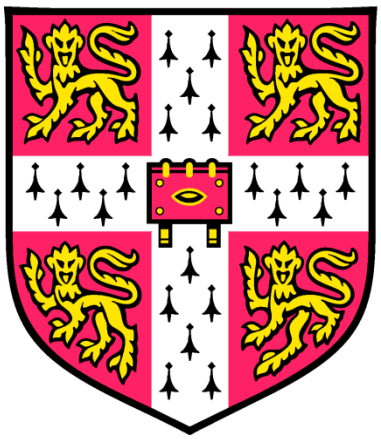
Computational Approaches to Predicting Drug Induced Toxicity

Richard Liam Marchese Robinson
King's College



This dissertation is submitted for the degree of Doctor of Philosophy

Abstract

Novel approaches and models for predicting drug induced toxicity *in silico* are presented. Typically, these were based on Quantitative Structure-Activity Relationships (QSAR). The following endpoints were modelled: mutagenicity, carcinogenicity, inhibition of the hERG ion channel and the associated arrhythmia - Torsades de Pointes.

A consensus model was developed based on Derek for Windows^{TM} and Toxtree and used to filter compounds as part of a collaborative effort resulting in the identification of potential starting points for anti-tuberculosis drugs.

Based on the careful selection of data from the literature, binary classifiers were generated for the identification of potent hERG inhibitors. These were found to perform competitively with, or better than, those computational approaches previously presented in the literature.

Some of these models were generated using Winnow, in conjunction with a novel proposal for encoding molecular structures as required by this algorithm. The Winnow models were found to perform comparably to models generated using the Support Vector Machine and Random Forest algorithms.

These studies also emphasised the variability in results which may be obtained when applying the same approaches to different train/test combinations.

Novel approaches to combining chemical information with Ultrafast Shape Recognition (USR) descriptors are introduced: Atom Type USR (ATUSR) and a combination between a proposed Atom Type Fingerprint (ATFP) and USR (USR-ATFP). These were applied to the task of predicting protein-ligand interactions - including the prediction of hERG inhibition. Whilst, for some of the datasets considered, either ATUSR or USR-ATFP was found to perform marginally better than all other descriptor sets to which they were compared, most differences were statistically insignificant. Further work is warranted to determine the advantages which ATUSR and USR-ATFP might offer with respect to established descriptor sets.

The first attempts to construct QSAR models for Torsades de Pointes using predicted cardiac ion channel inhibitory potencies as descriptors are presented, along with the first evaluation of experimentally determined inhibitory potencies as an alternative, or complement to, standard descriptors. No (clear) evidence was found that 'predicted' ('experimental') 'IC-

descriptors' improve performance. However, their value may lie in the greater interpretability they could confer upon the models.

Building upon the work presented in the preceding chapters, this thesis ends with specific proposals for future research directions.

Acknowledgements

First and foremost, many thanks go to my supervisors, Dr John Mitchell and Professor Robert Glen, for their valuable guidance, feedback and support throughout my doctoral studies. Thanks are also owed to many colleagues, both past and present, at the Unilever Centre for Molecular Science Informatics. In particular, I owe gratitude to the following individuals: Dr Florian Nigsch, for discussions regarding Winnow and initial guidance; Dr Pedro Ballester, for discussions regarding USR and 3D-QSAR; Dr Hamse Mussa, for his explanations of Machine Learning and statistics and Rob Lowe, for invaluable programming advice and many enjoyable discussions inside and outside of work.

Dr Ed Cannon is thanked for providing Python code which was used as a starting point for implementing the descriptors presented in Chapter 5.

Dr Chris Marchese, Dr Susana Tomasio, Dr Johannes Kirchmair, Shardul Paricharak, Florian Roessler, Sonia Liggi and Matt Grayson are thanked for proofreading.

I thank the following researchers for helpfully discussing their work and/or providing data used in the work presented in this thesis: Drs Khac-Minh Thai, Elodie Dubus, Ismail Ijjali, Olivier Taboureau, Chris Swain, Chun Wei Yap, Andreas Bender and Munikumar Doddareddy. Valuable feedback was also received from scientists from the Chemical Computing Group, Ideaconsult, ChemAxon, Accelrys and Lhasa Limited.

I am indebted to the administrative support provided by Susan Begg and Emma Graham, as well as the technical support provided by the computer officers at the Department of Chemistry.

My friends and family are also owed immense thanks for their support and encouragement, particularly in recent years.

Unilever and the Engineering and Physical Sciences Research Council are thanked for funding this work.

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated.

This thesis does not exceed the specified word limit (60,000) as defined by the Degree Committee of the Faculty of Physics and Chemistry.

Publications

Sections of this thesis are based upon (contributions to) the following journal articles and communications.

Chapter 2

The sections on QSAR modelling methods, figures of merit, model validation and the applicability domain are based on this author's contributions to:

Gleeson, M. P.; Modi, S.; Bender, A.; Marchese Robinson, R. L.; Kirchmair, J.; Promkakaew, M.; Hannongbua, S.; Glen, R. C. *Curr. Pharm. Des.* **2012**, *18*, 1266–1291.

Chapter 3

An abridged description of this author's contribution to this collaboration is presented in:

Ballester, P.J..; Mangold, M.; Howard, N.I.; Marchese Robinson, R.L.; Abell, C.; Blumberger, J.; Mitchell, J.B.O. *J. R. Soc. Interface* Article In Press.

Chapter 4

Marchese Robinson, R. L.; Glen, R. C.; Mitchell, J. B. O. *Mol. Inf.* **2011**, *30*, 443–458.

Marchese Robinson, R. L.; Glen, R. C.; Mitchell, J. B. O. *J. Cheminf.* **2012**, *4(Suppl 1)*, O6.

Contents

Chapter 1 Introduction 1

1.1 Determination of Drug Induced Toxicity 2
1.2 The Value of Computational Approaches 3
1.3 Drug Induced Toxicity Endpoints Modelled in this Thesis 7
1.4 The Intention of this Thesis 21

Chapter 2 Computational Toxicology and Quantitative Structure-Activity Relationships 22

2.1 Defining Computational Toxicology 22
2.2 Historical Background 22
2.3 Approaches to Predicting Toxicology *In Silico* 23
2.4 Read Across 24
2.5 Expert Systems 25
2.6 Quantitative Structure-Activity Relationships 27

Chapter 3 Screening for Mutagenicity and Carcinogenicity in the Context of a Prospective Virtual Screen 50

3.1 Overview of Collaboration 50
3.2 Approach Developed for Toxicity Screening 55
3.3 Empirical Validation of Toxicity Models 59
3.4 Results Obtained from Application of Toxicity Models 62
3.5 Conclusions 68

Chapter 4 Development and Assessment of Binary Classifiers for Identifying Potent hERG Inhibitors 69

4.1 Introduction 69
4.2 Datasets 70
4.3 Model Development and Validation 74
4.4 Comparisons with Models Developed by Thai and Ecker and by Dubus et al. 85
4.5 Avoiding Feature Selection Bias: Calculating Overlap between Different Datasets 88
4.6 Results and Discussion 88
4.7 Conclusions 106

Chapter 5 Development of Novel 3D Descriptors 108

5.1 Introduction 108
5.2 Proposed Methodology 112
5.3 Evaluation of Methodology 117
5.4 Results and Discussion 131
5.5 Conclusions 150

Chapter 6 Predicting Drug Induced Torsades de Pointes Using Biological Descriptors 152

6.1 Introduction 152
6.2 Modelling Approaches Employed Here 154
6.3 Datasets 158
6.4 Summary of TdP Model Comparisons 167
6.5 Results and Discussion 169
6.6 Conclusions 182
Chapter 7 Conclusions and Future Work 183
7.1 Conclusions 183
7.2 Future Work 185
Bibliography 189
Appendix A. Supplementary Files 217
Appendix B. Performance of Toxicity Models Previously Reported in the Literature 218
Appendix C. Additional Computational Details 220

List of Figures

Figure 1.1 Overview of empirical (yellow) and possible *in silico* (red) approaches to toxicity assessment in a typical discovery testing scheme for an orally available pharmaceutical. Grey arrows indicate the requirement for empirical data to build most *in silico* models. 5

Figure 1.2 Overview of empirical (yellow) and possible *in silico* (red) approaches to toxicity assessment in drug development and post-marketing surveillance for a typical orally available pharmaceutical. Grey arrows indicate the requirement for empirical data to build most *in silico* models. 6

Figure 1.3 The standard model for the potential consequences of hERG inhibition. Here, 'B' denotes a hERG blocker. This image is based on that presented by Crumb and Caverro.[59] The induction of, and degeneration into fibrillation of, Torsades de Pointes is more completely discussed in section 1.3.3. 10

Figure 2.1 Two of the structural alerts for genotoxic carcinogenicity used in the "hybrid" expert system software program Toxtree. Here, R denotes any atom/group except OH or SH.[37] 26

Figure 2.2 An overview of SVM classification. (A) A linearly separable dataset in the feature space, and a possible separating hyperplane. (B) The SVM solution for such a dataset: the maximum margin hyperplane. (C) A non-linearly separable dataset; highlighted are two misclassified instances and two further instances lying inside the margin, with their corresponding slack-variables. (D) A conceivable corresponding decision boundary (as shown in (C)) in the descriptor space (supposing the feature space is a higher dimensional projection of the descriptor space); only the two misclassified instances are highlighted. N.B.: These images are for illustrative purposes only. 39

Figure 3.1 Overview of the workflow undertaken to identify novel, experimentally verified inhibitors of type II DHQase. This author's contributions to the study are circled. 54

Figure 3.2 Results obtained from validating carcinogenicity predictions of all 112 models on the ISSCAN database. All equivocal carcinogens were considered non-carcinogens. 63

Figure 3.3 Results obtained from validating carcinogenicity predictions of all 112 models on the ISSCAN database. All equivocal carcinogens were considered carcinogens. 64

Figure 3.4 Results obtained from validating mutagenicity predictions of all 112 models on the ISSCAN database. All equivocal mutagens were considered non-mutagens. 64

Figure 3.5 Results obtained from validating mutagenicity predictions of all 112 models on the ISSCAN database. All equivocal mutagens were considered mutagens. 65
Figure 4.1 Two examples of topological substructures encoded as features, for an example molecule, by a generic circular fingerprint considering environments extending up to two bonds from the central atom. 78
Figure 4.2 The assignment of 'discretized descriptor features' corresponding to descriptor D. 81
Figure 4.3 The procedure employed to generate orthogonal sparse bigrams (OSBs) as additional features for a generic molecule and generic original set of features (monograms) using a window size of three.....83
Figure 4.4 A summary of the generation of the 94 feature sets – comprising fingerprint features, discretized descriptor features or both – evaluated in the current work.....84
Figure 4.5 Distribution of Int-Set (red) and ExtTest-Set (blue) within the plane defined by the first two principal components (PCA plot) for the P_VSA descriptor set (computed as per section 4.3.2.1). Principal components were calculated from the combined Int-Set and ExtTest-Set, using the prcomp( ) function in R[225] - with scale=TRUE. 94
Figure 4.6 PCA plots generated as per Figure 4.5 (training set: red, 'external' test set: blue), for all splits of the Literature-368 dataset; from top left to bottom right: original split, random:1, random:2 and random:3. 95
Figure 4.7 PCA plots (generated as per Figure 4.5) for the Diverse Subset partitions of the Thai-313 (LHS) and Dubus-203 (RHS) datasets for which results are presented in Table 4.3. 96
Figure 4.8 Mean MCC values for externally validated (i.e. only 'unbiased' feature sets were considered, where relevant) selected Winnow models (generated using multiple training cycles, where relevant), compared to QuaSAR-Classify (QC) mean MCC values and Binary QSAR (BQ) MCC values when trained and tested on the same data. 97
Figure 4.9 Mean MCC values for externally validated Winnow models (as per Figure 4.8), and the corresponding (mean) MCC values for the externally validated SVM and RF models. 98
Figure 5.1 Computation of all three (A) USR and (B) supplementary ATUSR descriptors (for atoms typed as H bond donors) corresponding to the distances (dotted lines) between the molecular centroid (black sphere) and the relevant atoms. N.B.: The [NH+] donor (RHS of

the molecular centroid; N = blue; H = white) is also typed as a cationic and ring atom. This figure was generated using VMD (version 1.9).[357,358] 116
Figure 5.2 Removal of inorganic structures from Doddareddy dataset in Pipeline Pilot, prior to standardization. The numbers reflect the total number of compounds, in all SMILES files, subsequent to assignment of unique compound IDs, presented by Dr Andreas Bender. The 2,644 compounds referred to were presented in a subset of these files. 121
Figure 5.3 Standard workflow used to process 'raw' structures in all datasets modelled in the work presented in this chapter. N.B.: The structures obtained subsequent to each of the different pre- CORINA processing steps were also parsed via Pybel.[177] 126
Figure 5.4 R^2 values obtained from 10-10CV (five RNG seeds for randomForest) on the hERG-196 dataset , with different descriptor sets calculated from structures obtained: (A) prior to Molecular Mechanics calculations, (B) from local minimisations, (C) from global minimisations, (D) from docking (ChemScore selections), (E) from docking (RF-Score selections). The black lines and circle centres denote the median and mean results respectively. 133
Figure 5.5 Corresponding MCC values (c.f. Figure 5.4) obtained on the hERG-196:Subset dataset. 135
Figure 5.6 Mean R^2 values obtained, for the hERG-196 dataset, using the following descriptors, computed from different 3D structures: (A) ATUSR, (B) USR-ATFP, (C) USR+MACCS, (D) USR. 137
Figure 5.7 Mean MCC values obtained, for the hERG-196: Subset dataset, using the following descriptors, computed from different 3D structures: (A) ATUSR, (B) USR-ATFP, (C) USR+MACCS, (D) USR. 138
Figure 5.8 Images of Imai *et al.*[305] docked structures for cisapride (A) and E-4031 (G) and corresponding aligned structures obtained via: STERGEN (B,H), local minimisation (C,I), global minimisation (D,J), ChemScore selection (E,K) and RF-Score selection (F,L). All molecular images were generated using VIDA (version 4.1.1).[378] 140
Figure 5.9 R^2 values obtained for: (A) hERG-196 dataset, (B) ThaiReg dataset. In both cases, the structures used for descriptor calculations were obtained from the standard workflow (see 5.3.4.1). The black lines and circles denote the median and mean results respectively. 142
Figure 5.10 MCC values obtained for: (A) hERG-196:Subset dataset, (B) ThaiReg: Subset dataset, (C) Doddareddy dataset and (D) Schattel dataset. In all cases, the structures used for

descriptor calculations were obtained from the standard workflow (see 5.3.4.1). The black lines and circles denote the median and mean respectively. 144
Figure 5.11 P(a) versus number of structures in the active cluster subset when clustering the following datasets using predicted logD: (A) Doddareddy (Test), (B) Doddareddy (Train), (C) ThaiReg:Subset, (D) hERG-196:Subset, (E) Schattel. N.B.: The red (blue) points denote values obtained for the real (activity permuted) data; the green points denote a theoretical upper bound to p(a) - computed as the total number of actives in the dataset divided by the current number of structures in the active cluster subset.[331] 147
Figure 5.12 Corresponding plots to those presented in Figure 5.11, based on clustering using USR. 148
Figure 6.1 Fragments identified in cathinone by Clark and co-workers that were likewise found (green) and not found (red) by this author's implementation of their bit-string descriptor. 156
Figure 6.2 SMARTS patterns matching a tertiary amine and the corresponding amide. 156
Figure 6.3 Overview of the procedures used to prepare the structures in the ion channel and TdP datasets from which InChIs and structural descriptors were computed. 166
Figure 6.4 MCC values (black line: median, circle: mean) obtained using predicted IC-descriptors, and the corresponding results obtained using structural-descriptors alone, across all cross-validation folds, repetitions and RNG seeds. Results obtained on the Yap-2004 (A,C,E) and Clark-2009 (B,D,F) datasets, using predicted IC-descriptors generated from all models (A,B), the putative most relevant set - i.e. KCNH2, KNCQ1, CACNA1C (C,D) and just the KCNH2 model (E,F). 175
Figure 6.5 MCC values (black line: median, circle: mean) obtained using experimental KCNH2 IC-descriptors, and the corresponding results obtained using structural-descriptors alone, across all cross-validation folds, repetitions and RNG seeds. Results obtained on subsets of the Yap-2004 (A,C) and Clark-2009 (B,D) datasets corresponding to all compounds for which KCNH2 experimental IC-descriptors were obtained (A,B) and subsequent to reducing the inconsistency in the experimental conditions used to obtain the underlying measurements (C,D). 178

List of Tables

Table 2.1 The confusion matrix for a binary classifier. 43
Table 3.1 Combinations of options considered for predicting compounds as mutagens/carcinogens based upon the output generated by Toxtree. If any one of the 'ticked' outcomes occurred, a positive prediction of mutagenicity/carcinogenicity was made. If these conditions were not met, compounds were deemed to be predicted non-mutagenic/non-carcinogenic by Toxtree. 61
Table 3.2 Combinations of options considered for predicting compounds as mutagens/carcinogens based upon the output generated by DfW. If any one of the 'ticked' outcomes occurred, a positive prediction of mutagenicity/carcinogenicity was made. If these conditions were not met, compounds were deemed to be predicted non-mutagenic/non-carcinogenic by DfW. 61
Table 3.3 Performance of the selected model. 66
Table 4.1 Numbers of hERG inhibitors and their distribution amongst the potency categories for the datasets modelled in this chapter. 73
Table 4.2 Performance of this author's selected models and literature models on 'external' test sets for all ('Int/Ext') partitions of the Literature-368 dataset into training and 'external' test sets. MCC values (mean MCC values for Winnow, Random Forest (RF) and QuaSAR-Classify) are presented. Values in parentheses are the maximum MCC values obtained across the 50 runs of the QuaSAR-Classify module. Missing values for Winnow (multiple cycles) signify that a single cycle gave the best 5CV result. 92
Table 4.3 MCC values obtained on 'external' test sets for all partitions of the Thai-313 and Dubus-203 datasets used to evaluate this author's modelling procedures. See Table 4.2 for presentational details. 93
Table 4.4 Some of the associations, learnt by the Winnow models, between 'important' ECFP_4 features and hERG blockade which were consistent with trends previously noted in the literature. N.B.: $< W(A) >$ denotes the arithmetic mean weight, across all scorers, assigned to the feature for class A etc. The median difference over all 50 training set orders is reported. All Dubus-203 training set compounds assigned feature 1430169877 were observed to possess a corresponding amide fragment. Occurrence ratios denote the fraction of compounds belonging to the class in question and containing the feature. All SMILES

patterns were visualised using MarvinView (version 5.5.0.1);[298] the 'A' symbols denote wildcard, i.e. undefined, heavy atom connections.[185,308] 102
Table 4.5 Results obtained on the Thai-313 and Dubus-203 datasets in the current work compared to those reported by Thai and Ecker[69] and Dubus *et al.*[91] The split of the Thai-313 dataset for which results are highlighted corresponds to the split with the highest MCC values obtained for both Binary QSAR models (i.e. the split used to evaluate this author's models). The partitions of the Thai-313 dataset used to generate results in the current work (training set :80 actives/160 inactives, test set: 20 actives/53 inactives) differed slightly from the single partition used by Thai and Ecker (training set: 81 actives/159 inactives, test set: 19 actives/54 inactives), as explained in section 4.4.5. All MCC values obtained by Dubus *et al.* and Thai and Ecker were estimated herein. Acc., Rec. and Prec. denote accuracy, recall and precision respectively. 105
Table 4.6 Chi-squared p-values corresponding to (mean, across 50 runs, for QuaSAR-Classify) test set MCC values obtained here for the partitions of the Literature-368, Dubus-203 and Thai-313 dataset used to assess this author's models. These p-values were computed using the CHIDIST( ) function in Excel 2007 (32-bit), supposing one degree of freedom (see Chapter 2, section 2.6.4.1), save where a negative MCC was obtained – for which it was supposed that the model must effectively be a random predictor and the p-value was set to one. 106
Table 5.1 Descriptor sets compared in this work. 118
Table 5.2 RMSD values (computed from heavy atoms) and USR similarities (computed from all atoms) upon alignment of the structures obtained here to the docking poses presented by Imai *et al.*[305] 141
Table 5.3 Numbers of clusters obtained, upon clustering using predicted logD (or USR descriptors), computed prior to the application of CORINA (or from structures processed via the standard workflow) with a distance cut-off of 0.2 (see section 5.3.6). For both descriptor sets, the classification and regression datasets were ranked in order of decreasing number of clusters - supposed to correspond to decreasing diversity. 145
Table 6.1 Total number of compounds used to generate ion channel models. 161
Table 6.2 Numbers of TdP dataset compounds with experimental IC-descriptors based on comparing names of compounds in TdP and ion channel datasets. 165
Table 6.3 Cross-validated results obtained on the entirety of the ion channel datasets derived for the generation of predicted IC-descriptors. 170

Table 6.4 Performance of ion channel models on the TdP datasets; the performance of those models for which only a single TdP dataset compound with an experimental IC-descriptor was obtained is omitted.....171
Table 6.5 Performance of KCNH2 model on TdP datasets, after removing problematic measurements and increasing the experimental consistency of the retained measurements. 171

Glossary

AC = Discretization method using all midpoints between adjacent, ordered, training set instances, with different class labels, as split points (see Chapter 4)

Acc. = Accuracy

AD = Applicability domain

ADME = Absorption, distribution, metabolism and excretion

ADR = Adverse drug reaction

AE = Adverse event

AERS = Adverse Event Reporting System

ANN(s) = Artificial Neural Network(s)

APD(s) = Action potential duration(s)

ArizonaCERT = Arizona Center for Education and Research on Therapeutics

ASP = Astex Statistical Potential

BQ = Binary QSAR

CA = ChemAxon descriptor set (see Chapter 4)

CART = Classification and Regression Trees algorithm

CV = Cross-validation

DfW = Derek for Windows™

Dubus-Rel = Descriptor set obtained via feature selection by Dubus and co-workers (see Chapter 4)

ECVAM = European Centre for the Validation of Alternative Methods

EPA = Environmental Protection Agency

FDA = Food and Drug Administration

FI = Fayyad and Irani's discretization method

FOM(s) = Figure(s) of merit

GLP = Good Laboratory Practice

GSK = GlaxoSmithKline

hERG = Human ether-à-go-go-related gene (or the corresponding ion channel)

HTS = High throughput screening

IC-descriptors = Descriptors corresponding to (predicted) pIC_{50} values for cardiac ion channels as described in Chapter 6

ICH = International Conference on Harmonisation

IK_r = Rapid delayed rectifier current

InChI = IUPAC International Chemical Identifier

'Int/Ext' = Dataset partition into a training and test set, where the test set was not directly used for model selection (see Chapter 4)

kNN = k-Nearest Neighbours

LDA = Linear Discriminant Analysis

LOOCV = Leave-one out CV

m_{try} = Number of descriptors randomly sub-sampled at each node when training a Random Forest model

MAE = Arithmetic mean absolute error

MCC = Matthews Correlation Coefficient

MCCV = Monte-Carlo CV

MLR = Multiple Linear Regression

n.o. = Number of

n_{tree} = Number of trees in a Random Forest model

NMEs = New molecular entities

OOB = Out-of-bag

OSB(s) = Orthogonal sparse bigram(s)

PC(s) = Principal component(s)

PCA = Principal components analysis

PDB = Protein Data Bank

PLS = Partial Least Squares

Prec. = Precision

QC = QuaSAR-Classify

(Q)SAR = (Quantitative) Structure-Activity Relationship

Rec. = Recall

RF = Random Forest

RF-Score = Ballester and Mitchell's docking scoring function built using Random Forest

RMSE = Root Mean Square Error

RNG = Random number generator

SA(s) = Structural alert(s)

SDF = Structure-Data file

SMILES = Simplified Molecular Input Line Entry System

SRS = Spontaneous Reporting System

STERGEN = Stereoisomer generator, integrated into the software program CORINA

SVM = Support Vector Machine

SVR = Support Vector Regression

TdP = Torsades de Pointes

TdP+/- = Label assigned to a compound with/without the potential to induce Torsades de Pointes

Thai-Rel = Descriptor set obtained via feature selection by Thai and Ecker (see Chapter 4)

Type II DHQase = Type II dehydroquinase

UMC = Uppsala Monitoring Centre

USR = Ultrafast Shape Recognition

XO = Xenopus Ooctytes

Chapter 1 Introduction

This thesis presents novel computational models and modelling approaches for predicting some of the most serious forms of pharmaceutical drug induced toxicity.

Toxicology may be broadly defined as the "study of the adverse effects of drugs and chemicals on living systems".[1] Pharmaceutical drugs are chemicals introduced to the human body to induce a desired therapeutic effect.[2] Principally, albeit not exclusively, these are 'small' molecules, with a majority being administered orally.[3] A variety of types of adverse effect, or toxicity "endpoints",[4] may be induced by pharmaceuticals.[5–7]

As of the year 2000, safety issues were one of the leading causes of drug attrition, accounting for more than 20% of failures.[8] Safety concerns also led to 26 drugs, deemed to be new molecular entities (NMEs) by the US regulatory body the Food and Drug Administration (FDA),[9] being withdrawn from the US market in the period 1980-2010.[10] As well as the resultant costs incurred by the pharmaceutical sector due to attrition,[8] and market withdrawal, failure to anticipate adverse drug reactions (ADRs)* prior to marketing approval may have lethal consequences for patients.[11–13]

Hence, there is a clear need for,[14] and corresponding focus in industry on,[6,7,15] the early identification of the toxic liabilities of potential drug candidates. As is emphasised in this thesis, *in silico* methods can play an important role in meeting this need.

This chapter discusses the experimental approaches used to anticipate drug induced toxicity, along with the approaches used to identify drug induced toxicity in the patient population. This provides the context for explaining the importance of computational approaches, such as those presented in this thesis, for predicting these toxic effects. This chapter proceeds with an explanation of the importance of the types of toxicity which were modelled in the studies described here and briefly discusses the key issues associated with modelling these endpoints:

---

*The World Health Organisation (WHO) defines ADRs as "any noxious, unintended, and undesired effect of a drug, which occurs at doses used in humans for prophylaxis, diagnosis, or therapy". The term “adverse drug event” may be considered to more broadly encompass any injury, including those resulting from overdose, induced upon administration of a drug.[11] However, the term “adverse drug reaction” may be used in a looser sense - also encompassing overdose related injuries.[6] In this thesis, the term 'adverse drug reaction' is used in this looser sense – i.e. interchangeably with 'adverse (drug) event'.

mutagenicity, carcinogenicity, hERG inhibition and Torsades de Pointes. This chapter concludes with an overview of the work to be presented in the rest of this thesis.

1.1 Determination of Drug Induced Toxicity

As schematically, and incompletely, illustrated in Figure 1.1 and Figure 1.2, a typical orally available drug is subject to various experimental toxicology, and - ultimately - human clinical safety, assessments prior* to product registration and launch.[7,16,17] To put these assessments into context, these figures also depict the typical testing stages associated with moving from a (virtual, i.e. existing only as computer records) compound library to hits, which show 'reasonable' activity against the therapeutic target, to leads, which have the potential to be optimised by medicinal chemists to yield acceptable drugs, and ultimately to a clinical candidate that is subject to human clinical trials. A fuller explanation of this process, and these terms, is provided in the referenced texts/articles.[7,17–19]

Regulatory agencies require data from certain preclinical *in vitro* (cell based) and *in vivo* (in animals) experimental toxicity assays, which must conform to Good Laboratory Practice (GLP), to be submitted for a clinical candidate.[7,17] For example, the *in vitro* assays currently called for by regulatory agencies include bacterial mutagenicity assessments and an assessment of inhibition of the potassium ion channel encoded by the human ether-à-go-go-related gene (hERG) - endpoints discussed in more detail in section 1.3.[17]

Prior to this, however, a variety of experimental toxicology assays are carried out during drug discovery. Early stage *in vitro* assays may be divided into "prospective" and "retrospective" assays. "Retrospective" assays, e.g. for phospholipidosis, are usually designed to identify target organ specific, "dose-limiting" toxicities (which would limit the therapeutic effect achievable by simply increasing the dose of a drug); these assays are usually carried out after the potential for such types of toxicity being relevant is indicated by short term *in vivo* assays. "Prospective" assays, e.g. for hERG inhibition, are designed to identify "development-limiting" toxicities which could force compound discontinuation if identified.[7]

* As previously noted, even these extensive testing regimes may fail to anticipate ADRs prior to product launch, which may only be identified during “post-marketing surveillance”. This is, by definition, particularly likely for unpredictable “idiosyncratic” ADRs.[7,10]

1.2 The Value of Computational Approaches

However, it would be far more cost effective if experimental toxicity assessment of potential drug candidates could be reduced. *In silico* approaches to toxicity assessment, which generate toxicity predictions for untested, and possibly unsynthesised, compounds, have the potential to reduce the time and money invested in toxicity assessment in drug discovery. This requires that their predictions are sufficiently reliable.[15] One should note that the error associated with the predictions of *in silico* models cannot be reduced beyond that inherent in the experimental measurements/observations made for the modelled endpoint. Indeed, these errors pose challenges for both the empirical generation[*] and assessment of *in silico* models.[20] The specific errors associated with observations made for each of the endpoints modelled in subsequent chapters of this thesis are discussed in section 1.3.

Various uses for such *in silico* models can be envisaged:

1. The most high throughput, computationally inexpensive models, could be used to remove predicted potent toxicants from (virtual) screening libraries (prior to synthesis).[15,20]
2. The models could be used to prioritise (synthesis of)[21] a lead series.[15,22]
3. If the models are interpretable to medicinal chemists, they could be used to guide structural modifications, to remove a toxic liability.[15,20,21]
4. Rather than acting as an alternative to experimental assays, for which it is commonly accepted the currently available models are typically insufficiently reliable,[20,22] they could be used to prioritise compounds for experimental testing.[15,20,22,23]
5. They could be used to assist regulatory decision making.[24]

Gavaghan and co-workers at AstraZeneca suggested that their in-house hERG inhibition model was sufficiently predictive and interpretable to be used, to some extent, for the first four purposes outlined above.[15] Davenport and co-workers at Evotec recently reported the successful application, in the context of a specific discovery project, of *in silico* models for hERG inhibition for the second and third of these tasks – considerably reducing the typical levels of hERG inhibition measured for newly synthesised compounds.[21] Indeed Müller *et al.* suggest that, as of the last decade, "computational tools have supported the optimisation of compounds from a specific chemical class with a small amount of experimental hERG data".

---

[*] See Chapter 2.

The *in silico* prediction of mutagenicity has also been deemed a success story, with Hoffman-La Roche Limited reporting a significant reduction in Ames positive (see section 1.3.1) test results subsequent to the introduction of an *in silico* pre-filter in the late 1990s.[6] Computational models are also increasingly being used to support regulatory decision making and guidelines specifying the requirements for models to be used in a regulatory environment have been published.[24,25]

In contrast to predicting the outcomes of *in vitro* assays, generating models for the direct prediction of human ADRs is deemed particularly challenging - due to the possibility of many, complex mechanisms, the limitations of available data for human ADRs and variable individual susceptibilities to ADRs, leading to "idiosyncratic" toxicity.[7,14,26] However, over the last decade, work has been ongoing to build models based on data for human ADRs.[27–30] Indeed, Pharmatrope is currently making its models for predicting human ADRs directly from chemical structure, built from post-marketing surveillance data, available to its clients.[31] Additional work has been undertaken to relate ADRs to interactions with specific protein targets[30,32,33] as well as gene expression data[14] - such linkages perhaps enabling the future anticipation of human ADRs from assay profiles. The publicly available databases, derived from post-marketing surveillance, for human ADRs, and the potential limitations with such data, were discussed by Clark and Wiseman[28] and, more recently, by Nigsch *et al.*[14] These issues are returned to in more detail in section 1.3.3.

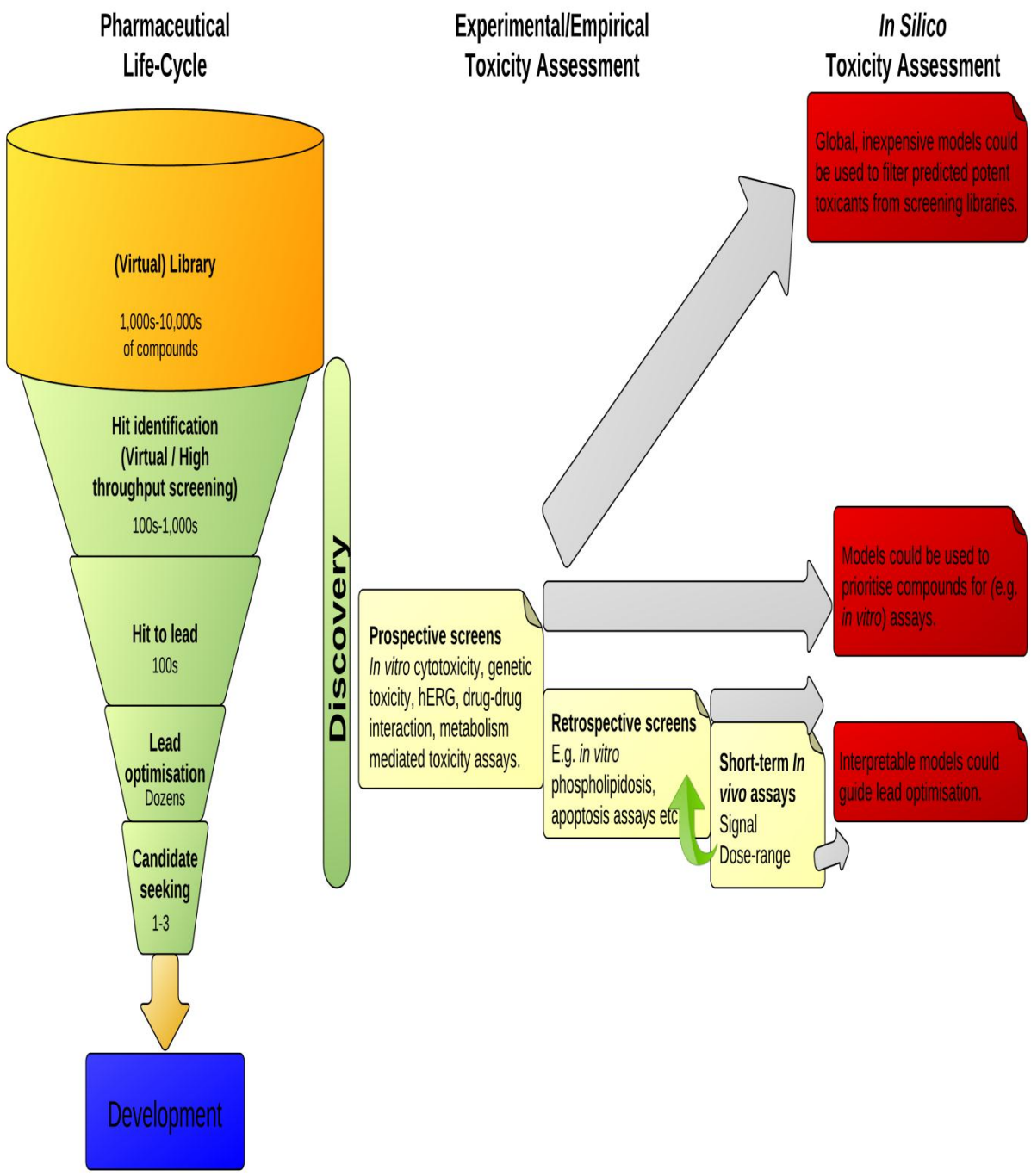


Figure 1.1 Overview of empirical (yellow) and possible *in silico* (red) approaches to toxicity assessment in a typical discovery testing scheme for an orally available pharmaceutical. Grey arrows indicate the requirement for empirical data to build most *in silico* models.

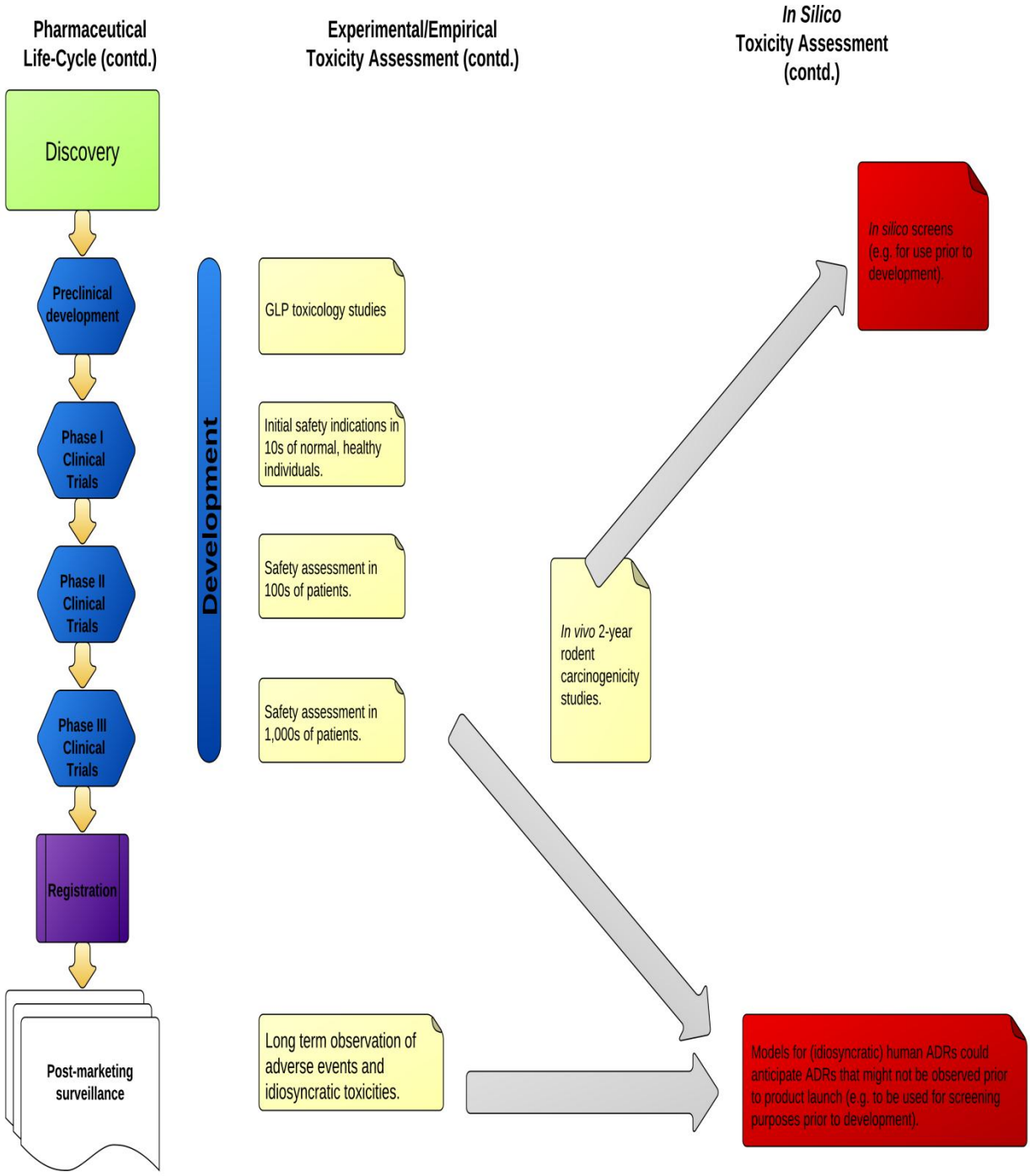


Figure 1.2 Overview of empirical (yellow) and possible *in silico* (red) approaches to toxicity assessment in drug development and post-marketing surveillance for a typical orally available pharmaceutical. Grey arrows indicate the requirement for empirical data to build most *in silico* models.

