P-
Enalapril	P-	P+	P-	P-	P-	P-
Encainide	P-	P-	P-	P+	P-	P-
Enflurane	P-	P-	P-	P-	P-	P-
Enoxacin	P-	P-	P-	P-	P-	P-
Entacapone	P+	P-	P+	P-	P+	P-
Epinastine	P-	P+	P-	P+	P-	P-
Epinephrine	P-	P-	P-	P-	P+	P-
Eplerenone	P+	P+	P-	P-	P-	P-
Eprosartan	P-	P-	P-	P-	P+	P-

Ergometrine	P-	P+	P-	P-	P-	P-
Ergotamine	P+	P+	P-	P-	P-	P-
Erythromycin	P+	P+	P+	P-	P-	P-
Escitalopram	P-	P+	P-	P+	P+	P-
Estradiol	P-	P+	P-	P+	P-	P+
Estrone	P-	P+	P-	P-	P-	P+
Ethanol	P-	P+	P-	P-	P+	P-
Ethinylestradiol	P+	P+	P-	P-	P-	P-
Ethosuximide	P-	P+	P-	P-	P-	P-
Ethotoin	P-	P-	P-	P-	P-	P-
Ethylbenzene	P-	P-	P-	P-	P-	P-
Ethylmorphine	P-	P+	P-	P+	P-	P-
Etomidate	P-	P-	P-	P-	P-	P-
Etonogestrel	P-	P+	P-	P-	P-	P-
Etoperidone	P-	P+	P+	P+	P-	P+
Etoposide	P+	P+	P-	P-	P-	P-
Etoricoxib	P+	P+	P+	P+	P+	P+
Everolimus	P-	P+	P-	P-	P-	P-
Exemestane	P-	P+	P-	P-	P-	P-
Ezlopitant	P-	P+	P+	P+	P-	P-
Fadrozole	P-	P-	P-	P-	P-	P-
Famotidine	P+	P-	P-	P-	P-	P-
Felbamate	P-	P+	P-	P-	P-	P-
Felodipine	P+	P+	P+	P-	P+	P-
Fenfluramine	P-	P-	P-	P+	P-	P-
Fenofibrate	P-	P+	P-	P-	P-	P-
Fentanyl	P+	P+	P-	P-	P-	P-
Fexofenadine	P-	P+	P+	P-	P-	P-
Finasteride	P-	P+	P-	P-	P-	P-

Flecainide	P-	P-	P+	P+	P-	P-
Flosequinan	P-	P+	P-	P-	P-	P-
Flucloxacillin	P-	P+	P-	P-	P-	P-
Fluconazole	P+	P-	P-	P-	P+	P-
Flufenamic acid	P-	P-	P-	P-	P-	P-
Flunarizine	P-	P-	P-	P+	P-	P+
Flunitrazepam	P-	P+	P-	P-	P-	P+
Fluorouracil	P-	P-	P-	P-	P+	P-
Fluoxetine	P+	P+	P+	P+	P+	P+
Fluparoxan	P-	P-	P-	P-	P-	P-
Fluperlapine	P-	P-	P+	P+	P-	P-
Fluphenazine	P+	P-	P+	P+	P-	P-
Flurazepam	P-	P-	P-	P-	P-	P-
Flurbiprofen	P-	P-	P-	P-	P+	P+
Flurithromycin	P+	P-	P-	P-	P-	P-
Flutamide	P+	P+	P-	P-	P-	P-
Fluticasone	P-	P+	P-	P-	P-	P-
Fluvastatin	P+	P+	P+	P+	P+	P+
Fluvoxamine	P+	P-	P+	P+	P+	P-
Fosphenytoin	P-	P-	P-	P-	P-	P-
Frovatriptan	P-	P-	P-	P-	P-	P-
Furafylline	P-	P-	P-	P-	P-	P-
Galantamine	P-	P+	P-	P+	P-	P-
Gallopamil	P+	P+	P-	P-	P-	P-
Ganaxolone	P-	P+	P-	P-	P-	P-
Gatifloxacin	P-	P-	P-	P-	P-	P-
Gefitinib	P-	P+	P-	P-	P-	P-
Gemfibrozil	P+	P+	P-	P-	P+	P-
Gepirone	P-	P+	P-	P+	P-	P-

Gestodene	P+	P+	P-	P-	P-	P-
Glibenclamide	P+	P+	P-	P-	P+	P+
Gliclazide	P-	P-	P-	P-	P-	P+
Glimepiride	P-	P-	P-	P-	P-	P+
Glipizide	P+	P-	P-	P-	P-	P+
Glyceryl trinitrate	P-	P+	P-	P-	P-	P-
Granisetron	P-	P+	P-	P-	P-	P-
Grepafloxacin	P+	P-	P-	P-	P-	P-
Griseofulvin	P-	P-	P-	P-	P-	P-
Guanabenz	P-	P-	P-	P-	P-	P-
Guanoxan	P-	P-	P-	P+	P-	P-
Halofantrine	P-	P+	P+	P+	P-	P-
Haloperidol	P+	P+	P+	P+	P-	P-
Halothane	P-	P+	P-	P+	P-	P+
Hexobarbital	P-	P-	P-	P-	P-	P+
Homochlorcyclizine	P+	P-	P-	P-	P-	P-
Hydralazine	P+	P-	P-	P-	P-	P-
Hydrocodone	P-	P+	P-	P+	P-	P-
Hydrocortisone	P-	P+	P-	P+	P-	P-
Hydroquinidine	P-	P-	P+	P-	P-	P-
Hydroxyamfetamine	P-	P-	P-	P+	P-	P-
Hydroxychloroquine	P-	P-	P+	P-	P-	P-
Hydroxyzine	P-	P-	P+	P-	P-	P-
Ibuprofen	P-	P-	P-	P-	P+	P+
Ibutilide	P-	P-	P-	P-	P-	P-
Ifosfamide	P+	P+	P-	P-	P-	P+
Iloperidone	P-	P+	P-	P+	P-	P-
Imatinib	P+	P+	P+	P-	P+	P-
Imipramine	P-	P+	P+	P+	P-	P+

Imiquimod	P-	P+	P-	P-	P-	P-
Indinavir	P+	P+	P+	P+	P+	P+
Indometacin	P-	P-	P-	P-	P+	P+
Indoramin	P-	P-	P-	P+	P-	P-
Ipriflavone	P+	P-	P-	P-	P+	P-
Irbesartan	P+	P-	P-	P-	P+	P+
Irinotecan	P+	P+	P-	P-	P+	P-
Isbogrel	P-	P-	P-	P-	P-	P-
Isoconazole	P-	P-	P-	P-	P-	P-
Isoflurane	P-	P-	P-	P-	P-	P-
Isoniazid	P+	P-	P+	P-	P+	P-
Isosorbide dinitrate	P-	P+	P-	P-	P-	P-
Isradipine	P-	P+	P-	P-	P+	P-
Itraconazole	P+	P+	P-	P-	P+	P-
Ivermectin	P-	P+	P-	P-	P-	P-
Josamycin	P+	P-	P-	P-	P-	P-
Ketamine	P-	P+	P-	P+	P-	P+
Ketoconazole	P+	P+	P+	P-	P+	P-
Ketoprofen	P-	P-	P-	P-	P+	P-
Labetalol	P-	P-	P+	P+	P-	P-
Lacidipine	P-	P+	P-	P-	P-	P-
Lansoprazole	P+	P+	P+	P-	P+	P+
Leflunomide	P-	P+	P-	P-	P+	P-
Lercanidipine	P+	P+	P+	P-	P-	P-
Letrozole	P-	P+	P-	P-	P-	P-
Levacetylmethadol	P-	P+	P-	P+	P-	P+
Levobupivacaine	P-	P+	P-	P-	P-	P-
Levofloxacin	P-	P-	P-	P-	P-	P-
Levomepromazine	P-	P-	P+	P-	P-	P-

Levonorgestrel	P-	P+	P-	P-	P-	P-
Levothyroxine sodium	P-	P+	P-	P-	P-	P-
Liarozole	P-	P-	P-	P-	P-	P-
Lidocaine	P-	P+	P+	P+	P-	P+
Lilopristone	P+	P+	P-	P-	P-	P-
Linezolid	P-	P-	P-	P-	P-	P-
Liothyronine	P-	P-	P-	P-	P-	P-
Lisofylline	P-	P+	P-	P-	P-	P-
Lisuride	P-	P+	P-	P+	P-	P-
Litoxetine	P-	P-	P-	P-	P-	P-
Lobeline	P-	P-	P+	P-	P-	P-
Lomefloxacin	P-	P-	P-	P-	P-	P-
Lomustine	P+	P-	P+	P-	P-	P-
Lopinavir	P+	P+	P+	P-	P+	P-
Loratadine	P+	P+	P+	P+	P+	P-
Lornoxicam	P-	P-	P-	P-	P+	P+
Losartan	P+	P+	P-	P-	P+	P+
Losigamone	P-	P-	P-	P-	P-	P-
Lovastatin	P-	P+	P+	P-	P+	P-
Lumefantrine	P-	P+	P-	P-	P-	P-
Malathion	P+	P-	P-	P-	P-	P-
Manidipine	P+	P+	P+	P-	P+	P-
Maprotiline	P-	P-	P+	P+	P-	P-
m-Chlorophenylpiperazine	P-	P-	P-	P+	P-	P-
Mebendazole	P-	P-	P-	P-	P-	P-
Medazepam	P-	P-	P-	P-	P-	P-
Medifoxamine	P-	P+	P-	P-	P-	P-
Medroxyprogesterone	P-	P+	P-	P-	P-	P-
Mefenamic acid	P-	P-	P-	P-	P+	P+