1.3 Drug Induced Toxicity Endpoints Modelled in this Thesis

1.3.1 Mutagenicity and Carcinogenicity

The potentially lethal effects of cancer, characterised by groups of abnormally proliferating cells, or tumours,[34] are well known, with more than 1.7 million cancer related deaths in Europe estimated in 2004.[35]

Current understanding divides chemical carcinogens* into two categories: genotoxic and non-genotoxic (epigenetic). Genotoxic carcinogens cause direct damage to DNA, whilst non-genotoxic carcinogens act via a variety of alternative mechanisms.[37,38] Mutagenicity, a form of genotoxicity leading to transmissable DNA damage,[38] is typically a cause for concern within the pharmaceutical industry because it is viewed as a surrogate for carcinogenicity.[5,39] Not all mutagens, however, are carcinogens[38,40] and the importance of mutagenicity as a toxicity endpoint in its own right, with a possible association with birth defects,[41] has also been stressed.[36,39]

1.3.1.1 Significance of Mutagenicity and Carcinogenicity for the Pharmaceutical Industry

Compounds exhibiting positive results (i.e. indicated to be mutagens) in mutagenicity assays are rarely progressed to clinical trials.[16,42] Non-genotoxic carcinogens, however, are usually negative in these assays.[38] Whilst regulatory agencies have recently accepted shorter duration assessments in some cases,[43,44] the standard two-year rodent bioassay for carcinogenicity traditionally called for may actually entail a study lasting four to five years.[16,43–45] Since a positive result from rodent bioassays can cause the costly abandonment of a drug candidate in late stage clinical trials, there is a major need to identify carcinogens during discovery or early development.[16]

1.3.1.2 Experimental Characterisation of Mutagenicity and Carcinogenicity

The *in vitro* Ames assay is used by almost every pharmaceutical company for short-term assessment of mutagenic (hence, carcinogenic) potential.[5] Originally developed by Ames,

* A carcinogen is "an agent which increases the rate of formation of tumors in a population".[36]

this assay uses a panel of mutant, histidine-requiring, strains of bacteria, located on a histidine deficient medium - typically in the presence of homogenised liver, to allow for mammalian metabolic activation.[5,20,46] Mutagenic compounds have the potential to induce mutations that restore histidine independence, allowing the formation of bacterial colonies. Hence, a positive result is obtained if a compound induces significant colony growth in one of the strains. Different versions of the Ames test may be performed with different strains and/or without metabolic activation.[47] If compounds have not been assessed in the presence of all recommended strains[20] and/or in the absence of metabolic activation, the assay may fail to detect mutagenicity. A GLP Ames assay is required before a candidate can enter clinical trials.[7]

The degree of interlaboratory concordance of the Ames assay has been reported as 80-85%,[38] or 70-87%.[47] Recent analysis by Sushko *et al.* suggested that the average per-compound accuracy (defined as the maximum number of positive/negative results, divided by the total number of decisive results for a compound) of this assay, primarily determined from repeated measurements in the same laboratory, could be 90-94%.[47]

Recent studies have presented an alternative "Ames II" assay, with similar levels of predictivity for rodent carcinogenicity test outcomes, which requires lower quantities of a compound for testing.[48] Other genotoxicity tests, such as the *in vitro* micronucleus assay, may also be used to screen for carcinogenicity in the pharmaceutical industry.[5]

Whilst an Ames positive result corresponds to rodent carcinogenicity 77-98% of the time,[40] 46% of the 149 carcinogens (or equivocal carcinogens) studied by Zeiger, which were identified in long term rodent assays, were found to be non-mutagens in the Ames assay.[49] Indeed, a 2005 study concluded that information from short-term studies was still not sufficiently predictive of the long-term rodent carcinogenicity assays to replace them.[44]

However, the ability of these traditional rodent bioassays to identify human carcinogens, which are traditionally identified via epidemiological studies, has been questioned. In addition to concerns regarding species extrapolation, Cohen also raised concerns regarding dose-extrapolation - i.e. traditional carcinogenicity studies are performed using doses which may considerably exceed the expected exposure levels for humans.[45,50]

1.3.1.3 Availability of Data for Mutagenicity and Carcinogenicity

A variety of datasets for both mutagenicity and carcinogenicity have been made freely available in an electronic format as required for modelling (see Chapter 2, section 2.6.2) by the US Environmental Protection Agency (EPA) and other organisations.[6,51] These include the ISSCAN database, with carcinogenicity and mutagenicity data from long-term rodent studies and the Ames assay respectively[51] and the IRISTR database, with carcinogenicity assessments derived in accordance with the EPA's cancer risk assessment guidelines - including, where available, data from human epidemiological studies.[52–54] The ISSCAN (version 3a) and IRISTR (version 1b) databases comprised 1,153 and 544 compounds respectively at the time of writing, albeit the same carcinogenicity/mutagenicity assessments had not been made for all compounds therein. Hansen *et al.* recently presented a freely available Ames mutagenicity dataset comprising 6,512 compounds.[55] To the best of this author's knowledge, this was the largest freely available mutagenicity dataset at the time of writing.

Mutagenicity and carcinogenicity datasets of 8,412 and 1,634 compounds respectively were, at the time of writing, commercially available from LeadScope.[56] Additional, proprietary, data is held by the pharmaceutical industry.[57]

1.3.2 hERG Inhibition

Drug inhibition of the hERG potassium ion channel, which carries the rapid delayed rectifier current (IK_r) in cardiac myoctyes,[58,59] i.e. heart muscle cells,[60] is of significance due to its potential relationship to Torsades de Pointes (TdP). TdP, a potentially fatal cardiac arrhythmia,[58*] is ultimately the toxicity endpoint of concern, with hERG inhibition considered as a typical surrogate in the pharmaceutical industry.[6,15] Further consideration is given to TdP, as an endpoint in its own right, in section 1.3.3.

In the standard model, drug blockade of the hERG ion channel leads to prolongation of the cardiac action potential,[†] which can lead to prolongation of the QT interval,[‡] followed by TdP (Figure 1.3).[59,61] However, levels of hERG inhibition resulting in considerable QT

[*] An abnormal heart rhythm.[60]

[†] The "pattern of electrical activity associated with excitable heart cells".[59]

[‡] A period on the electrocardiogram (ECG), which measures electrical activity across the heart as a whole.[59]

prolongation for some drugs, may not lead to observable prolongation in others.[62] Furthermore, some compounds exhibiting modest QT prolongation have been significantly associated with TdP, whilst amiodarone induces larger increases in QT interval, yet is associated with negligible incidence of TdP.[63] Indeed, it is not unknown for TdP to occur in the presence of QT shortening.[64]

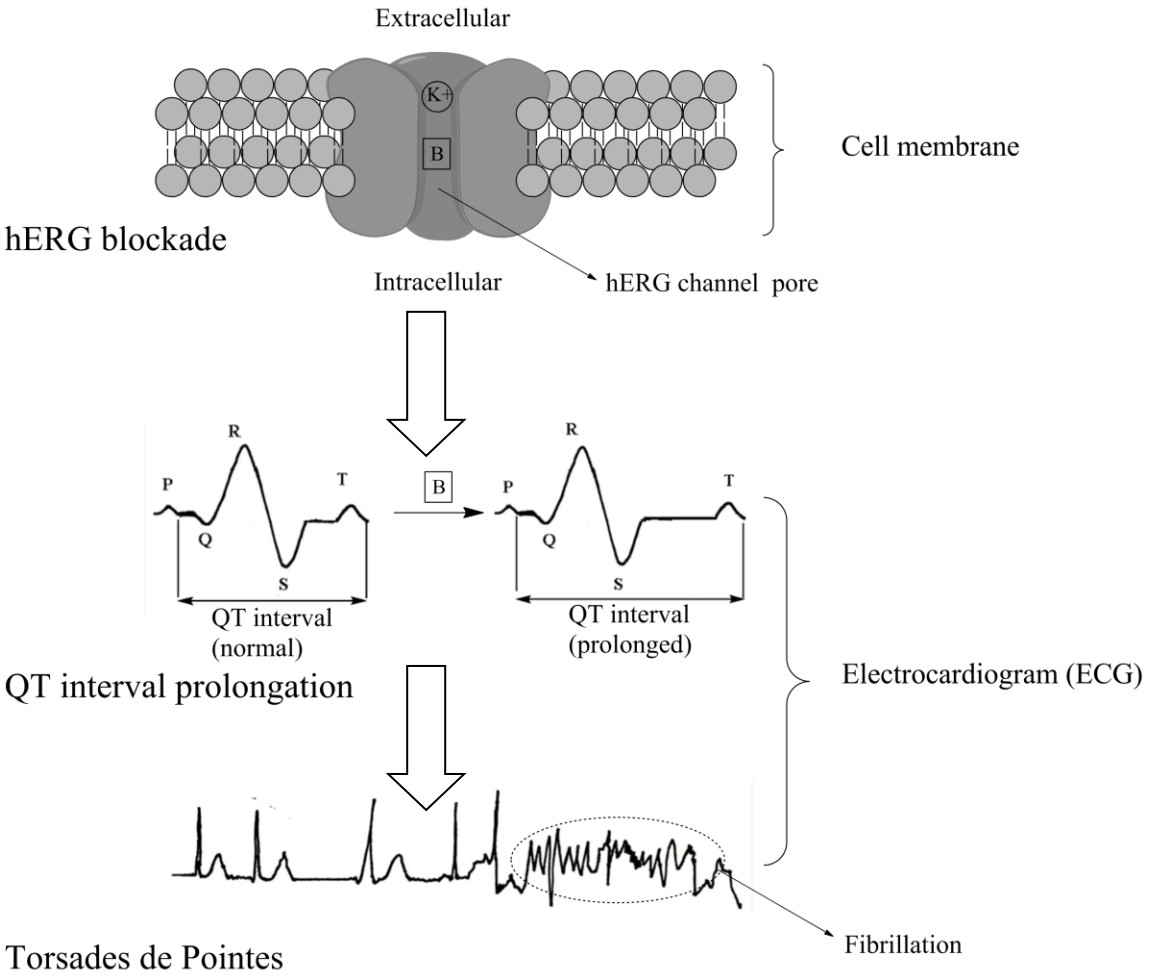


Figure 1.3 The standard model for the potential consequences of hERG inhibition. Here, 'B' denotes a hERG blocker. This image is based on that presented by Crumb and Caverro.[59] The induction of, and degeneration into fibrillation of, Torsades de Pointes is more completely discussed in section 1.3.3.

1.3.2.1 Significance of hERG Inhibition for the Pharmaceutical Industry

In spite of the simplistic nature of the 'standard model', virtually all drugs known to induce QT prolongation inhibit hERG[63] and hERG inhibition is the most common mechanism for drug prolongation of the QT interval.[65] Moreover, the link between drug induced QT prolongation and the risk of TdP was of sufficient concern to regulators by the year 2000, that

a working committee of the International Conference on Harmonisation (ICH) was charged with drafting international guidance for regulators[66] that ultimately called for *in vitro* assessment of I_{Kr} inhibition for clinical candidates.[65] Even prior to the finalisation of these guidelines in 2005, pharmaceutical companies were being requested to provide hERG inhibition data to regulators and experimental assessment of hERG inhibition had become a common part of the drug discovery process.[15,66]

A correlation between hERG inhibition and QT prolongation/Torsadogenic potential (the potential of a compound to induce TdP) has been determined in various studies. Redfern *et al.* suggested that a ratio of more than 30:1 between hERG IC_{50}* and effective therapeutic plasma concentration could be taken as indicative of a lack of TdP causing potential.[68] Similarly, Yao *et al.* proposed a safety margin of 300:1 (hERG IC_{50} : maximum unbound plasma concentration) to separate drugs causing TdP and/or QT interval prolongation from those inducing neither cardiac disorder.[62]

Heuristics have been developed in the pharmaceutical industry directly relating hERG IC_{50} values to their typical toxicological significance. For example, the experience of Yao and co-workers within GlaxoSmithKline (GSK) was that compounds with $IC_{50} < 1$ μM typically prolonged the QT interval in *in vivo* assays, whereas those with $IC_{50} > 10$ μM typically did not and compounds lying above their proposed safety margin usually had IC_{50} values above 10 μM. Hence, in early drug development, compounds with hERG $IC_{50} < 1$ μM were typically not progressed, whereas those with hERG $IC_{50} > 10$ μM were typically progressed without *in vivo* assessment of their potential to prolong the QT interval.[62] Similar heuristics are accepted elsewhere in the industry: compounds with $IC_{50} > 10$ μM are commonly considered safe, whilst the development of potent inhibitors with $IC_{50} < 1$ μM[15] is commonly discontinued.[69]

1.3.2.2 Experimental Characterisation of hERG Inhibition

A variety of experimental assays are used to assess hERG inhibition.[67] These may be grouped into binding and functional assays. Binding assays estimate the affinity with which a putative hERG blocker binds to the channel, yet do not directly assess a compound's ability to inhibit

* The half-maximal inhibitory concentration,[67] as discussed in section 1.3.2.2.

hERG current.[67] Functional assays estimate the extent to which compounds block the hERG channel.[70]

Binding affinity may be quantified by K_d, defined in Equation 1.1.[71]

$$K_d = \frac{[L][R]}{[L-R]} \qquad 1.1$$

In Equation 1.1, [L], [R] and [L-R] represent the equilibrium concentrations of unbound ligand, unbound receptor and bound receptor respectively. In the present context, the ligand and receptor correspond to a putative hERG inhibitor and hERG channel respectively. Whilst K_d values can be determined directly via saturation-binding experiments,[71] they are typically estimated from competitive binding assays.[67,70,72,73] In these assays, the ability of the putative hERG blocker to displace a previously assessed, labelled hERG binder - typically, a potent hERG blocker - is determined. The "IC_{50}" obtained from these experiments corresponds to the unbound concentration of putative blocker at which 50% of the labelled binder has been displaced; this "IC_{50}" is dependent upon the amount of labelled binder used in the assay and, hence, is not an absolute measure of binding affinity. An *estimate* (K_i) for the K_d of the putative blocker can be obtained using the Cheng-Prusoff equation (Equation 1.2).[71,72,74,75]

$$K_i = \frac{IC_{50}}{1 + \frac{[L]}{K_{d,L}}} \qquad 1.2$$

In Equation 1.2, [L] and $K_{d,L}$ respectively denote the equilibrium unbound concentration, corresponding to 50% displacement of the labelled binder, and K_d for the labelled binder.

In principle, binding assays suffer from fundamental limitations. They are unable to distinguish between binding which inhibits or activates the hERG channel and may fail to detect blockers which bind in an alternative binding site to the labelled binder.[67] Other limitations may arise from failing to assay hERG channels in whole cells.[73] Nonetheless, good experimental correlation has been obtained, in a number of studies, between the K_i values obtained from common competitive binding assays and the IC_{50} values obtained from electrophysiological (see below) functional assays. However, inhibitory potency may be underestimated to varying degrees for different compounds.[70,72,73]

High throughput functional assays, such as the rubidium efflux and membrane potential sensitive dye based assays, which indirectly estimate the effect of a compound on hERG current, may also be used for initial screening purposes.[67,70]

However, none of these assays are deemed to be a substitute for electrophysiological assays - which directly measure current flowing through ion channels at specified transmembrane potentials.[67,70,76]

hERG channels transition between different states, commonly denoted closed, inactivated and open, in response to changes in transmembrane potential - with current flow only possible in the open state.[58,67,76] In principle, electrophysiological assays allow for the assessment of state dependent block.[76]

Whilst traditionally corresponding to low throughput techniques, work on, higher throughput, automated electrophysiology assays was initiated in the late 1990s.[77] In the last decade, these became widely available[77] and work was undertaken to minimise the discrepancies between these assays and conventional electrophysiology.[78,79]

Whilst IC_{50} values may be estimated from the percentage inhibition at a single concentration, more reliable estimates are obtained from a concentration-response curve.[80,81] Typically, inhibition measurements are fitted to a version of the Hill equation - which commonly,[62,82,83] though not always,[80] can be expressed in the following form (Equation 1.3).

$$Y = \frac{1}{1 + \left(\frac{IC_{50}}{[B]}\right)^{n_H}} \qquad 1.3$$

In Equation 1.3, Y corresponds to the fractional block of hERG current, [B] corresponds to the unbound concentration of blocker B for which the IC_{50} is determined, and n_H corresponds to the Hill coefficient which is, usually,[84] also estimated from fitting. Y may be estimated as $(I_C - I_B)/I_C$, where I_B denotes the estimated hERG current in the presence of a given concentration of blocker B and I_C corresponds to the estimated hERG current in the absence of B.[83] This expression supposes that 100% current inhibition is achievable in the limit that [B] tends to infinity, such that the IC_{50} corresponds to the concentration required for 50% channel blockade.[83] Minor deviations from 100% maximum inhibition may be taken into account via minor modification of the Hill equation.[80] In some studies, however, maximum block of less than 50% has been suggested, requiring alternate IC_{50} estimations.[85,86]

Commonly, hERG inhibitory potency is expressed in terms of the pIC_{50} ($-\log_{10} IC_{50}$(mol litre^{-1})) calculated from the IC_{50} obtained in electrophysiological assays.[15,78]

Reliable estimates of endogenous IK_r inhibition are technically challenging.[87] In the pharmaceutical industry,[62,78] it is common practice to heterologously express hERG, though

not potential auxiliary subunits or other modulating proteins,[87] in non-cardiac mammalian expression systems.[67,87]

However, inhibitory potencies obtained from cells natively expressing IK_r may differ appreciably from those obtained in heterologous expression systems and potencies may vary between different heterologous expression systems.[67,82,87] Notably, inhibitory potencies obtained from heterologous expression in Xenopus Ooctytes (XO) are typically lower than those obtained in mammalian systems.[67,87,88]

A variety of additional factors - including temperature, voltage protocol etc. - can considerably affect the IC_{50} value determined from electrophysiological measurements. However, variations in these factors do not necessarily lead to statistically significant differences in estimated IC_{50} values.[82]

Nonetheless, the challenges posed to empirical model generation due to the variability in literature derived pIC_{50} estimates - typically obtained under different conditions[67] - have been widely remarked upon.[89–93] Above and beyond these systematic variations (for a given compound), replicate electrophysiological pIC_{50} measurements have been reported to have an error of 0.1-0.5 log units.[15]

1.3.2.3 Availability of Data for hERG Inhibition

The pharmaceutical industry has access to databases comprising thousands of compounds screened for hERG inhibition in the same assay and, in the last decade, the generation and validation of *in silico* models on these databases has been reported.[*] To date, most publicly available datasets are considerably smaller.

At the time of writing, data for 1,960 compounds screened in the same membrane potential dye based assay were freely available from PubChem.[95] Larger datasets/databases comprising measurements obtained using a variety of assays/experimental protocols, are also freely available - including, at the time of writing, literature curated hERG activity data for 5,972 compounds from the ChEMBL database.[96] Doddareddy *et al.* recently presented a dataset of 2,644 compounds, which they claimed to be the largest publicly available dataset at that

[*] Modelling studies based on nearly 9,000 and 60,000 compounds, assessed in a single in-house assay, were respectively presented in 2007 by Gavaghan *et al.* (Astrazenca)[15] and in 2005 by O'Brien and de Groot (Pfizer).[94]

time.[93] A recent publication[97] noted a variety of free access online repositories containing hERG/I_{Kr} inhibition data, curated from the literature for 100-600 compounds.[98–101]

Literature curated datasets might also be purchased from commercial providers, such as Aureus Sciences (formerly Aureus Pharma)[102] and Sunset Molecular.[103]

It should be re-emphasised, however, that not all the measurements in these literature derived datasets/databases will have been derived from the same cell type or assays etc. Modelling studies presented in the recent literature, including work presented in Chapter 4 of this thesis, have made available datasets incorporating electrophysiological measurements for a few hundred compounds.[92,104]

A recent publication presented HERGCentral - a freely available repository of screening results obtained in a high-throughput electrophysiological assay for more than 300,000 compounds.[97,105]

1.3.3 Torsades de Pointes

Torsades de Pointes (TdP) is a rare ventricular tachycardia,[*] which may degenerate into ventricular fibrilation[†] - which, left untreated,[60] leads to sudden death.[87]

Understanding of the mechanisms underlying drug induced TdP is still incomplete.[106,107] Nonetheless, the induction of TdP is believed to be related to an increase in the heterogeneity of ventricular repolarisation. This process, and the underlying molecular interactions, are very briefly discussed below. The interested reader is referred to the reviews by Fenichel *et al.*,[87] Gupta *et al.*[108] and Varró and Baczkó.[106]

Repolarising currents, of which I_{Kr} carried by hERG is one, carry cations to the cellular exterior - restoring resting transmembrane potential. The converse is true for depolarising currents, carried by different cardiac ion channels. The cycle of depolarisation and repolarisation constitutes the action potential,[58] and a reduction in repolarising currents lengthens the action potential duration;[59] prolongation of ventricular action potential durations (APDs) is reflected by elongation of the QT interval (see above).[87]

---

[*] An unusually rapid heart rhythm originating in one of the ventricles (the lower chambers of the heart).[60]

[†] Rapid, chaotic beating of the ventricles which prevents the heart from effectively pumping blood to the body.[60]

The differential distribution of cardiac ion channels amongst different cell types leads to slight differences in APDs across normal ventricular tissue. Drug induced reduction in ventricular repolarisation can exacerbate differences in APDs which may, albeit not necessarily,[109] promote heterogeneity of repolarisation - creating the conditions for TdP.[87]

The potential for a drug to reduce ventricular repolarisation, and increase heterogeneity of repolarisation, depends, in principle, upon its ability to perturb the currents (both depolarising and repolarising) carried by multiple cardiac ion channels.[87,106]

As well as direct inhibition of cardiac ion channels, indirect perturbation of ion channel currents may occur via other mechanisms, which may entail interactions with other, regulatory, proteins, such as the inhibition of hERG trafficking to the cell membrane.[61]

There is clearly a need to also take pharmacokinetic and physicochemical factors, such as might affect a compound's cellular/tissue distribution and plasma concentrations, into account when considering a compound's potential to induce TdP[61,62] - i.e. its Torsadogenic potential.

In addition to considering the intrinsic Torsadogenic potential of a drug, it is also important to consider the potential for the risk of TdP induction to be increased due to drug-drug interactions and other risk factors, e.g. female gender.[12,108] Some analyses suggest that incidents of drug induced TdP almost never occur with non-cardiac medications in the absence of such risk factors.[115] These additional complications both complicate the situation with regard to the prescription, and hence marketing, of potentially Torsadogenic drugs (section 1.3.3.1)[87] as well as the assessment of Torsadogenic potential (section 1.3.3.2).

1.3.3.1 Significance of Torsades de Pointes for the Pharmaceutical Industry

Given that the significance of hERG inhibition ultimately stems from concerns regarding the Torsadogenic liability of potential drug candidates, the significance of TdP for the pharmaceutical industry is largely reflected in the significance of hERG inhibition (see section 1.3.2).

The potentially lethal effects of TdP make it highly important to identify the Torsadogenic potential of drug candidates prior to market approval. Whilst only around 1 in 120,000 patients prescribed the antireflux[110] drug cisapride suffered subsequent incidents of TdP, or other arrhythmias, between 1993 and 1995,[12] the drug was (inconclusively) linked to around 80 deaths in Canada and the US.[110] Cisapride was eventually withdrawn from the market in

the year 2000[10,110] due to its association with TdP.[10,111] However, the consequences of drug induced TdP are less likely to result in market withdrawal for drugs used to treat life threatening conditions.[58] For example, the antiarrhythmic quinidine, with a higher incidence of drug induced TdP than cisapride,[58] remained on the market as of 2010.[112]

Even if a drug is not withdrawn from the market, it may be assigned a warning label - with a "boxed warning", or "black box warning", designed to indicate risk of the most serious, e.g. life-threatening, ADRs.[113] These may negatively impact the revenue generated by a drug, albeit an analysis of the effect of black-box warnings issued for cisapride prior to its withdrawal suggested they had no significant effect on prescriptions.[114]

The rarity of TdP induction by non-antiarrhythmic drugs, typically with < 1 case per 10,000-100,000 exposures, makes it virtually impossible to detect in clinical trials. Hence, there is a particularly pressing need for methods - either experimental or *in silico* - which could identify TdP inducing drugs prior to market launch.[87]

1.3.3.2 Approaches to Defining the Torsadogenic Potential of Drugs

Torsadogenic potential is commonly assessed preclinically by considering surrogate indicators which can be measured in *in vitro* or *in vivo* assays.[62,115] Indeed, various pre-clinical assays of both kinds are proposed by regulatory guidelines - including single cell assessments of I_{Kr} inhibition (see section 1.3.2), changes in action potential duration in multicellular test systems and QT interval alterations in laboratory animals.[65]

In light of the limitations with the 'standard model' outlined in section 1.3.2, assessment of I_{Kr} inhibition (as indicated above, the effects of this may be offset or enhanced by inhibition/activation of other cardiac ion channels etc.) and QT interval prolongation cannot be considered as direct indicators of Torsadogenic potential. Regarding *in vivo* assessments of QT interval prolongation, various other ECG parameters have been suggested as potentially more appropriate indicators.[87,108] Moreover, measurements of the QT interval are technically difficult. Indeed, what is actually measured are changes in the QT interval corrected for various co-factors (such as heart rate); appropriate correction is considered challenging - particularly in animal models.[87]

Given the limitations of these surrogate measurements, work has been undertaken to directly assess the Torsadogenic potential of drugs in animals treated to increase their predisposition to TdP. The development of arrhythmia has also been studied in isolated whole animal hearts.

Measurements of the electrical activity across an isolated section of an animal heart (the "wedge preparation") may also be able to detect pseudo-ECG patterns characteristic of TdP and could be used to determine Torsadogenic potential.[87,116] Very recently, a human ventricular slice preparation was proposed for experimental assessment of arrhythmic risk.[117]

Ideally, Torsadogenic potential would be assessed based upon human data. However, using drug induced TdP as an end point in prospective studies is not always feasible for ethical reasons.[118]

In clinical trials, regulatory guidelines emphasise the importance of detecting the risk of TdP induction via assessing prolongation of the QT interval.[119] However, given how rare drug induced TdP is, direct assessment of the risk of TdP induction in humans is not expected to be feasible until after clinical trials.[87,115,119]

Assessment of risk from post-marketing surveillance is hugely complicated. Firstly, a considerable number of possible confounding factors exist. These include drug-drug/drug-food interactions, variable individual/demographic susceptibilities (and biases towards different patient demographics), over-reporting (e.g. in response to elevated public concern regarding a drug's safety), under-reporting (e.g. for supposed obvious toxicities) as well as variable tendencies of physicians to report adverse events.[14,28,33] An additional confounding factor is inconsistent drug usage (e.g. overdoses).

Secondly, a reported case of an adverse event being caused by a drug might be based on a physician's impression,[28] and hence be somewhat subjective - or based upon questionable evidence.[118] Other potential problems include the use of multiple terms to categorise the same ADR,[14] multiple names being used for the same drug, the submission of multiple reports for the same incident and the increased chance of observing adverse events with increasing usage of the drug (e.g. as time spent on the market increased).[28]

Efforts have been made by the FDA to ensure that standardized terms are used for adverse event reports compiled in its Adverse Event Reporting System (AERS),[14,28] and that resources are available for normalisation of drug names.[28] Given the potential for chance identification of a drug as the cause of an incident of TdP,[28,118] various researchers have proposed only considering a drug to cause TdP based upon statistical tests for association. For example, it has been supposed that reports of TdP induction were due to chance co-occurrence, with this null-hypothesis not being rejected unless the number of reports was statistically significantly higher than the number of reports expected from chance.[28,33] For a

given adverse event (E), and a given drug (D), the number of reports expected by chance (R') could be calculated as follows (Equation 1.4).

$$R' = N(D) \times P(E) \qquad 1.4$$

In Equation 1.4, $N(D)$ is the number of pills/injections of D administered, for all patients, and $P(E)$ is the estimated probability of E co-occurring with any drug (Equation 1.5).

$$P(E) = \frac{\sum_{D'} M(D')}{\sum_{D'} N(D')} \qquad 1.5$$

In Equation 1.5, $M(D')$ is the number of times E co-occurred with the drug D'.

$N(D')$ could be estimated from prescription numbers or shipping volumes.[33,120] However, when using databases of adverse event reports, a more readily available estimate[33] would be the total number of reports involving D' in the database - as recently used to estimate R' for cardiotoxic adverse events by Matthews and Frid[33] as well as Clark and Wiseman.[28] This is not without its limitations. For example, a disproportionately large number of reports for a particular drug or adverse event would artificially inflate R', requiring a larger value of $M(D')$ for a statistically significant association between D' and E to be identified.[33]

In addition to statistical analysis of adverse event databases, different types of evidence might be obtained from a variety of sources and used by experts to assess a drug's ability to induce TdP in humans. This approach is employed by the Arizona Center for Education and Research on Therapeutics (ArizonaCERT).[121]

Assessments based upon human data will usually consider a drug to pose no risk of inducing TdP unless evidence of (the potential for) drug induced TdP is available.[27,28] In some cases, an assessment of Torsadogenic potential might be based upon evidence for drugs within the same therapeutic class.[118] Given the low incidence of TdP with TdP causing drugs (e.g. cisapride, see above), this could result in a number of 'false negative' assignments.

It should be noted that, whilst it is common[27–29,122–128] to conceive of Torsadogenic potential in terms of TdP causing and non-TdP causing agents, this is somewhat of a simplification: some TdP causing drugs might have a greater potential to cause TdP than others. Clark and Wiseman suggested they, unsuccessfully, tried to predict a continuous indicator of

Torsadogenic potential quantifying the statistical association between a compound and reported incidents of TdP induction.[28]

1.3.3.3 Availability of Data on the Torsadogenic Potential of Drugs

In addition to isolated case reports/clinical incidents reported in the literature,[12] publications by De Ponti *et al.*,[118] Redfern *et al.*[68] as well as Fermini and Fossa,[129] presented lists of drugs categorised according to their Torsdaogenic potential.

ArizonaCERT makes its lists of (potentially) TdP causing drugs, grouped into different categories of Torsadogenic risk, freely available online.[130] Commercially available resources which present an assessment of drug safety based upon expert review of the literature, and have been used to identify TdP causing agents,[27,33] include Micromedex[131] and Meyler's Side Effects of Drugs.[132]

The US FDA's AERS presents case reports of adverse events received from 1996-present; these are either voluntary submissions by patients and healthcare professionals or mandatory submissions from drug manufacturers.[133] In principle, a case report details the names of all drugs administered to the patient, the drug(s) suspected of causing the adverse event(s), details of administration (including dosage), the observed adverse drug event(s) and the outcome for the patient.[28] AERS data is freely available from the FDA's website.[133] The AERS replaced the older Spontaneous Reporting System (SRS) which compiled adverse event case reports from 1969-1997, yet has been suggested to be less well curated.[28]

The VigiBase database system compiles adverse event case reports, in standardized language, from across the globe. Access to VigiBase is controlled by the Uppsala Monitoring Centre (UMC).[134,135]

Data might also be obtained from official warnings/labels, in spite of their limitations - e.g. they may not be based on the most up to date evidence, and might be based upon supposed risk - such as assignments based upon a drug's therapeutic class.[118] Nonetheless, this information is largely publicly available,[136] e.g. from the FDA's freely accessible MedWatch data files,[137] and may indicate the frequency with which an ADR has been observed in patients taking a particular drug.[136]

Studies by Yap *et al.* [27] and Xue *et al.*[122] presented datasets comprising the names of a few hundred drugs categorised as TdP causing, using data from ArizonaCERT amongst other sources, and non-TdP causing. Clark and Wiseman recently presented a dataset comprising

more than 1,000 drugs, including structures, categorised as either TdP causing or non-TdP causing, based upon statistical analysis of the FDA's AERS.[28]

1.4 The Intention of this Thesis

To re-iterate, the aim of the work presented in this thesis was the development of new *in silico* models and/or modelling approaches which could be used to predict drug induction of some of the most serious toxicity endpoints: mutagenicity and carcinogenicity, hERG inhibition and Torsades de Pointes. Ideally, these approaches would be chemically interpretable, so as to guide lead optimisation/library design.

Chapter 2 presents an overview of the computational approaches which may be used for toxicity prediction, with a particular focus on those approaches which were explored in the research presented in subsequent chapters. Chapter 3 presents work undertaken to screen for mutagenic and carcinogenic drug-like compounds in the context of a collaboration resulting in the *in silico* identification of experimentally verified inhibitors of type II dehydroquinase (DHQase). Chapter 4 describes the development of models to identify potent inhibitors of the hERG ion channel and compares their performance to literature approaches on different datasets. Chapter 5 reports novel approaches to encoding 3D molecular structure for *in silico* modelling; their merits, compared to existing approaches to encoding molecular structure, are assessed by generating models for hERG inhibition and another biological endpoint. Chapter 6 presents an assessment of novel approaches to identifying TdP causing drugs based upon mechanistically relevant biological information.

Finally, Chapter 7 presents the conclusions drawn from the work presented in preceding chapters and offers suggestions for future research.

Additional results referred to in this thesis (omitted for brevity), along with source code and data files generated in the course of the research presented, have been made available on the DVD attached to the inside cover (see Appendix A). Appendix B summarises the performance of toxicity models previously presented in the literature, with which the models presented in this thesis are compared. Additional computational details, including versions of software employed where not specified in the main text, are provided in Appendix C.

Chapter 2 Computational Toxicology and Quantitative Structure-Activity Relationships

As discussed in Chapter 1, there is a clear need for *in silico* methods to predict the potential toxicological effects of pharmaceuticals and guide the efforts of medicinal chemists in reducing toxic liabilities. Computational toxicology is a field which seeks to develop such methods.

This chapter defines 'computational toxicology', presents a brief historical background to the field, followed by an overview of contemporary approaches to predicting toxicology *in silico*. Arguably, approaches which seek to predict the toxicological effects of compounds solely from their molecular structure are most desirable, and all computational work presented in subsequent chapters of this thesis is almost exclusively concerned with such methods - both expert systems and, in particular, Quantitative Structure-Activity Relationships (QSARs). The remainder of this chapter describes these approaches, with a focus on QSARs, in some detail.

2.1 Defining Computational Toxicology

Computational toxicology may broadly be defined as the application of "mathematical and computer models to predict adverse effects and to better understand the mechanism(s) through which a given chemical causes harm".[138] More narrowly, the field may be conceived as one which employs "computer technology and information processing (informatics) to analyse, model, and estimate chemical toxicity based upon structure activity relationships (SAR)",[139] albeit - as touched upon in section 2.3 - a wider range of *in silico* approaches are also currently employed for toxicity prediction.

2.2 Historical Background

The notion that chemical composition is a key determinant of toxicological properties is not new and was articulated by Aleksandr Borodin as early as 1858.[140] By 1868, Brown and Fraser were seeking to empirically determine relationships between molecular structure and biological effects,[141] a paradigm underpinning many of the methods employed in contemporary computational toxicology. Various studies in the late 19th century also sought to relate toxicological characteristics to experimentally determined properties of compounds.[140,142] In 1944, Lazarev published work which quantitatively related both the equilibrium constant for the partitioning of organic compounds between water and olive oil

and aqueous solubility to toxicological properties. Lazarev further presented limited attempts to relate changes in these physicochemical parameters to functional group contributions - as would allow for the prediction of toxicological effects from molecular structure alone. The rationale underpinning the empirical focus of Lazarev's work - that the complexity of the induction of toxicological effects obscured the mechanism of action[143] - remains a justification for empirical *in silico* models today.[19]

The seminal work of Hansch *et al.*,[144–146] in the early 1960s, presented, to the best of this author's knowledge, the first derivations of equations for bioactivities, including toxicological properties such as carcinogenic potency, based on contributions from *multiple* structural parameters.[146] Notably, they employed computers to derive these relationships,[144] and began to construct estimates of their, otherwise experimentally derived, substituent constants for compounds for which direct measurements were not obtained.[146,147] Around the same time, Free and Wilson employed the newly introduced electronic computers to derive regression models for the LD_{50}* values of chemical analogues, using the presence of specific side chains as predictors.[148] Whilst not coining the phrase, these studies arguably mark the advent of computational toxicology.

Whilst initial studies sought to generate relationships between toxicity and chemical structure for small sets of congeneric molecules, belonging to a "well-defined family of molecular structures .... [with different] substituents attached to a common molecular skeleton",[149] the development of software for predicting the toxicological effects of large, heterogeneous sets of compounds commenced in the late 1970s.[140] Today, a wide variety of predictive toxicology software packages exist which may be used to generate predictions for structurally diverse sets of compounds.[150]

2.3 Approaches to Predicting Toxicology *In Silico*

A variety of computational approaches are currently used to predict toxicity.[14,20,150] Some of these (typically) make predictions based solely on the molecular structure of the toxicant - for example, read across,[20,151,152] (Q)SARs[20,25,153] and expert systems.[42,140,154] Some require knowledge of the receptor via which the toxicant acts - for example, proteochemometrics,[155] as well as the "docking" of putative toxicants into the (putative) binding sites of 3D receptor models.[156,157] Others typically require various properties of the toxicant to be experimentally

* The median lethal dose.[140]

determined - for example, physiologically based pharmacokinetic and pharmacodynamic modelling.[104,107,158,159] Some predictive toxicology programs employ more than one approach to arrive at predictions; for example, so-called "hybrid" expert systems may also employ QSAR components.[154]

Detailed discussion of all the approaches currently employed in, and all the issues associated with, computational toxicology is beyond the scope of this thesis. The interested reader is directed to a number of recent reviews/texts for further discussion of these approaches, and other issues that are not covered herein.[14,19,20,150] Since this author was most interested in approaches based on toxicant molecular structure alone, read across, expert systems and QSARs are discussed in greater detail below.

2.4 Read Across

The "read across" approach supposes that chemically similar molecules should exhibit similar toxic effects for a given endpoint.[151] Indeed, this reflects the more generally supposed "molecular similarity principle", which holds that similar molecules should exhibit similar properties;[2] this principle informs the widespread use of similarity searching, in which molecules are ranked according to their similarity to "query" molecules, known to be desirably active against a protein target of interest, and the top ranking compounds are deemed most likely to be active against the same target.[2,20] Likewise, read across assesses a compound of interest for possible toxicity by considering the most similar molecule(s) known to be toxic.[20,151]

It must be noted, however, that the molecular similarity principle is not always observed to hold in practice. Indeed, whether or not this principle holds may be contingent on the manner in which molecules are determined to be similar.[2]

This emphasises that the molecular properties used to assess similarity are critical: read across commonly employs "obvious" chemical similarities which should be mechanistically relevant for the endpoint of interest.[20,151,152] For example, similarity might be assessed based on the common presence of particular functional groups.[20] Based on the importance of electrophilic reactivity for skin sensitization, Enoch *et al.* proposed a read across approach for assessing skin sensitization potency based on computing similarity in terms of an "electrophilicity index".[151]

2.5 Expert Systems

Expert systems "seek to emulate a human expert making predictions".[154] They comprise a "knowledge base", which typically encodes the generalized knowledge of human experts as a collection of rules and an "inference engine" governing how these rules are used to arrive at predictions for specific cases.[140,154]

The rules used in an expert system vary in their complexity. In the context of computational toxicology, a simple rule might be: if a molecule exhibits a given structural property, the compound will induce a specific endpoint. Rules of this kind may be used by the Toxtree program to make carcinogenicity predictions on the basis of 'structural alerts' for which there are no “modulating factors” (see below).[37] More complex rules are usually employed by the program Derek for Windows™,[160] for which the rules take the following general form:

If [grounds] is [threshold] then [proposition] is [force]

Here, "grounds" would refer to some variable providing evidence in favour of the "proposition" - with the value ("level of belief", e.g. "plausible", "doubted"') of the "grounds" denoted by "threshold", and the level of belief in the "proposition" denoted by "force". An example rule of this form is presented below.

The rules used to make predictions are commonly based on the presence of "structural alerts", or simply "alerts" (sets of substructures associated with a particular toxicity endpoint), as illustrated in Figure 2.1.[37,154,160] The presence of these substructures may be ignored, i.e. the alert will not "fire", in the presence of "modulating factors" - structural features deemed to diminish or abolish the toxicity conferred by the alert.[37] It is important to note that these alerts are solely capable of generating positive predictions for toxicity, and the absence of an alert is not an indication of a lack of toxicity - rather, a relevant alert for the compound's structural class may simply not have been compiled yet.[20,37]

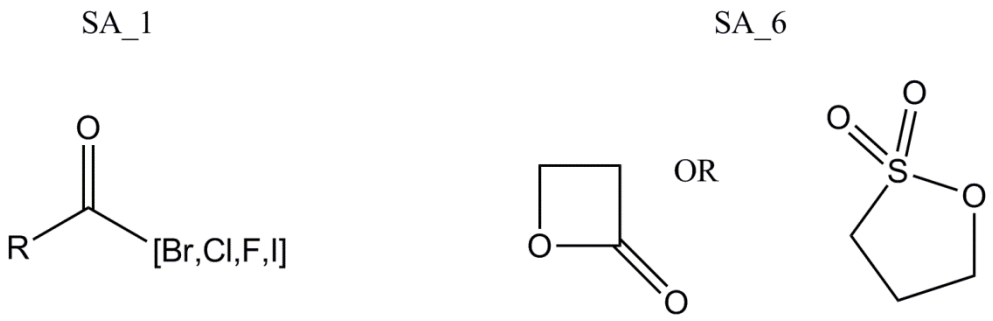


Figure 2.1 Two of the structural alerts for genotoxic carcinogenicity used in the "hybrid" expert system software program Toxtree. Here, R denotes any atom/group except OH or SH.[37]

These rules may also incorporate predicted physicochemical properties (e.g. the coefficient of skin permeability, K_p), as well as the predictions for mechanistically related toxicity endpoints. For example,* a possible rule employed in the expert system software program Derek for Windows™ (DfW) [160] would be:

If [logK_P < -5] is [certain] then [skin sensitization] is [species dependent variable 6]

An example of the type of processes applied by an inference engine would be the system of "argumentation" employed within DfW. Argumentation is used, for example, to propagate arguments along a chain, undercut arguments and resolve multiple arguments about the same proposition. In propagating arguments along a chain, uncertainties in the "grounds" (see above) are propagated as uncertainties in the proposition - e.g. a rule declaring that the "force" for a proposition was "certain" when a given "grounds" was "certain" could be used to infer that the force was "plausible" when the same "grounds" was "plausible". Undercutting arguments entails the introduction of context dependency into the strength of an argument in favour of a proposition - e.g. the "force" associated with a proposition based upon the threshold for a specific "grounds" might be dependent upon the species for which the toxic effects of the compound were of interest. In resolving multiple arguments for and against a proposition, DfW would combine the "forces" generated by multiple arguments to arrive at a single "force" - for example, arguments yielding conclusions of "probable" and "doubted" would yield a final conclusion of "plausible". Fuller explanation of this system, and relevant examples of its application, is provided in the referenced articles.[160,161]

* Also, see Chapter 3, section 3.2.2.1.

A variety of computational toxicology software programs employing expert systems were commercially (including Derek for Windows™)[160] and freely (including Toxtree[162,163] and OncoLogic™)[164] available at the time of writing.

Additional explanation of how the programs DfW and Toxtree make toxicity predictions, specifically for mutagenicity and carcinogenicity, is provided in Chapter 3.

2.6 Quantitative Structure-Activity Relationships

In this thesis, the term Quantitative Structure-Activity Relationship (QSAR) is used to denote any empirically derived relationship between molecular structure, encoded as a set of continuous or discrete numbers (descriptors), and bioactivity - either a continuous (e.g. pIC_{50}) or categorical value (e.g. toxic vs. non-toxic). Whilst this broad definition is consistent with the description of QSAR modelling recently presented by Tropsha,[153] and Worth,[25] this term may also be used more specifically, to denote approaches which seek to determine how bioactivity changes within an "island" of "chemical space" defined by some set of descriptors.[19]

The relationship expressed by a QSAR is determined (validated) using a training (test) set, requiring a set of molecules, encoded via descriptors, with observed bioactivities. The rest of this chapter will discuss the key issues associated with training and validating QSARs, along with examples of the approaches employed for these tasks - with particular emphasis upon those which were used in the work presented in subsequent chapters of this thesis.

2.6.1 Dataset Preparation

The first step in developing a QSAR model is the acquisition of a dataset comprising molecular structures (represented in an appropriate electronic format, as briefly discussed in section 2.6.2.1) and associated biological observations. Whilst researchers in the pharmaceutical industry may have access to experimentally consistent datasets comprising thousands of compounds,[15,94] QSAR researchers commonly work with published data[165] and there may be a necessary trade-off between the consistency of biological measurements and the requirement for a sufficiently large* dataset for robust model generation and validation.

* If building a "global" model (see section 2.6.3.1), this dataset must also be structurally diverse.

Datasets may be compiled from primary or secondary literature sources, and various 'benchmark' QSAR datasets have been made freely available in recent publications[55,166] (with the latter comprising measurements derived using the same experimental protocol).[166] QSAR datasets may also be downloaded in electronic formats from various* online repositories.[167] Various public databases also exist which serve as repositories of structures and/or associated biological data.[168–171]

Recently, however, concerns have been raised regarding the accuracy of the structural data contained within various public and private databases, with estimated error rates ranging between 0.1-10+%,[165,172,173] potentially having a significant effect on estimated QSAR performance.[165,172]

Another concern is the possibility of duplicate entries when combining data from different sources, which may overlap considerably;[174] unless these are identified and removed, the associated redundancy may skew QSAR analysis. If stereochemically indifferent descriptors are used, stereoisomers are effectively structural duplicates and should arguably be treated as such.[165]

The 'standardization' of chemical structures is an important step in the preparation of a QSAR dataset.[32,165] Different representations of the same (sub)structures are a problem for two reasons: (a) they may obstruct the identification of duplicate structures within the dataset and (b) the descriptors calculated for the same (or similar) structures will be inconsistent.[32,165]

One possible complication is the treatment of salts and multi-fragment compounds; whilst it is standard practice to remove counterions and, more generally, retain the largest fragment in multi-component dataset entries,[32,165] this may not be appropriate if there are reasons to believe either of the fragments could be significant for the observed bioactivity or when the bioactivities of salts differ significantly from their corresponding neutral form.[165,172]

The practice of removing additional compounds from the dataset, e.g. inorganics/organometallics, for which some software programs may have problems calculating descriptors,[165,175] and/or[175] compounds violating Lipinski's rules,[176] has also been

* Those linked to from the webpage cited here do not represent an exhaustive compilation.

advocated in the literature.* However, some types of descriptors may be computed for some types of inorganics/organometallics, and analysis a few years ago by Overington *et al.* suggested that around 25% of small molecule drugs (not including "biological drugs") were not compliant with Lipinski's rules.[3] This further serves to emphasise that the most appropriate pre-processing of structures within and removal of entries from a QSAR dataset is non-trivial - different procedures may be appropriate under different circumstances.[165]

Additional dataset preparation steps will also depend upon the descriptors to be calculated. For example, the calculation of 3D descriptors (as discussed in section 2.6.2.2) requires - presuming these are not already pre-computed or experimentally available - the generation of an appropriate 3D structure.

2.6.2 Computer Representation of Chemical Structures

2.6.2.1 Chemical Structure Formats

Prior to the computation of descriptors, chemical structures must be represented in an electronic format. A plethora of such formats exist.[177] Those which were principally employed in the work presented in this thesis were: the IUPAC International Chemical Identifier (InChI),[178] the Simplified Molecular Input Line Entry System (SMILES),[179] the Structure-Data file[180] (SDF) and Tripos MOL2 file formats.[181]

2.6.2.2 Descriptors

A plethora of molecular descriptors, or numerical encodings of molecular structure "obtained by a well-specified algorithm applied to a defined molecular representation or a well-specified experimental procedure", exist; many of these were reviewed in the year 2000 by Todeschini and Consonni[182] and the number of descriptors continues to grow.[183] Indeed, much of the work presented in this thesis – from the development of a new approach to encoding descriptors for use with Nigsch's version of the Winnow algorithm[184] in Chapter 4,

* Originally designed to alert chemists to compounds with potentially undesirable absorption or permeation, in a drug discovery or development context,[176] compliance with Lipinski's rules may also be used as a surrogate for "drug-likeness"[175] (see Chapter 3, section 3.1).

the 3D descriptors proposed in Chapter 5 and the *in silico* biological descriptors presented in Chapter 6 – is concerned with novel descriptors.

Computationally derived descriptors are commonly grouped into the following categories. Those which can be computed solely from the molecular formula, such as molecular weight and elemental atom counts, may be denoted "0D descriptors". The term "1D descriptors" may be used to denote descriptors encoding the list of substructures present in a molecule, with "2D descriptors" corresponding to those based upon the precise connectivity (i.e. topology) of the entire molecular structure.[182] Alternatively, as is accepted in this thesis, descriptors computed from the molecular formula and all descriptors calculated from molecular connectivity (i.e. including the presence/absence of substructures) may be termed 1D and 2D descriptors respectively.[20] Some 2D descriptors may be augmented with stereochemical information, even in the absence of 3D geometrical information;[185] the term 'topological descriptors' is used in this thesis to denote 2D descriptors which do not encode such information. Descriptors calculated from the fixed three-dimensional co-ordinates of a single molecular conformation are termed "3D descriptors".[182,186] The so-called "4D descriptors" are based upon an ensemble of conformations;[20,186] again, this label is defined differently elsewhere.[182]

Recent studies have proposed using the output of short-term biological assays as descriptors for predicting long-term biological responses.[187–189] Further recent studies have proposed that *in vitro* assay information,[128,190] or protein target predictions,[191] be used to inform purely *in silico* QSAR models of *in vivo* toxicity. Indeed, the use of predicted *in vitro* bioactivities as descriptors for predicting Torsades de Pointes is explored in Chapter 6 of this thesis.

2.6.3 QSAR Modelling Methods

2.6.3.1 General Concepts

Regression and Classification

QSAR models are designed to predict continuous or categorical measures of bioactivity, denoted regression or classification models respectively;[192] a further distinction may be made between "hard point" classifiers, which solely aim to map an instance (i.e. a molecule encoded via descriptors) onto a class label, and probabilistic classifiers which estimate

probabilities of class membership. Here, probability may be viewed in a Bayesian sense - i.e. as a quantification of the degree of confidence in a particular belief.[193,194]

Probabilistic classifiers are advantageous when the costs of incorrect class assignment are class dependent and subject to change;[193] for example, it is arguably the case that, in early drug development, an incorrect prediction of toxicity is worse than incorrectly predicting a candidate as non-toxic, whereas the converse is true in a regulatory environment.[195]

It must be noted, however, that a number of QSAR modelling methods may effectively be used for both regression and classification - for example, Random Forest (RF),[192] Artificial Neural Networks (ANNs)[196] and Partial Least Squares (PLS).[192]

Linear vs. Non-Linear Models

In linear models, the output of the model (either the predicted bioactivity for regression models or the output of a classifier - thresholds for which are used for class assignment), takes the following general form[193,197] (Equation 2.1).

$$y_{pred} = \sum_{m=1}^{M} w_m x_m + w_0 \qquad 2.1$$

In Equation 2.1, y_{pred} is the model output, x_m and w_m the mth descriptor (out of a total of M) and corresponding coefficient (weight) respectively, and w_0 an offset term.

Linear modelling methods have certain potential advantages over non-linear methods which do not assume this functional relationship. Firstly, the contributions of individual descriptors to the predictions are readily apparent, making them potentially more interpretable. Secondly, the computational overhead of linear methods may be lower than corresponding non-linear approaches - for example, when considering the choice of kernel function (see below) for a Support Vector Machine (SVM).[198] However, many structure-activity relationships, particularly those based on multiple mechanisms,[199] are likely to be non-linear - hence purely linear modelling methods could yield suboptimal predictivity. However, some non-linear relationships may be captured by linear modelling methods via generating additional descriptors based upon combinations of the original descriptors,[184] or by raising the original descriptors to the power n ($n \geq 2$).[146]

Global vs. Local Models

"Local" models are based on structurally non-diverse sets of compounds, often on a single congeneric series,[200] and are expected to have limited applicability.[201] "Global" models, as developed in the work presented in this thesis, are based on a diverse range of chemical series and are required for screening diverse chemical libraries.[15]

Overfitting

This phenomenon may be defined as a scenario under which a model "fits the idiosyncrasies of a particular training set at the expense of the predictivity of a similar set of molecules"[202] - for example, by learning the noise (random experimental error) as well as the signal in the training set.

Hyperparameters

QSAR modelling methods may be associated with adjustable "hyperparameters" - parameters which control the distribution of other parameters;[193] in the current context, they control the relationship between bioactivity and molecular descriptors learnt from the training set. For some methods (e.g. the Support Vector Machine, described below), these must be carefully selected.[203]

2.6.3.2 Examples

A plethora of methods may be employed to generate a QSAR model. Some commonly employed examples are: Multiple Linear Regression (MLR), Partial Least Squares (PLS),[197] Recursive Partitioning (or Decision Tree),[91,192,204,205] Artificial Neural Networks (ANNs),[193,196,206] k-Nearest Neighbours (kNN),[207,208] Support Vector Machines (SVMs),[193,203,209,210] Random Forest[192,204,211] and Naive Bayes.[193,212–214] It should be noted that this list is far from exhaustive - additional methods are presented below and new approaches are regularly proposed in the literature.[194,215] The following presents a detailed discussion of those methods which were used in the work presented in subsequent chapters of this thesis.

Winnow

The Winnow algorithm, originally developed by Littlestone,[216] was recently adapted for QSAR generation by Nigsch.[184] The following description refers to the version of this algorithm implemented by Nigsch[184,213] which was, with the minor addition of multiple training cycles (see below), employed in the work presented in Chapter 4 of this thesis.

Instances are presented to Winnow as a set of 'features' – text strings which may either be present or absent in an instance.* For each class (C), Winnow holds S independently trained weight vectors (scorers) with positive elements (W_C^m) for all M features seen during training, used to generate S scores (S_C^i) (Equation 2.2), for a given instance (i), for each class. The predicted class is the class with the highest arithmetic mean score.

$$S_C^i = \sum_{m=1}^{M} W_C^m \cdot F_m^i \qquad 2.2$$

In Equation 2.2, F_m^i is 1 (0) if the mth feature is present (absent) in instance i, and all M values of F_m^i may be considered a set of binary descriptors. The features themselves could correspond to the 'on-bits' of a fingerprint descriptor (as proposed by Nigsch)[184,213] or, as proposed in Chapter 4 of this thesis, to discretized descriptors.

The training instances are presented sequentially to Winnow, in a predefined yet randomised order, with S training instances held in memory, and randomly distributed amongst the scorers, for a given training iteration; for the next training iteration, the first training instance which entered the "first-in first-out cache" is removed from memory and the next training instance added. Since only a subset of the training data is held in memory at any point in time, this "on-line" algorithm is highly memory efficient during training;[184] however, as illustrated in Chapter 4 of this thesis, this does mean the exact model learnt by Winnow is dependent upon the exact order in which the training set instances are presented.

Nigsch originally proposed presenting each training set instance once, for inclusion in the cache, such that training ceased after the last training set instance was added to the cache and all scorers were subsequently updated (i.e. after a 'single training cycle').[184] In the work presented in Chapter 4 of this thesis, however, the repeated presentation of the training set (in the same order), i.e. multiple training cycles, was also considered. The presentation of the training set in different randomised orders, as commonly used to train ANNs,[196] would arguably yield better results, but this is harder to implement computationally.

* More generally, the term "features" may be used interchangeably with "descriptors" in the QSAR literature, and acquires a distinct meaning again in the context of discussing SVM – see below.[209]

The classwise feature weights for each scorer, initially set to unity, are updated via the following "error-driven" procedure.[184] For each scorer, upon encountering instance *i* during training, S_C^i is calculated, for all classes. The weights of all features in instance *i* are decreased when S_{C*}^i, the score for the wrong class ($C*$), exceeds $\sum_m^M F_m^i$ (the number of features in instance *i*) and when S_{C*}^i does not exceed this threshold, yet $|S_{C*}^i - \sum_m^M F_m^i| \leq \varepsilon \cdot \sum_m^M F_m^i$. The weights of all features in instance *i* are increased when S_{C**}^i, the score for the correct class ($C**$), does not exceed the threshold and when S_{C**}^i does exceed the threshold, yet $|S_{C**}^i - \sum_m^M F_m^i| \leq \varepsilon \cdot \sum_m^M F_m^i$, where $0 \leq \varepsilon < 1$. A non-zero value for ε, forcing weight updates even when the correct class of the current training instance is relatively highly scored, may make the algorithm more robust.[184]

The weights of all features in instance *i* are increased via multiplying by a "promotion" factor (p), where $1 < p < p_{max}$, or decreased by multiplying using a "demotion" factor (d), where $d_{min} < d < 1$. Nigsch initially proposed variable promotion and demotion factors, with p_{max} and d_{min} being positive constants,[184] albeit constant positive values may also be considered as per the work presented in Chapter 4 of this thesis.

Earlier work by Nigsch indicated that this algorithm, under some circumstances, could perform comparably to Random Forest[184] and a Laplacian modified Naive Bayes classifier.[213] The work presented in Chapter 4 of this thesis indicates that, under some circumstances, this algorithm may perform comparably to SVMs as well.

Whilst a linear algorithm, Nigsch's implementation is able to capture non-linearity in the original feature space. This is achieved by constructing additional features, termed "orthogonal sparse bigrams" (OSBs), via non-exhaustive pairing of the original features present in a molecule as previously described by Nigsch.[184,213]

Linear Discriminant Analysis

Linear classifiers may also be derived from Linear Discriminant Analysis (LDA). A variety of approaches to determining a linear discriminant separating two classes (i.e. a model with the form of Equation 2.1 along with a threshold delimiting predictions for one class from predictions for the other) exist.[193] However, "LDA" is usually used[217] to denote the determination of a linear discriminant via adjusting the coefficients of the descriptors in order to maximise the Fisher criterion[193,217,218] – designed to maximise the ratio of between-class-variance to within-class-variance within the training set.[193,217,219]

Binary classifiers derived from canonical discriminant analysis, for mutagenicity and carcinogenicity are incorporated into the Toxtree progam,[37] and hence were considered when generating predictions for mutagenicity and carcinogenicity in the work presented in Chapter 3. For binary classification, since the class means are co-linear, canonical discriminant analysis corresponds to LDA.[220]

Naive Bayes

This approach estimates the (relative) *posterior* probabilities of class membership (i.e. the probabilities given the descriptor values), assuming the class conditional descriptor distributions, denoted $p(x_1, x_2, \dots, x_M|C_k)$, are given by the product of the class conditional distribution of the individual descriptors, denoted $p(x_m|C_k)$. From Bayes' theorem, it follows that this assumption allows for the *posterior* probability of class C_k, denoted $p(C_k|x_1, x_2, \dots, x_M)$, to be expressed as follows (Equation 2.3):[193]

$$p(C_k|x_1, x_2, \dots, x_M) \propto p(C_k) \prod_m^M p(x_m|C_k) \qquad 2.3$$

In Equation 2.3, $p(C_k)$ is the *prior* probability of membership of class C_k and the constant of proportionality is the same across all classes.

The Binary QSAR methodology, developed for binary classification,[214] was proposed by Thai and Ecker[69] for generating predictive models for hERG inhibition, and their approach was compared to novel approaches in the work presented in Chapter 4. Essentially, this methodology employs a Naive Bayes approach to estimate the posterior probability of a compound belonging to the "active" class.[214] The methodology can work with continuous descriptors - specifically, the descriptors used are principal components derived from the original descriptors (see section 2.6.6) - via estimating $p(x_m|C_k)$ for the mth retained[69] principal component. The priors are estimated from the training data using modified "active"/"inactive" occurrence counts, which do not tend to zero in the limit of there being few "actives"/"inactives" in the training set.[214]

Support Vector Machines

A Support Vector Machine (SVM) determines a linear "decision boundary" (or "hyperplane") which is designed to separate molecules, represented as instances located in a “feature

space", belonging to one class from molecules belonging to the other possible class. Hence, this decision boundary may be used for binary classification. The "feature space" referred to may correspond to the space defined by the original descriptors, or a non-linear projection of this space – generating a linear or non-linear classifier respectively.

Mathematically, the decision boundary is defined as $y(\underline{x}) = 0$, with molecules assigned to, say, the 'toxic' ('non-toxic') class when $y(\underline{x}) > 0$ $(y(\underline{x}) < 0)$ with $\underline{x}$ representing a descriptor vector, where $y(\underline{x})$ is computed as per Equation 2.4.

$$y(\underline{x}) = \sum_{m}^{M} w_m \emptyset_m(\underline{x}) + b = \underline{W} \cdot \underline{\emptyset}(\underline{x}) + b \quad 2.4$$

In Equation 2.4, $\underline{\emptyset}(\underline{x})$ represents a vector in the feature space, b represents a positive or negative scalar (the "bias"), and $\underline{W}$ is the normal to the hyperplane.

If the training set is linearly separable - i.e. all training set data points in, say, the 'toxic' class can be perfectly separated from all training set data points in the 'non-toxic' class by some hyperplane(s) - *in the feature space*, an SVM finds the "maximum margin" hyperplane in order to minimise overfitting;* here, the margin is defined as the perpendicular distance between the hyperplane and the closest training set data point.[193,209,210]

If the dimensions of the feature space are scaled such that, by definition,

$|\underline{W} \cdot \underline{\emptyset}(\underline{x}) + b| = 1$ (i.e. $y(\underline{x}) = \pm 1$) for the training set instances lying closest to the perfectly separating hyperplane, the margin is given by $|\underline{W}|^{-1}$ and maximisation of the margin corresponds to minimisation of $|\underline{W}|^2$.[193]

If the training set is not linearly separable in the feature space, SVMs allow for some degree of misclassification by introducing the "slack-variables". Training set instances associated with non-zero slack-variables may be misclassified and the margin is now defined in terms of the closest training set instances associated with zero-valued slack-variables. During training, the need to maximise the margin, in order to limit overfitting, is balanced against the need to

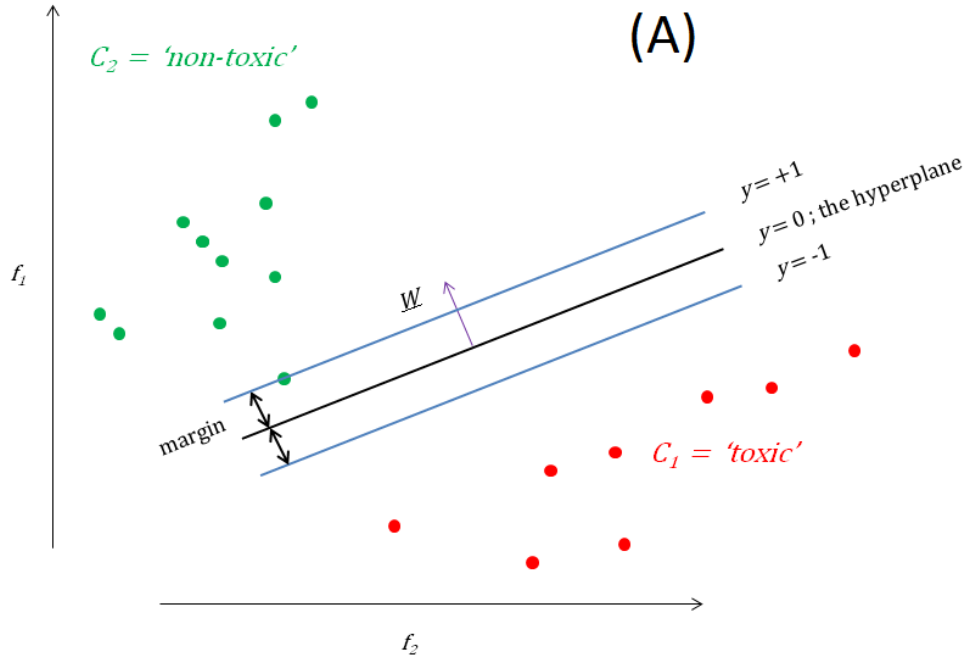
* If the feature space corresponds to a highly non-linear projection of the descriptor space, considerable overfitting may nonetheless occur. The degree of non-linearity of this projection is typically controlled by a kernel parameter, emphasising the importance of SVM hyperparameter selection (see below).[203,209]

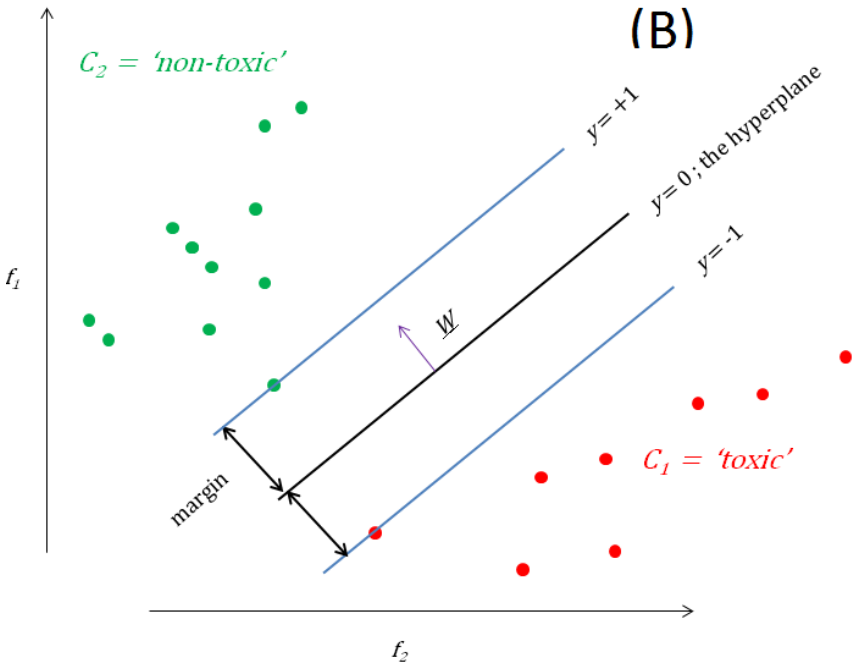
minimise the extent of training set misclassification, such that minimisation of the following expression (Equation 2.5) is attempted.[193,209,210]

$$\min_{\underline{W},b,\underline{\varepsilon}} \frac{1}{2}\left|\underline{W}\right|^2 + C\sum_{i}^{N}\varepsilon_i \qquad 2.5$$

In Equation 2.5, ε_i is the slack-variable for the ith training set compound (a positive scalar which increases as the compound moves further away from the 'wrong side' of the hyperplane), $\underline{\varepsilon}$ the vector of all slack-variables, and C is the "regularization constant",[210] a hyperparameter[203] which determines the trade-off between the minimisation of misclassification (RHS of Equation 2.5) and the maximisation of the margin (LHS of Equation 2.5) during training.[210]

Figure 2.2 presents an overview of this classification procedure.


(A)
C_2 = 'non-toxic'
f_1
$y = +1$
$y = 0$; the hyperplane
$y = -1$
W
margin
C_1 = 'toxic'
f_2


(B)
C_2 = 'non-toxic'
f_1
$y = +1$
$y = 0$; the hyperplane
$y = -1$
W
margin
C_1 = 'toxic'
f_2

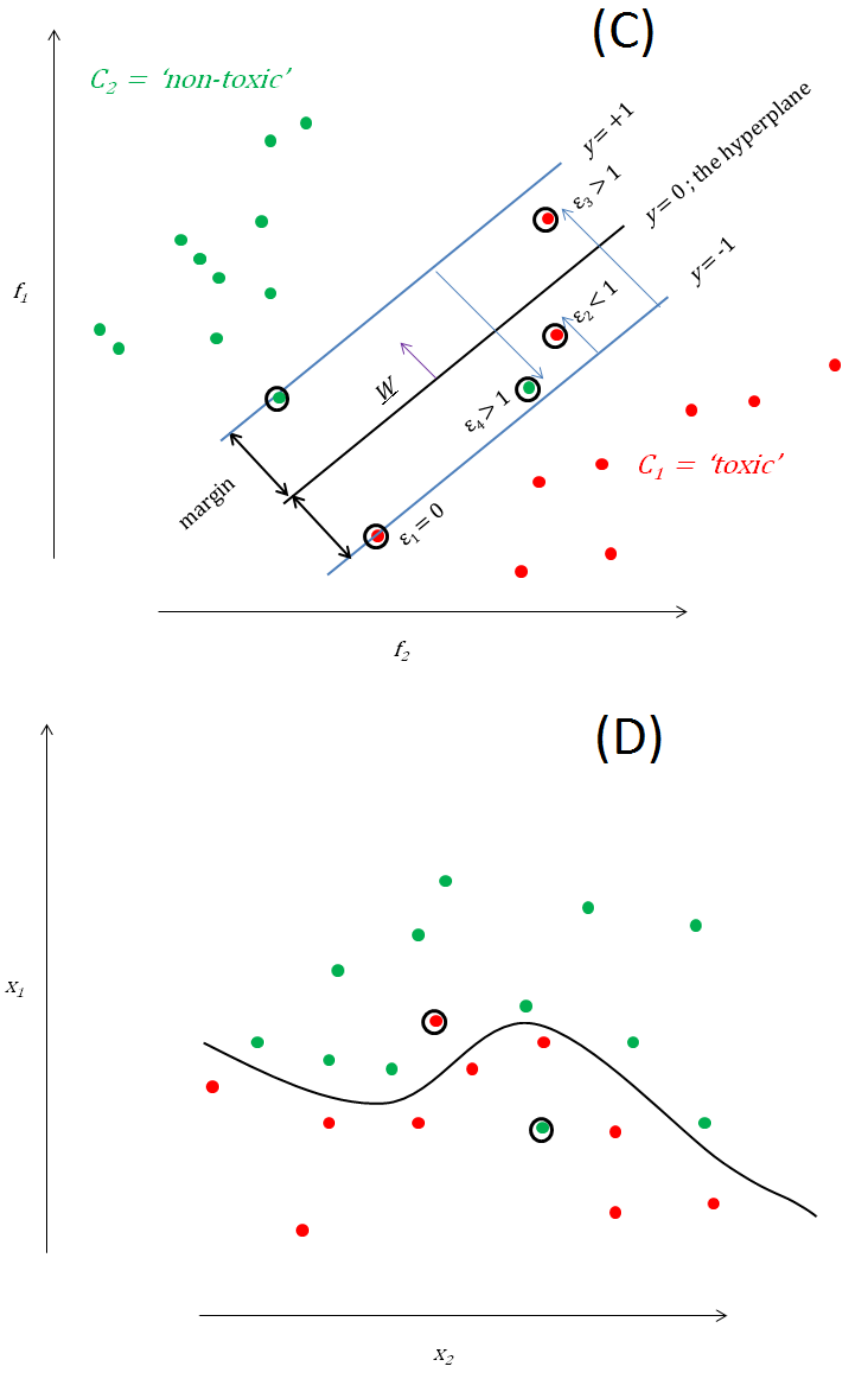


Figure 2.2 An overview of SVM classification. (A) A linearly separable dataset in the feature space, and a possible separating hyperplane. (B) The SVM solution for such a dataset: the maximum margin hyperplane. (C) A non-linearly separable dataset; highlighted are two misclassified instances and two further instances lying inside the margin, with their corresponding slack-variables. (D) A conceivable corresponding decision boundary (as shown in (C)) in the descriptor space (supposing the feature space is a higher dimensional projection of the descriptor space); only the two misclassified instances are highlighted. N.B.: These images are for illustrative purposes only.

Obtaining the SVM solution of Equation 2.5, and generating predictions via Equation 2.4, does not require explicit generation of the feature space. Rather, all that is required is the determination of dot-products in this space, which may be computed (Equation 2.6) from corresponding vectors in the descriptor space using a kernel function.[210]

$$\underline{\emptyset}(\underline{x}) \cdot \underline{\emptyset}(\underline{x}') = k(\underline{x}, \underline{x}') \qquad 2.6$$

In Equation 2.6, $\underline{x}$ and $\underline{x}'$ are vectors in the descriptor space and $k( )$ is the kernel function. The linear kernel, $k(\underline{x}, \underline{x}') = \underline{x} \cdot \underline{x}'$, corresponds to a dot-product in the descriptor space and yields a linear classifier.[193] A popular choice of non-linear kernel in the QSAR community[209] is the Gaussian Radial Basis Function (RBF) kernel (Equation 2.7).[221] This kernel was used to generate non-linear SVM models in the work presented in Chapter 4 of this thesis.

$$k(\underline{x}, \underline{x}') = exp\left(-\gamma |\underline{x} - \underline{x}'|^2\right) \qquad 2.7$$

In Equation 2.7, γ is an additional hyperparameter controlling the degree of non-linearity (i.e. the degree of "flexibility")[203] of the SVM model.

Different variants of the basic SVM classification procedure outlined above exist,[193,209,210] as well as an adaptation of the SVM approach for regression: Support Vector Regression (SVR).[193,209,210,222]

Recursive Partitioning

This approach is used to construct a single Decision Tree model from the training set. Starting from the entire training set (the "root node"), each descriptor is searched for "cutpoints" which partition the training set compounds at the current "parent node" into K "daughter nodes", such that the separation in the experimental bioactivities of the subgroups of the data passed to the daughter nodes is maximised according to some measure of separation.[204] For classification, this measure might be the mean decrease in Gini impurity,[223] or a t-test might be used for continuous bioactivities.[224]

Commonly, $K = 2$ - i.e. only one split criterion is sought per descriptor (e.g. $x_1 > A$, or $x_1 \leq A$).[204] Cutpoints might also be selected for linear or non-linear combinations of descriptors.[205]

Partitioning continues until some stopping criterion (e.g. all compounds in the current node belong to the same class) is met. Predictions are generated by passing compounds through the

tree, and assigning the majority class or the average bioactivity value for training set compounds in the final ("leaf") node.[204]

Recursive Partitioning is notably prone to overfitting;[205] even small changes to the training set could yield changes to one cutpoint and have the subsequent effect of appreciably changing the structure of the Decision Tree.[204] Overfitting may be limited to some extent via pruning[204] - removing branches from the fully grown tree, with the optimal depth of the tree determined using internal validation (see section 2.6.5) on the training set.[91]

The QuaSAR-Classify approaches proposed by Dubus *et al.*[91] for the generation of hERG blocker classifiers, which were further evaluated in the work presented in Chapter 4, are based on a Recursive Partitioning algorithm.

Random Forest

As proposed by Breiman,[211] this method grows a forest of unpruned decision trees, each trained on independent, random *bootstrap* samples (i.e. select *N* from *N* with replacement) of the training set, with the cutpoints at each node selected from independently, and randomly, chosen subsets of the descriptors.

For a new molecule, the forest makes predictions via aggregating the predictions for each tree (either via majority voting for classification or via averaging for regression). Since, approximately 1/3 of training set instances are expected not to be selected for the generation of any given tree, an estimation of the forest's predictivity may be made on the training set by only aggregating the predictions, for a given training set molecule, of those trees for which the molecule was "out-of-bag" (OOB), i.e. not selected.[192]

Whilst other decision tree algorithms could be employed, the standard Random Forest algorithm (as implemented in the randomForest package in the R Statistical Programming language)[225] generates decision trees using the Classification and Regression Trees (CART) algorithm.[192] CART uses a weighted sum of the Gini impurity[223] (variance)[226] over both daughter nodes for split selection when employed for classification (regression).[192,223,227,228]

Importantly, Random Forest commonly performs well (albeit, not necessarily optimally) 'off-the-shelf' - i.e. with a set of default hyperparameters. Here, the hyperparameters refer to the number of trees in the forest (n_{tree}) and the number of descriptors randomly sub-sampled at each node (m_{try}).[192]

The behaviour of the forest is expected to converge[192,204,211] as n_{tree} increases (indeed, Breiman proved convergence of the expected error rate for Random Forest classifiers as n_{tree} tends to infinity),[211] with Svetnik *et al.* suggesting that 500 trees are usually sufficient for many purposes.[192]

Due to its ability to capture many non-linear relationships, and approximate linear decision boundaries,[204] its good performance 'off-the-shelf', its inbuilt OOB estimates of model performance and its readily computed[192,223] - albeit, imperfect[227,229] - variable importance measures, the randomForest implementation of Random Forest was used extensively in the work presented in subsequent chapters of this thesis to generate both classification (Chapter 4, Chapter 5 and Chapter 6) and regression (Chapter 5 and Chapter 6) QSARs. For the versions of randomForest employed in the work presented in this thesis, the default m_{try} ($M/3$ and $\sqrt{M}$ for regression and classification respectively, where M is the number of descriptors, rounded down to the nearest nonzero integer) and n_{tree} (500) values corresponded to those proposed by Svetnik *et al.*[192]

2.6.4 Quantifying the Predictive Power of QSAR Models

In order to summarise the predictive performance of a QSAR model, a variety of statistics, or figures of merit (FOMs), may be employed - each of them with their own strengths and weaknesses. N.B.: Often, in assessing the predictive performance of a modelling approach, multiple values might be computed for a figure of merit (FOM) and summarised using the arithmetic mean; all references to mean values for a figure of merit in subsequent chapters should be understood to refer to the arithmetic mean.

2.6.4.1 Figures of Merit for Classification

The starting point for the assessment of QSAR classifiers is typically the "confusion matrix" ($\underline{C}$),* comprising elements C_{kl} denoting the number of compounds in class k predicted to belong to class l.[230] The following figures of merit are all based on the confusion matrix.

A commonly reported measure of overall performance is "accuracy"[126,128,231] (or the "fraction of correct predictions"),[184] the number of correct predictions divided by the total number of

* Alternatively, this matrix being an example of a "contingency table",[94] the "contingency matrix".[230]

compounds for which predictions are made.[231] However, this is a poor measure of performance - particularly when the compounds to be predicted belong disproportionately to one class (i.e. the data is "unbalanced"); for example, if 99% of compounds belong to a single class, then a totally non-discriminative classifier which predicted everything to belong to that class would appear to be highly predictive - with an accuracy of 99%.[232]

For binary classification, with a 'positive' and 'negative' class, the confusion matrix simplifies as follows (Table 2.1).

		Observed	
		Positive	Negative
Predicted	Positive	Number of true positives (TP)	Number of false positives (FP)
	Negative	Number of false negatives (FN)	Number of true negatives (TN)

Table 2.1 The confusion matrix for a binary classifier.

The Matthews Correlation Coefficient (MCC) is an appropriate FOM for assessing the overall discriminative ability of a binary classifier (Equation 2.8), which assigns equal credit (equally penalises) correct (incorrect) predictions for either class. The MCC takes values between negative one (all predictions are incorrect) and one (all predictions are correct), whilst random assignment of the class labels would have an expectation value of zero - this value also being obtained in the limit that a classifier predicted all compounds to belong to a single class. Hence, the MCC is more appropriate than accuracy for ranking classifiers predicting unbalanced validation data and may be considered a measure of the extent to which the model would exceed the performance of a random predictor.[230]

Indeed, as noted by Baldi,[230] the MCC is related to the chi-squared test-statistic (χ^2) for independence (the "null hypothesis", corresponding here to a random predictor) for a 2 ×2 contingency table (Equation 2.9).[233] Assuming a chi-squared distribution, and supposing the validation data is large enough for the degrees of freedom to be taken as one, a "p-value" (i.e.

the probability of obtaining a test-statistic at least as large given the null hypothesis)* may readily be calculated.[233]

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} \; - \; \mathrm{FP} \times \; \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \qquad 2.8$$

$$|\mathrm{MCC}| = \sqrt{\frac{\chi^2}{N'}} \qquad 2.9$$

In Equation 2.9, N' is the total number of compounds used for model validation.