Mefloquine	P+	P+	P-	P-	P-	P-
Meloxicam	P-	P+	P-	P-	P+	P+
Mepacrine	P-	P-	P+	P-	P-	P-
Mephenytoin	P-	P-	P-	P-	P+	P+
Mepyramine	P-	P-	P+	P-	P-	P-
Mequitazine	P+	P-	P-	P+	P-	P-
Mestranol	P-	P-	P-	P-	P-	P+
Metamfetamine	P-	P-	P-	P+	P-	P-
Methadone	P+	P+	P+	P+	P-	P+
Methaqualone	P-	P+	P-	P-	P-	P-
Methiocarb	P-	P+	P-	P+	P-	P-
Methomyl	P-	P-	P-	P-	P-	P-
Methoxsalen	P+	P-	P+	P-	P+	P-
Methoxyflurane	P-	P+	P-	P-	P-	P-
Methylergometrine	P-	P+	P-	P-	P-	P-
Methylphenidate	P-	P+	P+	P-	P-	P-
Methylphenobarbital	P-	P-	P-	P-	P-	P-
Methylprednisolone	P+	P+	P-	P-	P-	P-
Metoclopramide	P+	P+	P+	P+	P-	P-
Metoprolol	P-	P-	P+	P+	P-	P-
Metronidazole	P+	P-	P-	P-	P+	P+
Metyrapone	P+	P-	P-	P-	P-	P-
Mevastatin	P-	P-	P-	P-	P-	P-
Mexazolam	P-	P+	P-	P-	P-	P-
Mexiletine	P-	P-	P-	P+	P-	P-
Mianserin	P-	P+	P+	P+	P-	P+
Mibefradil	P+	P+	P+	P-	P-	P-
Miconazole	P+	P+	P+	P-	P+	P-
Midazolam	P+	P+	P+	P-	P+	P+

Midecamycin	P+	P-	P-	P-	P-	P-
Mifepristone	P+	P+	P+	P-	P-	P-
Milameline	P-	P-	P-	P+	P-	P-
Milnacipran	P-	P-	P-	P-	P-	P-
Minaprine	P-	P-	P-	P+	P-	P-
Minoxidil	P-	P-	P-	P-	P-	P-
Mirtazapine	P+	P+	P+	P+	P-	P+
Mitoxantrone	P+	P-	P-	P-	P-	P-
Mizolastine	P+	P+	P+	P+	P+	P-
Moclobemide	P-	P-	P+	P-	P+	P-
Modafinil	P-	P+	P-	P-	P+	P-
Molindone	P-	P-	P-	P+	P-	P-
Montelukast	P-	P+	P-	P-	P-	P+
Moracizine	P-	P-	P-	P-	P-	P-
Morphine	P-	P+	P-	P+	P-	P+
Mosapride	P-	P+	P-	P-	P-	P-
Nafcillin	P-	P-	P-	P-	P-	P-
Nalidixic acid	P-	P-	P-	P-	P-	P-
Naproxen	P-	P-	P-	P-	P-	P+
Nateglinide	P-	P+	P-	P-	P+	P+
Nefazodone	P+	P+	P+	P-	P-	P-
Nefiracetam	P-	P+	P-	P-	P-	P-
Nelfinavir	P+	P+	P+	P+	P+	P+
Nevirapine	P+	P+	P-	P+	P-	P+
Nicardipine	P+	P+	P+	P+	P+	P-
Niclosamide	P-	P-	P+	P-	P-	P-
Nicotine	P-	P+	P-	P+	P-	P+
Nifedipine	P+	P+	P+	P+	P+	P-
Niludipine	P-	P+	P-	P-	P-	P-

Nilutamide	P-	P-	P-	P-	P-	P-
Nilvadipine	P+	P+	P-	P-	P-	P-
Nimodipine	P+	P+	P-	P-	P+	P-
Nimorazole	P-	P-	P-	P-	P-	P-
Nisoldipine	P+	P+	P-	P-	P-	P-
Nitrendipine	P+	P+	P-	P-	P+	P-
Nitrosamine	P-	P-	P-	P-	P-	P-
Norcodeine	P-	P-	P-	P+	P-	P-
Nordazepam	P-	P+	P-	P-	P-	P-
Norethisterone	P-	P-	P-	P-	P-	P-
Norfloxacin	P+	P-	P-	P-	P-	P-
Norfluoxetine	P+	P-	P+	P+	P+	P-
Nortriptyline	P-	P+	P+	P+	P-	P-
Ofloxacin	P-	P-	P-	P-	P-	P-
Olanzapine	P+	P-	P+	P+	P+	P-
Olopatadine	P-	P+	P-	P-	P-	P-
Oltipraz	P+	P-	P-	P-	P-	P-
Omapatrilat	P-	P-	P-	P-	P-	P-
Omeprazole	P+	P+	P+	P+	P+	P+
Onapristone	P+	P+	P-	P-	P-	P-
Ondansetron	P-	P+	P+	P+	P+	P+
Opipramol	P-	P-	P-	P+	P-	P-
Orphenadrine	P+	P+	P+	P+	P+	P-
Oxamniquine	P-	P-	P+	P-	P-	P-
Oxcarbazepine	P-	P-	P-	P-	P-	P-
Oxiconazole	P+	P-	P-	P-	P-	P-
Oxodipine	P-	P+	P-	P-	P-	P-
Oxomemazine	P-	P-	P-	P-	P-	P-
Oxprenolol	P-	P-	P+	P-	P-	P-

Oxybutynin	P+	P+	P+	P-	P-	P-
Oxycodone	P-	P-	P-	P+	P-	P-
Paclitaxel	P+	P+	P-	P-	P-	P-
Pantoprazole	P+	P+	P-	P-	P-	P-
Papaverine	P+	P-	P-	P+	P-	P-
Paracetamol	P+	P+	P-	P+	P-	P+
Parathion	P+	P+	P-	P+	P+	P-
Paraxanthine	P-	P-	P-	P-	P-	P-
Parecoxib	P-	P-	P-	P-	P+	P-
Pargyline	P-	P-	P+	P-	P-	P-
Paroxetine	P+	P-	P+	P+	P+	P-
Pefloxacin	P-	P-	P-	P-	P-	P-
Penbutolol	P-	P-	P-	P+	P-	P-
Pentamidine	P-	P-	P+	P-	P-	P-
Pentazocine	P-	P-	P-	P+	P-	P-
Pentobarbital	P-	P-	P-	P-	P-	P-
Pentoxifylline	P-	P-	P-	P-	P-	P-
Pergolide	P+	P-	P+	P-	P-	P-
Perhexiline	P-	P-	P+	P+	P-	P-
Perospirone	P-	P+	P-	P+	P-	P-
Perphenazine	P+	P+	P+	P+	P+	P+
Pethidine	P-	P-	P-	P+	P-	P-
Phenacetin	P-	P+	P-	P+	P-	P+
Phenazone	P-	P+	P-	P-	P-	P+
Phencyclidine	P+	P+	P-	P-	P-	P-
Phenformin	P-	P-	P-	P+	P-	P-
Phenobarbital	P-	P-	P-	P-	P-	P+
Phenol	P-	P-	P-	P-	P-	P-
Phenprocoumon	P-	P-	P-	P-	P+	P+

Phensuximide	P-	P-	P-	P-	P-	P-
Phenylbutazone	P-	P-	P-	P-	P+	P+
Phenytoin	P-	P-	P-	P+	P+	P+
Pilocarpine	P+	P-	P-	P-	P-	P-
Pilsicainide	P+	P-	P-	P-	P-	P-
Pimobendan	P-	P+	P-	P-	P-	P-
Pimozide	P+	P+	P+	P-	P-	P-
Pinacidil	P-	P+	P-	P+	P-	P-
Pindolol	P-	P-	P+	P+	P-	P-
Pioglitazone	P+	P+	P-	P-	P+	P-
Pipemidic acid	P-	P-	P-	P-	P-	P-
Piroxicam	P+	P-	P-	P-	P+	P+
Plomestane	P-	P-	P-	P-	P-	P-
Pramipexole	P-	P-	P+	P-	P-	P-
Pranidipine	P-	P+	P-	P-	P-	P-
Prasterone	P-	P+	P-	P-	P-	P-
Pravastatin	P+	P+	P+	P-	P+	P-
Praziquantel	P-	P+	P+	P+	P-	P+
Prednisolone	P+	P+	P-	P-	P-	P-
Prednisone	P+	P+	P-	P-	P-	P-
Pregnenolone	P-	P+	P-	P+	P-	P-
Primaquine	P+	P+	P+	P+	P-	P-
Primidone	P-	P-	P-	P-	P-	P-
Proadifen	P+	P-	P+	P-	P+	P-
Probenecid	P-	P-	P-	P-	P+	P-
Procainamide	P-	P-	P-	P+	P-	P-
Progesterone	P-	P+	P-	P+	P+	P+
Proguanil	P-	P+	P-	P-	P-	P+
Promazine	P-	P+	P-	P+	P-	P+

Promethazine	P-	P-	P+	P+	P-	P-
Propafenone	P-	P+	P+	P+	P-	P-
Propanol	P-	P-	P-	P-	P+	P-
Propofol	P+	P+	P+	P+	P+	P+
Propranolol	P-	P+	P+	P+	P-	P-
Pyrantel	P-	P-	P-	P+	P-	P+
Pyridostigmine bromide	P-	P-	P-	P-	P-	P-
Pyrimethamine	P-	P-	P+	P-	P+	P-
Quercetin	P+	P+	P+	P-	P+	P-
Quetiapine	P-	P+	P-	P+	P+	P+
Quinelorane	P+	P-	P-	P-	P-	P-
Quinine	P+	P+	P+	P-	P-	P-
Quinupristin	P+	P-	P-	P-	P-	P-
Rabeprazole	P+	P+	P-	P-	P+	P-
Raloxifene	P+	P-	P-	P-	P-	P-
Ranitidine	P+	P-	P+	P+	P-	P-
Ranolazine	P-	P+	P-	P-	P-	P-
Rebamipide	P-	P+	P-	P-	P+	P-
Reboxetine	P+	P+	P+	P-	P-	P-
Remacemide	P+	P-	P-	P-	P-	P-
Remoxipride	P-	P-	P-	P+	P-	P-
Repaglinide	P-	P+	P-	P-	P-	P+
Reserpine	P+	P-	P-	P-	P-	P-
Resiquimod	P-	P+	P-	P-	P-	P-
Rifabutin	P-	P+	P-	P-	P-	P-
Rifampicin	P-	P+	P-	P-	P-	P+
Rifamycin	P-	P-	P-	P-	P-	P-
Rifapentine	P-	P-	P-	P-	P-	P-
Riluzole	P-	P-	P+	P-	P-	P-