Given the useful properties of the MCC outlined above, this figure of merit was extensively used to summarise the *overall* performance of the binary classifiers developed and/or assessed in the work presented in subsequent chapters of this thesis. However, any single FOM loses information with respect to the complete confusion matrix - i.e. there is not a one-to-one mapping between the confusion matrix and any FOM. Moreover, the MCC may be relatively high when few compounds are predicted to belong to one class.[230]

Cohen's kappa[234] has similar properties to the MCC and may be used to assess classifier predictivity for an unlimited number of classes. Gorodkin also presented an extension of the MCC to multiple classes.[232]

The performance of the model for individual classes may be assessed in terms of the recall and precision (which may further be combined into a single measure - the F Measure). These are defined below[175] (for the 'positive' class).† For the binary classification problem toxic vs. non-toxic, recall of toxic compounds and non-toxic compounds may be termed sensitivity and specificity respectively[195]

---

* As noted in section 2.6.4.2, p-values can be computed for other figures of merit (given the null hypothesis of a random predictor). Comparison of these p-values (lower p-values offering stronger grounds for rejecting the null hypothesis) might be more appropriate than direct comparison of these figures of merit when comparing the performance of models assessed on validation data comprising different numbers of compounds.

† Clearly, the expressions for any other class are analogous.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad 2.10$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad 2.11$$

$$\text{F Measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \qquad 2.12$$

2.6.4.2 Figures of Merit for Regression

For regression models, the Pearson's correlation coefficient (*r*), the coefficient of determination (R^2) and the Root Mean Square Error (RMSE) are commonly computed.[235] The common expressions,[236] used throughout in the work presented in this thesis, for these are presented in the following equations. Spearman's rank-correlation coefficient (ρ) was also computed (Equation 2.15).[226,237]

$$r = \frac{\sum_i^{N'} \left(y_{i,experi} - \overline{y_{i,experi}}\right) \left(y_{i,model} - \overline{y_{i,model}}\right)}{\sqrt{\sum_i^{N'} \left(y_{i,experi} - \overline{y_{i,experi}}\right)^2 \sum_i^{N'} \left(y_{i,model} - \overline{y_{i,model}}\right)^2}} \qquad 2.13$$

$$R^2 = 1 - \frac{\sum_i^{N'} (y_{i,experi} - y_{i,model})^2}{\sum_i^{N'} \left(y_{i,experi} - \overline{y_{i,experi}}\right)^2} \qquad 2.14$$

$$\rho = 1 - \frac{6 \sum_i^{N'} d_i^2}{N' \left(N'^2 - 1\right)} \qquad 2.15$$

$$\text{RMSE} = \sqrt{\frac{\sum_i^{N'} (y_{i,experi} - y_{i,model})^2}{N'}} \qquad 2.16$$

In these preceding equations, N' denotes the number of compounds for which predictions are made, $y_{i,experi}$ and $y_{i,model}$ denote the experimentally measured and predicted bioactivity for the ith compound with $\overline{y_{i,experi}}$ and $\overline{y_{i,model}}$ denoting their respective means. Finally, d_i denotes the difference in ranks assigned to compound i when the compounds used for validation are ranked according to their predicted and observed bioactivities.

These statistics have their advantages and disadvantages.[186] For example, the RMSE may appear low (i.e. the model may misleadingly appear predictive) if the model simply predicts all compounds to have the mean bioactivity of the training set (supposing this is close to the mean bioactivity of the test set and the test set's bioactivity values are distributed over a narrow range).[238] Additionally, neither the Pearson's nor Spearman's correlation coefficients take account of non-zero model bias (i.e. when $\frac{1}{N'}\sum_i^{N'}\left(y_{i,experi} - y_{i,model}\right) \neq 0$).[239] However, for the Pearson's and Spearman's correlation coefficients, statistical tests exist for determining the probability of observing a correlation coefficient at least as positive* as that obtained given the null hypothesis of random predictions[226,237] (i.e. a p-value).†

2.6.5 Model Validation

It is important that unbiased estimations of model performance are made. The compounds used for model validation should not have been used to train the model.[240] When validation data is not used to directly train the model, yet has been used to select hyperparameters or descriptors (selection of the latter commonly termed "feature selection"),[69,91] model performance is expected to be overestimated.[200,207,241–243] The unbiased assessment of a model on data not used for training nor model selection (for example, hyperparameter or descriptor selection) may be termed "external" validation (as opposed to "internal" validation).[153] In this

* N.B.: Here, the alternative hypothesis (accepted when the null hypothesis is rejected) is that the correlation coefficient expected for the "population" (in our case, all possible validation set molecules) is greater than zero (the value expected for the null hypothesis).[226] This contrasts with the statistical test proposed for the MCC in section 2.6.4.1, for which - strictly speaking - this null hypothesis can only be rejected in favour of concluding that the classifier is not a random predictor.

thesis, 'internal' validation is used to denote model assessment on data not directly used for training, but still used to guide model development via hyperparameter or feature selection.

As well as using a single distinct test set (a "holdout" sample) to validate a model, cross-validation, based on repeatedly training a model on a subset of the original training set and making predictions for the remaining compounds (see below for details), might yield more robust estimates of model performance if few compounds are available for the holdout sample.[243] Cross-validation is often viewed, by definition, as an "internal" validation method.[186,200] However, if all steps of model generation (including hyperparameter and feature selection) are repeated for each training set, cross-validation estimates of model performance need not be optimistically biased;[243] under such circumstances, one may speak of "external cross-validation".[153,189]

Arguably, a further distinction may be made here between attempts to estimate, without bias, the performance of:

1. A single model.
2. A modelling approach (e.g. a specific QSAR modelling method and/or descriptor set).

In the first scenario, "external cross-validation" may be carried out on the training set eventually used in its entirety to build the final model. Hence, one may consider assessment on a holdout sample to be external in a profounder sense. In the second, one is not interested in pooling all the data used for cross-validation to train a single model, hence this consideration does not apply.

Cross-validation (CV) typically, albeit not always,[244] randomly partitions a dataset into several training and test sets and estimates the performance of a modelling approach across the partitions. In K-fold CV, a dataset is partitioned into K disjoint sets (folds) of comparable size, and each fold is used in turn to validate the model, with all remaining instances used for training; leave-one out CV (LOOCV) entails setting K to N - the size of the dataset - i.e. one compound is used for model validation, and the rest for training, at a time. The performance statistics commonly used are the mean FOMs across all K folds,[240,245] albeit the predictions might be pooled across all folds prior to computing a FOM.[246] In Monte-Carlo CV (MCCV), a dataset is repeatedly, independently randomly partitioned into a training and test set (using a fixed percentage of the data for testing) and the performance of the modelling approach is assessed by computing the mean FOMs, obtained on the test set, across all repetitions.[245] In stratified CV, an attempt is made to maintain the ratio between the classes in the entire

dataset across all partitions.[240,246] Alternative approaches for partitioning the dataset into training and validation data may also be considered.[240]

As well as assessing model predictivity on the original, available data, one might also employ "y-randomisation", or "(y-)scrambling" for model assessment.[153,186,247] This entails the randomisation (typically, via permutation) of the bioactivities within a dataset, and an assessment of the *expected* performance of models (estimated from multiple repetitions of the procedure) obtained by applying the original *modelling procedure* (including hyperparameter and feature selection) to the permuted dataset.[247] Model performance on the training and/or test data may be compared[186,189] with some authors having applied a model trained on randomised data to non-randomised test data.[222] and a reduction in estimated model performance after scrambling may be interpreted as indicating that the QSAR model(s) are not based on capturing chance correlations within the dataset.[189]

2.6.6 Applicability Domain

Since the training set used to derive a model represents an incomplete coverage of "chemical structure space", the model may only be "successfully" applied to a finite variety of chemical structures.[248] The "applicability domain" (AD) of a model is widely understood to define the range of chemical structures to which the model is "applicable". More precisely, a report for the European Centre for the Validation of Alternative Methods (ECVAM) defined the applicability domain as: "the response and chemical structure space in which the model makes predictions with a given reliability".[249] Whilst this may be interpreted as a range of chemical structures for which the expected model performance is well characterised,[25] the applicability domain is commonly interpreted as a region of chemical structure space in which the model is known to exhibit desirable predictivity.[47,202,248,250,251]

A distinction may be made between those approaches which simply try to categorise compounds as "inside AD/outside AD" and those which seek to directly assess the expected performance of the model for a particular compound.[47]

Examples of the former include approaches based on molecular fragments, such as categorising compounds with fragments not seen in the training set as outside the AD,[249–251] as well as those which are informed by an understanding of the molecular mechanisms responsible for the bioactivity of interest.[249]As an example of the latter, skin sensitizers may be categorised as belonging to different "reaction mechanistic domains" and models developed that are specific to one such domain.[252]

Other "inside AD/outside AD" approaches are based upon the location of compounds with respect to the training set within the space defined by a particular set of descriptors. The simplest of these are the range based methods, which categorise compounds as inside/outside the AD if (some of) their descriptor values lie inside/outside the ranges observed for the training set. Alternatively, range based methods may be applied after transforming to a new co-ordinate system based on the principal components (see below). [249]

A principal components analysis (PCA) plot may be used for rapid, qualitative assessment of whether or not a (set of) compound(s) lie inside the chemical space of the training set[69] - i.e. whether or not they may be considered to lie inside the applicability domain of the model. These types of plots were used for qualitative assessment of the separation in chemical space between the training and test sets for the models presented in Chapter 4 of this thesis.

PCA starts with the computation of the principal components - linear, orthogonal combinations of the original descriptors. The principal components (PCs) are the M eigenvectors of the covariance matrix ($\underline{\underline{X}}^T\underline{\underline{X}}$), computed from the $N \times M$ matrix ($\underline{\underline{X}}$) with elements X_{im} denoting the value of the mth descriptor for the ith molecule. The corresponding eigenvalues measure the relative, independent, contribution to the variance associated with the original descriptors for a given PC.[*197] The approximate distribution of the data within the descriptor space may then be visualised by plotting the data in the plane defined by the two PCs with the largest eigenvalues - i.e. a PCA plot.[189,193]

In recent years, a number of studies have sought to benchmark measures designed to estimate the predictive performance of a QSAR model for a new compound.[47,202,248,251] The studies by Shusko *et al.*[47] and Dragos *et al.*,[248] in keeping with earlier publications,[251] advocated the use of measures based on the variation in predictions across an ensemble of predictors as a metric for discriminating between 'well predicted' and 'poorly predicted' compounds.

[*]The scaling of descriptors, to remove 'artificially larger' contributions, e.g. due to different units, to the structural variation captured by the descriptors may be appropriate prior to PC computation.[197]

Chapter 3 Screening for Mutagenicity and Carcinogenicity in the Context of a Prospective Virtual Screen

This chapter presents this author's contribution to a prospective virtual screening project which identified experimentally verified inhibitors of type II dehydroquinase (DHQase).[253,254] This project was a collaboration between members of the Mitchell, Blumberger and Abell research groups of the Department of Chemistry, University of Cambridge. This author's contribution was the development and application of a toxicity filter, used to remove putative inhibitors predicted to induce particularly important toxicity endpoints (see Chapter 1) during the initial stages of this project.

This chapter starts (section 3.1) by situating this author's contribution within the context of this collaboration – summarising the workflow employed and the key findings of the project. The nature of the toxicity filter, along with the motivation for applying a toxicity filter of this kind, is fully explained in section 3.2. The empirical validation of various modelling options, which informed the options selected for the toxicity filter, is presented in sections 3.3 and 3.4. This chapter ends by discussing the implications of the toxicities predicted for the experimentally determined inhibitors.

3.1 Overview of Collaboration

The aim of the project was to identify structurally novel, 'drug-like' inhibitors of type II DHQase. This enzyme catalyses the reversible dehydration of 3-dehydroquinate to give 3-dehydroshikimate, a key step in the essential shikimate pathway in *Streptomyces coelicolor, Helicobacter pylori* and *Mycobacterium tuberculosis* bacteria[253–255] - the latter two being pathogenic.[253,255,256] However, this pathway is not present in mammals.[253–255]

Consequently, inhibitors of the *H. pylori* or *M. tuberculosis* isoforms may serve as antibiotics - indeed, inhibitors of the latter may serve as anti-tuberculosis drugs[253,255] - making the identification of new inhibitors of type II DHQase an important step towards discovering structurally novel drugs which could circumvent resistance to existing antibiotics.[255,257]

The workflow used in this project is schematically illustrated in Figure 3.1.

Firstly, a pre-defined "drug-like"* subset of the freely available ZINC (ZINC8) database was downloaded on April the 14th 2009. This contained 8,784,580 commercially available compounds.[258–260]

The similarity of the available conformers was compared to three ligands co-crystallised with type II DHQase (CA2, RP4 and GAJ† from Protein Data Bank (PDB)[261,262] entries 2BT4, 2CJF and 2C4W respectively) using Ballester and Richards' Ultrafast Shape Recognition (USR) methodology.[263,264] For each template (i.e. co-crystallised ligand), conformers with a similarity below a pre-defined threshold (0.90) were discarded. This resulted in three sets of compounds with 2,963 (CA2 template), 918 (RP4 template) and 498 (GAJ template) unique, non-overlapping ZINC codes (i.e. 4,379 unique ZINC codes in total, corresponding[260] to 4,379 compounds). These sets of compounds were presented to this author in SDF format.

As described in section 3.2, these files were parsed to generate combined predictions for mutagenicity and carcinogenicity. Compounds predicted to be both mutagenic and carcinogenic were removed. SDF entries corresponding to ZINC codes which were duplicated in the original files generated by this toxicity filter were also removed.‡ Ultimately, three SDF files, comprising entries with unique ZINC codes, were obtained. These contained 406 (GAJ template), 847 (RP4 template) and 2,655 (CA2 template) SDF entries respectively, i.e. 3,908 compounds, corresponding to 3,908 unique ZINC codes, in total.

---

*A "drug-like" compound has "sufficiently acceptable ADME properties and sufficiently acceptable toxicity properties to survive through the completion of human Phase I clinical trials".[258] As is common practice,[258] the maintainers of ZINC define this in terms of filters based upon - calculated - molecular properties.[259]

†These codes denote the following ligands: (1S,3R,4R,5S)-1,3,4-trihydroxy-5-(3-phenoxypropyl)cyclohexanecarboxylic-acid (CA2), (1S,4S,5S)-1,4,5-trihydroxy-3-[3-(phenylthio)phenyl]cyclohex-2-ene-1-carboxylic acid (RP4), N-tetrazol-5-yl 9-oxo 9H-xanthene-2 sulphonamide (GAJ).

‡Given that the maintainers of ZINC note[260] that they enumerate possible stereoisomers for compounds with incompletely defined stereochemistry, it is possible that these compounds may have been associated with stereochemical ambiguity - which could have resulted in misleading findings in the subsequent docking studies.

These 3,908 compounds were docked, using GOLD,[265] into five PDB crystal structures of type II DHQase (PDB codes: 2BT4, 2C4W, 2CJF, 1GU1 and 1H0R). These corresponded to the *S. coelicolor* (1GU1, 2BT4, 2CJF), *H. pylori* (2C4W) and *M. tuberculosis* (1H0R) isoforms respectively. Docking poses were generated using ChemScore, and re-scored using GoldScore, Astex Statistical Potential (ASP) and RF-score.[266,267]

Based on the scores obtained, three protocols were employed to generate overall rankings for all 3,908 compounds. Protocol 1 (a consensus scoring strategy) sorted all poses, per protein structure, generated from docking into the 2BT4, 2CJF and 2C4W structures by their average rank (according to ChemScore, GoldScore and ASP). Protocol 2 ranked the poses according to RF-score. In both cases, the top ranking 100 compounds for each target were selected for further consideration. Protocol 3 considered all five sets of poses, sorted according to the average ChemScore, GoldScore and ASP ranks; the compounds ranked within the top 500 for all five protein targets were selected for further consideration.

For protocol 4, all 4,379 compounds initially passing the USR screen were considered. All such compounds were ranked according to their USR similarity score with respect to the RP4 template; top ranking compounds were selected for further consideration.

Given financial constraints and redundancy between the compounds selected by these protocols, a small number of compounds, prioritised via these protocols, was finally purchased for experimental testing.

All purchased compounds were tested for their ability to inhibit conversion of the natural substrate (i.e. 3-dehydroquinate) by both *S. coelicolor* and *M. tuberculosis* isoforms of type II DHQase in a kinetic assay and K_i values estimated from the measured IC_{50} values, using the original form of the Cheng-Prusoff equation, which is valid for reversible, competitive inhibition.[75]

Of the 148 compounds tested, 89 (91) were confirmed as inhibitors of the *M. tuberculosis* (*S. coelicolor*) isoform with K_i < 500 μM. Median K_i values were 115 and 108 μM for the *M. tuberculosis* and *S. coelicolor* isoforms respectively. The most potent of these inhibitors had K_i values of 23 and 4 μM for the *M. tuberculosis* and *S. coelicolor* isoforms respectively. Arguably, this compares favourably to the recent discovery of a novel inhibitor via high throughput screening (HTS), entailing experimental assessment, of 150,000 compounds against the *H. pylori* isoform, which had a K_i of 20 μM and 230 μM for the *H. pylori* and *S. coelicolor* isoforms respectively.[268]

The two sets of inhibitors were separately combined with inhibitors of the corresponding isoform previously published in the literature, as identified from the ChEMBL database.[269] For both expanded sets, clustering analysis,[270] and manual inspection of the structures of compounds in each cluster, indicated that the new inhibitors discovered in this virtual screening workflow were structurally distinct from those previously published in the literature.

At the time of writing, this work had been submitted for publication. This publication will present full details of the computational and experimental procedures employed, and results obtained in this study - beyond the description of the toxicity filter presented below.

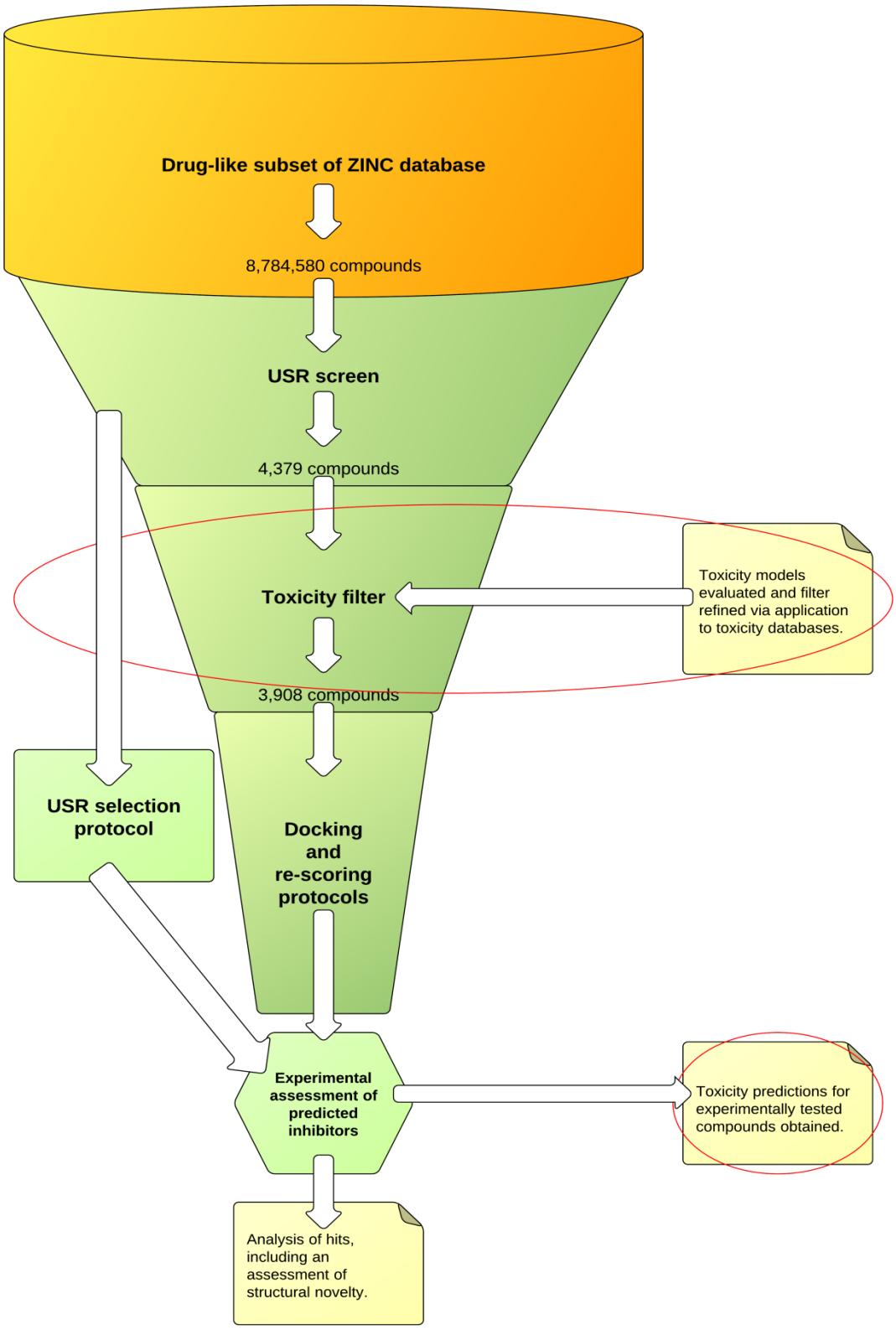


Figure 3.1 Overview of the workflow undertaken to identify novel, experimentally verified inhibitors of type II DHQase. This author's contributions to the study are circled.

3.2 Approach Developed for Toxicity Screening

3.2.1 General Overview

Combined predictions were generated for both the mutagenicity and carcinogenicity endpoints using Derek for Windows™ (version 11.0),[160,271] made available by Lhasa Limited, and the freely available Toxtree (version 1.51)[159–161] software. For brevity, Derek for Windows™ is referred to as DfW.

As previously noted by Simmon-Hettich *et al.*, predictive toxicology filters applied early on in the pharmaceutical life cycle should prioritise specificity over sensitivity (see Chapter 2, section 2.6.4) - i.e. they should focus on minimising the loss of potential candidates due to incorrect predictions of toxicity, even at the expense of retaining some toxic compounds, when attrition costs are low.[195] Hence, only compounds predicted to be both mutagenic and carcinogenic were selected for removal, to reduce losses due to incorrect predictions for either endpoint.

The focus on these endpoints was due to the particular importance of identifying both mutagens and carcinogens early on in the pharmaceutical lifecycle, as explained in Chapter 1, section 1.3.1. Of course, these are not the only endpoints which require early identification to reduce attrition costs (e.g. see Chapter 1, sections 1.3.2 and 1.3.3); however, removal of compounds deemed to exhibit additional toxicities would either have resulted in an accumulated loss of incorrectly predicted toxic compounds - if compounds were removed if predicted to either induce mutagencity and carcinogenicity or another endpoint - or an increased retention of toxic compounds if compounds were only removed if they were predicted to exhibit many types of toxicity.

The focus on prioritising specificity over sensitivity also informed the selection of options used to generate both the predictions made using both programs individually and the manner in which combined predictions were generated. Prior to selecting the options presented below, a range of options were considered and their predictive performance assessed on an experimentally derived mutagenicity and carcinogenicity database (see section 3.3). In order to explain the options selected, it is necessary to briefly review the manner in which DfW and Toxtree generate predictions for mutagenicity and carcinogenicity. Further details may be obtained from the documentation for these programs and the referenced articles and texts.

3.2.2 Generating Predictions using DfW

3.2.2.1 Background

As discussed in Chapter 2, section 2.5, DfW makes predictions on the basis of rules that are based on various types of evidence - structural alerts, physicochemical properties, toxicological properties as well as conclusions arrived at for related toxicity endpoints. For example, the following rules might be used to build up an argument in favour of observing carcinogenicity in rats.[160]

If [structural alert = beta-O/S-substituted carboxylic acid or precursor] is [certain] then [peroxisome proliferation] is [plausible (in rats)]

If [peroxisome proliferation (in rats)] is [plausible] then [carcinogenicity] is [plausible (in rats)]

These example rules also emphasises some important characteristics of DfW. In contrast to Toxtree,[37] DfW generates conclusions of varying strengths (as opposed to simple binary predictions) as well as species specific conclusions - e.g. carcinogenicity is "certain" in rodents, yet "plausible" in humans.[160]

3.2.2.2 Specific Options Selected

However, for the purposes of the project presented in this chapter, binary predictions (carcinogen/mutagen vs. non-carcinogen/non-mutagen) were required. Hence, 'positive' predictions for carcinogenicity/mutagenicity were generated when DfW concluded that carcinogenicity in humans/*in vitro* mutagenicity in *Salmonella typhimurium* (the test organism used in the standard Ames test)[38,46] was "certain", "probable" or "plausible", with DfW predictions being otherwise *interpreted* as 'negative'.[*]

[*]Whilst in keeping with literature precedence,[272] and required for the purpose of generating an automated toxicity filter herein, digitising the output of DfW in this fashion clearly loses information that would be of considerable value in, say, prioritising compounds for experimental assessment. The specific choice of cutoff is, also, to some extent arbitary;[154] however, internal analysis within Lhasa Limited has suggested that, for several endpoints, compounds classified as "equivocal" - which ranks immediately below "plausible"[160] had a 50:50 chance of being toxic (private correspondence with Dr Chris Barber).

3.2.3 Generating Predictions using Toxtree

3.2.3.1 Background

Toxtree was used to make predictions for mutagenicity/carcinogenicity on the basis of the Benigni/Bossa rulebase.[37] Using this rulebase, the program reports all genotoxic and nongenotoxic carcinogenicity structural alerts (SAs) - see Chapter 2, section 2.5 - that were deemed to have "fired" for the compound in question - i.e. were found in the molecular structure in the absence of any possible modulating factors supposed to abolish the toxic potential conferred by the SA.

If an SA substructure corresponding to an aromatic amine or an α,β-unsaturated aldehyde is identified (even if it does not fire), binary classification, canonical discriminant analysis QSARs may be applied (see Chapter 2, section 2.6.3.2). These QSARs were trained to discriminate between: 1. aromatic amines inducing, or not inducing, mutagenicity in *Salmonella typhimurium* TA100 (including metabolic activation), 2. aromatic amines exhibiting, or not exhibiting, carcinogenicity in rodents, 3. α,β-unsaturated aliphatic aldehydes inducing, or not-inducing, mutagenicity in *Salmonella typhimurium* TA100 (without metabolic activation). These QSARs are denoted: 1. QSAR6, 2. QSAR8 and 3. QSAR13 respectively.

3.2.3.2 Specific Options Selected

Toxtree was interpreted as predicting mutagenicity when one of its genotoxic SAs fired, with carcinogenicity also predicted in these cases and when a positive prediction for carcinogenicity was obtained from QSAR8, with negative predictions ignored. Ignoring negative predictions for the Toxtree QSAR models could help avoid a scenario in which toxicity conferred by another SA, other than that for which the QSAR was applied, was ignored due to a negative QSAR prediction. When mutagenicity/carcinogenicity was not predicted, Toxtree was *interpreted* as predicting the compound to be a non-mutagen/non-carcinogen.

3.2.4 Combined Predictions

3.2.4.1 Background

The value of making toxicity predictions based upon the combined output of different *in silico* models was stressed in the recent literature.[273,274] The manner in which overall predictions are generated when the models do not agree can be adjusted to increase sensitivity (specificity), at the cost of reduced specificity (sensitivity).[273]

Hence, given the need to maximise specificity for the toxicity filter employed in this project, the used of a combined prediction strategy was attractive.

It should be noted that, for such combined models to add value, they should be complementary - i.e. they should not consistently make identical predictions - so that greater confidence may be assigned when a positive (i.e. toxic) prediction is generated by multiple models.[274] The DfW and Toxtree programs clearly employ different prediction paradigms (see above). Moreover, they *may* employ complementary SAs: the SA resulting in a prediction for carcinogenicity by DfW, as noted in section 3.2.2.1, is not documented as corresponding to a SA employed by Toxtree for carcinogenicity prediction.[37,160] However, it was not possible to fully assess the extent to which the SAs employed by Toxtree (version 1.51) and DfW differed, since access to the DfW knowledge base was not permitted by the terms under which Derek for Windows™ (version 11.0) was licensed to us. Ultimately, the complementary nature of the two programs is evidenced by the differences in specificity and sensitivity obtained when considering conflicting predictions to be positive or negative (see section 3.3).*

3.2.4.2 Specific Options Selected

Based on the 'yes/no' (or 'positive/negative') predictions generated for carcinogenicity/mutagenicity via the interpretation of the output of DfW and Toxtree presented above, compounds were predicted to be mutagenic/carcinogenic if 'positive' predictions were generated via both programs; compounds were otherwise predicted to be non-mutagenic/non-carcinogenic.

---

*Whether similar results would be obtained with the current versions of these programs is an open question.

3.2.5 Practical Implementation

All SDF files provided to this author, were batch processed via the Toxtree (version 1.51) GUI.* Ultimate generation of toxicity predictions, and generation of filtered SDF files, entailed parsing of SDF files using Python[275] scripts, largely via the Python module Pybel.[177] Parsing of the SDF files included writing the titles (ZINC codes) of each entry to an SDF field (<ZINCID>), as required for generating predictions and filtering compounds predicted to be mutagenic and carcinogenic via parsing the SDFs, updated with Toxtree output, using an additional set of Python scripts. These scripts also triggered the generation of Derek for Windows™ (version 11.0) predictions; DfW predictions were matched to the corresponding Toxtree predictions using the <ZINCID> field.

Clearly documented versions of these Python scripts, and the DfW configuration file, are made available with this thesis (Appendix A). The original versions of these scripts, used to generate (combined) toxicity predictions for the ZINC database subsets, were essentially identical to those used to generate Toxtree and DfW predictions for the toxicity database considered in the next section - save for using a database specific field, specified below, to match Toxtree and DfW predictions.

3.3 Empirical Validation of Toxicity Models

3.3.1 Overview

The ISSCAN[51] (version 3a) database (see Chapter 1, section 1.3.1.3) was downloaded in SDF format and used to assess the predictive performance of various mutagenicity and carcinogenicity binary classifiers constructed using DfW and/or Toxtree. The performance of these models informed the selection of the specific options used for toxicity filtering (as presented in section 3.2).

3.3.2 Options Assessed

In Table 3.1 and Table 3.2, the 12 and four sets of options for generating positive predictions for mutagenicity and carcinogenicity using Toxtree and DfW respectively are presented. As

* For the CA2 set, Toxtree was applied prior to addition of the <ZINCID> field; for the GAJ and RP4 sets, Toxtree was applied following this step. DfW was applied to the SDF files generated following <ZINCID> addition to the SDF files originally provided to this author.

well as generating predictions using Toxtree and DfW individually, all possible combined predictions were generated. Two options were considered for making combined positive predictions (for both endpoints): 1. an overall positive prediction was generated if either submodel made a positive prediction, 2. an overall positive prediction was only generated if both sub-models made a positive prediction. These options for generating combined predictions were expected to maximise 1. sensitivity or 2. specificity respectively. Ultimately, this yielded 112 (4+12+[12×4×2]) models to be evaluated. In the absence of positive predictions, all models were *interpreted* as yielding negative predictions.*

*No claim is being made here that when these models do not predict a compound to be toxic, this actually indicates the compound is non-toxic; indeed, save when a QSAR analysis can be applied, Toxtree is inherently incapable of making negative predictions,[37] and not all assessments made by DfW suggest the presence or absence of toxicity (e.g. an "equivocal" or "open" prediction),[160] with a "nothing to report" outcome generated in the absence of either a structural alert firing or the activation of any other relevant rule in the knowledge base.[276] However, for the purposes of the current work, a toxicity filter was required that would discard compounds anticipated to be toxic whilst retaining all others (i.e. make a *de facto* non-toxic prediction for the latter compounds).

Toxtree Model Number	Outcomes Treated as Triggering a Positive Mutagenicity Prediction			Corresponding Outcomes Treated as Triggering a Positive Carcinogenicity Prediction		
	Genotoxic SA fired	QSAR6 positive prediction	QSAR13 positive prediction	Positive mutagenicity prediction	Nongenotoxic SA fired	QSAR8 positive prediction
1	✓			✓		
2	✓			✓	✓	
3	✓	✓		✓		
4	✓	✓		✓	✓	
5	✓		✓	✓		
6	✓		✓	✓	✓	
7	✓			✓		✓
8	✓			✓	✓	✓
9	✓	✓	✓	✓		
10	✓	✓	✓	✓	✓	
11	✓	✓	✓	✓		✓
12	✓	✓	✓	✓	✓	✓

Table 3.1 Combinations of options considered for predicting compounds as mutagens/carcinogens based upon the output generated by Toxtree. If any one of the 'ticked' outcomes occurred, a positive prediction of mutagenicity/carcinogenicity was made. If these conditions were not met, compounds were deemed to be predicted non-mutagenic/non-carcinogenic by Toxtree.

DfW Model Number	Outcomes Treated as Triggering a Positive Mutagenicity Prediction		Corresponding Outcomes Treated as Triggering a Positive Carcinogenicity Prediction	
	In vitro mutagenicity in humans is certain, probable or plausible	Mutagenicity in *Salmonella typhimurium* is certain, probable or plausible	Carcinogenicity in humans is certain, probable or plausible	Carcinogenicity in rodents is certain, probable or plausible
1	✓		✓	
2	✓			✓
3		✓	✓	
4		✓		✓

Table 3.2 Combinations of options considered for predicting compounds as mutagens/carcinogens based upon the output generated by DfW. If any one of the 'ticked' outcomes occurred, a positive prediction of mutagenicity/carcinogenicity was made. If these conditions were not met, compounds were deemed to be predicted non-mutagenic/non-carcinogenic by DfW.

3.3.3 ISSCAN Validation

Subsequent to generating Toxtree predictions using the GUI, and DfW predictions, by parsing the downloaded SDF, the final toxicity predictions, based upon all 112 options, were

obtained via parsing the output files via Python scripts as per section 3.2.5. The <Substance ID> field was used to match Toxtree and DfW predictions for the same molecule. It was not possible to obtain any predictions for 12 out of the 1,153 entries in this database.

Experimentally derived mutagenicity and carcinogenicity classes were determined by considering the summary assessments for carcinogenicity and mutagenicity based upon rodent studies and the Ames test, and recorded in the SDF fields <Canc> and <SAL>, respectively. By default, all performance estimates were generated by considering all equivocal results as corresponding to non-mutagens/non-carcinogens. On this basis, there were 387 and 440 experimentally assessed mutagens and non-mutagens respectively, out of the 1,141 compounds for which predictions could be generated. The corresponding numbers of carcinogens and non-carcinogens were 713 and 428 respectively.

Upon counting compounds with equivocal results for these endpoints as mutagens/carcinogens, the number of mutagens and non-mutagens was changed to 399 and 428 (out of the same 1,141). The corresponding numbers of carcinogens and non-carcinogens were 781 and 360 respectively.

For each endpoint, performance estimates were only generated based upon those compounds for which toxicity predictions and experimentally derived classes could be obtained.

3.4 Results Obtained from Application of Toxicity Models

3.4.1 ISSCAN Validation

Detailed results, obtained from validation of all 112 models on the ISSCAN database, are made available with this thesis (Appendix A). These detailed results include the MCC and corresponding chi-squared p-value (as proposed in Chapter 2, section 2.6.4.1), computed using the CHIDIST( ) function in Excel 2007 (32-bit).

The performance of all 112 models is summarised in the Receiver Operating Characteristic (ROC) graphs[277] presented in Figure 3.2, Figure 3.3, Figure 3.4 and Figure 3.5. The line denoting $\text{sensitivity} = 1 - \text{specificity}$ corresponds to the expected performance of a random predictor, better than random predictions lying above this line.[277]

As is indicated by these graphs, the estimated performance of the models was not changed considerably by changing the class label assigned to compounds with equivocal mutagenicity/carcinogenicity assessments. The maximum (median) absolute difference in

carcinogen sensitivity was 0.04 (0.02), with all other sensitivity and specificity values changed by at most 0.01 (2dp) upon changing the class labels assigned for equivocal compounds.

The combination of prediction options selected (see section 3.2.2.2, 3.2.3.2 and 3.2.4.2) corresponded to one of two combinations with the maximum sensitivity, out of those models with the maximum specificity, when applied to discriminating carcinogens from non-carcinogens. This was the case irrespective of whether compounds with equivocal experimental assessments were deemed carcinogens or non-carcinogens.

Save for combinations of options with sensitivity values below 0.10, the selected model was one of those with the highest specificity value for separating mutagens and non-mutagens - again, irrespective of the class label assigned to compounds with equivocal experimental findings.

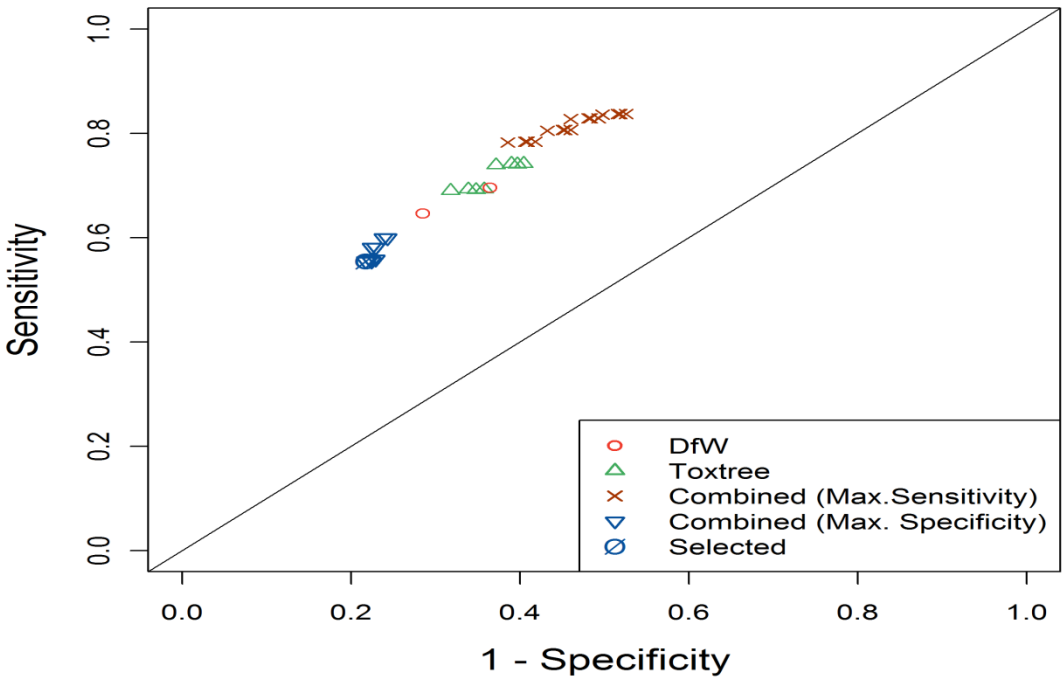


Figure 3.2 Results obtained from validating carcinogenicity predictions of all 112 models on the ISSCAN database. All equivocal carcinogens were considered non-carcinogens.

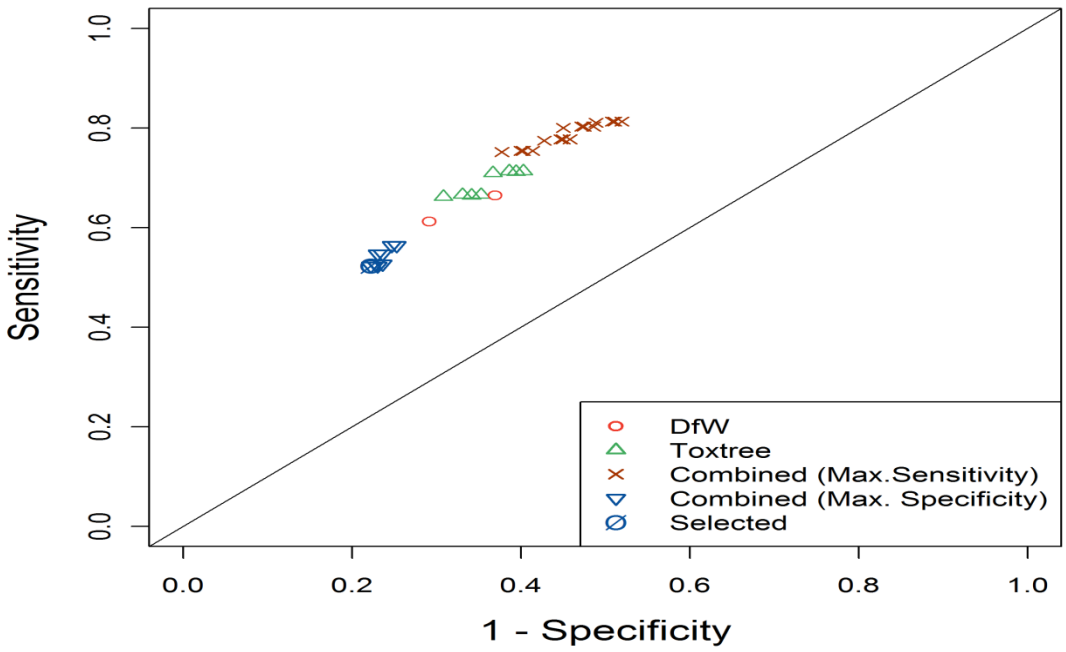


Figure 3.3 Results obtained from validating carcinogenicity predictions of all 112 models on the ISSCAN database. All equivocal carcinogens were considered carcinogens.

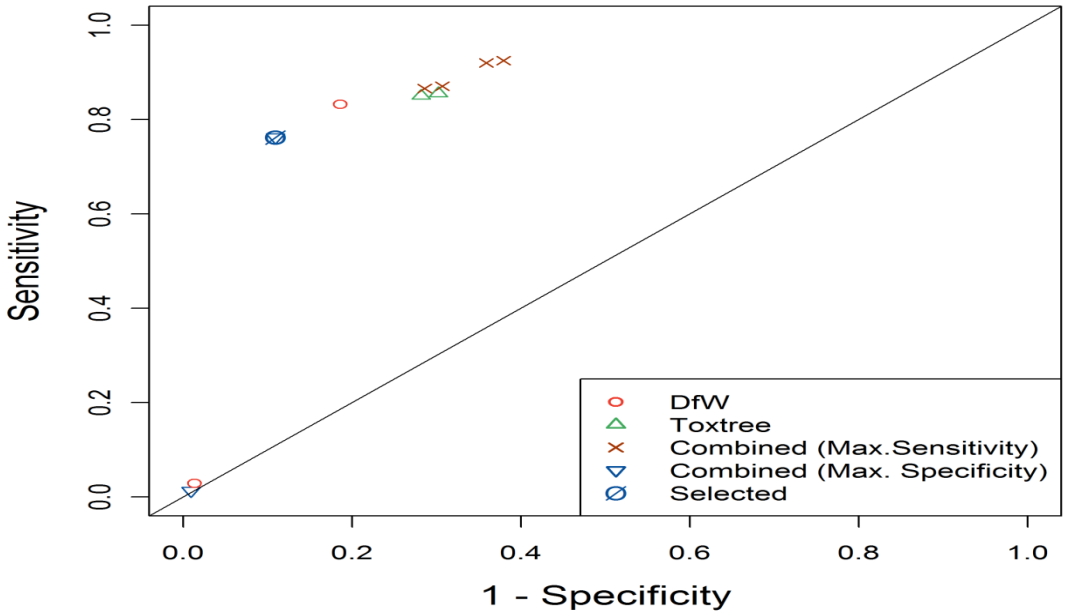


Figure 3.4 Results obtained from validating mutagenicity predictions of all 112 models on the ISSCAN database. All equivocal mutagens were considered non-mutagens.

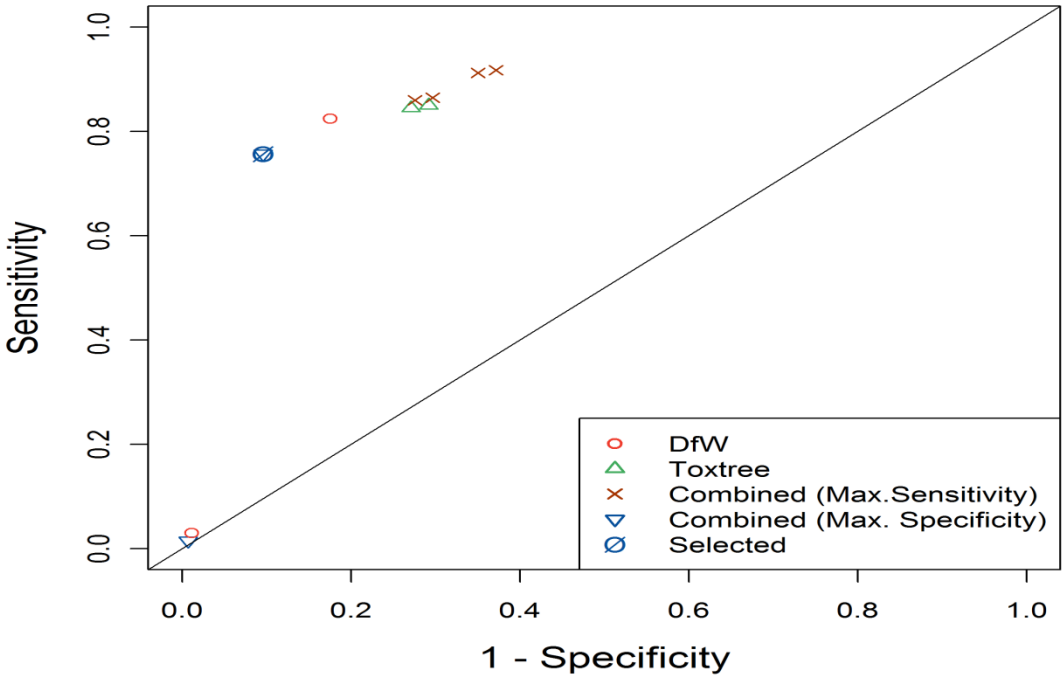


Figure 3.5 Results obtained from validating mutagenicity predictions of all 112 models on the ISSCAN database. All equivocal mutagens were considered mutagens.

3.4.1.1 Discussion of the Performance of the Selected Model

Table 3.3 summarises the performance of the selected model, used for filtering the subsets of the ZINC database provided to this author. In terms of the MCC values obtained for discriminating between mutagens and non-mutagens in the ISSCAN database, the selected model was one of the top performing models. Whilst this was not the case for discriminating carcinogens from non-carcinogens in this database, all models with higher MCC values corresponded to lower specificity values - and the minimisation of false positives was our priority in this work (see section 3.2.1).

These results suggest the choice of selected model was reasonable.

Endpoint	MCC	P-Value	Sensitivity	Specificity
Mutagenicity	0.66	1.1E-80	0.76	0.89
Mutagenicity (Equivocal mutagens considered mutagens)	0.67	7.6E-83	0.76	0.90
Carcinogenicity	0.33	5.5E-29	0.56	0.78
Carcinogenicity (Equivocal carcinogens considered carcinogens)	0.28	1.1E-21	0.52	0.78

Table 3.3 Performance of the selected model.

3.4.2 Application to Filtering of ZINC Datasets

The original SDF files passed to this author contained a cumulative total of 4,454 entries, corresponding to 4,379 unique ZINC codes. Following the application of the toxicity filter described in section 3.2, a cumulative total of 4,035 SDF entries remained, corresponding to 3,970 unique ZINC codes - i.e. 409 (9%) of the originally provided compounds were predicted to be both mutagenic and carcinogenic, hence were discarded.

Of the 3,970 ZINC codes corresponding to those SDF entries which were not immediately discarded following application of the toxicity filter, 30 (113) ZINC codes corresponded to predicted mutagens (carcinogens). Hence, 439 (10%) of the 4,379 compounds present in the SDF files originally provided to this author were predicted mutagens, and 522 (12%) were predicted carcinogens. It was not possible to experimentally validate these predictions.

3.4.3 Toxicity Predictions for Experimentally Tested Compounds

Of the 148 compounds experimentally tested for type II DHQase inhibition, 10 were predicted carcinogenic by the model used for toxicity filtering (see section 3.2), with only one compound predicted to be mutagenic. This supposed mutagen was also predicted to be carcinogenic, since this was amongst those selected from the original USR screen, prior to application of the toxicity filter.

Considering all toxicity predictions generated by Toxtree and DfW, 32 of the experimentally assessed compounds were not predicted to be associated with any potential toxic liability by

either program. For these 32 compounds, no Toxtree structural alerts for mutagenicity/carcinogenicity were fired, nor did any QSAR assessments indicate mutagenicity/carcinogenicity. Across all endpoints considered, any assessments made by DfW were that the toxic effect was "doubted", "improbable" or "impossible". However, for many of these compounds, DfW only generated an assessment for a few endpoints; this could mean that they were not covered by the rules in the DfW knowledge base for other endpoints, or that a structural alert did not fire due to modulating factors - as discussed in Chapter 2, section 2.5.

Of these 32 compounds, the lowest K_i (for the *M. tuberculosis* isoform) was 40 μM; the median K_i (for the *M. tuberculosis* isoform), across all 20 of these compounds for which a K_i was determined, was 120.5 μM - slightly higher than the median value of 115 μM obtained across all tested compounds for which a K_i was determined for this isoform.

Considering the most potent inhibitor of the *M. tuberculosis* isoform (K_i = 23 μM), DfW deemed phototoxicity and skin sensitization in rodent, hamster or human to be "plausible" (i.e., on balance, the evidence was deemed to support the proposition that this compound induced these toxicities).[160] Otherwise, no positive toxicity predictions were generated by DfW, with no Toxtree structural alerts fired or QSAR assessments indicating any toxic liability.

The mechanisms which underpin skin sensitization[278] may also lead to allergic responses, or drug hypersensitivity, which may be severe, in response to oral drugs.[279] However, given the potentially lethal consequences of untreated tuberculosis, many adverse effects are commonly deemed acceptable in conjunction with anti-tuberculosis drugs.[280] It should also be noted that phototoxic side effects, such as those induced by the anti-tuberculosis drug sparfloxacin, may be avoided if patients adhere to instructions to avoid sunlight.[281] Hence, supposing that these predicted toxicities were to be experimentally verified, and lead optimisation was unable to remove them, the most potent *M. tuberculosis* isoform inhibitor found in our study *could*[*] still be a viable lead candidate.

The toxicity predictions obtained for all experimentally tested compounds are summarised in an Excel file made available with this thesis (Appendix A).

---

[*] A variety of additional requirements exist, however, for a viable lead candidate - including the synthetic accessibility of more potent derivatives with acceptable ADME properties.[18]

3.5 Conclusions

Starting from a database of more than 8×10^6 drug-like compounds, a collaborative hierarchical virtual screening protocol was used to select 148 compounds for experimental assessment. Of these, 89 (91) were found to inhibit the *M. tuberculosis* (*S. coelicolor*) isoform of type II dehydroquinase with $K_i < 500$ μM. In a much more cost and resource effective manner, the strategy developed yielded comparable results to recent experimental high throughput screening programs designed to identify inhibitors of this enzyme.[268]

A toxicity filter, developed by this author using commonly employed (hybrid)* expert systems (Derek for Windows™ and Toxtree), was used to identify compounds with potential mutagenic and carcinogenic liabilities, whilst minimising the loss of false 'positives' (due to incorrectly identified toxicants) during this early stage discovery effort. An assessment of different modelling options on the ISSCAN database indicated that the software programs were complementary, such that models could be developed with greater sensitivity or, in our case, greater specificity than was achievable with either program alone. ROC graph analysis and consideration of the MCC, and its associated chi-squared p-value, indicated the selected models were more discriminative between mutagens/carcinogens than would be expected for a random predictor.

One potential limitation of these assessments, however, is the lack of mutagenicity/carcinogenicity validation data derived from human studies.

Considering both the inhibitory potencies with respect to the *M. tuberculosis* isoform and the lack of/nature of the anticipated toxicities for some of the identified inhibitors, it is *possible* that anti-tuberculosis lead candidates may have been identified. Further experimental work, including experimental toxicity screening assays - particularly for those toxicities anticipated *in silico*, is warranted to explore this possibility.

* See Chapter 2, section 2.5.

Chapter 4 Development and Assessment of Binary Classifiers for Identifying Potent hERG Inhibitors

This chapter describes work undertaken to develop models for binary classification of inhibitors of the human ether-à-go-go-related gene (hERG) potassium ion channel, according to toxicologically relevant potency thresholds proposed by the pharmaceutical industry. The aim here was to make predictions based upon computationally inexpensive, 2D descriptors and the approaches were compared to those previously presented in the literature based upon these kinds of descriptors. The novel approaches were found to perform comparably to, or better than, the performance of the methods previously reported in the literature. It was discovered that the performance of some of the modelling approaches could vary dramatically. This variation occurred, for example, when training and validating the models on different datasets and, in the case of the pseudo-stochastic methods considered here, for different training and testing runs on the same data.

4.1 Introduction

As discussed in detail in Chapter 1, section 1.3.2, there is a clear need for rapid and reliable computational approaches which can discriminate between potent (IC_{50} < 1 μM) hERG inhibitors, the development of which is often discontinued, and weaker inhibitors. Ideally, these models would also be comprehensible to medicinal chemists to facilitate designing out hERG inhibition from a compound series.

Whilst some models[69,91,282,283] have been developed to discriminate between potent inhibitors (IC_{50} < 1μM) and 'non-inhibitors' (IC_{50} > X μM, where X ≥ 10), these would be of limited value in drug discovery, where many compound series exhibit 'moderate' inhibition – with IC_{50} values in the range of 1-10 μM.[15]

Consequently, this author was interested in ligand-based, binary classifiers capable of separating 'blockers' (IC_{50} < 1 μM) from both moderate (1 μM ≤ IC_{50} < 10 μM) and weak (10 μM ≤ IC_{50}) inhibitors. Various studies in recent years have sought to develop binary classifiers of this kind. Both Li *et al.*[90] and Tobita and co-workers[284] developed classifiers based on Support Vector Machines (SVMs). Whilst Li *et al.* calculated 3D GRIND descriptors, after docking the structures into a homology model of the ion channel pore domain, Tobita *et al.* calculated computationally inexpensive MACCS keys and other 2D descriptors using the Molecular Operating Environment (MOE) software program.[285]

Likewise, Dubus *et al.*[91] and Thai and Ecker[69,231] developed models based on 2D descriptors computed using the same software.

In this chapter, novel approaches to generating binary classifiers of this kind are presented and their predictive performance directly compared, on various (partitions of) datasets, to approaches presented by Thai and Ecker[69] and Dubus *et al.*[91]

4.2 Datasets

The modelling approaches developed by this author, as well as implementations of those presented by Thai and Ecker[69] and Dubus *et al.*,[91] were assessed using the following three datasets. Summary statistics for the Literature-368, Dubus-203 and Thai-313 datasets are presented in Table 4.1.

The latter two datasets were primarily used to validate this author's implementations of the binary classifiers proposed by Dubus *et al.* and Thai and Ecker, via (inexact) replications (see 4.4.5) of the Diverse Subset training and test partitions presented in their respective publications.[69,91] The approaches developed here were also assessed on some of these training and test partitions respectively to provide additional comparisons between the different methods and assess dataset dependencies of performance estimates.

Copies of the Literature-368 and Thai-313 datasets are presented, in SDF format, in the files made available with this thesis (Appendix A); the Dubus-203 dataset is available on request from Dr Elodie Dubus of Aureus Sciences.[102]

Literature-368 Dataset

Disjoint sets of 220 and 148 compounds were compiled from literature sources by this author. Since, initially, all models were trained using (subsets of, where specified) the former and evaluated on the latter, these are referred to as the Int-Set and ExtTest-Set, respectively. The ExtTest-Set was compiled to ensure that it contained no compounds in common with those used for model development by Thai and Ecker,[69] nor by Dubus *et al.*[91] This allowed the ExtTest-Set to be used as a truly external test of this author's models and theirs.

For binary classification, compounds with (arithmetic mean) $pIC_{50} > 6$ were assigned to class A ('Active') and those with $6 \geq$ (arithmetic mean) pIC_{50} were assigned to class I ('Inactive').

In total, the Int-Set contained 55 compounds in class A (strong inhibitors) and 165 compounds in class I, with the ExtTest-Set comprising 24 class A and 124 class I compounds. The Int-Set class I compounds included 70 moderate inhibitors with $6.00 \geq$

(arithmetic mean) $pIC_{50} > 5.00$ and 95 weak inhibitors with $5.00 \geq$ (arithmetic mean) pIC_{50}. The numbers of moderate and weak inhibitors amongst the class I compounds in the ExtTest-Set were 50 and 74 respectively.

The starting point for the derivation of the Int-Set was the list of hERG inhibitors presented by Nisius and Göller in 2009.[286] Compounds were initially removed from this set where the available, relevant hERG inhibition measurements (see below) indicated assignment to more than one of the potency categories defined above – i.e. strong, moderate and weak inhibitors. In order to maintain desirably high numbers in each potency category, these compounds were supplemented with hERG inhibitors presented in the primary literature as obtained from a Scopus search.

Another Scopus search was subsequently employed, in late 2009, to identify additional hERG inhibitors for inclusion in the ExtTest-Set. In order to avoid adding compounds which were included in the earlier studies by Thai and Ecker[69] and Dubus *et al.*[91] (see above), this Scopus search was (in contrast to before) restricted to the primary literature published in 2008-2009. Additional compounds were identified via consulting a recent review of the primary literature[67] in addition to correspondence with Dr Chris Swain of Cambridge MedChem Consulting.

Where possible, all measurements from secondary sources were checked against their primary literature references and only retained where these indicated that the experimental conditions criteria outlined below were not violated. Otherwise, where the secondary sources indicated that these criteria were not violated, measurements from these sources were usually retained. However, non-unique pIC_{50} values – for a given compound - from secondary sources were additionally discarded if it could not be determined that they corresponded to genuinely distinct measurements.

In order to maximise the validity of the classes assigned, additional inhibition measurements – from the primary literature or patents – were sought for all compounds in this dataset via SciFinder Scholar 2007[287] and recorded if they met the experimental criteria presented below. Due to time constraints, only sources published prior to 2010 were considered.

The selection of hERG inhibition measurements for each compound was designed to estimate the expected pIC_{50} range to which the compounds would be assigned if electrophysiologically assessed under the conditions typically employed by the pharmaceutical industry – since the

1 μM (pIC_{50}=6) threshold delimiting class A and class I compounds was derived from such assessments.[15,62]

In order to minimise systematic errors, all inhibition measurements were, to the best of this author's knowledge, obtained from electrophysiological assays in mammalian, heterologous expression systems. Values which were clearly not derived from inhibition of the tail current elicited in these assays,[67] were excluded. These conditions are commonly used within the pharmaceutical industry.[62,78]

IC_{50} values which were explicitly noted to have been estimated from single concentration measurements were discarded, as these are typically less reliable (see Chapter 1, section 1.3.2.2). However, in the absence of IC_{50} values for a compound, single concentration inhibition measurements, which categorically established pIC_{50} values as > 6.00 or ≤ 5.00 - '40% inhibition at 10μM', say - were also used for class assignment.

Some measurements were obtained using the medium-throughput QPatch,[288] PatchXPress®[79] and IonWorks™[78] automated patch-clamp assays. IonWorks™ measurements which appeared to be potency underestimates[289] were discarded in favour of higher potency measurements. This author inferred that all IonWorks™ measurements were obtained using a similar protocol to that employed by Bridgland-Taylor *et al.*, which does not consistently underestimate or overestimate potency.[78]

In spite of taking care to minimise experimental inconsistencies, appreciable variability in the pIC_{50} values retained for some compounds is still evident. However, the median absolute difference in pIC_{50} values – not including measurements where the literature references simply indicated $pIC_{50} > 6$ or $pIC_{50} < 5$ – recorded for the same compound is only 0.35 log units (upper and lower quartiles: 0.69, 0.15). Moreover, for only five compounds were inhibition measurements, meeting the stringent criteria described above, obtained which indicated assignment to different classes or potency categories. In these cases, the mean pIC_{50} was used for class assignment (three compounds), and for additional categorisation of class I compounds as moderate or weak inhibitors (two compounds). The topology of all structures was checked using either SciFinder Scholar 2007[287] or literature sources. No pair of compounds in this dataset are stereoisomers. Whilst the measurements presented for some dataset entries (i.e. compounds) were derived for a specific enantiomer, some entries correspond to a racemate – with all measurements assumed to be derived for the racemate, where applicable, unless explicitly stated otherwise in the literature references.

All pIC_{50} estimates used for class assignments are presented, along with literature references, in a Word document (Literature_368_References.doc) made available with this thesis (Appendix A). This document also presents the references consulted for the structures presented in SDF format (Appendix A) for this dataset.

Dubus-203 Dataset

This dataset, previously used by Dubus *et al.*, and comprising 96 (107) compounds with $IC_{50} \leq 1$ μM ($IC_{50} > 1$ μM), also determined using electrophysiological measurements in mammalian cells,[91] was kindly provided to this author in SDF format by Dr Elodie Dubus. Here, the classes are referred to as 'A' and 'I', and all class A compounds referred to as strong inhibitors, as per the Literature-368 dataset. The class I compounds comprised 48 "moderate" inhibitors (1 μM < IC_{50} < 10 μM) and 59 "weak" inhibitors ($IC_{50} \geq 10$ μM),[91] essentially defined as per the Literature-368 dataset.

Thai-313 Dataset

The structures in this dataset, previously used by Thai and Ecker,[69] were in part (285 compounds) kindly provided to this author in SDF format by Dr Khac-Minh Thai, with the remaining structures obtained using SciFinder Scholar 2007[287] or the primary literature references provided by Thai and Ecker.[69] This dataset contained 100 (213) compounds in class A (I) - as defined for the Literature-368 dataset - with class labels assigned using the IC_{50} values provided by Thai and Ecker.[69] Based on these IC_{50} values, the class I compounds comprised 97 moderate and 116 weak inhibitors – as defined for the Literature-368 dataset.

Dataset	Total n.o. Compounds	Class A	Class I	Class I: Moderate Inhibitors	Class I: Weak Inhibitors
Literature-368	368	79	289	120	169
Dubus-203	203	96	107	48	59
Thai-313	313	100	213	97	116

Table 4.1 Numbers of hERG inhibitors and their distribution amongst the potency categories for the datasets modelled in this chapter.

4.3 Model Development and Validation

As detailed in sections 4.3.1 and 4.3.2, a variety of different 2D descriptor sets were considered for building the models developed in this work, along with a number of different Machine Learning algorithms. Descriptors and hyperparameters were selected using stratified, five-fold cross-validation (5CV) on the training set (see Chapter 2, section 2.6.5). In order to avoid optimistic bias in their estimated performance, all these models were validated on external test sets - i.e. datasets containing none of the instances used to select the descriptor sets or hyperparameters used to build the models and obviously containing none of the data used to train the models – after selecting the hyperparameters and descriptor sets (see Chapter 2, section 2.6.5).

The selected descriptor sets and hyperparameters were those which maximised the mean value (across all five validation sets) for the Matthews Correlation Coefficient (MCC) - defined in Chapter 2, section 2.6.4.1.* All learning algorithms, in conjunction with the selected descriptor sets and hyperparameters, were then trained on the entire training set, or subsets thereof, where specified, and evaluated on the corresponding external test set.

4.3.1 Machine Learning Algorithms Considered

The Machine Learning algorithms utilised to generate the models developed in this work are presented below. Save where otherwise specified, all algorithmic implementations were used with their default settings. Detailed descriptions of these algorithms and the meaning of the selected hyperparameters is provided in Chapter 2, section 2.6.3.2.

Winnow

Initially, models were developed using Nigsch's adaptation of Littlestone's Winnow algorithm.[184] Instances are presented to Winnow as a set of 'features' – text strings which may either be present or absent in an instance.† In the current context, the instances are compounds and the features correspond to descriptors, discussed in section 4.3.2.

Here, four distinct hyperparameter combinations were initially considered, in conjunction with 94 different feature sets. These hyperparameter combinations entailed:

---

* If multiple possibilities maximised this, a single, arbitrary, set was chosen.

† When not discussing this author's models, the term features is used more generally to refer to explanatory variables.

1. One scorer, combined with the original promotion (p) and demotion (d) factors (collectively 'update factors') proposed by Nigsch.[184]
2. Five scorers and the original update factors.
3. One scorer and constant (p=1.1 and d=0.9) update factors.
4. Five scorers and constant update factors.

When the feature sets included "orthogonal sparse bigrams" (OSBs), generated from the initial feature sets ("monograms"), as described by Nigsch, [184] ε (see Chapter 2, section 2.6.3.2) was set to 0.15, as per Nigsch,[213] with ε set to zero for all feature sets composed solely of monograms.

All combinations were evaluated by presenting the 5CV training sets to Winnow in a common, single order.

Winnow Based Feature Selection

The top ranking* hyperparameter-feature set combination out of the 376 (4×94) possibilities, defined the top feature set, used to generate predictive models using all subsequent learning algorithms. For all algorithms other than Winnow (see below), feature sets were encoded as bit-strings, with 1 (0) assigned when a training set feature was present (absent) in an instance.

Multiple Winnow Training Cycles

An additional Winnow model was generated using multiple training cycles, in conjunction with the aforementioned top ranking hyperparameter-feature set combination. For each training cycle, the training set instances were presented to Winnow in the same order. The number of training cycles was increased from one to 100, in steps of one, and the top ranking number of cycles selected to build the final model.

The implementation of Winnow, in C++, used here was largely identical to that previously used by Nigsch and is made available with this thesis (Appendix A).

Estimated Winnow Performance

All external test set results presented for Winnow are the mean over 50 different training set presentation orders.

*Based on the maximum training set 5CV mean MCC – as per all model selection discussed at the start of section 4.3.

Support Vector Machine

The Support Vector Machine (SVM) algorithm was used with the Radial Basis Function (RBF) kernel (see Chapter 2, Equation 2.7). In addition to the kernel hyperparameter γ, the regularisation constant (C) was also varied (see Chapter 2, Equation 2.5). As proposed by Morik *et al.*,[290] the use of the following non-uniform cost-factor (Equation 4.1) was considered (a 'uniform cost-factor' means: $C_+ = C_-$), in order to reduce bias in favour of the majority class in the training set:

$$\frac{C_+}{C_-} = \frac{n_-}{n_+} \qquad 4.1$$

Here, C_+ (C_-) is the relative weight given to the sum of the slack-variables (see Chapter 2, Equation 2.7) in the 'positive' ('negative') class and n_+ (n_-) the number of training set instances in this class. Here, class A was considered the 'positive', and class I the 'negative', class.

Joachims' SVM^{light} (version 6.02) implementation of SVM was used.[291] Efforts were undertaken to optimise γ, C and the cost-factor [SVM^{light} options: -g, -c, -j], using an adaptation of the grid-search procedure recommended by Hsu *et al.*[221] Initially, all combinations of $\log_2 C \in \{-5, -3, .., 13, 15, \log_2 C_{def}\}$ - where C_{def} denotes the SVMlight default – and $\log_2 \gamma \in \{-15, -13, .., 1, 3\}$, in conjunction with a uniform and non-uniform cost-factor, were considered. Additional adjustment of C and γ was subsequently carried out, in conjunction with the cost-factor used in the initially top ranking parameter set ($C = C*$, $\gamma = \gamma*$). If $C* \neq C_{def}$, all combinations corresponding to varying $\log_2 C$ and $\log_2 \gamma$, in steps of 0.25, between $\log_2 C* \pm 1.75$ and $\log_2 \gamma* \pm 1.75$, respectively, were considered. When $C* = C_{def}$, only γ was varied in this fashion, and the top ranking parameter set for which $C \neq C_{def}$ was also considered for optimisation in the usual fashion.

Random Forest

The R[225] implementation of Random Forest (randomForest) was used here. Since Random Forest (RF) is noted to perform well 'off the shelf',[192] the default hyperparameters were used throughout (see Chapter 2, section 2.6.3.2).

Estimated Random Forest Performance

All results presented with this algorithm are the mean of the results obtained from training and validating the model with five different random number generator (RNG) seeds.

4.3.2 Feature Sets Available to the Models Developed in this Work

The feature sets considered by this author for generating models included, one or both of, two types of features: circular fingerprint features and 'discretized descriptor features' corresponding to numerical 2D descriptors (discussed below). The computational workflow (see section 4.3.2.1) used to generate these meant that only molecular connectivity, and not inconsistently available stereochemical information, was encoded by these features.

Circular Fingerprint Features

Three circular fingerprints, encoding topological molecular substructures, were calculated: CFP2,[213] ECFP_4 and FCFP_6,[185] the former a version of the MOLPRINT 2D fingerprint developed by Bender *et al.*[292] and the latter two developed by SciTegic.[94,213] The former two were previously used to build protein-target prediction models using Winnow by Nigsch *et al.*,[213] whilst the latter was used to build predictive models for hERG inhibition by O'Brien and de Groot.[94]

The constituent features of these fingerprints, calculated for a given molecule, correspond to a set of possible[*] non-predefined atom centred topological substructures present in the molecule; each feature is centred on an atom in the molecule and encodes information about atoms located within a (maximum) number of bonds from the central atom (Figure 4.1). The size of the encoded topological substructures, the type of information and manner in which this information is encoded, varies between the three fingerprints.

[*] As illustrated in Table 4.4, for an example ECFP_4 feature, similar substructures may map onto the same circular fingerprint features.

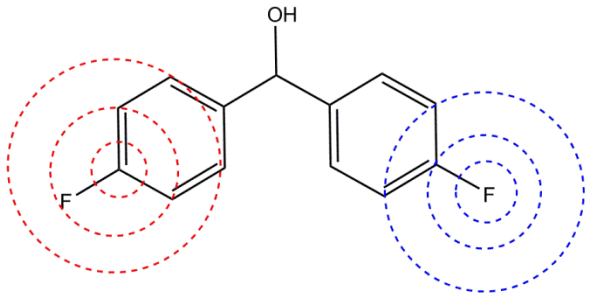


Figure 4.1 Two examples of topological substructures encoded as features, for an example molecule, by a generic circular fingerprint considering environments extending up to two bonds from the central atom.

Discretized Descriptor Features

These were generated using in-house Python scripts,[275] a clearly documented version of which is made available with this thesis (Appendix A). For every numerical descriptor calculated, a compound is assigned a feature corresponding to the range, out of a predefined set, that its value lies within. These ranges were delimited by split points chosen using the distribution of the classes with respect to the training data ordered by their values for the descriptor under consideration. This procedure is illustrated in Figure 4.2.

Two different *discretization* methods were considered for choosing the split points:

i. For all changes in class label going along the ordered training set, the midpoint between the descriptor values of instance K with class C_K and instance J with class C_J is chosen as a split point.

ii. Fayyad and Irani's method,[293] as implemented via the Orange module (version 1.0).[294]

The latter method also considers the values where there is a change in class for the ordered training set as possible split points for a given descriptor. Split points are, however, selected recursively: at most, a single split point is chosen for the original training set and the same procedure is then applied to the partitioned subsections of the training set until none of the possible split points is deemed 'valid'. Possible split points, for a given (subset of the) training set, are only deemed 'valid' if the following condition holds (Equation 4.2). Out of all such 'valid' split points, the split point selected is the one which maximises the first term (LHS) in this expression. This first term, the "information gain", is a measure of the extent to which the uncertainty in the class label for a randomly selected instance in the (subset of the) training set is reduced upon knowing which side of the split point it lies. The overall form of

this expression is designed to prevent excessive partitioning of the training set – whereby the complexity of specifying a split point is insufficiently justified by the extent to which the partition separates instances in different classes.[293]

$$Information\ Gain > Threshold \qquad 4.2$$

In Equation 4.2, the 'Information Gain' and 'Threshold' are given by equations 4.3 and 4.4 respectively. In the following equations, T, T_1 and T_2 denote the total number of classes represented in the original (subset of the) training set, comprising N instances, and the two subsets created by the split point respectively – where the split point corresponds to a value (S) of the descriptor of interest (D). The fraction of instances in the original (subset of the) training set lying either side of this split point are denoted by $p(D \leq S)$ and $p(D > S)$. The fraction of instances in the original (subset of the) training set belonging to class c is given by $p(c)$. Out of those instances assigned to one of the newly created susbsets, the fraction belonging to class c is given by $p(c|D \leq S)$ or $p(c|D > S)$.

$$-\sum_{c}^{T} p(c)\, log_2\, p(c) \qquad 4.3$$

$$+$$

$$p(D \leq S) \sum_{c}^{T_1} p(c|D \leq S)\, log_2\, p(c|D \leq S)$$

$$+$$

$$p(D > S) \sum_{c}^{T_2} p(c|D > S)\, log_2\, p(c|D > S)$$

$$\frac{log_2(N-1) + \Delta}{N} \qquad 4.4$$

where

$$\Delta = log_2(3^T - 2) + \left[T \sum_c^T p(c)\, log_2\, p(c) - \right. \qquad 4.5$$

$$T_1 \sum_c^{T_1} p(c|D \leq S)\, log_2\, p(c|D \leq S) -$$
$$T_2 \sum_c^{T_2} p(c|D > S)\, log_2\, p(c|D > S)\big]$$

For brevity, feature sets generated using the former and latter discretization methods are denoted by AC and FI respectively.

Four different numerical, 2D, descriptor sets were used to generate 'discretized feature sets'. These were:

i. The P_VSA descriptors, proposed by Labute to be widely applicable for QSAR modelling,[295] and successfully applied to hERG inhibition modelling by Dubus *et al.*[91] and by Thai and Ecker.[69,231,296] As per these earlier studies, these were calculated via MOE.
ii. The 23 "relevant" MOE descriptors used by Dubus *et al.* in their "model 1".[91]
iii. The 11 "relevant" MOE descriptors selected by Thai and Ecker.[69]
iv. A set of estimated physicochemical properties (logP, maximum basic pK_a, minimum acidic pK_a) and a topological index (Wiener Index),[182,297] calculated using ChemAxon's *cxcalc* tool.[298] Where no basic (acidic) pK_a values were reported, the descriptor values were set to -10 (+20), the default minimum (maximum) reported value.

The constitution of descriptor set iv was informed via existing understanding of molecular properties relevant for hERG inhibition. The significance of variations in logP, basic pK_a and the incorporation of acidic moieties (i.e. variations in acidic pK_a) for hERG inhibitory potential, along with mechanistic rationales, was previously highlighted by Jamieson *et al.*[299] The Wiener Index increases with the number of atoms, and - for a constant number of atoms - is maximised for a linear structure,[182] i.e. may partially capture molecular shape, including size, which has previously been indicated to be important for hERG inhibition.[300]

For brevity, descriptor sets ii, iii and iv, and all feature sets generated from them, are denoted Dubus-Rel, Thai-Rel and CA respectively.

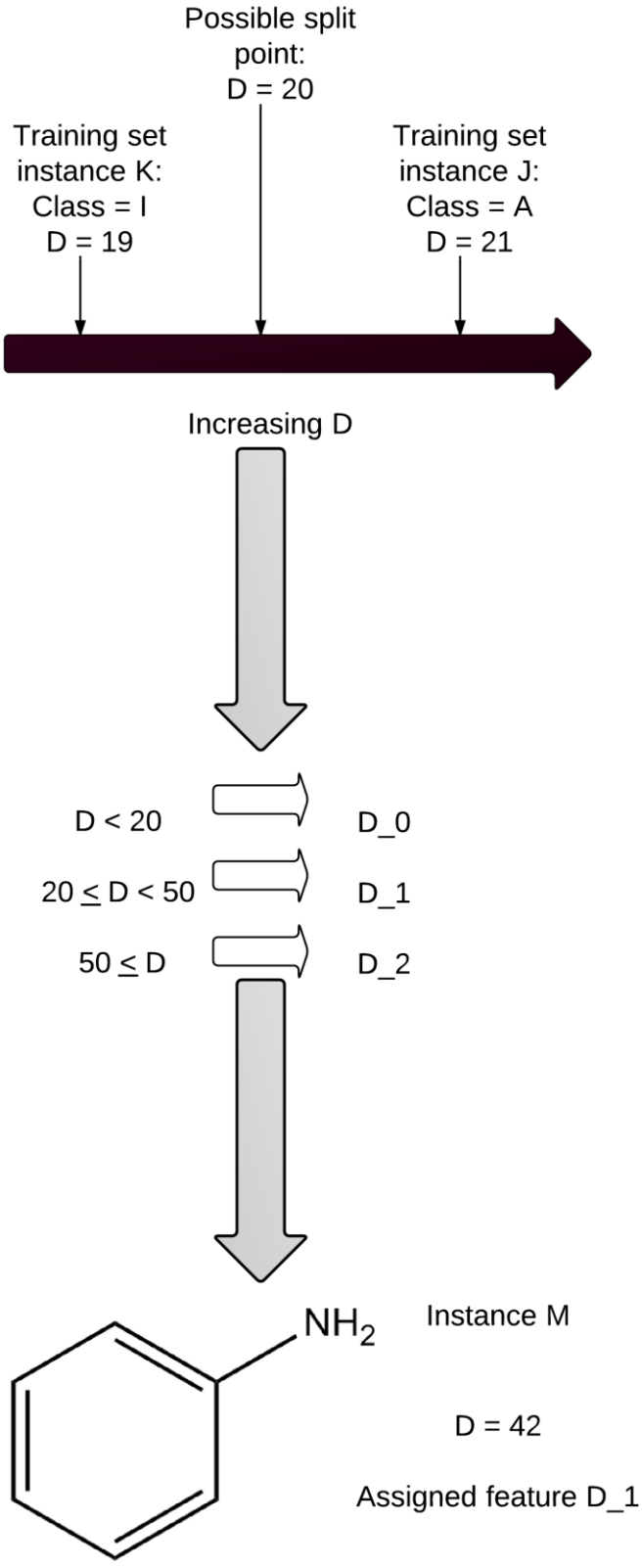


Figure 4.2 The assignment of 'discretized descriptor features' corresponding to descriptor D.

4.3.2.1 Descriptor Calculation

Prior to calculating ChemAxon descriptors, all molecular structures were standardized using ChemAxon's *standardizer* tool* and retained in SDF format.

ChemAxon's *molconvert* tool was then used to generate MOL2 files from the standardized SDFs, from which CFP2 fingerprints, as implemented by Nigsch,[213] were calculated.

Prior to SciTegic fingerprint calculation, using the Molecular Fingerprint component in Pipeline Pilot Student Edition,[301] and MOE descriptor calculation, using MOE's *sddesc* tool, explicit hydrogens were added to the standardized structures using ChemAxon's *standardizer* tool.

All structures in the Dubus-203 and Thai-313 datasets were updated with new, unique, IDs prior to standardizing, using the Pybel module, with all Thai-313 structures being similarly updated prior to generating Binary QSAR models (see section 4.4.2).[177] One structure in the Dubus-203 set, which was not correctly standardized by ChemAxon's *standardizer* tool, was manually standardized† prior to calculating CFP2 and SciTegic fingerprints. In 14 cases, explicit hydrogens were not fully removed when standardizing structures in the Thai-313 dataset.