Risperidone	P+	P+	P+	P+	P-	P-
Ritonavir	P+	P+	P+	P+	P+	P+
Rofecoxib	P-	P-	P-	P-	P-	P-
Rogletimide	P-	P-	P-	P-	P-	P-
Rokitamycin	P+	P-	P-	P-	P-	P-
Ropinirole	P-	P+	P+	P-	P-	P-
Ropivacaine	P-	P+	P-	P+	P-	P-
Roquinimex	P-	P+	P-	P-	P-	P-
Rosiglitazone	P+	P-	P+	P-	P+	P+
Rosuvastatin	P-	P-	P-	P-	P-	P+
Roxatidine	P-	P-	P-	P-	P-	P-
Roxithromycin	P+	P+	P-	P-	P-	P-
Rupatadine	P-	P+	P-	P-	P-	P-
Salbutamol	P+	P-	P-	P-	P-	P-
Salicylic acid	P-	P+	P-	P-	P-	P-
Salmeterol	P-	P+	P-	P-	P-	P-
Saquinavir	P+	P+	P+	P+	P+	P-
Secobarbital	P-	P-	P-	P-	P-	P-
Selegiline	P+	P+	P+	P+	P+	P+
Seratrodast	P+	P+	P+	P-	P+	P+
Sertindole	P+	P+	P+	P+	P-	P-
Sertraline	P+	P+	P+	P+	P+	P+
Sevoflurane	P-	P+	P-	P-	P-	P-
Sibutramine	P-	P+	P-	P-	P-	P-
Sildenafil	P+	P+	P-	P+	P+	P+
Simvastatin	P+	P+	P+	P-	P+	P-
Sirolimus	P+	P+	P-	P-	P-	P-
Sparteine	P-	P-	P+	P+	P-	P-
Spironolactone	P+	P-	P-	P-	P-	P-

Stiripentol	P+	P-	P+	P-	P+	P-
Styrene	P-	P-	P-	P-	P-	P-
Sufentanil	P-	P+	P-	P-	P-	P-
Sulconazole	P+	P-	P+	P-	P+	P-
Sulfadiazine	P-	P+	P-	P-	P+	P+
Sulfadimethoxine	P-	P-	P-	P-	P+	P-
Sulfadimidine	P-	P-	P-	P-	P+	P-
Sulfadoxine	P-	P-	P-	P-	P+	P-
Sulfafurazole	P-	P-	P-	P-	P+	P-
Sulfamerazine	P-	P-	P-	P-	P+	P-
Sulfamethizole	P+	P-	P-	P-	P+	P-
Sulfamethoxazole	P+	P+	P-	P-	P+	P+
Sulfamoxole	P-	P-	P-	P-	P+	P-
Sulfanilamide	P-	P-	P-	P-	P+	P-
Sulfaphenazole	P+	P-	P-	P-	P+	P-
Sulfapyridine	P-	P-	P-	P-	P+	P-
Sulfasalazine	P-	P-	P-	P-	P-	P-
Sulfatroxazole	P-	P-	P-	P-	P+	P-
Sulfinpyrazone	P+	P+	P-	P-	P+	P+
Sulindac	P-	P-	P-	P-	P+	P-
Sulpiride	P+	P-	P-	P-	P-	P-
Suprofen	P-	P-	P-	P-	P+	P+
Tacrine	P-	P-	P-	P-	P-	P-
Tacrolimus	P+	P+	P-	P-	P-	P-
Tadalafil	P-	P+	P-	P-	P-	P-
Tamoxifen	P+	P+	P+	P+	P+	P+
Tamsulosin	P-	P+	P-	P+	P-	P-
Tasosartan	P-	P+	P-	P-	P-	P-
Tauromustine	P-	P+	P-	P+	P-	P+

Tazanolast	P+	P-	P-	P-	P-	P-
Tazofelone	P-	P+	P-	P-	P-	P-
Tecastemizole	P+	P+	P-	P-	P-	P-
Tegafur	P-	P-	P-	P-	P-	P+
Tegaserod	P+	P-	P+	P+	P+	P-
Telmisartan	P-	P-	P-	P-	P-	P-
Temazepam	P-	P+	P-	P-	P-	P+
Teniposide	P+	P+	P-	P-	P+	P-
Tenofovir	P-	P-	P-	P-	P-	P-
Tenoxicam	P-	P-	P-	P-	P+	P+
Terbinafine	P-	P+	P+	P-	P-	P+
Terfenadine	P+	P+	P+	P+	P+	P-
Terguride	P-	P+	P-	P+	P-	P-
Testolactone	P-	P-	P-	P-	P-	P-
Testosterone	P+	P+	P-	P+	P-	P+
Tetracycline	P+	P-	P-	P-	P-	P-
Tezosentan	P-	P-	P-	P-	P+	P-
Thalidomide	P-	P-	P-	P-	P-	P-
Theobromine	P-	P-	P-	P-	P-	P-
Theophylline	P+	P+	P-	P+	P-	P+
Thiamazole	P+	P-	P+	P-	P+	P-
Thioridazine	P-	P-	P+	P+	P-	P-
Thiotepa	P-	P-	P-	P-	P-	P-
Tiabendazole	P-	P-	P-	P-	P+	P-
Tiagabine	P-	P+	P-	P+	P-	P-
Tiaramide	P-	P-	P-	P-	P-	P+
Ticlopidine	P+	P+	P+	P-	P+	P-
Tienilic acid	P-	P-	P-	P-	P+	P+
Timolol	P-	P-	P+	P+	P-	P-

Timoprazole	P+	P-	P-	P-	P-	P-
Tinidazole	P-	P+	P-	P-	P-	P-
Tioconazole	P+	P-	P+	P-	P+	P-
Tiotixene	P-	P-	P-	P-	P-	P-
Tipranavir	P-	P+	P-	P-	P-	P-
Tirilazad	P-	P+	P-	P-	P-	P-
Tocainide	P-	P-	P-	P-	P-	P-
Tolbutamide	P-	P-	P-	P-	P+	P+
Tolcapone	P-	P+	P-	P-	P+	P-
Tolperisone	P-	P-	P+	P+	P-	P-
Tolterodine	P-	P+	P-	P+	P-	P+
Toluene	P-	P-	P-	P-	P-	P-
Topiramate	P-	P-	P-	P-	P-	P-
Torasemide	P-	P-	P-	P-	P-	P+
Toremifene	P-	P+	P-	P-	P-	P+
Tramadol	P-	P+	P-	P+	P-	P-
Tranylcypromine	P-	P-	P-	P-	P-	P-
Trapidil	P-	P-	P-	P-	P-	P-
Trazodone	P-	P+	P-	P+	P-	P-
Tretinoin	P-	P+	P-	P-	P-	P-
Triazolam	P-	P+	P-	P-	P-	P-
Trichloroethylene	P-	P+	P-	P-	P-	P-
Trichloromethane	P-	P-	P-	P+	P-	P+
Trifluoperazine	P-	P-	P-	P-	P-	P-
Trifluperidol	P-	P-	P-	P+	P-	P-
Trimetazidine	P-	P-	P-	P-	P-	P-
Trimethadione	P-	P+	P-	P-	P-	P+
Trimethoprim	P+	P+	P+	P-	P+	P+
Trimetrexate	P-	P+	P-	P-	P-	P-

Trimipramine	P-	P+	P+	P+	P-	P-
Tripelennamine	P-	P-	P+	P-	P-	P-
Triprolidine	P-	P-	P+	P-	P-	P-
Trofosfamide	P-	P+	P-	P-	P-	P-
Troglitazone	P+	P+	P+	P+	P+	P+
Troleandomycin	P+	P+	P-	P-	P-	P-
Tropisetron	P-	P+	P+	P+	P-	P+
Trospium chloride	P-	P-	P+	P-	P-	P-
Valdecoxib	P+	P+	P+	P+	P+	P+
Valproic acid	P+	P-	P+	P-	P+	P+
Valsartan	P-	P-	P-	P-	P+	P-
Valspodar	P+	P+	P-	P-	P-	P-
Vanoxerine	P-	P+	P-	P-	P-	P-
Venlafaxine	P+	P+	P+	P+	P-	P+
Verapamil	P+	P+	P+	P-	P+	P+
Vesnarinone	P-	P+	P-	P-	P-	P-
Vinblastine	P+	P+	P+	P-	P-	P-
Vincristine	P+	P+	P+	P-	P-	P-
Vindesine	P+	P+	P-	P-	P-	P-
Vinorelbine	P+	P+	P+	P+	P-	P-
Voriconazole	P+	P+	P-	P-	P+	P+
Vorozole	P-	P-	P-	P-	P-	P-
Warfarin	P-	P+	P-	P+	P+	P+
Yohimbine	P-	P+	P+	P+	P-	P-
Zafirlukast	P+	P-	P+	P-	P+	P+
Zaleplon	P-	P+	P-	P-	P-	P-
Zaltoprofen	P+	P-	P-	P-	P+	P+
Zatosetron	P-	P+	P-	P-	P-	P-
Zidovudine	P-	P+	P-	P-	P-	P+

Zileuton	P+	P+	P+	P-	P+	P+
Ziprasidone	P+	P+	P+	P-	P-	P-
Zolmitriptan	P-	P-	P-	P-	P-	P-
Zolpidem	P+	P+	P+	P+	P+	P+
Zonisamide	P-	P+	P-	P-	P-	P-
Zopiclone	P-	P+	P-	P-	P-	P+
Zotepine	P-	P+	P-	P+	P-	P-
Zoxazolamine	P-	P-	P-	P-	P-	P-
Zuclopenthixol	P-	P-	P-	P+	P-	P-

Table 9: Total clearance dataset.