4.3.2.2 Feature Sets Searched

In total, 94 feature sets were generated. Each fingerprint type corresponded to one feature set, as did all eight combinations of each of the four numerical descriptor sets with both discretization methods. All combinations of discretized feature sets with fingerprint feature sets (3×8) yielded 24 additional feature sets.

All the constituent features of these feature sets can be considered "monograms".[184,213] For each monogram set, "orthogonal sparse bigrams" (OSBs) – non-exhaustive combinations of pairs of monograms – were generated, as previously described by Nisgch,[184,213] using a window-size of three (Figure 4.3). Thus, an additional 35 feature sets were generated based on the combination of all 35 monogram feature sets with their corresponding OSBs. Finally,

* The *standardizer* configuration files are made available with this thesis (Appendix A).

† Full details are provided in FINAL_Dubus203_README-FT.txt, which is made available with this thesis (Appendix A).

all combinations of the three new (monograms and OSBs) fingerprint feature sets with all eight (monogram only) discretized feature sets yielded 24 further feature sets.

The presence of OSBs in the feature sets should allow for synergies between the monograms to be taken into account by linear classifiers, such as Winnow.[213]

The generation of all 94 feature sets evaluated using Winnow is summarised in Figure 4.4.

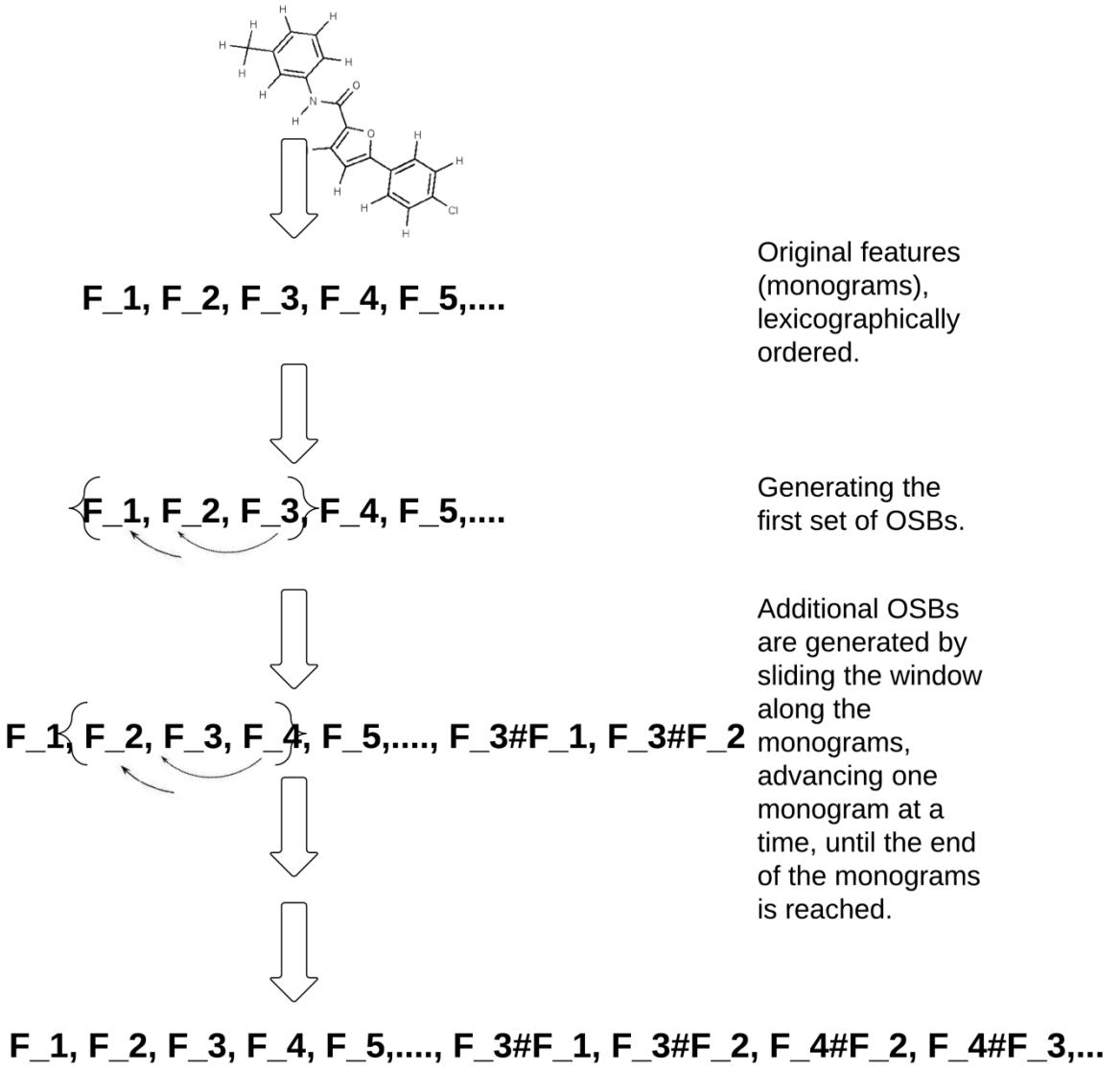


Figure 4.3 The procedure employed to generate orthogonal sparse bigrams (OSBs) as additional features for a generic molecule and generic original set of features (monograms) using a window size of three.

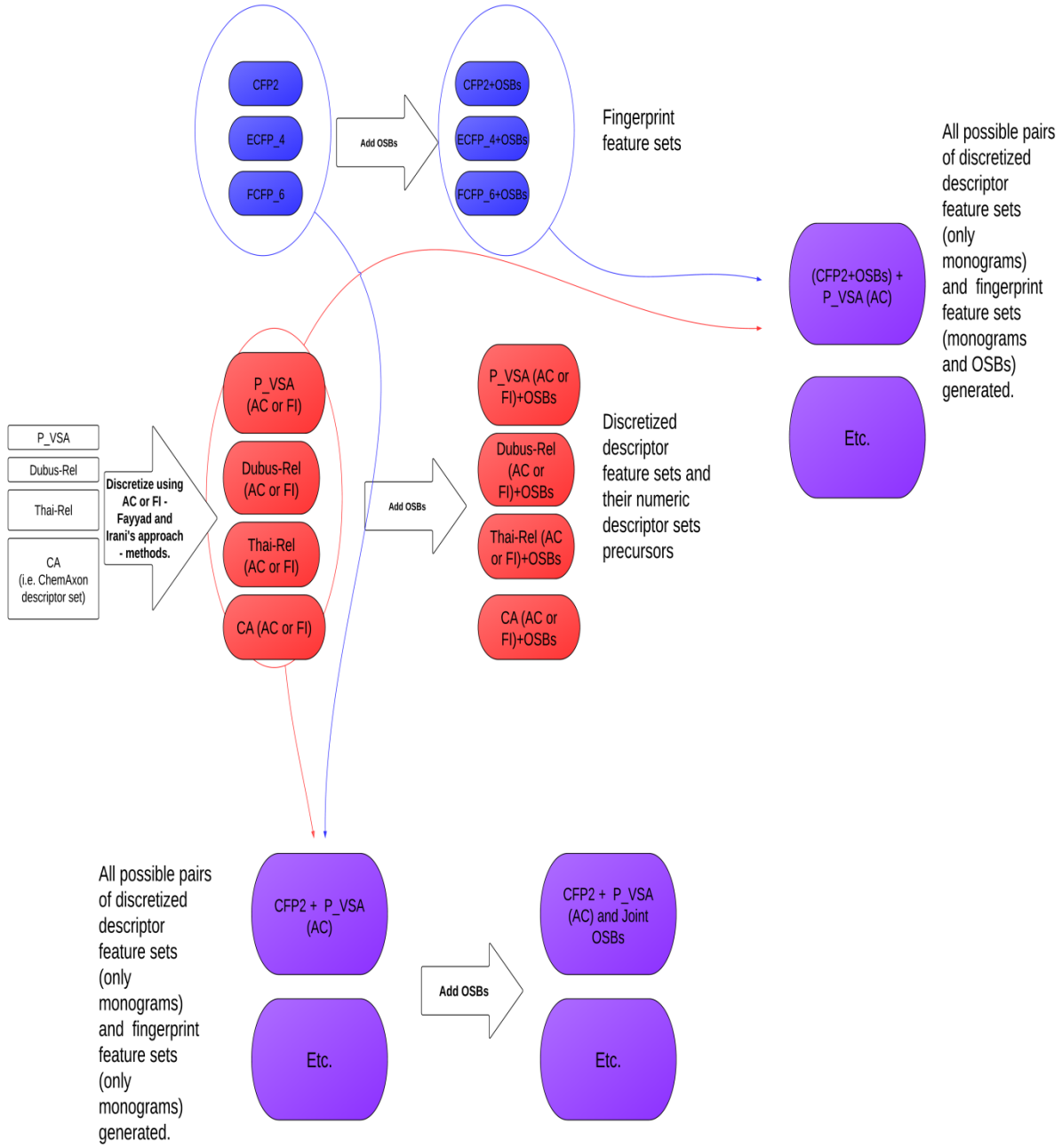


Figure 4.4 A summary of the generation of the 94 feature sets – comprising fingerprint features, discretized descriptor features or both – evaluated in the current work.

4.4 Comparisons with Models Developed by Thai and Ecker and by Dubus et al.

4.4.1 Background

Recent studies by Dubus *et al.*[91] and Thai and Ecker[69] indicated that their binary classification models, using the same threshold as here, were highly predictive. Thus, this author was keen to see how well the models developed here compared with theirs when trained and externally validated on the same datasets.

Dubus and co-workers developed models using MOE's QuaSAR-Classify tool, using the P_VSA and Dubus-Rel descriptor sets referred to above (section 4.3.2).[91] Similarly, Thai and Ecker developed Binary QSAR models based on the aforementioned (section 4.3.2) P_VSA and Thai-Rel descriptor sets.[69] This author implemented all four models using MOE.

4.4.2 Additional Protocols for Descriptor Calculation

In order to more fairly make direct comparisons with the modelling approaches proposed by Dubus *et al.* and Thai and Ecker and better assess dataset dependencies of model performance, this author calculated the descriptors used to build these models using computational procedures more closely reflecting those employed by them (private correspondence with Dr Ismail Ijjaali and Dr Khac-Minh Thai respectively). These procedures started with the structures available prior to any previously described standardization and are referred to, here, as the 'Dubus Standardization' and 'Thai Standardization' procedures.[*]

- Dubus Standardization: ChemAxon's *standardizer* tool was used to remove extraneous molecular fragments and add explicit hydrogens. Subsequently, ChemAxon's *cxcalc* tool was used to calculate major protonation states (pH 7.4), with explicit proton assignment [cxcalc majormicrospecies -H 7.4 -f sdf:H]. The resulting structures, in SDF format, were imported into a MOE database and descriptors were calculated using the GUI, then stored in SDF format.
- Thai Standardization: All structures, post-removal of extraneous fragments and neutralization, if necessary, using ChemAxon's *standardizer* tool, were imported into a MOE database and Energy Minimized – followed by descriptor calculation using the GUI.

[*] The ChemAxon *standardizer* configuration files are made available with this thesis (Appendix A).

4.4.3 Comparisons with QuaSAR-Classify

As described in the MOE documentation, QuaSAR-Classify employs a recursive partitioning algorithm (see Chapter 2, section 2.6.3.2) to grow a single decision tree from the training set. The Node Split Size and Max Tree Depth were set to five and 15 respectively, with random five-fold cross-validation on the training data being used for pruning, in keeping with Dubus *et al.* [91]

All QuaSAR-Classify experiments were run using SVL scripts. All QuaSAR-Classify models were run 50 times within the same MOE session, and, unless stated otherwise, the results presented are the mean over these runs. Since Dubus *et al.* suggested that their models would perform sub-optimally using 'unbalanced' training sets, [91] when the training sets used to select the models developed by this author were unbalanced, subsets of these training sets were generated for training models subsequently evaluated on the corresponding external test sets.

These 'balanced' subsets were generated using MOE's Diverse Subset tool to remove the least diverse compounds from the majority class until the number of remaining compounds equalled the number of compounds in the minority class (i.e. class A). Diverse Subset rankings were generated independently for each class, using (non-bit packed) MACCS keys and the Tanimoto coefficient similarity metric, after importing the structures available after calculation of MOE descriptors, for the models developed by this author, into a MOE database.

For computational expediency, only the previously selected Winnow models were re-evaluated after regenerating all discretized features using the balanced training sets.

4.4.4 Comparisons with Binary QSAR

All models were trained with the Component Limit set to 8, as per Thai and Ecker.[69]

4.4.5 Diverse Subset Partitions of the Thai-313 and Dubus-203 Datasets

The Diverse Subset train/test splits of the Dubus-203[91] and Thai-313[69] datasets, respectively used by Dubus *et al.* and Thai and Ecker to evaluate their models, were sought - principally in order to validate the implementations of their models used in the current work, which employed a later version of MOE (Appendix C).

For the Thai-313 dataset, Diverse Subset calculation utilised all 184 2D MOE descriptors calculated, using the 'Thai Standardization' procedure and bioactivity values ($-\log_{10}(IC_{50}(\mu M))$) calculated from the IC_{50} values provided by these authors.[69] Values of 4.0 and -4.0 were assigned when they reported the IC_{50} as << 1 µM, and >> 10 µM respectively.

Diverse Subset calculation was run from the MOE GUI, using the Euclidean Distance, scaled to unit variance, to assign diversity rankings, with the 240 most diverse compounds selected for the training set. This procedure conformed to that used by Thai and Ecker (inferred from private correspondence with Dr Khac-Minh Thai). Since the outcome of the Diverse Subset protocol is dependent on the identity of the first compound in the database, this author generated all four possible Diverse Subset splits corresponding to a test set of 20 actives and 53 inactives (the closest match to the split presented by Thai and Ecker).[69] These were identified by computing all 313 possible splits using an in-house Python implementation of the Diverse Subset algorithm; 309 of these 313 conceivable splits were excluded, as they did not correspond to a 20:53 division.

For the Dubus-203 dataset, Diversity rankings were computed using (non-bit packed) MACCS keys and the Tanimoto coefficient similarity metric. The Diverse Subset procedure was separately applied to the 'Dubus Standardization' processed structures in classes "A" and "I" (referred to as "High" and "Weak" by Dubus *et al.),*[91] to select the 80 maximally diverse structures in each class for training, with the remaining compounds being used as a test set.

The models developed by this author were also trained and selected on the Dubus-203 and one of the Thai-313 Diverse Subset training sets,* and then 'externally' validated on the corresponding test sets. As explained below, models generated using the Dubus-Rel and Thai-Rel descriptor sets may not have been truly externally validated using this test protocol.

* The train/test partition of the Thai-313 dataset used being the partition for which the highest test set MCC values were obtained using both Binary QSAR models; this enabled a direct comparison between the modelling approaches proposed in this work and the best performance achieved, by this author, with the Binary QSAR models.

4.5 Avoiding Feature Selection Bias: Calculating Overlap between Different Datasets

The Thai-Rel and Dubus-Rel descriptor sets were selected using the entirety of the Thai-313 and Dubus-203 datasets respectively (private correspondence with Dr Khac-Minh Thai and Dr Elodie Dubus). This author prefers the practice of only using the training set for feature selection, to avoid any risk of an optimistic bias on the test set arising from its having been included in the feature selection process.[*]

No structures were identified in common between the ExtTest-Set and Int-Set datasets as well as between the ExtTest-Set and the Thai-313 and Dubus-203 datasets. This prevented the possibility of incurring feature selection bias when the ExtTest-Set was used for model validation. In practice, such bias could always be avoided for the modelling approaches proposed in the current work, when the 'external' test sets overlapped with the Dubus-203 (Thai-313) dataset, if the feature sets based on the Dubus-Rel (Thai-Rel) descriptors were not selected to build the final models.

The structures in common between different datasets were identified by comparing them using stereochemically indifferent, canonical SMILES[302] generated using ChemAxon's *molconvert* tool (molconvert smiles:0), from the structures used for SciTegic fingerprint calculation.[†]

This procedure also identified 107 (87) compounds in the Int-Set with structures in the Thai-313 (Dubus-203) dataset. This procedure was used to identify all dataset overlaps referred to below.

4.6 Results and Discussion

4.6.1 Overall Model Performance

Table 4.2 and Table 4.3 summarise the MCC values obtained using the approaches developed in this work when validated on various 'external' test sets - i.e. sets of compounds neither used to directly select the hyperparameters or features employed by, nor train, the models. In

[*] The good Diverse Subset test set results obtained by both sets of authors using their P_VSA models, and all results obtained by Thai and Ecker on an external test set of 58 compounds,[69,231] are unaffected by such issues.

[†] These comparisons may not fully capture the overlap between datasets due to the presence of different tautomeric forms etc.

addition to presenting results for the original Int-Set/ExtTest-Set split, results are presented for randomly generated 'Int/Ext' splits of the Literature-368 dataset to provide a more robust assessment of the expected performance of the modelling approaches. All random splits were generated such that the same numbers of strong, moderate and weak inhibitors were maintained in each Int/Ext split as per the Int-Set/ExtTest-Set. The MCC values obtained with this author's implementation of the Binary QSAR and QuaSAR-Classify models on these test sets (when trained on the same data as the models developed in this work) are presented for comparison.

The 'external' test sets for the random partitions of the Literature-368 dataset and the Diverse Subset test sets for both the Dubus-203 and Thai-313 datasets used to 'externally' validate this author's models, overlapped with both the Thai-313 and Dubus-203 datasets. The Dubus-203 (Thai-313) test set contained approximately 17 (15) compounds in the Thai-313 (Dubus-203) dataset. For the random partitions of the Literature-368 dataset - random:1, random:2 and random:3 - there were approximately 40 (34), 41 (29) and 38 (29) 'external' validation set compounds, respectively, in the Thai-313 (Dubus-203) dataset. Thus, when 'biased' features based on the Dubus-Rel or Thai-Rel descriptor sets were selected, additional models were generated based on an 'unbiased' feature selection protocol – excluding any such feature sets from selection. The consistently better results obtained on these test sets with this author's models, including when 'unbiased' feature sets were selected, suggest that the poorer results obtained for the original Int-Set/ExtTest-Set split were not due to the absence of feature selection bias.

Rather, as illustrated in Figure 4.5, these poor results appear to reflect extrapolation beyond the applicability domain of the models: many compounds in the original ExtTest-Set appear to lie outside of the chemical space occupied by the Int-Set, yet this is not apparent for the random Int/Ext splits of the Literature-368 dataset (Figure 4.6). Figure 4.7 likewise indicates that the Diverse Subset test sets, for which results are presented in Table 4.3, lay within the applicability domain of the models trained on the corresponding training sets.

The compositions of all train/test partitions are presented in CSV files made available with this thesis (Appendix A), along with additional details required to exactly reproduce results

obtained for the pseudo-stochastic methods used to develop models in this work: Winnow and Random Forest.*

The range of MCC values obtained on the 'external' validation sets, with either feature selection protocol, is comparable with those previously reported in the literature (Appendix B, Table B.1). Save for two cases, all models developed by this author, trained and validated on the same data, achieved higher (mean) MCCs than all Binary QSAR and QuaSAR-Classify models, even when truly external validations of this author's models are presented and not of these others (see section 4.5).

Figure 4.8 graphically summarises the relative performance of the Winnow models (based on an 'optimised' number of training cycles), on truly external data, and all Binary QSAR and QuaSAR-Classify models trained and validated on the same data. As later discussed in detail, these comparisons also highlight the considerable variation in estimated performance of these approaches when trained and validated on different datasets. Figure 4.9 graphically summarises the relative performance of these linear Winnow models compared to the performance of the corresponding SVM and Random Forest (RF) models. The comparable, or better, performance of Winnow to these more sophisticated, non-linear Machine Learning algorithms is interesting: Nigsch *et al.* previously found that, when models were built using the same features, Random Forest consistently outperformed Winnow, in terms of the MCC.[184]

Detailed (mean) figures of merit, including 5CV results, obtained for selected models, are presented, along with the corresponding selected hyperparameters, in a set of CSV files made available with this thesis (Appendix A). Therein, results are also presented (Thesis_Chapter4_Table_S4k.xlsx) which compare the (mean) MCC values presented in Table 4.2 and Table 4.3 to the (mean) MCC values quantifying the performance of these models when applied to the separation of strong (mean $pIC_{50} > 6.00$) from moderate ($5.00 <$ mean $pIC_{50} \leq 6.00$), or weak (mean $pIC_{50} \leq 5.00$), inhibitors in the corresponding test sets. Thesis_Chapter4_Table_S4k.xlsx also presents corresponding Bonferroni corrected[303] p-values.† The corrected p-values for all truly externally validated models proposed by this

* Such files are not provided for the Dubus-203 dataset, which is available from Aureus Sciences.[102]

†The uncorrected p-values were calculated (see Chapter 2, section 2.6.4.1) from the MCC value (or mean value for the pseudo-stochastic methods – Winnow, RF, QuaSAR-Classify)

author, when trained and tested on the random:1 and random:2 partitions of the Literature-368 dataset, are less than 0.05 - which strongly suggests that these models were capable of discriminating moderate from strong inhibitors, as would be of value in the pharmaceutical industry.[15] Indeed, save for one SVM model, all models developed by this author performed better, in terms of the (mean) MCC, than expected for a random predictor (MCC equal to zero) when applied to the separation of strong and moderate inhibitors in truly external test sets.

4.6.2 The Effect of Unbalanced Training Sets

The use of unbalanced training sets (i.e. with more compounds in class I than A) appears to bias the Winnow models in favour of predicting compounds as non-blockers (i.e. compounds in class I). For all corresponding 'externally' validated Winnow models, the mean numbers of true positives (here, blockers are 'positives' and non-blockers are 'negatives') and false positives (true negatives and false negatives) decreased (increased) upon moving from a balanced to an unbalanced training set. The introduction of bias towards the majority class in the training set has long been recognised as a problem for many learning algorithms.[304]

Interestingly, however, *all* models trained here using unbalanced data had higher (mean) recall and, save for one Random Forest model validated on the Thai-313 dataset, higher (mean) precision for class I than class A when 'externally' validated. The combination of higher precision *and* higher recall, for the 'external' validation sets, suggests these models were better able to identify non-blockers than blockers in these test sets.

---

obtained when training and testing the selected modelling approaches on a given train/test partition. They denote the conditional probability of obtaining a (mean) MCC value with at least the magnitude observed on a given test set, supposing a random predictor had been built on the corresponding training set. The Bonferroni correction provides an upper bound, equal to the value below which the corrected p-values are deemed statistically significant, to the conditional probability, supposing an unknown number of approaches performed like random predictors on average, of erroneously declaring that any model (or random selection from a set of corresponding models for the pseudo-stochastic algorithms) built, in this work, on one of the training sets would perform no differently, in terms of the mean MCC across various test sets, to a random predictor - based on the (mean) MCC value observed on the single corresponding test set.

Int/Ext Split	Feature Set Selected (Winnow, SVM, RF)	Model							
		Winnow		SVM	RF	Binary QSAR		QuaSAR-Classify	
		Single cycle	Multi-ple cycles			P_VSA	Thai-Rel	P_VSA	Dubus-Rel
Original	CFP2 and Thai-Rel(AC)	0.26	--	0.08	0.34	0.16	0.08	--	--
Random:1	ECFP_4 and CA(FI)	0.53	0.59	0.55	0.52	0.29	0.12	--	--
Random:2	ECFP_4 and CA(AC)	0.46	0.51	0.47	0.52	0.34	0.17	--	--
Random:2	ECFP_4 and Thai-Rel(AC)	0.44	0.42	0.40	0.54	0.34	0.17	--	--
Random:3	ECFP_4 and CA(AC)+ Joint OSBs	0.43	--	0.40	0.44	0.03	-0.04	--	--
Original (Balanced training set)	CFP2 and Thai-Rel(AC)	0.10	--	--	--	--	--	-0.03 (0.04)	0.15 (0.18)
Random:1 (Balanced training set)	ECFP_4 and CA(FI)	0.42	0.49	--	--	--	--	0.29 (0.37)	0.35 (0.38)
Random:2 (Balanced training set)	ECFP_4 and CA(AC)	0.26	0.27	--	--	--	--	0.03 (0.06)	0.22 (0.26)
Random:2 (Balanced training set)	ECFP_4 and Thai-Rel(AC)	0.28	0.29	--	--	--	--	0.03 (0.06)	0.22 (0.26)
Random:3 (Balanced training set)	ECFP_4 and CA(AC)+Joint OSBs	0.37	--	--	--	--	--	0.22 (0.24)	0.25 (0.26)

Table 4.2 Performance of this author's selected models and literature models on 'external' test sets for all ('Int/Ext') partitions of the Literature-368 dataset into training and 'external' test sets. MCC values (mean MCC values for Winnow, Random Forest (RF) and QuaSAR-Classify) are presented. Values in parentheses are the maximum MCC values obtained across the 50 runs of the QuaSAR-Classify module. Missing values for Winnow (multiple cycles) signify that a single cycle gave the best 5CV result.

Int/Ext Split	Feature Set Selected (Winnow, SVM, RF)	Model							
		Winnow		SVM	RF	Binary QSAR		QuaSAR-Classify	
		Single cycle	Multi-ple cycles			P_VSA	Thai-Rel	P_VSA	Dubus-Rel
Dubus-203 Diverse Subset	ECFP_4+OSBs and CA(FI)	0.84	--	0.83	0.87	--	--	0.39 (0.65)	0.45 (0.65)
Dubus-203 Diverse Subset	CFP2+OSBs and Dubus-Rel(AC)	0.78	--	0.83	0.79	--	--	0.39 (0.65)	0.45 (0.65)
Thai-313 Diverse Subset	CFP2 and CA(FI)+Joint OSBs	0.80	--	0.79	0.83	0.60	0.68	--	--
Thai-313 Diverse Subset	CFP2 and Dubus-Rel(AC)+Joint OSBs	0.81	--	0.79	0.82	0.60	0.68	--	--

Table 4.3 MCC values obtained on 'external' test sets for all partitions of the Thai-313 and Dubus-203 datasets used to evaluate this author's modelling procedures. See Table 4.2 for presentational details.

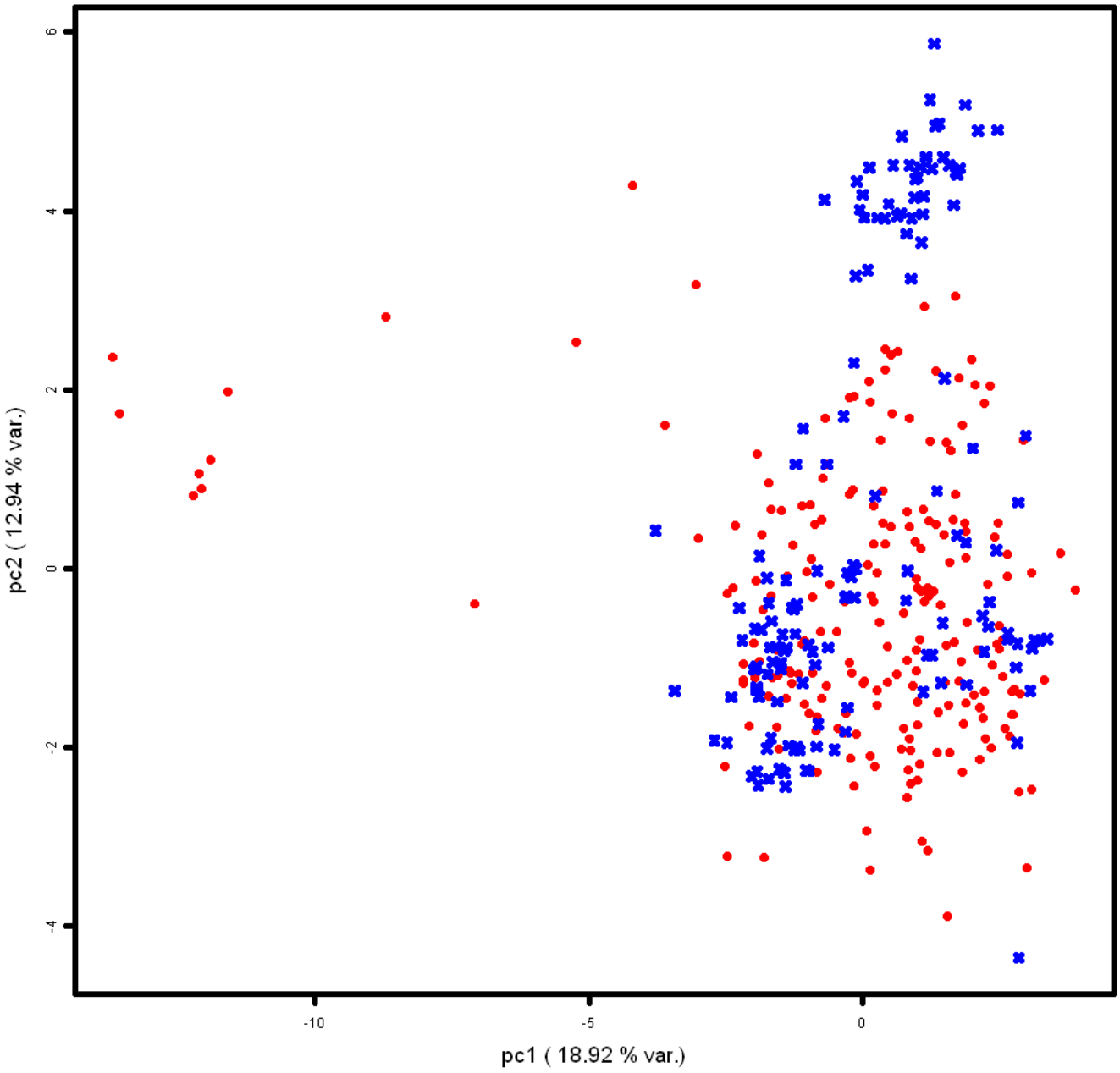


Figure 4.5 Distribution of Int-Set (red) and ExtTest-Set (blue) within the plane defined by the first two principal components (PCA plot) for the P_VSA descriptor set (computed as per section 4.3.2.1). Principal components were calculated from the combined Int-Set and ExtTest-Set, using the prcomp( ) function in R[225] - with scale=TRUE.

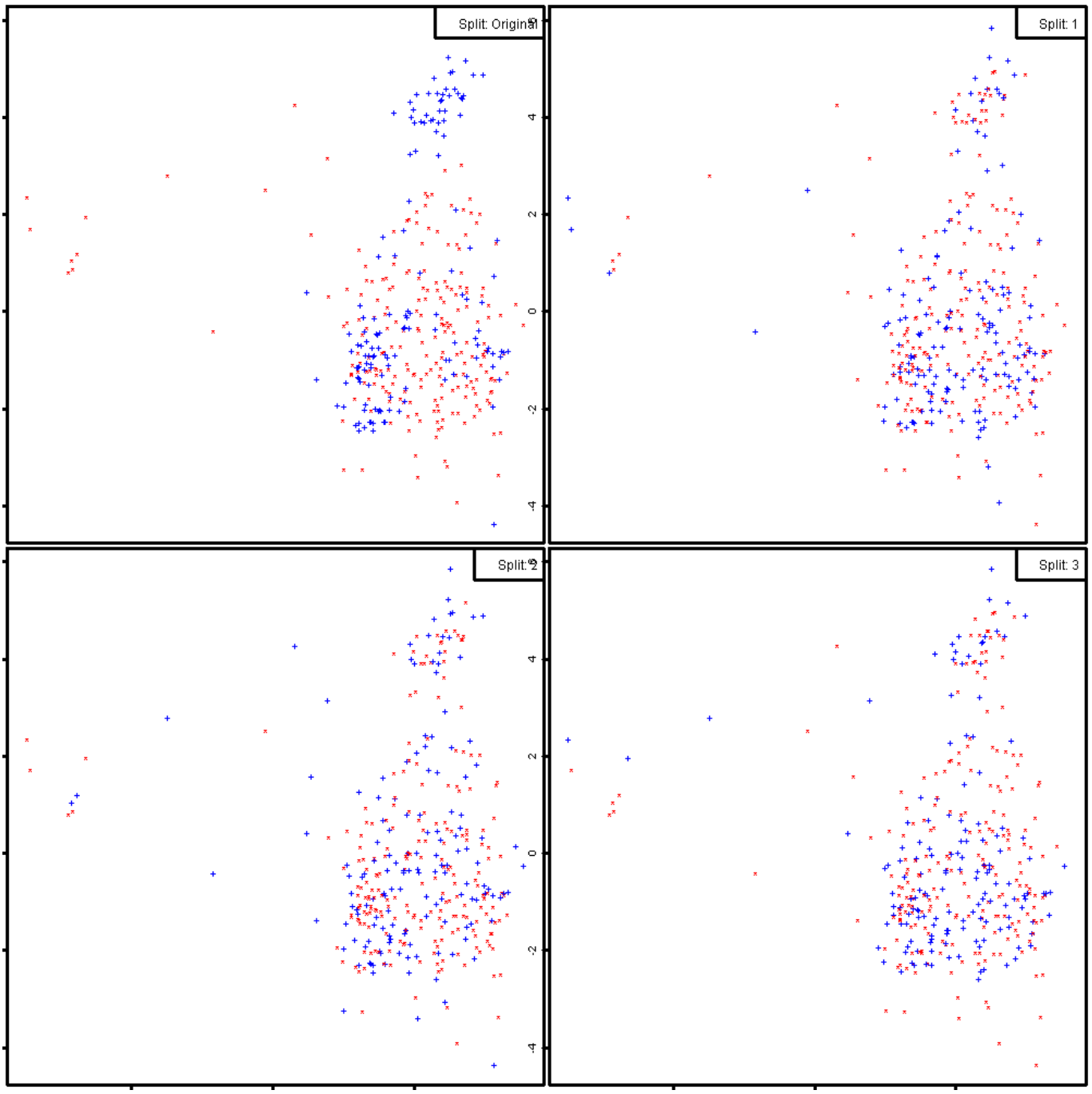


Figure 4.6 PCA plots generated as per Figure 4.5 (training set: red, 'external' test set: blue), for all splits of the Literature-368 dataset; from top left to bottom right: original split, random:1, random:2 and random:3.

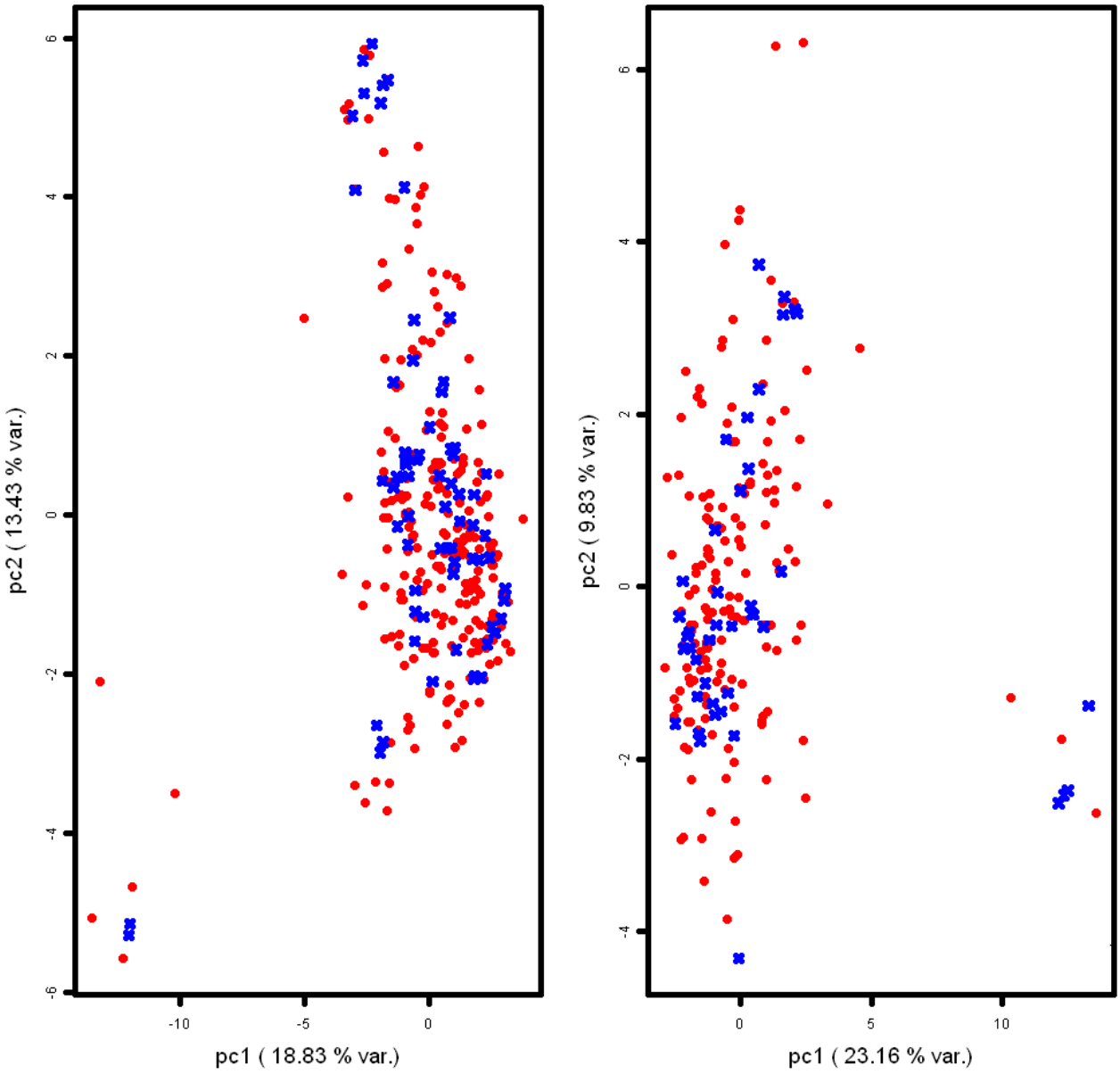


Figure 4.7 PCA plots (generated as per Figure 4.5) for the Diverse Subset partitions of the Thai-313 (LHS) and Dubus-203 (RHS) datasets for which results are presented in Table 4.3.