Compound	Total clearance (ml/min/kg)	Set
Chlorpropamide	0.04	Training
Droxicam	0.05	Training
Benoxaprofen	0.07	Training
Tenidap	0.09	Training
Meloxicam	0.11	Training
Azapropazone	0.14	Training
Dutasteride	0.14	Training
Nordazepam	0.17	Training
Ethosuximide	0.19	Training
Delorazepam	0.21	Training
Ceftriaxone	0.23	Training
Sulfasalazine	0.24	Training
Liothyronine	0.25	Training
Cefpiramide	0.28	Training
Sulfafurazole	0.28	Training
Tolbutamide	0.29	Training
Fluconazole	0.30	Training
Sulfamethoxazole	0.31	Validation
Olmesartan	0.32	Training
Cefonicid	0.33	Training
Amobarbital	0.35	Training
Benazeprilat	0.35	Training
Flurbiprofen	0.35	Training
Topiramate	0.37	Training
rhein	0.38	Training
Ertapenem	0.40	Training
Secnidazole	0.40	Training

Mefloquine	0.41	Training
Levocabastine	0.43	Training
Borocaptate	0.43	Training
Fosinoprilat	0.46	Training
Phenazone	0.46	Training
Phenytoin	0.46	Training
Mecillinam	0.48	Training
Ketorolac	0.50	Training
Lamotrigine	0.51	Training
Doxycycline	0.52	Training
Glipizide	0.52	Training
Ceforanide	0.56	Training
Diazepam	0.56	Training
Rufloxacin	0.57	Training
Rosiglitazone	0.58	Training
Tinidazole	0.58	Training
Raltitrexed	0.59	Validation
Brodimoprim	0.60	Training
Dapsone	0.60	Training
Tamsulosin	0.62	Training
Cefotetan	0.63	Validation
Nevirapine	0.63	Training
Etoricoxib	0.70	Training
Cefodizime	0.71	Training
Torasemide	0.71	Training
Theophylline	0.73	Training
Vigabatrin	0.74	Training
Cetirizine	0.75	Training
Glimepiride	0.76	Validation

Acivicin	0.78	Training
Bromazepam	0.82	Training
Sulfinpyrazone	0.82	Training
Cefazolin	0.84	Training
Alprazolam	0.86	Training
Modafinil	0.88	Training
Minocycline	0.89	Training
Trimetrexate	0.89	Training
Tiaprofenic acid	0.91	Validation
Trimazosin	0.94	Training
Terodiline	0.95	Training
Cefoperazone	0.96	Training
Levetiracetam	0.96	Training
Procyclidine	0.97	Training
Tegafur	0.98	Training
Prednisolone	1.00	Training
Chlortalidone	1.02	Training
Aprepitant	1.07	Training
Oxazepam	1.08	Training
Edetate	1.09	Training
Lorazepam	1.10	Validation
Terazosin	1.10	Training
Carbamazepine	1.11	Training
Ibuprofen	1.12	Training
Mizolastine	1.15	Training
Iohexol	1.17	Training
Carbenicillin	1.18	Training
Pidotimod	1.19	Training
Ketoprofen	1.20	Training

Pemetrexed	1.20	Training
Toremifene	1.21	Validation
Alendronic acid	1.22	Validation
Ifosfamide	1.23	Training
Metronidazole	1.26	Training
Temazepam	1.27	Training
Azosemide	1.29	Training
Cefixime	1.30	Training
Glibenclamide	1.30	Training
Piracetam	1.30	Training
Tolmetin	1.30	Training
Latamoxef	1.32	Training
Iobitridol	1.33	Validation
Zoledronic acid	1.33	Training
Flunitrazepam	1.34	Training
Tobramycin	1.35	Training
Indometacin	1.40	Training
Tamoxifen	1.40	Training
Ticarcillin	1.41	Validation
Flufenamic acid	1.43	Training
Pioglitazone	1.43	Training
Valdecoxib	1.43	Validation
Ioxilan	1.44	Training
Levocarnitine	1.45	Training
Trovafloxacin	1.46	Training
Linezolid	1.49	Training
Candesartan	1.50	Training
Iopromide	1.53	Validation
Acetylcysteine	1.55	Training

Clonazepam	1.55	Validation
Foscarnet	1.58	Training
Tolcapone	1.58	Training
Temozolomide	1.59	Training
Cefozopran	1.60	Training
Dicloxacillin	1.60	Training
Gabapentin	1.60	Validation
Hydroxycarbamide	1.60	Training
Zanamivir	1.60	Training
Tranexamic acid	1.61	Training
Furosemide	1.66	Training
Bezafibrate	1.67	Training
Tetracycline	1.67	Training
Trandolaprilat	1.67	Training
Pentostatin	1.68	Training
Chlorphenamine	1.70	Validation
Gemfibrozil	1.70	Training
Remoxipride	1.70	Training
Rofecoxib	1.70	Training
Etanidazole	1.73	Training
Methadone	1.77	Training
Tiagabine	1.78	Training
Ceftizoxime	1.79	Training
Fleroxacin	1.79	Training
Cefepime	1.83	Training
Cefmetazole	1.84	Validation
Tertatolol	1.86	Validation
Ceftazidime	1.88	Training
Amiodarone	1.90	Validation

Caffeine	1.90	Training
Melagatran	1.90	Training
Trimethoprim	1.90	Training
Trospectomycin	1.90	Validation
Aztreonam	1.91	Validation
Lamifiban	1.91	Validation
Cefsulodin	1.92	Training
Pefloxacin	1.93	Validation
Cefpirome	1.98	Validation
Lincomycin	1.98	Training
Cefetamet	2.00	Validation
Rosoxacin	2.00	Training
N-acetylhomocysteine	2.06	Validation
Bromfenac	2.10	Training
Cefuroxime	2.10	Validation
Methotrexate	2.10	Validation
Trazodone	2.10	Training
Troxacitabine	2.13	Training
Anagrelide	2.14	Training
Pheniramine	2.14	Training
Sotalol	2.14	Training
Cefclidin	2.15	Training
Clodronic acid	2.20	Training
Thalidomide	2.20	Validation
Piritrexim	2.21	Training
Cefmenoxime	2.22	Validation
Ibandronic acid	2.29	Validation
Cidofovir	2.30	Training
Finasteride	2.30	Training

Pirmenol	2.34	Validation
Nolatrexed	2.36	Training
Cefpodoxime	2.38	Training
Eplerenone	2.38	Training
Cinoxacin	2.50	Training
Irbesartan	2.51	Validation
Azlocillin	2.52	Validation
Norfloxacin	2.52	Training
Donepezil	2.53	Training
Atenolol	2.54	Training
Cyclophosphamide	2.54	Validation
Pantoprazole	2.57	Training
Levofloxacin	2.58	Validation
Amoxicillin	2.60	Validation
Chlorambucil	2.60	Training
Tocainide	2.60	Validation
Thiamphenicol	2.62	Validation
Bumetanide	2.64	Training
CI-921	2.67	Training
Isosorbide mononitrate	2.67	Validation
Urapidil	2.69	Training
Disopyramide	2.70	Training
Mezlocillin	2.71	Training
Sparfloxacin	2.71	Training
Baclofen	2.72	Training
Brotizolam	2.74	Training
Biapenem	2.77	Training
Bosentan	2.81	Training
Cefamandole	2.82	Training

Chloramphenicol	2.82	Training
Gatifloxacin	2.85	Training
Cefoxitin	2.89	Validation
Cefadroxil	2.90	Training
Tazobactam	2.91	Training
Nadolol	2.94	Training
Tenofovir	2.97	Training
Cerivastatin	2.98	Training
Cilastatin	3.00	Validation
Prazosin	3.00	Training
Thiopental	3.02	Training
Fludarabine	3.06	Training
Imipenem	3.08	Training
Frovatriptan	3.09	Training
Efavirenz	3.10	Validation
Ampicillin	3.11	Training
Temafloxacin	3.19	Training
HI-6	3.20	Validation
Risperidone	3.20	Training
Acecainide	3.22	Training
Lomefloxacin	3.30	Validation
Cefprozil	3.31	Validation
Moexiprilat	3.31	Validation
Cilazaprilat	3.33	Training
Dronabinol	3.33	Training
Imatinib	3.33	Validation
Tianeptine	3.43	Training
Ganciclovir	3.46	Training
Drotaverine	3.47	Training

Meropenem	3.49	Training
Ofloxacin	3.50	Training
Bisoprolol	3.55	Training
Glycerol	3.55	Validation
Lumefantrine	3.57	Training
Moxifloxacin	3.57	Validation
Pirenzepine	3.57	Training
Clavulanic acid	3.60	Training
Hexobarbital	3.60	Training
Prednisone	3.60	Validation
Atomoxetine	3.63	Training
Bicalutamide	3.67	Training
Piperacillin	3.68	Training
2-fluoro-arabinoside-A	3.70	Training
Adefovir	3.72	Validation
Cadralazine	3.79	Validation
Trapidil	3.79	Training
Esomeprazole	3.81	Training
Sematilide	3.86	Training
Quinidine	3.87	Training
Tirofiban	3.87	Training
Lisinopril	3.89	Training
Terbutaline	4.00	Validation
Fosfomycin	4.02	Training
Levacetylmethadol	4.09	Validation
Pimozide	4.10	Training
Enprofylline	4.16	Training
Amantadine	4.17	Validation
Flomoxef	4.17	Training

Midazolam	4.17	Validation
Busulfan	4.18	Training
Cisapride	4.21	Training
Pimagedine	4.27	Training
Cefuzonam	4.28	Training
Hydrocortisone	4.29	Training
Cefalexin	4.30	Validation
Amsacrine	4.33	Training
Dexrazoxane	4.42	Training
Loracarbef	4.45	Validation
Recainam	4.47	Validation
Chlorothiazide	4.50	Training
Spiraprilat	4.52	Training
Thiotepa	4.60	Training
Argatroban	4.70	Validation
Betaxolol	4.70	Validation
Cefradine	4.80	Training
Enalaprilat	4.90	Training
Lamivudine	4.95	Validation
Paracetamol	5.00	Training
Ribavirin	5.00	Training
Clindamycin	5.05	Training
Loprazolam	5.09	Validation
Phencyclidine	5.17	Training
Melphalan	5.20	Training
Citalopram	5.21	Validation
Dofetilide	5.23	Training
Nifedipine	5.23	Training
Levosimendan	5.29	Training

Cefotiam	5.32	Training
Galantamine	5.37	Training
Zolpidem	5.41	Training
Clozapine	5.49	Training
Methylprednisolone	5.49	Training
Isoniazid	5.55	Training
Flecainide	5.60	Validation
Zalcitabine	5.61	Training
Metrifonate	5.67	Training
Clonidine	5.70	Training
Sultopride	5.71	Validation
Doxapram	5.75	Training
Ketanserin	5.86	Training
Amlodipine	5.90	Validation
Diclofenac	5.90	Training
Rabeprazole	6.00	Training
Rolipram	6.00	Validation
Sildenafil	6.00	Training
Tolterodine	6.03	Training
Cefaclor	6.10	Training
Oxacillin	6.10	Validation
Milrinone	6.17	Training
Perindopril	6.19	Training
Lansoprazole	6.23	Training
Mexiletine	6.30	Training
Moexipril	6.30	Training
Naratriptan	6.35	Training
Cefapirin	6.36	Training
Avitriptan	6.40	Training