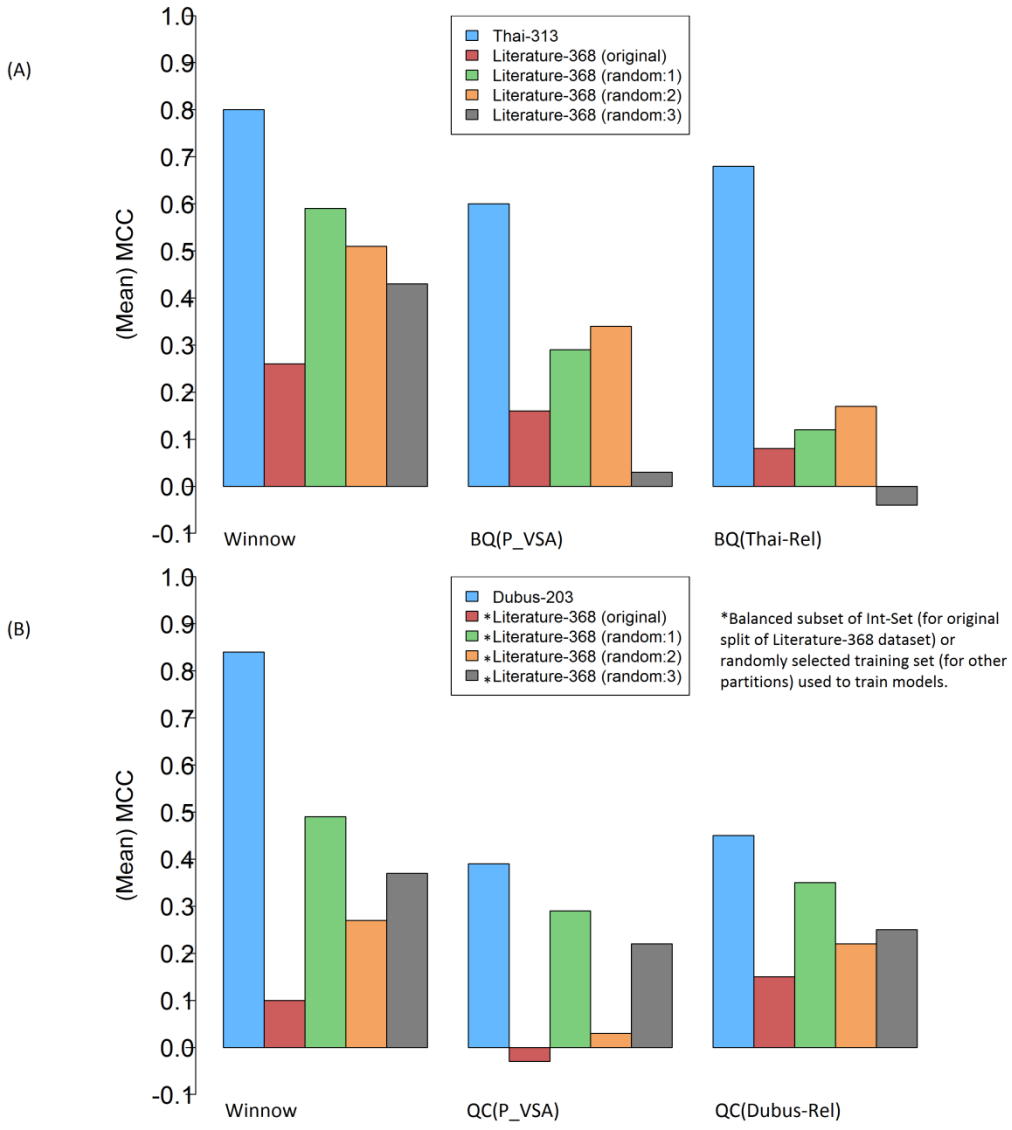


Figure 4.8 Mean MCC values for externally validated (i.e. only 'unbiased' feature sets were considered, where relevant) selected Winnow models (generated using multiple training cycles, where relevant), compared to QuaSAR-Classify (QC) mean MCC values and Binary QSAR (BQ) MCC values when trained and tested on the same data.

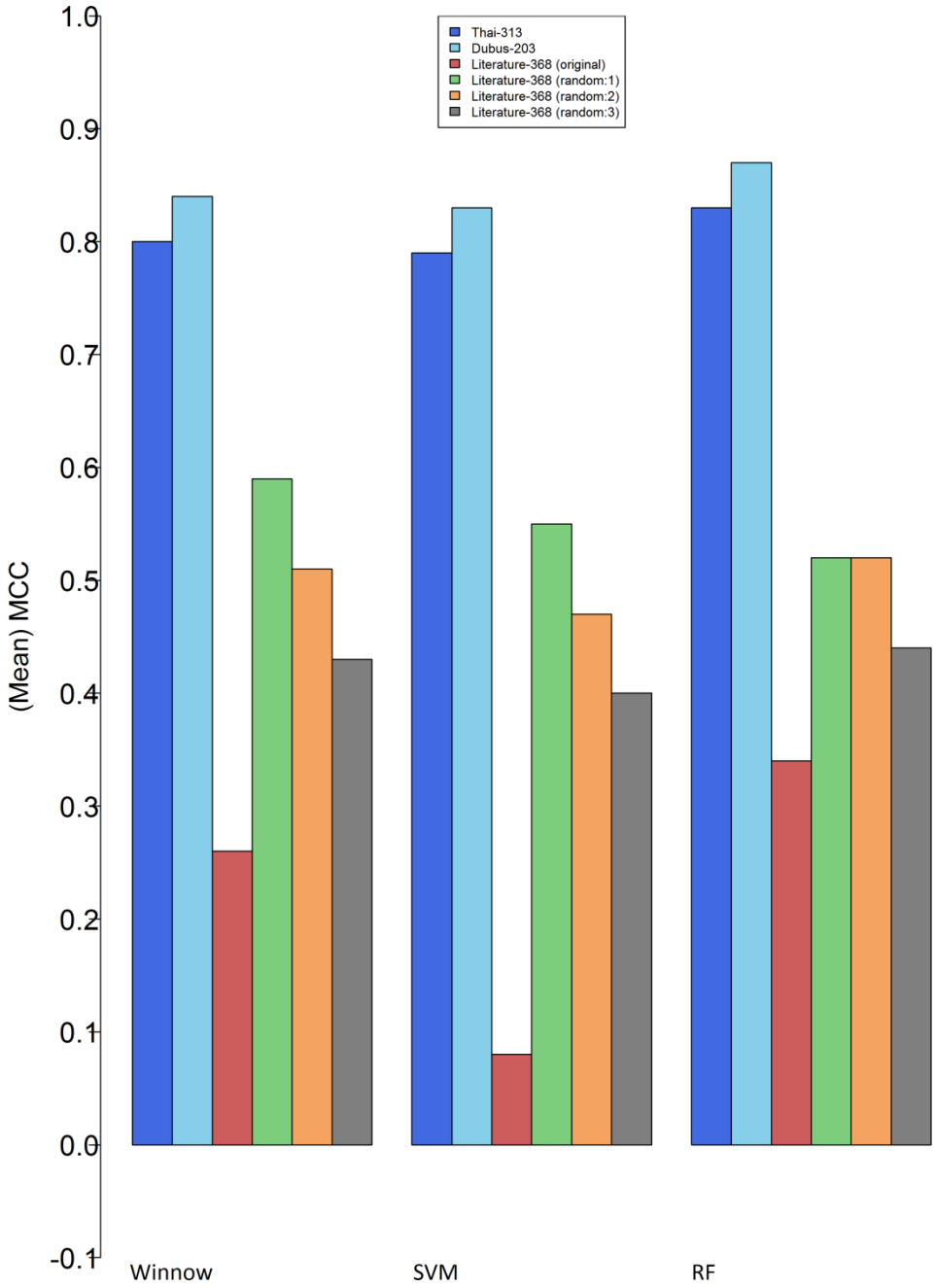


Figure 4.9 Mean MCC values for externally validated Winnow models (as per Figure 4.8), and the corresponding (mean) MCC values for the externally validated SVM and RF models.

4.6.3 Features Emphasised by the Models Developed in this Work

To better understand the models, estimates were computed of the importance of each feature in all 'externally' validated Random Forest and Winnow models – where computationally feasible.* For Winnow, the C++ code was modified to print the arithmetic mean, across all scorers, weights assigned to each feature for both classes and the feature importance was estimated as the absolute difference in these mean weights. For Random Forest, the Gini importance measure was used.[223] Overall feature importance was assigned on the basis of the median rank (more important features having lower values) across all 50 (five) runs of Winnow (Random Forest). For the purposes of the following discussion, the 'important features' are the top ranking 1% of features for the model in question.

Evidently, the modelling methods differ in the manner in which they make use of the features: in contrast to the SVM and Random Forest models, the Winnow algorithm can only model linear relationships between the classes and the features and only make predictions on the basis of features present in a compound. This is reflected in the incomplete overlap between the important features for each model: for all, evaluated, Random Forest models trained on subsets of the Literature-368 dataset, less than 50% of the important features were also deemed important for the Winnow models trained on the same data. Except for the 'random:1' split of this dataset, the common important features included fingerprint and discretized descriptor features, suggesting the distinct types of descriptor were adding comparably important information to both types of models.

This author was interested as to whether these models might highlight novel molecular substructures associated with high/low potency hERG inhibition. The median (across all training cycles) signed difference in the mean weights assigned to the important Winnow features was taken as *indicative* of the 'global'† association of those features with hERG active/inactive compounds. This difference was defined as positive for features with a higher mean weight for class A than for class I. Those 'important features' corresponding to CFP2 or ECFP_4 fingerprint components were closely inspected; SMILES representations of the substructure(s) mapping onto the ECFP_4 features (across the whole of the Literature-368 or Dubus-203 dataset for the corresponding models) were generated using Pipeline Pilot. An

* This was not the case for the Random Forest models based on Dubus-Rel features.

† In addition to the caveats raised below, since the training sets cannot be claimed to span all of chemical space, any trends are, of course, only indicative.

important difference between the CFP2 and ECFP_4 features is that the latter encode a specific set of allowed, non-hydrogen, attachment points for the corresponding substructures.[185] It is important to note, however, that there is not a one to one mapping between molecular substructures and either type of fingerprint features. Multiple substructures *may* correspond to the same feature, and, more rarely, the same ECFP_4 feature may correspond to multiple substructures.[185] Furthermore, the aforementioned SMILES representations for subtly different substructures may be equivalent (personal communication from Accelrys Limited).

The ECFP_4 based Winnow models appear to exhibit consistency with trends previously noted in the literature, as illustrated in Table 4.4.[*] In keeping with Song and Clark's analysis of their linear Support Vector Regression model,[222] the highlighted Winnow models associated a feature encoding a fluoro-phenyl moiety with strong hERG inhibition and a feature encoding an amide associated with weaker inhibition. The association of a feature (2131972380) encoding a tertiary amine, located away from the molecular periphery, with potent hERG blockade, learnt by all ECFP_4 based Winnow models and highlighted for one such model in Table 4.4, is consistent with the well-known association between such tertiary amines and hERG inhibition.[222,299,305] Various mechanistic rationales have been proposed for this: protonation of the basic amine possibly leads to pi-cation interactions or facilitates non-classical hydrogen bonds with hERG Tyr652 residues.[299,305]

The latter of these findings demonstrates that the ECFP_4 based Winnow models were capable of capturing chemically meaningful relationships (i.e. relationships which were unlikely to be artefacts of the dataset) between hERG inhibition and molecular structure. However, in general, care must be taken when seeking to interpret the contributions of different features to these types of models.

Firstly, chance correlations between mechanistically important moieties and chemically unimportant moieties may reveal spurious relationships between hERG blockade liability and chemical substructures. This caveat is perhaps most notably illustrated by the fact that a feature (863188371) corresponding to an ethyl group was deemed the 5th most important for

[*] All features presented in Table 4.4 were amongst the important features for these models. All these features, where contributing to a Winnow model, were exclusively associated with one or more of the SMILES presented in Table 4.4 and a median difference in mean weights with the same sign as presented here.

the Winnow model built on the training set of the random:1 split of the Literature-368 dataset using multiple training cycles.

It must also be noted that Winnow appears to have assigned questionably high significance to some features - given the relative number of active and inactive training set compounds associated with the feature in question. For example, the fluoro-phenyl moiety encoded by feature -296909061, highlighted as an important hERG blocker feature in Table 4.4, was associated with a higher median difference in mean weights than the 2131972380 feature (see Table 4.4) which was more clearly associated with active training set compounds.

In addition, Winnow, as a linear algorithm, does not consider the contextual significance of these features and the effect of adding/removing a molecular fragment on hERG inhibitory potency may be dependent on the remaining chemical functionality as well as the position of the substitution.

One final point to bear in mind when seeking to chemically interpret these models is the possibility that hERG inhibition *may* arise from binding to different (open, closed or inactivated)[58] states of the ion channel and, indeed, different binding sites.[306] The exact nature of the interactions, hence the chemical significance of molecular substructures, *may* be different for blockers binding to different states of/sites within the ion channel.[300,306,307]

The median difference in mean weights is presented, along with corresponding SMILES where relevant, for the important Winnow features in all assessed models, in an Excel file made available with this thesis (Appendix A).

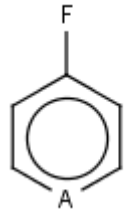
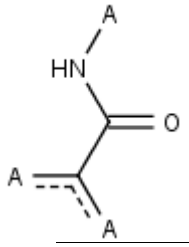
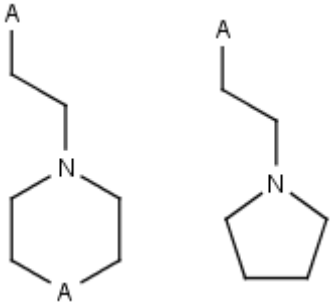
Feature	Fragment SMILES	Winnow Model	Median $\langle W(A)\rangle - \langle W(I)\rangle$	Training Set Occurrence Ratio Per Class
-296909061		Training set: Random:1 Multiple training cycles	1.0196	Class A: 18/55 (0.33) Class I: 17/165 (0.10)
1430169877		Training set: Dubus-203 Feature set: ECFP_4+OSBs and CA(FI))	-2.4957	Class A: 1/80 (0.01) Class I: 7/80 (0.09)
2131972380		Training set: Random:1 Multiple training cycles	0.9675	Class A: 17/55 (0.31) Class I: 5/165 (0.03)

Table 4.4 Some of the associations, learnt by the Winnow models, between 'important' ECFP_4 features and hERG blockade which were consistent with trends previously noted in the literature. N.B.: $\langle W(A)\rangle$ denotes the arithmetic mean weight, across all scorers, assigned to the feature for class A etc. The median difference over all 50 training set orders is reported. All Dubus-203 training set compounds assigned feature 1430169877 were observed to possess a corresponding amide fragment. Occurrence ratios denote the fraction of compounds belonging to the class in question and containing the feature. All SMILES patterns were visualised using MarvinView (version 5.5.0.1);[298] the 'A' symbols denote wildcard, i.e. undefined, heavy atom connections.[185,308]

4.6.4 Dataset Performance Dependencies

This author was surprised to have obtained considerably worse results on the Literature-368 dataset with Thai and Ecker's and with Dubus *et al.*'s binary classifiers compared with those that they previously reported.[69,91] However, as extensively detailed in Table 4.5, the results obtained with this author's implementations of their methods on their datasets indicate only marginal differences between the different implementations of their binary classifiers: the differences between the maximum test set MCCs obtained here and the corresponding test set MCCs calculated from their previous results are either less than or marginally greater than the

range of MCC values obtained using this author's implementations. Furthermore, it is clear that this author was unable to exactly reproduce Thai and Ecker's test set exactly – presumably due to using a later version of MOE (Appendix C) – making it impossible to obtain a direct comparison between the different implementations of their binary classifiers.

These findings suggest that the poorer results obtained on the Literature-368 dataset primarily reflect differences in the data used to train and validate these models.

The considerable range in (mean) results obtained using the Binary QSAR and QuaSAR-Classify models, for different train/test partitions of the Literature-368 dataset (Table 4.2), indicates that the consistently lower results obtained on this dataset are not merely a reflection of the reduction in training set size (compared to the Thai-313 and Dubus-203 datasets). The consistently lower results obtained on the Literature-368 dataset, compared to the Leave One Out cross-validated (LOOCV) results on the Thai-313 dataset (Table 4.5), suggest the lower results obtained, with the Binary QSAR models, do not merely reflect the absence of a Diverse Subset train/test partition of the Literaure-368 dataset.[*]

It could be argued that the variation in overall binary classification performance of these models is more appropriately assessed in terms of the chi-squared p-values corresponding to the (mean) MCC values, rather than the MCC values themselves, since these may allow fairer comparisons between the results obtained on differently sized test sets (see Chapter 2, section 2.6.4.1). Table 4.6 presents such p-values for all partitions of the Literature-368, Thai-313 and Dubus-203 datasets used to externally validate this author's models. The p-values obtained for the Binary QSAR (QuaSAR-Classify) approaches on the Thai-313 (Dubus-203) test set are consistently lower (lower in 50% of cases) than those obtained on the Literature-368 dataset. This suggests that the (mean) MCC values obtained on the Literature-368 test set (148 compounds) were not lower than those obtained on these other test sets (Dubus-203: 43 compounds, Thai-313: 73 compounds) as an artefact of the test set size having increased.[†]

---

[*] The Diverse Subset algorithm, as described in the MOE documentation, is designed to ensure pairs of similar molecules are split between the training and test set when partitioning a dataset.[285]

[†] The probability of a random predictor obtaining an MCC value greater than or equal to a given size, i.e. the p-value, is smaller for a larger test set - i.e. the chance of a random predictor obtaining a large MCC value decreases with increasing test set size. Hence, if it

Since the Thai-313 LOOCV effective test set size (240 compounds, i.e. the number of compounds in the training set) is larger than the Literature-368 test set size, the decrease in MCC values obtained on the Literature-368 dataset using the Binary QSAR methods cannot be an artefact of test set size.

---

were supposed that these two approaches effectively yielded random predictors, the MCC would be expected to decrease for larger test set.

(i) Binary QSAR Results (Diverse Subset Splits of Thai-313 Dataset)									
Source	Validati-on	Descrip-tors	Acc.	Rec. (A)	Rec. (I)	Prec. (A)	Prec. (I)	MCC	MCC Range (Across All Splits)
Current work (Example split)	Training Set - LOOCV	P_VSA	0.79	0.51	0.93	--	--	0.50	0.51-0.49
		Thai-Rel	0.81	0.61	0.91	--	--	0.55	0.59-0.55
	Test Set	P_VSA	0.85	0.60	0.94	0.80	0.86	0.60	0.60-0.48
		Thai-Rel	0.88	0.70	0.94	0.82	0.89	0.68	0.68-0.57
Thai and Ecker[69]	Training Set - LOOCV	P_VSA	0.82	0.58	0.94	--	--	0.57	--
		Thai-Rel	0.80	0.61	0.91	--	--	0.56	--
	Test Set	P_VSA	0.89	0.68	0.96	0.87	0.90	0.70	--
		Thai-Rel	0.93	0.79	0.98	0.94	0.93	0.82	--

(ii) QuaSAR-Classify Test Set Results (Diverse Subset Split of Dubus-203 Dataset)						
Source	Descrip-tors	Acc.	Rec. (A)	Rec. (I)	MCC	MCC Range (Across All Runs)
Current work (Highest MCC from 50 runs)	P_VSA	0.79	1.00	0.67	0.65	0.65-0.36
	Dubus-Rel	0.79	1.00	0.67	0.65	0.65-0.43
Dubus *et al.*[91]	P_VSA	0.81	0.94	0.74	0.66	--
	Dubus-Rel	0.74	0.94	0.63	0.56	--

Table 4.5 Results obtained on the Thai-313 and Dubus-203 datasets in the current work compared to those reported by Thai and Ecker[69] and Dubus *et al.*[91] The split of the Thai-313 dataset for which results are highlighted corresponds to the split with the highest MCC values obtained for both Binary QSAR models (i.e. the split used to evaluate this author's models). The partitions of the Thai-313 dataset used to generate results in the current work (training set :80 actives/160 inactives, test set: 20 actives/53 inactives) differed slightly from the single partition used by Thai and Ecker (training set: 81 actives/159 inactives, test set: 19 actives/54 inactives), as explained in section 4.4.5. All MCC values obtained by Dubus *et al.* and Thai and Ecker were estimated herein. Acc., Rec. and Prec. denote accuracy, recall and precision respectively.

Int/Ext Split	Model			
	Binary QSAR (P_VSA)	Binary QSAR (Thai-Rel)	QuaSAR-Classify (P_VSA)	QuaSAR-Classify (Dubus-Rel)
Original	5.5E-02	3.1E-01	--	--
Random:1	3.5E-04	1.5E-01	--	--
Random:2	3.3E-05	3.8E-02	--	--
Random:3	7.4E-01	1.0E+00	--	--
Dubus-203 Diverse Subset	--	--	1.1E-02	3.3E-03
Thai-313 Diverse Subset	3.0E-07	6.6E-09	--	--
Original (Balanced training set)	--	--	1.0E+00	7.8E-02
Random:1 (Balanced training set)	--	--	3.4E-04	2.1E-05
Random:2 (Balanced training set)	--	--	6.7E-01	6.7E-03
Random:3 (Balanced training set)	--	--	8.0E-03	2.6E-03

Table 4.6 Chi-squared p-values corresponding to (mean, across 50 runs, for QuaSAR-Classify) test set MCC values obtained here for the partitions of the Literature-368, Dubus-203 and Thai-313 dataset used to assess this author's models. These p-values were computed using the CHIDIST( ) function in Excel 2007 (32-bit), supposing one degree of freedom (see Chapter 2, section 2.6.4.1), save where a negative MCC was obtained – for which it was supposed that the model must effectively be a random predictor and the p-value was set to one.

4.7 Conclusions

The work presented here shows the considerable variability in results that may be obtained by training and validating the same hERG blocker classifiers on different datasets. This is highlighted by the considerable reduction in model performance observed on the Literature-368 dataset with the models proposed by Dubus *et al.* [91] and by Thai and Ecker[69] compared to both previously published validations of these models, as well as validations of these approaches in the current work, on previously published datasets.

This author's Winnow models perform comparably to models, based on the same descriptors, employing the more sophisticated, non-linear Machine Learning algorithms: Support Vector Machines and Random Forest. The use of multiple training cycles was found to sometimes improve the performance of the Winnow models. Features based upon numerical molecular properties, an advance on earlier work with Winnow in QSAR research, were found to make positive contributions to the models.

Moreover, the contributions of some features to these Winnow models are interpretable in terms of known relationships between molecular structure and hERG blockade potency. However, considerable caution is advocated when seeking to infer novel structure-activity relationships from these kinds of models.

Having taken care to minimise possible model and feature selection bias, all binary classifiers developed in the current work almost always yielded results on external validation sets as good as or better than the classifiers proposed by Dubus *et al.* and by Thai and Ecker. Once possible extrapolation beyond the applicability domain of the model was removed, this author's approaches perform well when compared to previously published models.

Chapter 5 Development of Novel 3D Descriptors

This chapter describes the conception, implementation and performance evaluation of novel 3D descriptors. Specifically, a chemical feature 'coloured' version of Ballester and Richards' Ultrafast Shape Recognition (USR) methodology, 'Atom Type USR' (ATUSR), was proposed. The performance of this descriptor set when used for the generation of Quantitative Structure-Activity Relationships (QSARs) for protein-ligand binding problems was assessed. As highlighted in earlier chapters, some protein-ligand binding events may mediate serious forms of toxicity, making method development for predicting the binding affinities and selectivities of ligands for toxicologically relevant proteins an important contribution to computational toxiciology. The performance of the ATUSR descriptor set was compared to (standard) descriptor sets proposed in the literature and an adaptation of ATUSR designed to investigate the value added by explicitly encoding the 3D distribution of chemical functionality. The effects of conformer generation and dataset variation upon the results were also investigated.

5.1 Introduction

The importance of molecular shape as a determinant of protein receptor mediated biological activity, and receptor selectivity in particular, has been highlighted in the recent literature.[263,264,309,310] The need for "shape complementarity" between a ligand (a 'small' molecule) and a protein binding site arises for three reasons:[311]

1. Ligands with certain shapes are simply unable to fit into the binding site.
2. Close contacts are required to form sufficiently enthalpically favourable intermolecular interactions.
3. Space filling of the receptor site may lead to the entropically favourable expulsion of bound water molecules.

Different approaches to characterising molecular shape *in silico* have been proposed. These methods may be divided into "alignment-based" and "alignment-free"[309] (or "superposition" and "non-superposition")[263] methods.

Early superposition methods treated molecules as sets of overlapping hard spheres. However, a more realistic and computationally efficient approach characterises shape in terms of a set of atom centred Gaussian functions and computes shape similarity in terms of the maximal obtainable overlap (corresponding to an 'optimal' alignment) between the Gaussian based

representations of two molecular conformers.[263,311–313] A variation on this approach is employed by OpenEye's Rapid Overlay of Chemical Structures (ROCS) software;[314,315] recently, freely available, open source software programs employing a variation on this approach have been described.[316,317]

However, the need for molecular alignment inherent to superposition methods is an intrinsic limitation to their computational efficiency[263] and, possibly, if the optimal alignment is not obtained, to their effectiveness for shape comparison.[263,316] Furthermore, they do not immediately offer descriptors that could be used for QSAR modelling.[*] Various non-superposition, or rotationally invariant, methods of characterising shape (similarity) have been proposed during the last decade, all of which encode molecular shape as a set of descriptors. These include shape signatures, Zernicke descriptors and alpha-shape descriptors, proposed by Zauhar *et al.*,[319] Mak *et al.*[320] and Wilson *et al.*[321] respectively and the Ultrafast Shape Recognition (USR) descriptors proposed by Ballester and Richards.[263,264]

This latter approach supposes that molecular shape is uniquely determined by the distances between all the atoms in a molecule. In principle, this is a simplification, since different atoms have different (effective) radii and the set of inter-atomic distances cannot encode chirality, i.e. USR descriptors distinguish between diastereomers but not enantiomers. However, the first of these simplifications, which is also made by ROCS, may be of limited consequence in practice.[316] Furthermore, USR can be adapted (or supplemented with an additional descriptor) to capture chirality - as shown by Armstrong *et al.*[322] (Zhou *et al.*).[323] As its name suggests, USR has been indicated to be particularly computationally efficient for shape similarity calculations, yet there is a debate in the literature regarding its effectiveness at capturing molecular shape compared to some other methods.[263,264,310,324,325] Nonetheless, prospective virtual screening, via similarity searching using USR, recently identified an

[*] Subsequent to alignment, however, 3D descriptors might be computed based upon the values of molecular properties sampled at grid-points defined with respect to a reference molecule to which all other molecules are aligned. Indeed, whilst not aligning conformers according to maximal shape similarity, this approach is employed to derive descriptors by the widely used Comparative Molecular Field Analysis (CoMFA) 3D QSAR method. CoMFA descriptors correspond to the calculated potential energies, at all grid-points, for a set of different probe atoms/groups, designed to characterise the favourability of forming different kinds of intermolecular interactions at different locations around the molecule of interest.[318]

outstandingly high proportion of actives when compared to the results of experimentally screening against the same target[326] (see also Chapter 3 of this thesis).

Whilst molecular shape undoubtedly plays an important role in receptor-specific biological activity, some limitations on the use of molecular shape alone to discriminate between active and inactive compounds have been noted. Firstly, more flexible proteins and proteins with multiple binding sites may be able to bind compounds with different shapes.[310] Nonetheless, given the inexhaustive number of shapes which a binding site might adopt and the limited number of possible binding sites, active compounds would still, arguably, belong to certain classes of shapes. Secondly, perfect shape complementarity for a more flexible ligand may be entropically disfavoured due to the loss of rotational degrees of freedom upon binding.[311,327] Thirdly, chemistry contributes to binding free energy[327] - i.e. bioactivity is not dependent upon molecular shape alone.[310,328,329]

The importance of encoding chemical information was emphasised in recent studies by Kirchmair *et al.*[329] and Cannon *et al.*[175] These latter authors combined a variant of USR with an implementation of the MACCS key descriptor set, which encodes the presence or absence of pre-defined 2D substructural features using a bit-string, and generated classification models using Random Forest.[175] Their results suggest that shape (as encoded by USR) and chemical feature information may be combined synergistically - for some bioactivity prediction tasks.[175]

Whilst 3D descriptors are expected to encode additional, spatial, information that 2D descriptors cannot, 3D descriptors may not necessarily yield better models than 2D descriptors.[19,324,330,331] This could be because of an inability to approximate the bioactive conformer - leading to 3D descriptors failing to encode additional, relevant, information compared to 2D descriptors. It has also been speculated that much of the spatial information regarding molecular structures is implicitly encoded by their 2D structure.[19,330]

Prior to considering the general validity of the claim that the bioactive conformer is likely to be poorly approximated, it is worth briefly touching upon the computational approaches used to obtain such an approximation. An initial set of 3D co-ordinates must first be assigned to all atoms. A variety of programs are available for this purpose,[332] such as the widely employed CORINA. As summarised in the manual, CORINA generates a single, "low energy" conformation based upon a combination of 'typical' bond lengths, bond angles and torsion angles – informed by statistical analysis of the conformational preferences of acyclic

molecular fragments in small molecule crystal structures – and computationally inexpensive calculations.[333]

The initial 3D structure, possibly following refinement using a Molecular Mechanics force-field, which uses an empirically parameterised analytical expression for the relative energies of different 3D structures for a given molecule,[197,332] may then be subjected to a "conformational search" in order to locate other thermodynamically plausible conformations.[332]

A variety of approaches may be applied for conformational searching. Usually, truly systematic exploration of all possibilities is computationally inconceivable and many of the conformers found during a systematic search would be rejected on energetic grounds in any case. A more computationally efficient approach is to to consider random changes in torsion angles, possibly subject to constraints, combined with criteria for selecting energetically plausible intermediate solutions for further adjustments. Monte Carlo and genetic algorithms both employ, distinctive, variations on this kind of approach. Other algorithms, such as simulated annealing, might also be employed. Detailed discussion of conformational search algorithms lies beyond the scope of this thesis, but the interested reader is directed to Folloppe and Chen for an excellent introduction.[332] Conformational searching may be carried out *in vacuo* or in a (simplified) solvation model[332] or – as per the "docking" approach[334] – by assessing the energetic favourability of different conformations inside an experimentally determined,[334] or computationally predicted,[305] protein binding site.

Assessing how well such approaches can 'reproduce' the bioactive conformer depends upon a suitable definition of a 'closely reproduced' bioactive conformer. Foloppe and Chen suggest that a root mean squared deviation (RMSD) between corresponding atoms, upon maximal alignment, of 1 Å or below may be considered a "good fit".[332] They recently found that three popular software programs (MOE, Catalyst and ConfGen) were able to locate more than 60% of experimentally determined bioactive conformers in a test set comprising more than 200 "drug-like" compounds.[335] They further claimed that most bioactive conformers can be reproduced within an RMSD of 2 Å across methods.[332]

Nonetheless, even if a conformational search can locate a good approximation of the bioactive conformer, this does not necessarily mean that the program can recognise the bioactive conformer – i.e. determining the bioactive conformer may still remain a problem.

Indeed, Folloppe and Chen suggest that appropriate assessment of conformational energetic preferences remains a key challenge.[335]

However, nonwithstanding the possibility that 3D descriptors might genuinely yield worse predictions due to an inability to locate the bioactive conformer, it might be the case that observed superior performance with 2D descriptors is an artefact of "analogue bias" - i.e. datasets may be disproportionately composed of topologically similar active compounds.[330,336] Alternatively, 2D methods could correlate strongly with simple physicochemical properties such as logP,[337] which may correlate with nonspecific contributions to binding affinity, rather than the specific contributions that are required for, desirable, compound selectivity[338] and may be better captured by 3D approaches. Hence, 2D descriptors may outperform/perform as well as 3D descriptors for "biased" datasets in which, say, the set of actives was disproportionately made up of topologically similar molecules and/or actives and inactives were well separated in terms of simple (physicochemical) properties,[336,339] even when the 3D descriptors would be valuable under circumstances where such conditions did not apply.

In summary, the literature to date suggests that, in developing a novel 3D descriptor set, such as those presented in this chapter, it is important to:

1. Encode both shape and chemical features.
2. Investigate the effects of conformational sampling.
3. Benchmark the more sophisticated descriptors against simpler, 2D descriptors.
4. Assess their performance across different datasets.
5. Consider whether or not dataset biases/property distributions might underpin the observed relative performance of the assessed methods.

These points were considered when assessing the novel 3D descriptor set proposed in section 5.2.

5.2 Proposed Methodology

5.2.1 Descriptors

The novel descriptor set proposed in this work is an extension of USR to encode the 3D arrangement of chemical functionality relevant for binding within a protein. For brevity, this

is termed the Atom Type USR (ATUSR) descriptor set. An explanation of how ATUSR descriptors are calculated first requires a detailed explanation of USR.

USR descriptors are an approximate encoding of the distances between all atoms in a molecule, and hence of molecular shape (see section 5.1). A representation of these distances is generated by approximately encoding the distributions of the pairwise distances between all atoms in the molecule and a set of four reference locations, chosen to reduce redundancy between the pairwise distances from any pair of reference locations. These reference locations are: the molecular centroid (ctd), the closest atom to ctd (cst), the farthest atom from ctd (fct) and the farthest atom from fct. The associated distributions of pairwise distances are approximately encoded by computing estimators (i.e. sample moments) of the first moment about the origin (the mean distance to the reference location, encoding molecular size), plus the second and third moments about the mean.[263,264]

The lth moments about the origin (Equation 5.1) and the mean (Equation 5.2) are computed using the following expressions. In these equations, d_n denotes the Euclidean distance between the reference location and the nth atom out of all N atoms in the molecule. For each moment, the corresponding descriptor is its lth root, such that all 12 descriptors (four reference points, three moments per point) have dimensions of distance.[263,264]

$$m_l = \frac{1}{N}\sum_{n}^{N} {d_n}^{l} \qquad 5.1$$

$$\mu_l = \frac{1}{N}\sum_{n}^{N} (d_n - m_1)^l \qquad 5.2$$

The speed with which USR descriptors can be computed,[264,326] and the simplicity of their implementation, motivated the development of a novel 3D descriptor set based upon these descriptors. These ATUSR descriptors include the USR descriptors described above, plus 11 groups of additional descriptors corresponding to the USR descriptors computed using the same reference locations, yet only including distances to atoms of a particular type. Hence, ATUSR descriptors comprise 144 (12+11×12) values in total. The 11 atom types were designed to discriminate between chemical features which might make different contributions to the free energy of binding for the molecule in question to a protein binding site.[327] These

atom types were: (weak) hydrogen bond acceptor (donor), ring member, aromatic ring member, hydrophobic, fluorine, halogen, anionic and cationic.

All atom types were defined using SMARTS[340,341] patterns written by this author, and atoms corresponding to these SMARTS identified using the Python Pybel module.[177] In principle, atoms could be assigned more than one type - e.g. all 'aromatic ring atoms' are also 'ring atoms'. However, a rule that hydrogen bond donors (acceptors) could not be additionally typed as anionic (cationic) was also hardcoded. Steps were taken towards making atom typing independent of the tautomeric form or formal charge state of the input structure; to achieve the latter aim, steps were taken towards categorising all atoms expected to be protonated (deprotonated), or expected to be associated with delocalised positive (negative) charge in protonated (deprotonated) functional groups, at physiological pH, as cationic (anionic).[342–345]

The definition of hydrogen bond donors and acceptors was largely based on the definitions used by Mills and Dean (who explicitly excluded donors and acceptors, e.g. C-H groups, deemed to be 'weak') - i.e. donors included any nitrogen/oxygen with an attached hydrogen, whilst acceptors included sp^2 sulfurs and all oxygen, and some nitrogen atoms, possessing a lone pair, save where this was expected to be 'extensively' delocalised into an aromatic/electron withdrawing pi-system.[346,347] All nitrogen atoms presenting lone pairs upon expected deprotonation at physiological pH, as found in tetrazole[344,348] and acyl sulfonamides,[345,349] were typed as hydrogen bond acceptors. Likewise, all nitrogen atoms expected to be protonated at physiological pH were categorised as donors.[342]

Weak donors were defined as CH groups with polarizing nitrogen, oxygen or fluorine substituents (or aliphatic CH groups with multiple chlorine substituents) and sp CH groups and aliphatic carbon bound SH groups.[327,350,351] Weak acceptors were defined as all atoms in an aromatic ring, a double or triple bond (pi-acceptors) or sp^3 sulfurs.[350,351]

In all cases, the heavy atom (X) of an X-H donor was assigned the donor atom type.

The rationale for encoding the positions of atoms in ring systems in general, and aromatic rings in particular, was to encode both the degree of molecular flexibility (ring systems having lower conformational entropy than chains,[342] making binding to a receptor more entropically favourable), and the potential for enthalpically favourable offset pi-pi stacking etc.[327] respectively.

Except for the exclusion of carbons indicated as making negative contributions to Wildman and Crippen's logP model,[352] and the inclusion of fluorine,[351–353] hydrophobic atom types were assigned to all non-polar carbon atoms, bromine, iodine and some sulfur atoms, as accepted in the RDKit (BaseFeatures.fdef, Q22010_1).[354]

The assignment of a distinct atom type to carbon bound fluorines reflects the distinctive effects of fluorine substituents. As the most electronegative element, it may have a particularly profound effect on pK_a values,[353] hence on the potential for strong hydrogen bonding[350] or electrostatic interactions more generally. Moreover, CF interactions with polar groups "frequently occur in the PDB [Protein Data Bank]"[327] and can lead to five-fold increases in binding affinities.[353] Fluorine substituents occupy smaller volumes than other polar groups, increase lipophilicity on average,[353] and do not offer any potential for 'halogen bonding'.[355,356]

All iodine and bromine atoms and chlorine atoms attached to certain substituents which were indicated to increase the chance of their involvement in halogen bonding, were designated as 'halogens'.[355,356]

Figure 5.1 illustrates the computation of USR and ATUSR descriptors for an example molecule.

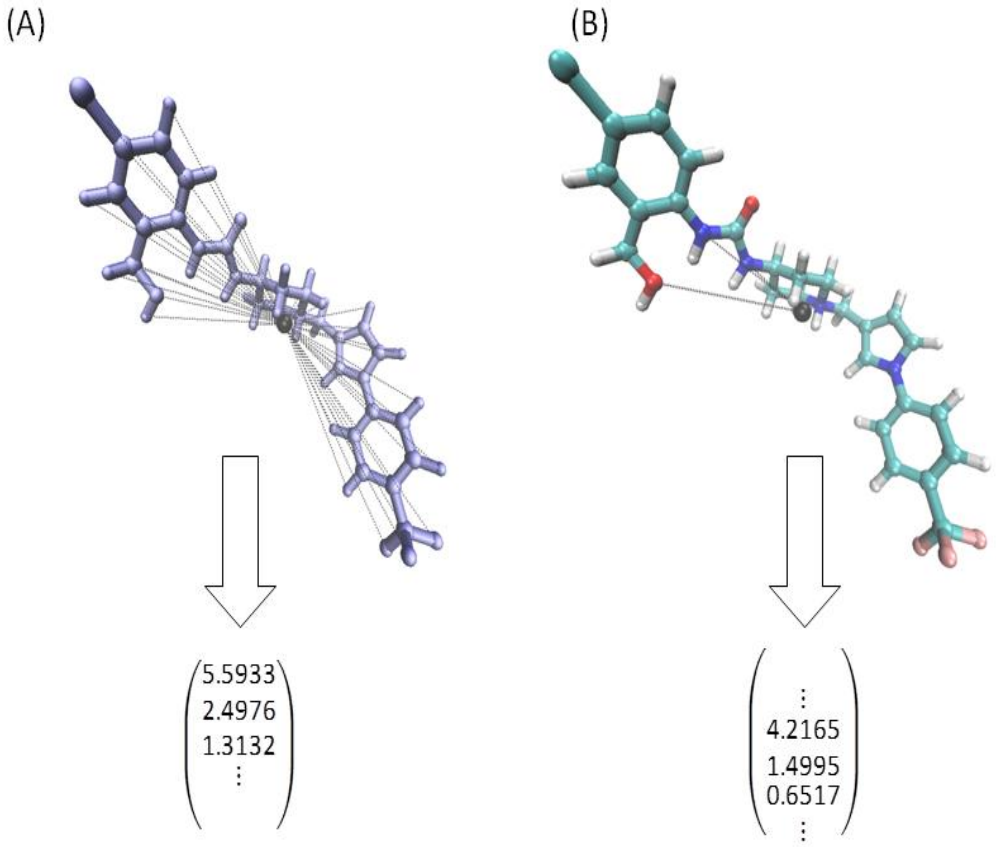


Figure 5.1 Computation of all three (A) USR and (B) supplementary ATUSR descriptors (for atoms typed as H bond donors) corresponding to the distances (dotted lines) between the molecular centroid (black sphere) and the relevant atoms. N.B.: The [NH+] donor (RHS of the molecular centroid; N = blue; H = white) is also typed as a cationic and ring atom. This figure was generated using VMD (version 1.9).[357,358]

5.2.2 Application of Descriptors to Bioactivity Prediction

The performance of the ATUSR descriptor set, compared to all other descriptor sets noted in section 5.3.1, when used to generate various regression and classification QSAR models using Random Forest (as implemented in the R[225] package randomForest) was assessed. Since the focus was on comparing different sets of descriptors, and randomForest is noted to perform well with its default parameters,[192] no hyperparameter selection was carried out. For all runs, randomForest( ) was called using n_{tree} = 501 (i.e. its default 500, adjusted to avoid random tie breaks for classification) and the default setting for m_{try} (see Chapter 2, section 2.6.3.2).