Buflomedil	6.43	Training
Alprenolol	6.47	Training
Olanzapine	6.48	Validation
Eletriptan	6.50	Validation
Bupivacaine	6.59	Training
Ciprofloxacin	6.62	Training
Alfuzosin	6.67	Training
Toborinone	6.67	Validation
Cefalotin	6.70	Validation
Rilmenidine	6.73	Training
Roxatidine	6.73	Training
Actisomide	6.79	Training
Paroxetine	6.80	Validation
Nafcillin	6.89	Training
Fluocortolone	7.00	Training
Chloroquine	7.09	Training
Metoclopramide	7.13	Training
Ropivacaine	7.14	Validation
Treprostinil	7.14	Training
Zileuton	7.14	Validation
Tasosartan	7.17	Training
Ritipenem	7.21	Training
Diphenhydramine	7.30	Training
Pseudoephedrine	7.33	Training
Talipexole	7.38	Training
Triazolam	7.55	Training
Celecoxib	7.59	Training
simvastatin beta-hydroxy acid	7.60	Training
Bendamustine	7.62	Validation

Pindolol	7.69	Validation
Bufuralol	7.70	Training
Amisulpride	7.80	Training
Salbutamol	7.89	Training
Metformin	7.91	Validation
Procainamide	7.98	Training
Atropine	8.00	Training
Famciclovir	8.00	Validation
Zatebradine	8.00	Training
Ethambutol	8.05	Training
Oxybutynin	8.10	Training
Timolol	8.10	Training
Pramipexole	8.20	Validation
Mitomycin	8.23	Training
Stavudine	8.24	Training
Mirtazapine	8.25	Training
Propiverine	8.32	Training
Azelastine	8.33	Validation
Domperidone	8.33	Validation
Losartan	8.34	Training
Carteolol	8.40	Training
Ketoconazole	8.40	Training
Crisnatol	8.44	Training
Pravastatin	8.50	Validation
Pyridostigmine	8.57	Validation
Cimetidine	8.63	Training
Ritodrine	8.67	Training
Carvedilol	8.70	Validation
Gusperimus	8.77	Training

Hydrocodone	8.82	Training
Doxofylline	8.93	Training
Famotidine	8.95	Training
Tezosentan	9.05	Training
Repaglinide	9.17	Training
Lidocaine	9.20	Training
Acetylsalicylic acid	9.30	Training
Penciclovir	9.36	Training
Methylphenidate	9.44	Training
Isradipine	9.52	Validation
Molsidomine	9.52	Validation
Fluoxetine	9.60	Training
Mitoxantrone	9.60	Training
Nizatidine	9.63	Training
Amiloride	9.70	Training
Chlorpromazine	9.80	Training
Tebufelone	9.88	Training
Allopurinol	9.90	Validation
Nitrofurantoin	9.90	Training
Fluphenazine	10.00	Training
Ketobemidone	10.00	Training
Dexmedetomidine	10.01	Training
Nedocromil	10.20	Training
Triflusal	10.71	Training
Tolamolol	10.80	Training
Perindoprilat	10.95	Training
Acebutolol	11.00	Validation
Codeine	11.00	Training
Diaziquone	11.00	Training

Mercaptopurine	11.00	Training
Propyphenazone	11.00	Validation
Metaclazepam	11.06	Training
Moclobemide	11.19	Training
Ropinirole	11.19	Validation
Entacapone	11.22	Training
Doxifluridine	11.43	Training
Oxycodone	11.43	Training
Eprosartan	11.55	Validation
Felodipine	11.59	Training
Amitriptyline	11.67	Training
Topotecan	11.78	Training
Haloperidol	11.80	Validation
Fentanyl	11.96	Training
Enoximone	12.02	Training
Nicorandil	12.29	Training
Trandolapril	12.38	Training
Moxonidine	12.50	Training
Emedastine	12.52	Validation
Sufentanil	12.70	Validation
Bromopride	12.86	Validation
Alosetron	12.93	Training
Abecarnil	13.00	Training
Cytarabine	13.00	Training
Abacavir	13.07	Training
Triamcinolone	13.07	Training
Captopril	13.33	Training
Methohexital	13.33	Training
Mebendazole	13.40	Training

Etilefrine	13.63	Validation
Propranolol	13.67	Training
Doxepin	14.00	Validation
Nalmefene	14.40	Training
Vinpocetine	15.00	Training
Didanosine	15.25	Training
Fluticasone	15.61	Validation
Etidocaine	15.86	Training
Trimipramine	15.90	Training
Terguride	16.00	Training
Fluvastatin	16.18	Training
Mosapride	16.19	Training
Ketamine	16.30	Validation
Doxorubicin	16.46	Training
Carmustine	16.70	Training
Neostigmine	16.70	Training
Pethidine	17.00	Training
Coumarin	17.30	Training
Dixyrazine	17.54	Training
Terbinafine	17.86	Training
Carbidopa	18.00	Validation
Lorcainide	18.00	Training
Methoxsalen	18.00	Training
Cetiedil	18.33	Validation
Tegaserod	18.33	Training
Nicotine	18.50	Training
Propantheline	18.86	Training
Rizatriptan	18.90	Training
Propafenone	19.00	Training

Quetiapine	19.06	Validation
Sumatriptan	19.15	Training
Hydromorphone	19.16	Training
Carbimide	20.00	Training
Fendiline	20.00	Validation
Mesna	20.50	Training
Fluvoxamine	21.40	Training
Midodrine	23.00	Training
Fenretinide	23.48	Training
Naloxone	23.50	Training
Perphenazine	23.81	Training
Morphine	24.50	Training
Rivastigmine	25.71	Training
Zidovudine	26.00	Training
Quinapril	26.43	Training
Nandrolone	26.67	Training
Labetalol	28.10	Training
Buspirone	28.30	Training
Emivirine	28.57	Training
Ibutilide	29.00	Training
Amineptine	29.55	Training
Nalbuphine	29.70	Training
Phenylephrine	29.90	Training
Cocaine	32.00	Training
Tizanidine	33.33	Training
Bupropion	36.00	Training
Dopexamine	36.00	Training
Sertraline	38.00	Training
Butorphanol	40.00	Training

Acadesine	41.67	Training
Oxaprozin	48.33	Training
Hydralazine	56.00	Training
Azathioprine	57.00	Training
Dobutamine	59.00	Training
Capecitabine	59.73	Training
Prilocaine	64.00	Training
N-(4-methoxyphenyl)retinamide	98.86	Training
Articaine	126.19	Training
Exemestane	145.00	Training
Encainide	177.14	Training
Misoprostol	240.00	Training

Table 10: GT+ compounds.

Acetaminophen	Methimazole	Trimetrexate	2-amino-6-ethoxybenzothiazole
Acrivastine	Methylphenidate	Troglitazone	3-acetyl-2,5-dichlorothiophene
Alendronate	Metronidazole	Warfarin	3-acetyl-2,5-dimethylthiophene
Almotriptan	Milrinone	Zaleplon	3-methyl-2-thiophenecarboxaldehyde
Amifostine	Moexipril	Ziprasidone	3-thiopheneacetonitrile
Aminocaproic acid	Morphine	Zolmitriptan	3-thiophenecarboxaldehyde
Aminosalicylic acid	Nabumetone	9,10-difluoro-2,3-dihydro-3-Me-7-Oxo-7h-pyrido-1,4-benzoxazine-6-cooh	5-chloro-2-thiophenecarboxaldehyde
Amitriptyline	Nalbuphine	6,9-dichloro-2-methoxyacridine	5-ethyl-2-thiophenecarboxaldehyde
Ampicillin	Naloxone	1-isoquinolinecarbonitrile	Ethyl 3-thiopheneacetate
Aspirin	Naproxen	9-anthraldehyde(P-tosyl)hydrazone	Thiophene-2-carbonitrile
Brinzolamide	Nitrofurantoin	1,N6-ethenoadenine	Methapyrilene
Bupropion	Nitroglycerin	(1-pyrrolidinylmethyl)benzotriazole	2-acetylthiophene
Chloramphenicol	Omeprazole	1-pyrenemethanol	2-nitrothiophene
Chloroquine	Oxcarbazepime	7-bromoindole	2,3,5-tribromothiophene
Chlorpheniramine	Pantoprazole	9-anthracenylmethyl 2,4,6-trimethylbenzoate	2-acetyl-5-chlorothiophene
Ciclopirox	Pentobarbital	7-chloroindole	3-methyl-2-

			thiophenecarboxylic acid
Ciprofloxacin	Pentostatin	9-anthryl-N,N-dimethylmethanamine hydrochloride	2,5-diiodothiophene
Citalopram	Pentoxifylline	7-chloro-1,2,3,4-tetrahydrocyclopent(B)indole	3-acetylthiophene
Clarithromycin	Peperacillin	2-aminoanthracene	2-thiopheneacetonitrile
Clofibrate	Pergolide	3-amino-9-ethylcarbazole	Ethyl 2-thiophenecarboxylate
Clomiphene	Permethrin	1-aminoanthracene	2,3-dibromothiophene
Dantrolene	Phenoxybenzamine	2-amino-1-methylbenzimidazole	2,5-dichlorothiophene
Dexrazoxane	Phenylephrine	5-aminoindole	2,5-dibromothiophene
Diazepam	Pilocarpine	9-aminophenanthrene	2,3-dihydrothieno-(3,4-B)-1,4-dioxin
Diflunisal	Podofilox	Bisbenzimidazole trihydrochloride	5-nitro-2-thiophenecarboxaldehyde
Dihydroergotamine	Praziquantel	Indoline	Di-2-thienyl ketone
Diphenhydramine	Procarbazine	Pipemidic acid	2,3-thiophenedicarboxaldehyde
Donepezil	Propranolol	4-aminoindole	3-(thianaphthen-3-yl)-L-alanine
Doxycycline	Pyrazinamide	8-aminoquinoline	4-keto-4,5,6,7-tetrahydrothianaphthene
Doxylamine	Pyrilamine	1-hydroxypyrene	2-acetyl-3-methylthiophene
Entacapone	Pyrimethamine	6-methylthiopurine	Dibenzothiophene-2,8-

			diylbis((N-carbonylmethylene) dimethylamine)
Epinephrine	Quetiapine	1,5-dihydroxynaphthalene	3-(diethylamino)phenol
Eprosartan	Quinidine	Indole	o-phenetidine
Erythromycin	Rabeprazole	1-naphthaldehyde	2-chloroaniline
Esomeprazole	Ranitidine	4,7-dichloroquinolinium	4-methoxy-2-methylaniline
Fosinopril	Remifentanil	Quinoxaline	p-anisidine
Fosphenytoin	Riluzole	Phthalazine	2-methoxy-5-methylaniline
Furazolidone	Risedronate	Quinazoline	Phenoxazine
Furosemide	Rivastigmine	4-fluoroindole	2-fluoroaniline
Grepafloxacin	Ropivacaine	Harmine	2,4-difluoroaniline
Griseofulvin	Rosiglitazone	2-phenylbenzimidazole	1,8-diaminonaphthalene
Haloperidol	Stavudine	3-indolylacetonitrile	m-phenetidine
Halothane	Sulfanilamide	7-methylindole	2-amino-4-tert-butylphenol
Hydrochlorothiazide	Sulfasalazine	6-methoxyindole	5-amino-2-methoxyphenol
Ibuprofen	Tacrine	Ethyl 5-hydroxy-2-methylindole-3-carb-oxylate	2,3-diaminotoluene
Ifosfamide	Tamoxifen	8-(trifluoromethyl)-4-quinolinol	6-amino-m-cresol
Imatinib	Tazobactam	8-sulfo-2,4-quinolinedicarboxylic acid	4-aminobenzyl cyanide
Imipramine	Temozolomide	2-(2-aminophenyl)indole	2,3-difluoroaniline
Indomethacin	Theophylline	5-amino-2-methylindole	2,5-dimethyl-1,4-phenylenediamine
Isoniazid	Thiotepa	N,R-	2-anilinopyridine

		diphenylbenzotriazolemetha namine	
Ketorolac	Tiagabine	N,methyl-N-phenylbenzotriazolemethanamine	2-amino-4-methylbenzonitrile
Lamivudine	Timolol	2-acetyl-5-bromothiophene	Enoxacin
Lansoprazole	Tolcapone	2-amino-4-methoxybenzothiazole	2,3-xylidine
Letrozole	Toremifene	2-amino-4-methylbenzothiazole	2-amino-4-chlorobenzonitrile
Levodopa	Tramadol	2-amino-6-flurobenzothiazole	3-amino-o-cresol
Loratidine	Travoprost	2-amino-6-methoxybenzothiazole	
Mebendazole	Trientine	2-bromo-5-chlorothiophene	
Melphalan	Trimethoprim	2-thiophenecarboxaldehyde	

Table 11: GT- compounds.

Acarbose	Miglitol	N-acetyl-L-tryptophan	4,6-diphenylthieno-(3,4-D)-(1,3)-dioxol-2-one-5,5-dioxide
Acebutolol	Mirtazapine	N-phenylbenzotriazolemethanamine	5-anilino-1,2,3,4-thiatriazole
Acitretin	Modafinil	9-anthraldehyde oxime	6-amino-2-mercaptobenzothiazole
Adapalene	Montelukast	N-(3-indolylacetyl)-L-phenylalanine	Cyclopropyl 2-thienyl ketone
Albendazole	Moricizine	1-methyl-L-tryptophan	N,N-dimethyl-N-((5-nitro-2-thienyl)methylene)-1,4-phenylenediamine
Alprazolam	Mupirocin	Anthrarobin	N,N-dimethyl-4-(6-methylbenzothiazol-2-yl)aniline
Alprostadil	Mycophenolate	2,8-quinolinediol	Thieno-(3,2-B)-pyridin-7-ol
Amantadine	Nafarelin	4-acridinol	Trans-3-(3-thienyl)acrylic acid
Amiloride	Nalidixic acid	5-carboxyfluorescein	Trans-2-(4-dimethylamino)styryl)benzothiazole
Aminolevulinic acid	Naltrexone	2-aminobenzimidazole	5,5-dibromo-2,2-biothiophene
Amiodarone	Naratriptan	2-mercaptobenzimidazole	3-phenylthiophene
Amlexanox	Nateglinide	4-H-cyclopenta(D,E,F)phenanthr	2,5-dibromo-3-hexylthiophene

		ene	
Amlodipine	Nedocromil	Acenaphthene	4,6-dimethyldibenzothiophene
Amphotericin B	Nefazodone	Acyclovir	2,2:5,2-terthiophene-5,5-dicarboxaldehyde
Amprenavir	Nelfinavir	Benomyl	5-nitrothiophene-2-carboxylic acid
Anagrelide	Nevirapine	Ellagic acid	D-R-(2-thienyl)glycine
Argatroban	Niacin	Etofylline	Cephalothin Sodium
Atenolol	Nicardipine	Ganciclovir	2-iodo-5-methylthiophene
Atorvastatin	Nicotine	1-isoquinolinamine	2-methylthianaphthene
Atovaquone	Nifedipine	1-aminopyrene	3-(2-thienyl)-L-alanine
Azelaic acid	Nilutamide	2,7-diaminofluorene	3-bromo-4-methylthiophene
Azelastine	Nimodipine	4-amino-pyrazolo(3,4-D)pyrimidine	P trans-2-(2-nitrovinyl)thiophene
Azithromycin	Nisoldipine	6-aminoindazole	3-butylthiophene
Aztreonam	Nizatidine	8-azaadenine	3-dodecylthiophene
Balsalazide	Norfloxacin	Acridine orange	2,5-dibromo-3-decylthiophene
Beclomethasone	Olanzapine	Aminopterin	2,5-dibromo-3-dodecylthiophene
Benazepril	Olopatadine	Bisbenzimide	2-bromo-5-methylthiophene
Benzoyl peroxide	Olsalazine	Coumarin, 7	L-R-(3-thienyl)glycine
Bepridil	Ondansetron	Harmane 1,2,3,4-tetrahydro-3-carboxylic acid	L-R-(2-thienyl)glycine
Bimatoprost	Orlistat	L-abrine	3-chloroacetylbenzo-(B)-thiophene

Bisoprolol	Oseltamivir	N-methyltryptamine	Perphenazine
Bitolterol	Oxaprozin	Sangivamycin hydrate	Tubercidin
Brimonidine	Oxiconazole	Sulfaquinoxaline	Disperse orange 11
Bromocriptine	Oxybutynin	1-pyrenecarboxaldehyde	1-(methylamino)anthraquinone
Buspirone	Pamidronate	2-ethylanthracene	1-aminoanthraquinone
Butaconazole	Paricalcitol	3-aminoquinoline	Ethyl 2-aminobenzoate
Butenafine	Paroxetine	1-ethylnaphthalene	2,4,6-trimethylaniline
Butorphanol	Pemirolast	6-aminoquinoline	2-amino-4-chlorobenzoic acid
Cabergoline	Perindopril	4,7-phenanthroline	o-anisidine
Caffeine	Phenylpropanolamine	4-amino-2-chloro-6,7-dimethoxyquinazoline	9-aminoacridine
Calcipotriene	Phenytoin	5-aminoquinoline	4,4¢-bis(diethylamino)benzophenone
Calcitriol	Phytonadione	Folic acid	N-phenylanthranilic acid
Candesartan	Pimozide	Indole-3-propionic acid	7-diethylamino-4-methylcoumarin
Carbamazepine	Pioglitazone	Nebularine	Ethoxyquin
Carisoprodol	Pirbuterol	4,8-dihydroxyquinoline-2-carboxylic acid	N,N-diethylaniline
Carteolol	Pramipexole	5-chlorobenzotriazole	2,4-diamino-6-phenyl-1,3,5-triazine
Carvedilol	Pravastatin	2-(trifluoromethyl)benzimidazole	2-chlorophenothiazine
Cefdinir	Primidone	5-fluoroindole	1-phenylpiperazine

Cefepime	Probenecid	Octaverine	Phenothiazine
Cefonicid	Procainamide	9-phenylacridine	Benzidine
Cefoperazone	Proguanil	5-methylindole	N-(4-hydroxyphenyl)-2-naphthylamine
Cefotaxime	Promethazine	5-methylbenzimidazole	1,1'-dianthrimide
Cefpodoxime	Propafenone	9-chloroanthracene	5-chloro-o-anisidine
Cefprozil	Propofol	2,6-dimethylquinoline	3,4-dimethylaniline
Ceftazidime	Pseudoephedrine	8,8-diquinolyl disulfide	2,5-dimethylaniline
Ceftibuten	Pyridostigmine	5-benzyloxyindole	3-dimethylaminophenol
Ceftizoxime	Quinapril	3-indoleglyoxylic acid	N-phenylglycine
Ceftriaxone	Raloxifene	9-bromoanthracene	4-bromoaniline
Cefuroxime	Ramipril	3,4,7,8-tetramethyl-1,10-phenanthroline	4-chloroaniline
Celecoxib	Repaglinide	Bathophenanthroline	3-chloroaniline
Cerivastatin	Rifabutine	Octrizole	Piperazine
Cetirizine	Rifapentine	1-methyl-2-phenylindole	2-(2-aminoethylamino)ethanol
Cevimeline	Rimantadine	2,9-Dimethyl-4,7-diphenyl-1,10-phenanthroline	Diethanolamine
Chirocaine	Risperidone	4-methoxyindole	Bis(2-methoxyethyl)amine
Chlorothiazide	Ritonavir	5-bromoindole	N-ethyl-1-naphthylamine
Chloroxine	Rizatriptan	2-ethyl-9,10-dimethoxyanthracene	N-methylanthranilic acid
Chlorthalidone	Rocuronium	N-(9-acridinyl)maleimide	o-tolidine
Cilastatin	Rofecoxib	9,10-bis(4-methoxyphenyl)-2-chloro-anthracene	3-(ethylamino)-p-cresol
Cisapride	Ropinirole	1-((phenylthio)methyl)-1h-benzotriazole	2-anilinoethanol
Clemastine	Salmeterol	2-biphenyl-4-yl-quinoline-	Sulfisoxazole

		4-carboxylic acid	
Clindamycin	Saquinavir	9,10-dimethoxy-1,2,3,4,5,6,7,8-octamethyl-anthracene	1,5-diaminoanthraquinone
Clonidine	Sertraline	4-hydroxy-6,7-diisobutoxy-quinoline-3-carboxylic acid ethyl ester	2,6-diaminoanthraquinone
Clopidogrel	Sibutramine	N-(9-anthracenylmethylene)-4-chloroaniline	4-phenoxyaniline
Clotrimazole	Sildenafil	N-(9-anthracenylmethylene)-2,4,6-trimethylaniline	2,6-diaminopyridine
Cromolyn	Simvastatin	9-benzoylanthracene	2,6-dimethylmorpholine
Cyclobenzaprine	Sotalol	4-(tert-butylthio)-7-chloroquinoline	Bis(hexamethylene)triamine
Cycloserine	Sulfamethoxazole	2-methyl-phenanthrene	p-phenetidine
Cyproheptadine	Sumatriptan	9-phenoxyacridine	2-aminopyridine
Dapsone	Tamsulosin	1,2,3,4-tetrahydro-9H-pyrido(3,4-B)indole	4-(dimethylamino)benzophenone
Desflurane	Tazarotene	1-deaza-2-chloro-N(6)-cyclopentyladenosine	N-methyldiphenylamine
Diclofenac	Telmisartan	2,4-diamino-6-hydroxymethylpteridine	N-allylaniline
Diltiazem	Terazosin	2-amino-5,6-dimethylbenzimidazole	3-aminophenol
Disopyramide	Terbinafine	2-aminoacridone	3-aminophenyl sulfone
Dolasetron	Terconazole	5-chloro-1,3-dihydro-1,3,3-	Triphenylamine

		trimethylspiro(indole-2,3-phenanthr(9,10-B)oxazine	
Dorzolamide	Tetracycline	6-methoxy-1,2,3,4-tetrahydro-9H-pyrido-(3,4-B) indole-1-carboxylic acid	4,4'-diaminobenzophenone
Doxazosin	Tiludronate	9,10-diaminophenanthrene	N-(2-carboxyphenyl)glycine
Dronabinol	Tirofiban	9-(methylaminomethyl)anthracene	N,N-dibutylaniline
Econazole	Tizanidine	9-amino-6-chloro-2-methoxyacridine	N-isopropylaniline
Efavirenz	Tobramycin	(S)-(-)-2,3,4,9-tetrahydro-1H-pyrido(3,4-B)indole-3-carboxylic acid	4-chloro-N-methylaniline
Eflornithine	Tocainide	9-bromo-2-methoxy-anthracene	1-(2-fluorophenyl)piperazine
Enalapril	Tolmetin	Acetazolamide	5-chloro-2-(methylamino)benzophenone
Enalaprilat	Tolterodine	Tenoxicam	10-methylphenothiazine
Epoprostenol	Topiramate	2,5-thiophenedicarboxaldehyde	1,2-diaminoanthraquinone
Eptifibatide	Torsemide	2-(4-aminophenyl)-6-methylbenzothiazole	N-cyclohexylaniline
Estazolam	Trandolapril	2-(dimethylaminomethyl)thiophene	Dipentylamine
Estramustine	Tretinoin	2-amino-4-	1-(4-

		chlorobenzothiazole	fluorophenyl)piperazine
Etodolac	Triamterene	2-amino-5,6-dimethylbenzothiazole	4-amino-3-methylbenzoic acid
Famotidine	Triprolidine	2-amino-6-chlorobenzothiazole	4-amino-m-cresol
Felbamate	Valproate	2-amino-6-methylbenzothiazole	2,3,5,6-tetramethyl-1,4-phenylenediamine
Felodipine	Valsartan	2-aminobenzothiazole	N,N-diisopropylaniline
Fenofibrate	Vancomycin	2-bromothiophene	2-(phenylsulfonyl)aniline
Fenoldopam	Venlafaxine	2-chlorothiophene	N,N'-bis(2-hydroxyethyl)ethylenediamine
Fentanyl	Verapamil	2-propylthiophene	2-(methylamino)pyridine
Fexofenadine	Zafirlukast	2-thiopheneacetic acid	4-(butylamino)benzoic acid
Flecainide	Zanamivir	2-thiophenecarboxamide	perhydroisoquinoline
Fluconazole	Zileuton	3-bromo-2-chlorothiophene	2-acetylphenothiazine
Flumazenil	Zoledronate	3-bromothianaphthene	2-benzylaminopyridine
Fluoxetine	Zolpidem	3-bromothiophene	Solvent blue 59
Fluticasone	Zonisamide	3-methoxythiophene	N-methyldodecylamine
Fluvastatin	6-carboxyfluorescein	3-thiopheneacetic acid	2-aminoterephthalic acid
Fluvoxamine	N-chloroacetyl-L-tryptophan	3-thiophenecarboxylic acid	4-(methylamino)benzoic acid
Formoterol	Dansyltryptamine	4-(2-thienyl)butyric acid	Tolfenamic acid
Fosfomycin	3-deazaadenosine	4-bromo-2-thiophenecarboxaldehyde	2-(propylamino)ethanol
Gabapentin	1,N6-etheno-2-deoxyadenosine	5-bromo-2-thiophenecarboxylic acid	4-(diethylamino)benzophenone

Galantamine	O6-methyl-2-deoxyguanosine	5-bromothiophene-2-carbaldehyde	6-methoxy-1,2,3,4-tetrahydro-9H-pyrido[3,4-b]indole
Glipizide	Ethyl-4-hydroxy-7-trifluoromethyl-3-quinolinecarboxylate	5-methyl-2-thiophenecarboxylic acid	2,2'-oxydianiline
Glyburide	9-(phenyliminomethyl) anthracene	Coumarin 6	1-(2-pyridyl)piperazine
Granisetron	2,5-dimethylindole	Dibenzothiophene-5,5-dioxide	6-norlysergic acid diethylamide
Guanadrel	1,3-diphenyl-benzo(F)quinoline	Ethyl 2-thiopheneacetate	5-phenyl-o-anisidine
Hexachlorophene	4,7-dihydroxy-1,10-phenanthroline	Ethyl 2-amino-4,5,6,7-tetrahydrobenzo-(B)-thiophene-3-carboxylate	2-amino-4-(ethylsulfonyl)phenol
Ibutilide	1-methyl-3-phenyl-benzo(F)quinoline	Suprofen	4'-piperazinoacetophenone
Imipenem	2-styryl-quinoline	2-(trifluoroacetyl)thiophene	N-ethyl-N-isopropylaniline
Imiquimod	2-phenylbenzo(H)quinoline	Thianaphthene	3,3',5,5'-tetramethylbenzidine
Indinavir	6,7-dihydro-5,8-dimethyldibenzo(B,J)(1,10)phenanthroline	Dibenzothiophene	N-(2-amino-4-chlorophenyl)anthranilic acid
Ipratropium	2-methylindole-3-carboxaldehyde	2-benzoylthiophene	6-Amino-2-naphthoic acid
Irbesartan	N-(9-	2,2-bithiophene	3-amino-1,2,4-triazole

	anthracenylmethylene)-P-toluidine		
Isosorbide	2-butoxy-7,10-dichloropyrido(3,2-B)quinoline	2-phenylthiophene	Cytarabine
Itraconazole	2-(methylsulfonyl)benzothiazole	2-thiopheneglyoxylic acid	1-naphthylamine-8-sulfonic acid
Ketoconazole	Acetic acid 10-acetoxymethyl-anthracen-9-ylmethyl ester	Diethyl 5-amino-3-methyl-2,4-thiophenedicarboxylate	2-trifluoromethylaniline
Ketoprofen	2,3-diphenyl-5,6-benzoquinoxaline	Methyl 3-amino-2-thiophene carboxylate	N-methylaniline
Ketotifin	1-(9-phenanthryl)-1-cyclohexanol	3-thiophenemalonic acid	N-phenyl-m-anisidine
Labetalol	2-(4-biphenylyl)quinoline	3,3-bithiophene	Formanilide
Lamotrigine	N-(9-anthracenylmethylene)-P-anisidine	5-methyl-2-thiophenecarboxaldehyde	Allylthiourea
Levalbuterol	N-(9-anthracenylmethylene)-M-anisidine	Nocodazole	1-naphthylamine-7-sulfonic acid
Levamisole	3-(trifluoroacetyl)indole	R-terthienyl	5-fluorouridine
Leviteracetam	Di-anthracen-9-yl-methanol	2,5-bis(5-tert-butyl-2-benzoxazolyl)thiophene	3,5-bis(trifluoromethyl)aniline

Levocarnitine	9,10-bis(4-methoxyphenyl)anthracene	3-iodothiophene	Sulfameter
Levomethadyl	R,R-diphenyl-2-quinolinemethanol	1-(2-thienyl)-1-propanone	3-methylxanthine
Lidocaine	2,4-diphenylbenzo(H)quinazoline	2-iodothiophene	5-dimethylamiloride
Lindane	2-chloro-4-(4-methoxyphenyl)-3-phenylquinoline	2-thiophenecarboxylic acid	5-(N,N-hexamethylene)amiloride
Linezolid	Pseudocoralyne	2-thiopheneethylamine	Guanazole
Lisinopril	1,N6-ethenoadenosine	4-methyldibenzothiophene	Hycanthone
Loracarbef	9-anthracenylmethyl 4-benzylphenyl ether	2-thiophenecarboxylic hydrazide	N-methylhomoveratrylamine
Losartan	9-anthracenylmethyl P-tolyl sulfide	2-bromo-3-methylthiophene	1,4,7-triazacyclononane
Lovastatin	9-anthryl trifluoromethyl ketone	2-(4-methoxybenzoyl)thiophene	Methyl 2-amino-5-chlorobenzoate
Mafenide	N-(3-indolylacetyl)-L-isoleucine	3,4-dibromothiophene	4-(diethylamino)benzoic acid
Mefloquine	9-anthracenylmethyl methyl sulfide	2-(3-thienyl)ethanol	3-amino-2-naphthoic acid
Meloxicam	1-pyreneacetic acid	2-(2-thienyl)ethanol	2',3'-dideoxyuridine
Meropenem	2-methyl-9-acridinecarboxaldehyde	Tetrachlorothiophene	N-methylglucamine

Mesalamine	3-(10-(2-carboxy-ethyl)anthracen-9-yl)propionic acid	3-octylthiophene	Proglumide
Mesna	2-mesitylquinoline	Cefoxitin sodium salt	2,4,6-triphenylaniline
Metformin	6,9-dichloro-2-methylacridine	â-(2-thienyl)-D-alanine	N,N'-bis(3-aminopropyl)ethylenediamine
Methyldopa	1,8-dichloro-9-methoxy-anthracene	2,2-thenil	3,5-dimethylanthranilic acid
Metolazone	4-methyl-2-(2-naphthyl)benzo(H)quinoline	2-((5-(dibutylamino)-2-thienyl)methylene)-1H-indene-1,3-(2H)-dione	(S)-(-)-1,1'-binaphthyl-2,2'-diamine
Metoprolol	3,6-bis(2-methyl-2-morpholinopropionyl)-9-octylcarbazole	2-amino-3,5-dinitrothiophene	N-(tert-butoxycarbonyl)-L-leucine methyl ester
Mexiletine	1-(4-biphenylyl)isoquinoline	2-chloro-5-(chloromethyl)thiophene	5-iodo-2',3'-dideoxyuridine
Midazolam	9-(1h-benzotriazol-1-ylmethyl)-9h-carbazole	2-amino-6-(methylsulfonyl) benzothiazole	1-[(2-hydroxyethyl)amino]-4-(methylamino)-9,10-anthracenedione
Midodrine	2-(2-benzo(B)thiophen-2-yl-vinyl)quinoline	3,6,9,14-tetrathiabicyclo-(9.2.1)-tetradeca-11,13-diened	

Table 12: TdP+ compounds.

Compound	Set	Compound	Set
Amantadine	Training	Ritodrine	Training
Amiodarone	Training	Salbutamol	Training
Azithromycin	Training	Salmeterol	Training
Bepridil	Training	Sibutramine	Training
Chloral hydrate	Training	Sotalol	Training
Chlorpromazine	Training	Sparfloxacin	Training
Cisapride	Training	Tacrolimus	Training
Clarithromycin	Training	Tamoxifen	Training
Cocaine	Training	Terbutaline	Training
Disopyramide	Training	Thioridazine	Training
Dobutamine	Training	Tizanidine	Training
Dofetilide	Training	Venlafaxine	Training
Domperidone	Training	Voriconazole	Training
Dopamine	Training	Ziprasidone	Training
Droperidol	Training	Adenosine phosphate	Validation
Ephedrine	Training	Ajmaline	Validation
Epinephrine	Training	Aprindine	Validation
Erythromycin	Training	Astemizole	Validation
Felbamate	Training	Atropine	Validation
Fenfluramine	Training	Azelastine	Validation
Flecainide	Training	Azimilide	Validation
Foscarnet sodium	Training	Chloroquine	Validation
Fosphenytoin	Training	Clindamycin	Validation
Gatifloxacin	Training	Diphenhydramine	Validation
Granisetron	Training	Emedastine	Validation
Halofantrine	Training	Grepafloxacin	Validation
Haloperidol	Training	Halothane	Validation

Hydrochlorothiazide	Training	Hydroquinidine	Validation
Ibutilide	Training	Ketanserin	Validation
Indapamide	Training	Maprotiline	Validation
Isoprenaline	Training	Mefloquine	Validation
Isradipine	Training	Mianserin	Validation
Levofloxacin	Training	Mibefradil	Validation
Mesoridazine	Training	Mizolastine	Validation
Methadone	Training	Olanzapine	Validation
Midodrine	Training	Papaverine	Validation
Moexipril	Training	Prenylamine	Validation
Moxifloxacin	Training	Probucol	Validation
Nicardipine	Training	Prochlorperazine	Validation
Norepinephrine	Training	Promethazine	Validation
Octreotide	Training	Quinine	Validation
Ondansetron	Training	Sematilide	Validation
Orciprenaline	Training	Sertindole	Validation
Pentamidine	Training	Spiramycin	Validation
Phentermine	Training	Sultopride	Validation
Phenylephrine	Training	Terfenadine	Validation
Phenylpropanolamine	Training	Terodiline	Validation
Pimozide	Training	Tiapride	Validation
Procainamide	Training	Trazodone	Validation
Pseudoephedrine	Training	Trimetaphan camsilate	Validation
Quetiapine	Training	Vasopressin	Validation
Quinidine	Training	Vincamine	Validation
Risperidone	Training	Zimeldine	Validation

Table 13: TdP- compounds.

Compound	Set	Compound	Set
Abacavir	Training	Methotrexate	Training
Acarbose	Training	Methoxsalen	Training
Acetazolamide	Training	Methyltestosterone	Training
Acetohydroxamic acid	Training	Methylthioninium chloride	Training
Acetylcysteine	Training	Miconazole	Training
Adapalene	Training	Mifepristone	Training
Albendazole	Training	Miglitol	Training
Alendronic acid	Training	Minocycline	Training
Alitretinoin	Training	Misoprostol	Training
Allopurinol	Training	Monoethanolamine oleate	Training
Amcinonide	Training	Montelukast	Training
Amifostine	Training	Nalidixic acid	Training
Amikacin	Training	Natamycin	Training
Amiloride	Training	Nelfinavir	Training
Aminosalicylic acid	Training	Neomycin	Training
Amoxicillin	Training	Nevirapine	Training
Ampicillin	Training	Nitrofural	Training
Anakinra	Training	Olsalazine	Training
Anastrozole	Training	Orlistat	Training
Azelaic acid	Training	Oxcarbazepine	Training
Aztreonam	Training	Oxytetracycline	Training
Bacampicillin	Training	Paracetamol	Training
Baclofen	Training	Paromomycin	Training
Balsalazide	Training	Perindopril	Training
Beclometasone	Training	Phenazopyridine	Training
Bendroflumethiazide	Training	Phenelzine	Training
Benzonatate	Training	Phytomenadione	Training

Betaine	Training	Piperacillin	Training
Bicalutamide	Training	Pramipexole	Training
Bumetanide	Training	Prednicarbate	Training
Butenafine	Training	Procarbazine	Training
Calcipotriol	Training	Pyrantel	Training
Calcium folinate	Training	Pyrazinamide	Training
Carbenicillin	Training	Raloxifene	Training
Carmustine	Training	Riboflavin	Training
Cefaclor	Training	Rifampicin	Training
Cefamandole	Training	Rifapentine	Training
Cefapirin	Training	Risedronic acid	Training
Cefazolin	Training	Ritonavir	Training
Cefdinir	Training	Salsalate	Training
Cefditoren	Training	Saquinavir	Training
Cefixime	Training	Secbutabarbital	Training
Cefoperazone	Training	Simvastatin	Training
Cefotetan	Training	Spectinomycin	Training
Cefoxitin	Training	Spironolactone	Training
Cefpodoxime	Training	Stanozolol	Training
Cefprozil	Training	Streptomycin	Training
Ceftazidime	Training	Streptozocin	Training
Ceftibuten	Training	Sulfadiazine	Training
Ceftizoxime	Training	Sulfafurazole	Training
Ceftriaxone	Training	Sulfamethoxazole	Training
Chlorzoxazone	Training	Sulfasalazine	Training
Cinoxacin	Training	Testolactone	Training
Clobetasol	Training	Tetracycline	Training
Clocortolone	Training	Theophylline	Training
Clopidogrel	Training	Thiamazole	Training

Colestipol	Training	Thiosulfate	Training
Colestyramine	Training	Thiotepa	Training
Cycloserine	Training	Tiabendazole	Training
Cytarabine	Training	Tiludronic acid	Training
Dacarbazine	Training	Tioguanine	Training
Danazol	Training	Tobramycin	Training
Dapsone	Training	Tolazamide	Training
Desmopressin	Training	Tolcapone	Training
Dexrazoxane	Training	Tolnaftate	Training
Diclofenamide	Training	Tranexamic acid	Training
Dicloxacillin	Training	Trichlormethiazide	Training
Dicoumarol	Training	Trientine	Training
Dicycloverine	Training	Trifluridine	Training
Diethylstilbestrol	Training	Troleandomycin	Training
Diiodohydroxyquinoline	Training	Trometamol	Training
Dirithromycin	Training	Unoprostone	Training
Dorzolamide	Training	Uramustine	Training
Eflornithine	Training	Ursodeoxycholic acid	Training
Estradiol	Training	Valrubicin	Training
Estrone	Training	Vincristine	Training
Ethambutol	Training	Vinorelbine	Training
Ethinylestradiol	Training	Warfarin	Training
Etidronic acid	Training	Zafirlukast	Training
Finasteride	Training	Zanamivir	Training
Floxuridine	Training	Zidovudine	Training
Fluconazole	Training	Zoledronic acid	Training
Flucytosine	Training	Alclometasone	Validation
Flunisolide	Training	Benzocaine	Validation
Fluocinolone acetonide	Training	Cefadroxil	Validation

Fluocinonide	Training	Cefalexin	Validation
Flutamide	Training	Cefradine	Validation
Fluvastatin	Training	Cefuroxime	Validation
Fosfomycin	Training	Clotrimazole	Validation
Furazolidone	Training	Demeclocycline	Validation
Furosemide	Training	Dienestrol	Validation
Gabapentin	Training	Doxycycline	Validation
Gemfibrozil	Training	Ethionamide	Validation
Gentamicin	Training	Ethosuximide	Validation
Glibenclamide	Training	Ethotoin	Validation
Glimepiride	Training	Guaifenesin	Validation
Glipizide	Training	Ketorolac	Validation
Griseofulvin	Training	Levonorgestrel	Validation
Guanabenz	Training	Lovastatin	Validation
Hydroflumethiazide	Training	Medroxyprogesterone	Validation
Hydroquinone	Training	Medrysone	Validation
Hydroxycarbamide	Training	Metirosine	Validation
Irinotecan	Training	Metronidazole	Validation
Isoniazid	Training	Mometasone	Validation
Kanamycin	Training	Nandrolone	Validation
Lactulose	Training	Nitrofurantoin	Validation
Lamotrigine	Training	Norethisterone	Validation
Letrozole	Training	Norgestrel	Validation
Levamisole	Training	Pemoline	Validation
Lincomycin	Training	Penicillamine	Validation
Liothyronine	Training	Primidone	Validation
Lomustine	Training	Propylthiouracil	Validation
Loracarbef	Training	Pyridoxine	Validation
Loteprednol	Training	Rimexolone	Validation

Mafenide	Training	Stavudine	Validation
Mebendazole	Training	Sulconazole	Validation
Meclocycline	Training	Sulfanilamide	Validation
Melphalan	Training	Testosterone	Validation
Mercaptopurine	Training	Ticarcillin	Validation
Mesalazine	Training	Trimethoprim	Validation
Methazolamide	Training	Zileuton	Validation
Methenamine	Training		