5.3 Evaluation of Methodology

The code for all descriptors implemented in this work, along with scripts used to carry out the modelling runs and analysis described in this section, is made available with this thesis (Appendix A).

5.3.1 Descriptor Sets Compared

The descriptor sets which were compared are presented in Table 5.1. Extensions of USR and USR-like descriptors based on descriptors computed using moments beyond the third moment about the mean were not considered. Whilst the inclusion of higher order moments might yield improved performance, albeit at greater computational cost,[175,359] the focus here was on comparing different ways of incorporating chemical information into USR-like descriptor sets and on comparing their performance relative to standard descriptor sets.

Comparisons were sought with MACCS[175,331,337,360] and P_VSA[69,91,214,296] 2D descriptors since both have been shown to be generally applicable to discriminating between bioactivity classes/bioactivity regression modelling. Indeed, the former set of descriptors was shown to outperform sets of USR-like descriptors for some classification tasks.[175] As noted previously, there is a clear need to benchmark the performance of 3D descriptor sets against sets of 2D descriptors, given both the lower computational overhead expected of the latter -which have no requirement to generate 3D conformers, a procedure which may also fail in some cases - and earlier studies which suggested that 3D representations may yield inferior predictive performance (see section 5.1).

This latter point was the rationale for also benchmarking the performance of ATUSR descriptors against USR-ATFP - a combination of USR and an 'Atom Type Fingerprint ' (ATFP) comprising the occurrence count for each of the 11 atom types used for the ATUSR descriptors (section 5.2.1). This corresponds to comparing two conceptually different approaches to adding chemical information to USR:

1. The distributions of potential macromolecular interaction sites, relative to the same reference points as USR, are explicitly encoded (as per ATUSR).
2. The presence/number of such sites is encoded in addition to encoding the 3D distribution of all atoms via USR (as per USR-ATFP).

The second of these approaches is analogous to the MACCS+USR-variant "hybrid descriptor" proposed by Cannon et al.[175] Likewise, just as Cannon et al. compared USR-

variants and their "hybrid descriptor" to MACCS alone, the ATFP descriptor set was also evaluated in isolation - giving an indication* of whether the encoding of molecular shape is actually adding predictivity.

Descriptor Set
ATUSR
USR-ATFP
ATFP
USR
USR+MACCS
MACCS
P_VSA

Table 5.1 Descriptor sets compared in this work.

5.3.2 Validation Protocol

For most datasets used in this work (see section 5.3.3), Random Forest models built using the aforementioned sets of descriptors (see section 5.2.2) were evaluated using 10 repetitions of 10-fold cross-validation. The partitions were designed to correspond to stratified folds, for classification tasks, or to independent sampling of percentile-based subgroups for continuous bioactivities - likewise intended to ensure that each fold was representative of the dataset as a whole.

However, for the larger Doddareddy dataset, a single train/test partition was considered for computational expediency. In all cases, model generation was repeated five times per train/test partition with a different seed for the randomForest (see section 5.2.2) random number generator (RNG).

5.3.3 Datasets

For most of the datasets considered to date, blockade of the hERG potassium ion channel was the modelled bioactivity. This was for the following reasons:

1. The considerable toxicological significance of hERG inhibition (see Chapter 1, section 1.3.2).

* To some extent, molecular size - which is encoded by USR - may correlate with these atom occurrence counts.

2. Earlier work having been undertaken by this author in both reviewing the literature for this toxicity endpoint, and in acquiring data from minimally experimentally inconsistent protocols (see Chapter 4).
3. Earlier work in the literature having highlighted the importance of molecular shape for potent hERG blockade.

Regarding this latter point, work by Chekmarev *et al.* had suggested that, on the basis of descriptors encoding molecular shape alone, strong and weak hERG inhibitors could be discriminated.[282] Moreover, docking of hERG inhibitors into a molecular dynamics informed model by Zachariae *et al.* suggested that, whilst hERG binding sites may be flexible, and compounds may bind in more than one site/to more than one channel state, specific conformational, size and peripheral structural requirements exist for potent hERG blockade.[300] This latter work suggested that, at least for discriminating potent from weaker blockade, a causative relationship between molecular shape and hERG inhibitory potency exists.

The Schattel benchmark dataset was of interest due to all biological measurements having being obtained using the same assay, within the same laboratory, and its having been previously modelled using 3D descriptors and Random Forest.[166]

hERG-196 Dataset

This comprised 196 compounds for which pIC_{50} estimates noted to have been derived using the IonWorks™ assay in CHO cells, using a similar protocol to Bridgland-Taylor *et al.*,[78] or otherwise inferred to have been obtained under similar conditions, or from manual, whole cell patch clamp measurements in CHO cells at room temperature, were obtained from the literature.[*] Where multiple pIC_{50} estimates were obtained for the same compound, the arithmetic mean was used for modelling.

The derivation of this dataset started by taking a subset of the Literature-368 dataset presented in Chapter 4, for which precise pIC_{50} values – i.e. excluding measurements merely indicating $pIC_{50} < 5$ or $pIC_{50} > 6$ – meeting the criteria noted above were available. Additional compounds, with hERG pIC_{50} values meeting these criteria, were obtained via a Scopus search of the recent literature in the summer of 2010. Where the compounds in the

[*] Bridgland-Taylor *et al.* indicated high correlation, with minimal systematic bias, between pIC_{50} estimates obtained under both sets of conditions.[78]

hERG-196 dataset corresponded to the original subset (see above) of the Literature-368 dataset, the same unique IDs were assigned.*

All structures, including stereochemistry, were manually checked using the experimental literature (where the compound was recently reported in a medicinal chemistry SAR study) or SciFinder Scholar 2007.[287] Care was taken to only specify stereo-centres where these were specified in the consulted sources, and that these were drawn as per the recommendations in the CORINA manual (see section 5.3.4.1) and that compounds for which all stereocentres were specified, yet only relative stereochemistry indicated, in these sources were noted (see section 5.3.4.2).

The 'raw' SDF and a CSV file containing all structures, prior to processing described below, and experimental pIC_{50} estimates respectively, along with references, are presented in the files made available with this thesis (Appendix A).

ThaiReg Dataset

This was a subset of the 248 compounds previously used to validate hERG regression models by Thai and Ecker.[296] Starting from the SDF containing the structures used in this study (kindly provided by Dr Khac-Minh Thai), a Python script was used to prepare a 'raw' SDF comprising 248 structures for which precise IC_{50} values were presented by Thai and Ecker.[296] Subsequent to removal of compounds which CORINA failed to parse, and additional compounds based on further analysis (see section 5.3.4.3), a dataset of 225 compounds with pIC_{50} values was obtained for regression modelling.

Since the number of compounds in common between, or with common stereoisomers in, the hERG-196 and ThaiReg datasets was estimated to be 73 or below,† it was suspected that these datasets might occupy different regions of chemical space. This, coupled with the fact

---

* These being the original IDs assigned to the Int-Set subset of the Literature-368 dataset, with the original ExtTest-Set subset IDs being advanced by 300 – as explained in Literature_368_References.doc (see Appendix A).

† Stereochemistry indifferent InChIs were computed, as per section 5.3.4.3 – save for the use of the SNon flag, for all structures in the 248 compound superset of the ThaiReg dataset and the hERG-196 dataset; 73 unique InChIs were found to be common to both sets of compounds. There were 246 and 196 unique InChIs associated with the superset of the ThaiReg dataset and hERG-196 dataset respectively.

that regression models were successfully built on the original superset of this dataset,[296] motivated the assessment of the descriptor sets of interest on this dataset as well.

Doddareddy Dataset

This was a subset of the hERG inhibition classification dataset (2,644 compounds) recently presented by Doddareddy *et al.*[93] Actives were compounds with IC_{50} < 10 μM and inactives were compounds with IC_{50} > 30 μM and FDA approved drugs (assumed inactives). Prior to the removal of structures subsequent to standardization (see section 5.3.4.3), inorganic compounds were removed (some of these having forced the generation of 3D structures using CORINA to abort) using Pipeline Pilot Student Edition (Figure 5.2),[301] giving 'raw' structures prior to subsequent processing. Ultimately, a dataset of 2,218 compounds was derived - with the training (1,187 inactives : 826 actives) and test (125 inactives : 80 actives) sets corresponding (post-removal of compounds) to the randomly selected split used by Doddareddy *et al.*

The rationale for modelling this dataset was that, due to its significant size compared to previously compiled hERG datasets, it was suspected to be significantly more diverse and offer a larger coverage of chemical space - characteristics which might affect the relative performance of the evaluated sets of descriptors.

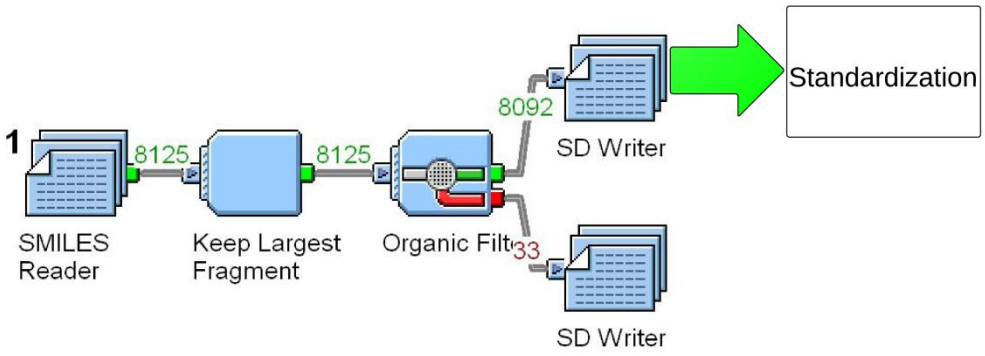


Figure 5.2 Removal of inorganic structures from Doddareddy dataset in Pipeline Pilot, prior to standardization. The numbers reflect the total number of compounds, in all SMILES files, subsequent to assignment of unique compound IDs, presented by Dr Andreas Bender. The 2,644 compounds referred to were presented in a subset of these files.

hERG-196:Subset and ThaiReg:Subset Datasets

In order to gain some insight into the effect that the large size of the Doddaredy dataset might have on the descriptor sets' relative performances, analogous active (IC_{50} < 10 μM) vs.

inactive (IC_{50} > 30 μM) classification datasets were derived from the hERG-196 and ThaiReg datasets, via excluding all compounds with (mean) pIC_{50} values not fulfilling the corresponding criteria. The hERG-196:Subset and ThaiReg:Subset datasets comprised 148 (122 active : 26 inactive) and 185 (155 active : 30 inactive) compounds respectively. Whilst the different active : inactive ratios, and assay types used to derive hERG activity, for the three hERG classification datasets would arguably impact upon the results, it was expected that the different sizes (and potentially different levels of diversity) would exert a greater effect on the *relative* performance obtained with different sets of descriptors.

Schattel Dataset

This was a dataset of 249 compounds provided by Schattel *et al.* in SDF format. This comprised compounds, with absolute stereochemistry characterised, determined as active (60 compounds) or inactive (189 compounds) with respect to inhibition of c-Jun N-terminal kinase-3 (JNK3).[166]

5.3.4 Dataset Processing and Descriptor Calculations

All scripts, and standardizer configuration files, specifying the exact options used with ChemAxon's tools,[298] CORINA (version 3.20) [333] and the Molecular Operating Environment (MOE),[285] screenshots completely detailing graphically set options for MacroModel,[361] as well as all files required to reproduce the GOLD[334,362] docking runs, and obtain the selected docking poses, are made available with this thesis (Appendix A).

5.3.4.1 Standard Processing of Structures

All 'raw' structures, presented in SDF format, in the (pre-cursors of) all datasets (see section 5.3.3) were processed via the workflow schematically illustrated in Figure 5.3 These operations were designed to yield one representative 3D structure, under typical assay conditions, per compound.

5.3.4.2 Structural Refinement of the hERG-196 Dataset

Initial results for the hERG-196 dataset, and results (see section 5.4) based on the standard workflow (Figure 5.3), indicated typically marginal differences between the performance of many of the 3D and 2D descriptor sets considered in this chapter (see Table 5.1). In order to investigate whether this reflected poor approximation of the bioactive conformer (see section

5.1), alternative workflows were followed (for the hERG-196 dataset) in order to obtain a conformer for each compound that might better approximate this.

For all workflows, the same processing of the 'raw' structures via ChemAxon's tools, and the conversion of all finally obtained SDF files to MOL2 format prior to MOE descriptor calculations, was carried out as per the standard workflow (Figure 5.3).

5.3.4.2.1 Preparatory Steps

For all workflows, initial 3D structures were generated, possibly more than one per compound, from the structures obtained subsequent to processing using ChemAxon's tools (Figure 5.3), using CORINA's stereoisomer generator ("STERGEN"); all stereoisomers generated were designed to correspond to different possibilities given the (possibly incomplete) stereochemistry specified for the original structures. Since the STERGEN protocol would not have generated multiple enantiomers for compounds with all stereocentres specified, yet for which absolute stereochemistry was not indicated (see section 5.3.3), enantiomers of the structures generated by CORINA for these compounds were obtained, by reflecting conformers in the x-y plane, using a Python[275] script employing the Pybel module.[177]

Taking into account those compounds for which multiple possible stereoisomers were generated, a total of 349 3D structures was obtained, for the 196 compounds in the hERG-196 dataset, subsequent to parsing the STERGEN generated structures via this final Python script. These 349 structures were processed using Molecular Mechanics (section 5.3.4.2.2) and docking (section 5.3.4.2.3) procedures.

For the structures obtained immediately prior to and subsequent to the application of Molecular Mechanics, an arbitrary stereoisomer was *consistently* selected per compound, for obtaining descriptors, where multiple possible stereoisomers were available for a given compound. However, since the highest scoring docking solution was selected per compound, a non-arbitrary possible stereoisomer was used for descriptor calculations subsequent to docking - this being the rationale for considering all possible stereoisomers for each compound.

5.3.4.2.2 Molecular Mechanics

Further processing of these 349 structures was carried out via MacroModel (version 9.7).[361] Two sets of calculations were set up:

1. Local minimisations.
2. Conformational searches, designed to find the global minimum for each structure.

In both cases, Maestro (version 9.1.207), [361] was used to set up the calculations, starting from the Pybel generated SDF, and extract MacroModel generated structures in SDF format.

Both sets of calculations employed the MMFF94s force-field,[363] and a "GB/SA" continuum model of aqueous solvation.[197,364] The "optimal" method was selected for all minimisations,* along with 10,000 maximum iterations and the default convergence criterion of 0.05 kJ mol^{-1} $Å^{-1}$.[361] Conformational searches employed the recommended[361] Monte Carlo Multiple Minimum (MCMM) approach,[366] with "intermediate" sampling of esters/amides, retention of mirror image conformations, 100 steps per rotatable bond, and a limit of 1,000 steps per molecule. Redundant conformers, identified based on a maximum atom deviation of 0.5 Å, were eliminated, along with any conformers higher than 21.0 kJ mol^{-1} above the current global minimum.

5.3.4.2.3 Docking

The Molecular Mechanics calculations were premised on the assumption that the global minimum of the free ligand in aqueous solution was a good approximation of the bioactive conformer (and, for the local minimisations, that this was the closest local minimum to the CORINA conformation). Indeed, a recent study suggested that a similar conformational search procedure to that described above could yield a global minimum which was a good approximation of the bioactive conformer.[367] However, the bioactive conformer could be substantially different to this global minimum.[19,332,368] Hence, as has elsewhere been proposed in the literature,[90,369] docking was employed to (hopefully) obtain a better approximation of the bioactive conformer.

All 349 locally minimised structures obtained for the hERG-196 dataset were docked into the hERG homology model presented by Imai *et al.*,[305] using GOLD (version 5.0.1),[334,362]

* In practice, this was a Truncated Newton Conjugate Gradient method.[365]

subsequent to preparing them in MOL2 format, with all atom and bond types assigned according to the GOLD conventions for functional groups, using CORINA. Docking runs started from the PDB format file presented by Imai *et al.*,[305] containing an experimentally supported cisapride-hERG complex, obtained from molecular docking.

This cisapride structure was extracted within GOLD and the binding site defined by all protein atoms (and associated residues) within 12 Å of this ligand. This was deemed appropriate since the binding site included the Ser624, Tyr652 and Phe656 residues (on all four subunits) highlighted by Imai *et al.*[305] as playing key roles in binding potent hERG inhibitors.

For each input structure, 40 docking solutions were obtained using ChemScore.[370,371] This strategy was based upon[*] the protocol employed by-Imai *et al.* to obtain experimentally supported docking solutions for hERG inhibitors.[305] The generation of diverse solutions, without early termination, was specified. Automatic genetic algorithm settings were employed using the default search efficiency. In keeping with the aforementioned conformational searches, rotations about amide bonds were not considered.

In addition to selecting the highest ranking solution *per compound* using ChemScore, the highest ranking solution upon re-scoring using RF-Score was also selected; in both cases, structures were extracted from the docking solution files written by GOLD using Python scripts[275] employing Pybel.[177] RF-Score was trained and validated on the same partition of the PDBbind dataset as Ballester and Mitchell,[266†] prior to applying to all docking solutions. Copies of these solutions were prepared as required for RF-Score using a Python script; the protein structure file required was obtained via reloading the protein data file specified in the GOLD configuration file and exporting in PDB format via Hermes version 1.4.1.[334]

Since the docking solutions written by GOLD were stripped of formal charges, the co-ordinates of the top ranking solutions (selected via ChemScore or RF-Score) were

---

[*] Unlike Imai *et al.*, for computational expediency, additional docking runs using GoldScore were not carried out.

[†] Save for some differences in software and hardware versions (see Appendix C), the instructions presented by these authors were followed. Minor changes were made to their code (see Appendix A); however, no differences were observed between the reported training and test set performances and those presented in Ballester and Mitchell's publication.[266]

transferred, via Python code, to the corresponding SDF structures obtained via local minimisation, prior to converting to MOL2 format as per the standard workflow (Figure 5.3).

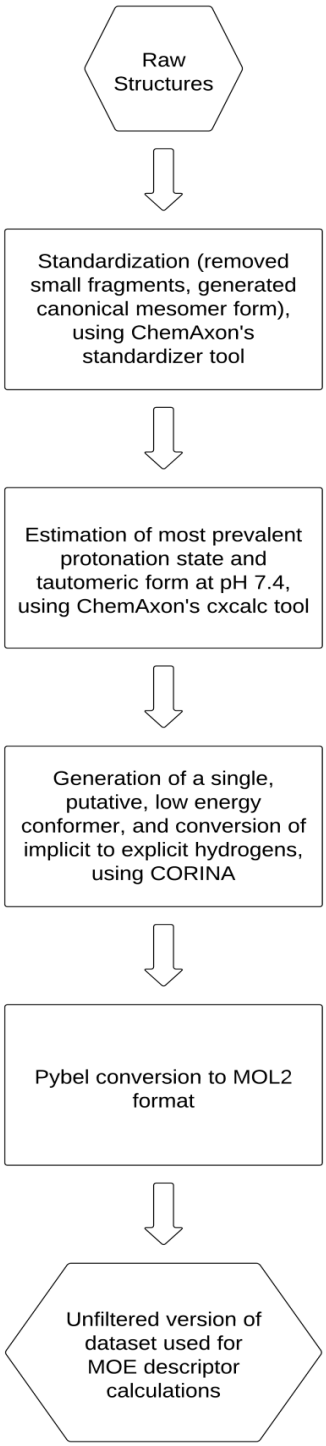


Figure 5.3 Standard workflow used to process 'raw' structures in all datasets modelled in the work presented in this chapter. N.B.: The structures obtained subsequent to each of the different pre-CORINA processing steps were also parsed via Pybel.[177]

5.3.4.3 Preparation of Files for Descriptor Calculations

All MOL2 files ultimately obtained from the workflows described above were used to compute MACCS keys and P_VSA descriptors (stored in SDF format), via MOE. Pybel[177] filtered versions of these MOE exported files and their MOL2 precursors were used to generate MACCS and P_VSA descriptors files and compute all other descriptor sets presented in Table 5.1 respectively.

An analysis of redundancy/potentially incorrect structures in these datasets was carried out via comparing stereochemistry encoding InChIs computed (from all SDF files obtained subsequent to processing via ChemAxon's tools, yet prior to CORINA processing) using the standalone Windows executable (InChI version 1.03) available from the International Union of Pure and Applied Chemistry (IUPAC).[178] The InChIs employed here for structure comparisons were generated using the default options. The structures were updated with these InChIs (as SDF fields) via a Python script employing Pybel.

All structures with non-unique InChIs,[*] and all those for which erroneous stereochemistry assignment was indicated by CORINA (based on warnings/error messages in the CORINA logs related to stereochemistry) were removed from the relevant datasets. In the course of this work, 15 erroneous structures were also removed from the ThaiReg dataset based upon consulting SciFinder[372] or the literature sources provided by Thai and Ecker.[296] The resulting MOL2 and SDF files used to compute descriptors are made available with this thesis (Appendix A).

5.3.5 Statistical Comparisons

All values for all figures of merit presented in section 5.4 were computed per randomForest RNG seed-test set combination. Hence, five results were obtained per descriptor set for the Doddareddy dataset and 500 for all other datasets evaluated via 10 times 10-fold cross-validation (10-10CV).

The variation observed for each set of descriptors motivated statistical evaluation of the results. The null hypotheses proposed were that, across all possible test set results that could

[*] Save for those active compounds in the Doddareddy dataset with InChIs only matching those computed from inactive compounds corresponding to the seven 'false non-blockers' noted by Doddareddy *et al.*[93] - these 'inactives' being separately specified for removal.

have been obtained for a dataset corresponding, in terms of its characteristics, to any given dataset which was actually evaluated:*

1. The overall mean difference in performance - as measured using a suitable figure of merit - for any pair of descriptor sets would have been zero.
2. For a given descriptor set, the overall mean performance would have been no better than a random predictor.

However, the variation in results obtained arises from two sources:

1. The selection of different training and test sets.
2. The internal randomness of Random Forest.

This author was unaware of any appropriate statistical test which would enable direct computation of p-values for either null hypothesis, given these distinct sources of variation.

Considering the first null hypothesis, for each set of results obtained for a given dataset using a single randomForest RNG seed, all 10-10CV results for a given pair of descriptors were compared using the "corrected repeated k-fold CV test" advocated by Bouckaert and Frank.[373] This test is designed to correct for the underestimation of p-values generated from the standard paired t-test[226] when applied to comparison of repeated cross-validation results - due to the lack of independent sampling of train/test partitions arising from the overlap of cross-validation training sets and, when repeated, test sets.[373] Supposing rejection of the null hypothesis for all RNG seeds, the overall mean difference was evaluated for statistical significance by applying a standard paired t-test to the cross-validation mean results, for each RNG seed, as it is reasonable to suppose these results were truly independent. For the Doddareddy dataset, only the latter test was applied; given the large size of the training and test set used for this dataset, it was assumed that the variation in results associated with different training and test set selections would be negligible. In all cases, p-values corresponding to a two-tail test[226] were computed. For all tests, the figures of merit considered for classification and regression problems were the MCC and R^2 respectively.

Considering the second null hypothesis, one-tail[226] versions of the statistical tests previously described were applied to the results obtained from 10-10CV.[373] Since, for a random

*Systematic differences were expected between the datasets, that could lead to genuine differences in the (relative) performances of the descriptors - hence, it was deemed appropriate to separately test these hypotheses for each dataset.

predictor, the mean MCC and Pearson's correlation coefficient values are expected to be zero,[226,230] these values were assessed for statistical significance for regression and classification problems respectively. For the Doddareddy dataset, the chi-squared p-values (Chapter 2, section 2.6.4.1) corresponding to the MCC were computed for each RNG seed. Supposing the (mean) results obtained for each RNG seed were all statistically significant, the overall mean MCC was assessed to be statistically significantly greater than zero based on a, standard, one-tail t-test.[226]

Ultimately, a considerable number of comparisons regarding the relative and absolute performance of the descriptor sets were made. Since a decision to reject a given null hypothesis was made upon the basis of observing a sufficiently low p-value (less than 0.05), the expected proportion of rejections which were erroneous (the "false discovery rate", or FDR)[374] could well exceed 0.05 if the original p-values were used.

It is appropriate to control the FDR for a set (or "family") of p-values, where any statistically significant finding amongst the family points towards an overall conclusion;[374] here the overall conclusions, depending upon the null hypothesis rejected, would be that there were genuine differences in the performance of the descriptor sets - or that a given set of descriptors yielded better than random performance - under some circumstances. Hence, all p-values obtained, across all datasets, from the tests applied for pairwise comparisons, were treated as a family, whilst only those p-values obtained for a given descriptor set were treated as a family for the purpose of rejecting the second null hypothesis.

In order to restrict the FDR to 0.05 or below, for a given family of comparisons, Benjamini and Yekutieli's method[375] was applied, using the R function p.adjust( ),[225] to generate adjusted p-values, with differences only declared statistically significant when associated with adjusted p-values less than 0.05.[303]

To summarise: multiple adjusted p-values were computed for any given comparison between pairs of descriptor sets or between the results obtained for a single descriptor set and the 'zero-baseline' expected for a random predictor. Differences in overall mean MCC, R^2 or Pearson's correlation coefficient values - for a given dataset - were only declared statistically significant if all relevant comparisons were deemed statistically significant (adjusted p-values < 0.05).

The results of all statistical comparisons, including a summary of all differences that were deemed statistically significant, are presented in CSV files made available with this thesis (Appendix A).

5.3.6 Dataset Evaluation

In order to facilitate interpretation of the relative performance of the assessed descriptor sets on the datasets presented (section 5.3.3), clustering analysis was employed as an *indication* of the diversity of the datasets (and, for the classification datasets, their separability into actives and inactives) in terms of molecular shape and lipophilicity - the latter commonly associated with nonspecific receptor binding.[338] Agglomerative hierarchical clustering[270] was employed: descriptor vector representations of all molecules in the dataset defined the first set of ("singleton")[331] clusters; the closest pair of these were replaced with a new cluster corresponding to the geometric mean of their descriptor vectors - this procedure being carried out iteratively.

USR (ChemAxon logD[*]) descriptors were employed to characterise the datasets in terms of molecular shape (lipophilicity). All descriptor calculations and clustering analyses were carried out using a combination of R[225] and Python[275] code made available with this thesis (Appendix A).

The number of clusters (excluding "singletons") formed prior to the smallest inter-cluster distance exceeding a specified cut-off was used as a measure of dataset diversity. The distance metric employed was one minus the similarity metric defined by Ballester and Richards for the comparison of USR descriptor vectors[263] - hence it took values between zero and one. Following Schreyer and Blundell, it was supposed that conformations with USR similarities below 0.8 could be considered 'significantly different';[368] the distance cut-off was set to 0.2 for both sets of descriptors for consistency.

The separability of classes was indicated via application of an analogous approach to that proposed by Brown and Martin:[331] for both the original dataset, and activity permuted

[*] In keeping with Shamovsky *et al.*, logD was considered a better measure of lipophilicity for ionisable compounds than logP.[338] The structures available immediately prior to processing via CORINA (see Figure 5.3) were used to compute logD predictions. It was not possible to compute logD estimates for nine of the 826 active entries in the Doddareddy training set; hence, these were excluded *when clustering based on logD*.

versions, the number of "active clusters" - when clustering was terminated upon formation of a specified minimum number of clusters - was defined as the number of ('non-singleton') clusters containing at least one active compound. The proportion (p(a)) of compounds assigned to these clusters (the "active cluster subset") which were active was computed, and compared to the proportion (p(0)) of active compounds in the entire dataset. If the p(a) values obtained, upon specification of increasingly larger numbers of clusters,[*] corresponding to increasingly fewer compounds in the "active cluster subset", for the real data exceed those obtained for the activity permuted data, this *indicates* that the descriptor set employed for clustering (partially) separates actives and inactives in the dataset.

5.4 Results and Discussion

All of the following plots were generated using R.[225] All box-and-whisker plots summarise the results across all RNG seeds and across all train/test partitions and were generated such that the whiskers extend to the data extremes, with the solid lines denoting the median and the box denoting the upper and lower quartiles.[376] The means are denoted by the centre of the circles superposed on the box and whisker plots, with error bars extending to plus and minus the standard error of the mean.[226]

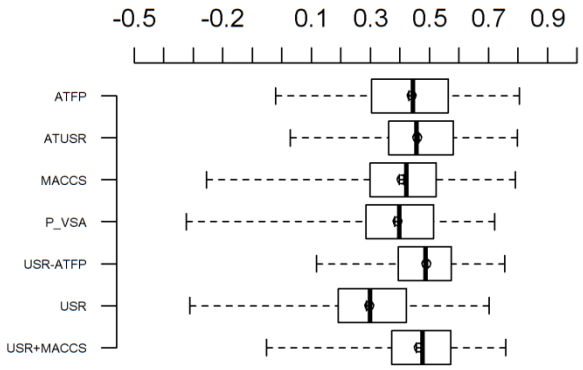
Unless otherwise noted, the coefficient of determination (R^2) and Matthews Correlation Coefficient (MCC) are used to summarise the overall performance of descriptor sets when used for regression and classification modelling respectively. More detailed results, referred to below, are made available with this thesis (Appendix A). (See Chapter 2, section 2.6.4 for definitions of these figures of merit.)

5.4.1 Effects of Structural Refinement

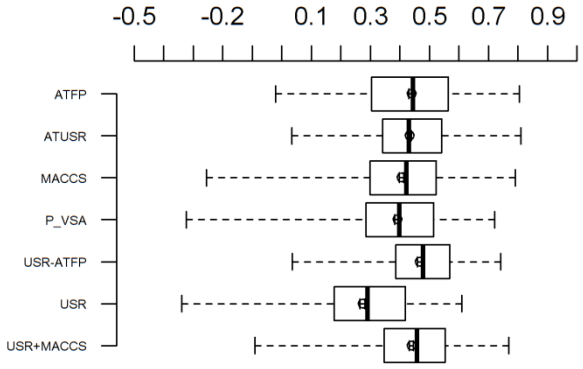
In the following plots, the overall performances of all descriptor sets for regression (Figure 5.4) and classification (Figure 5.5) modelling of the hERG-196 and hERG-196: Subset datasets respectively are summarised for all sets of 3D structures generated via the different workflows described in section 5.3.4.2.

[*] Here, the minimum number of clusters was set to one, then increased - insofar as was possible - from 30 in steps of 30, for all datasets.

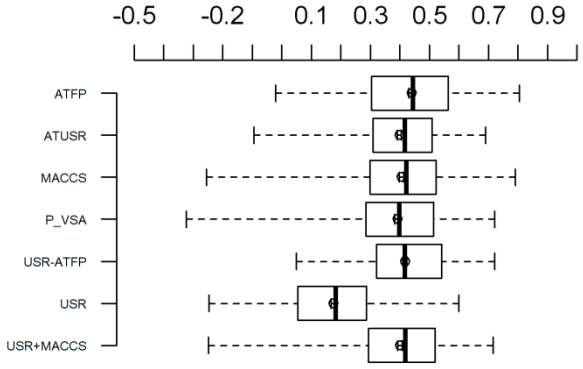
(A) R^2 (hERG-196, STERGEN Conformers)



(B) R^2 (hERG-196, Following Local Minimisation)



(C) R^2 (hERG-196, Following Global Minimisation)



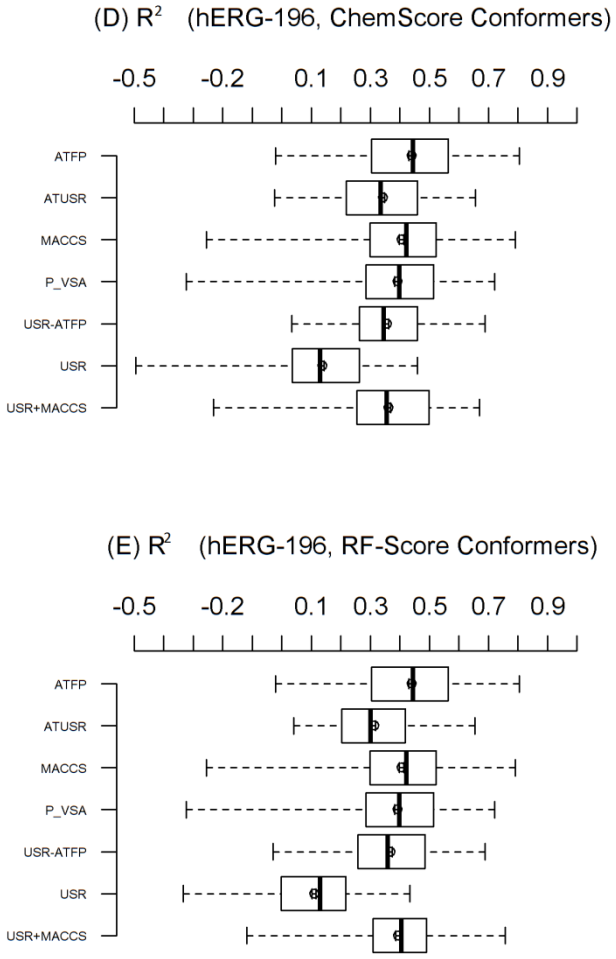
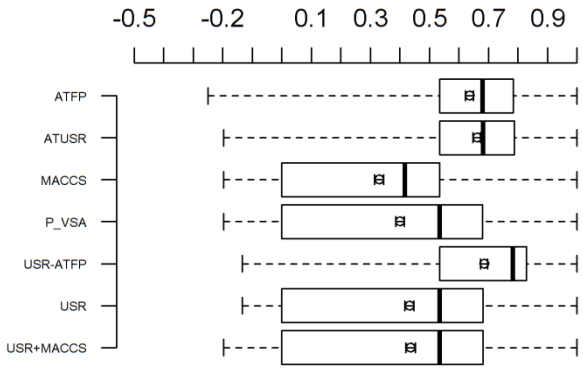
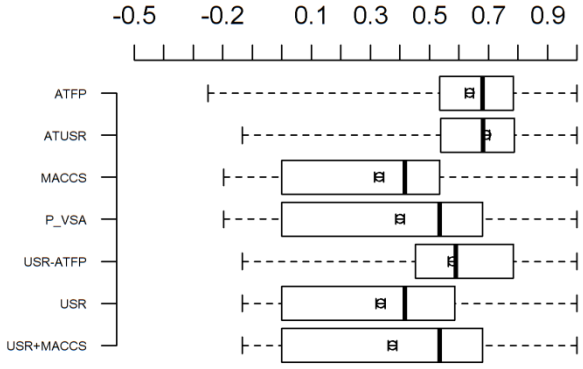


Figure 5.4 R^2 values obtained from 10-10CV (five RNG seeds for randomForest) on the hERG-196 dataset , with different descriptor sets calculated from structures obtained: (A) prior to Molecular Mechanics calculations, (B) from local minimisations, (C) from global minimisations, (D) from docking (ChemScore selections), (E) from docking (RF-Score selections). The black lines and circle centres denote the median and mean results respectively.

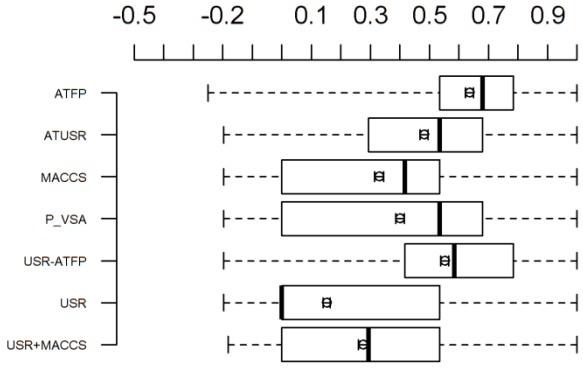
(A) MCC (hERG-196:Subset, STERGEN Conformers)



(B) MCC (hERG-196:Subset, Following Local Minimisation)



(C) MCC (hERG-196:Subset, Following Global Minimisation)



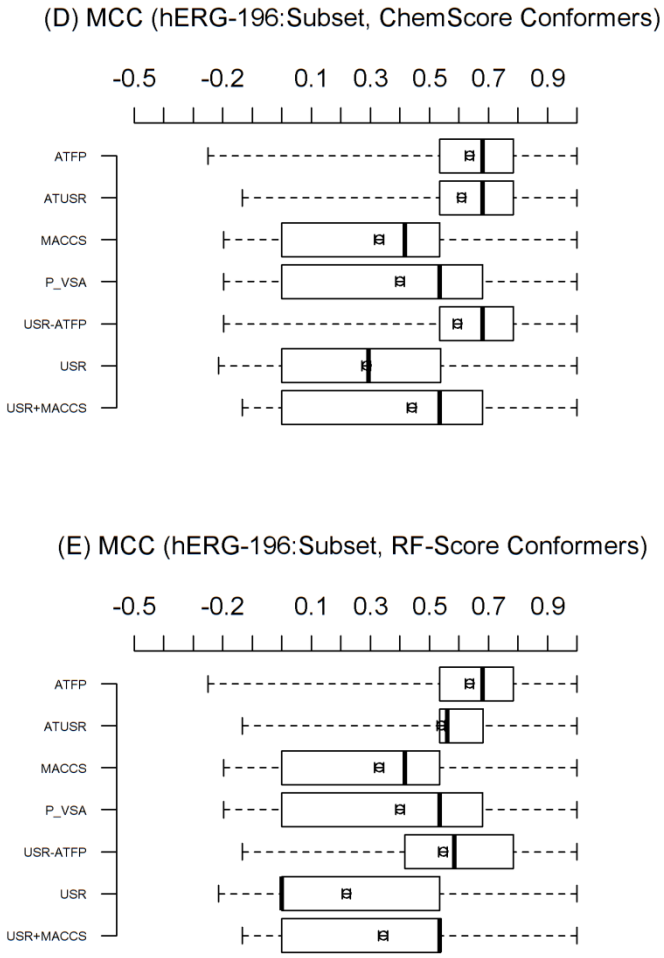


Figure 5.5 Corresponding MCC values (c.f. Figure 5.4) obtained on the hERG-196:Subset dataset.

Considering the hERG-196 regression results, both the results presented in Figure 5.4, and all corresponding plots for Pearson's (Spearman's) correlation coefficient, the Root Mean Square Error (RMSE) and the mean absolute error (MAE), indicate that the USR descriptor set is the worst. The USR descriptor set consistently yielded the lowest overall mean R^2, with all statistically significant pairwise differences in these values involving USR results.

This is not so surprising. Firstly, the importance of chemical information was highlighted in recent classification and similarity searching studies.[329,377] Secondly, it is reasonable to suppose that changes in molecular shape might be more useful in discriminating between clearly separated actives and inactives, rather than in controlling more subtle graduations in bioactivity that would need to be captured by a regression model.

In contrast, the classification results (Figure 5.5) obtained using USR are not consistently ranked as worst performing in terms of either their overall median (or mean) MCC - which may reflect the ability (as suggested above) of molecular shape to specifically discriminate actives from inactives.

Neither the computation of structures obtained from Molecular Mechanics nor docking yielded systematic improvements in the performance of the 3D descriptor sets.* Indeed, in most cases, the 'refinement' of the original structures obtained from STERGEN led to lower overall mean R^2 (MCC) values being obtained using the 3D descriptor sets (see Figure 5.6 and Figure 5.7). Furthermore, whilst the performance of all descriptor sets for both classification and regression was otherwise deemed statistically better than random, the mean MCC value for USR (and USR+MACCS) was not deemed statistically significantly better than random when descriptors were computed from ChemScore or RF-Score selected (and globally minimised) structures.

* Those descriptor sets (USR+MACCS and USR-ATFP) comprising both 3D and 2D components were considered 3D descriptor sets.

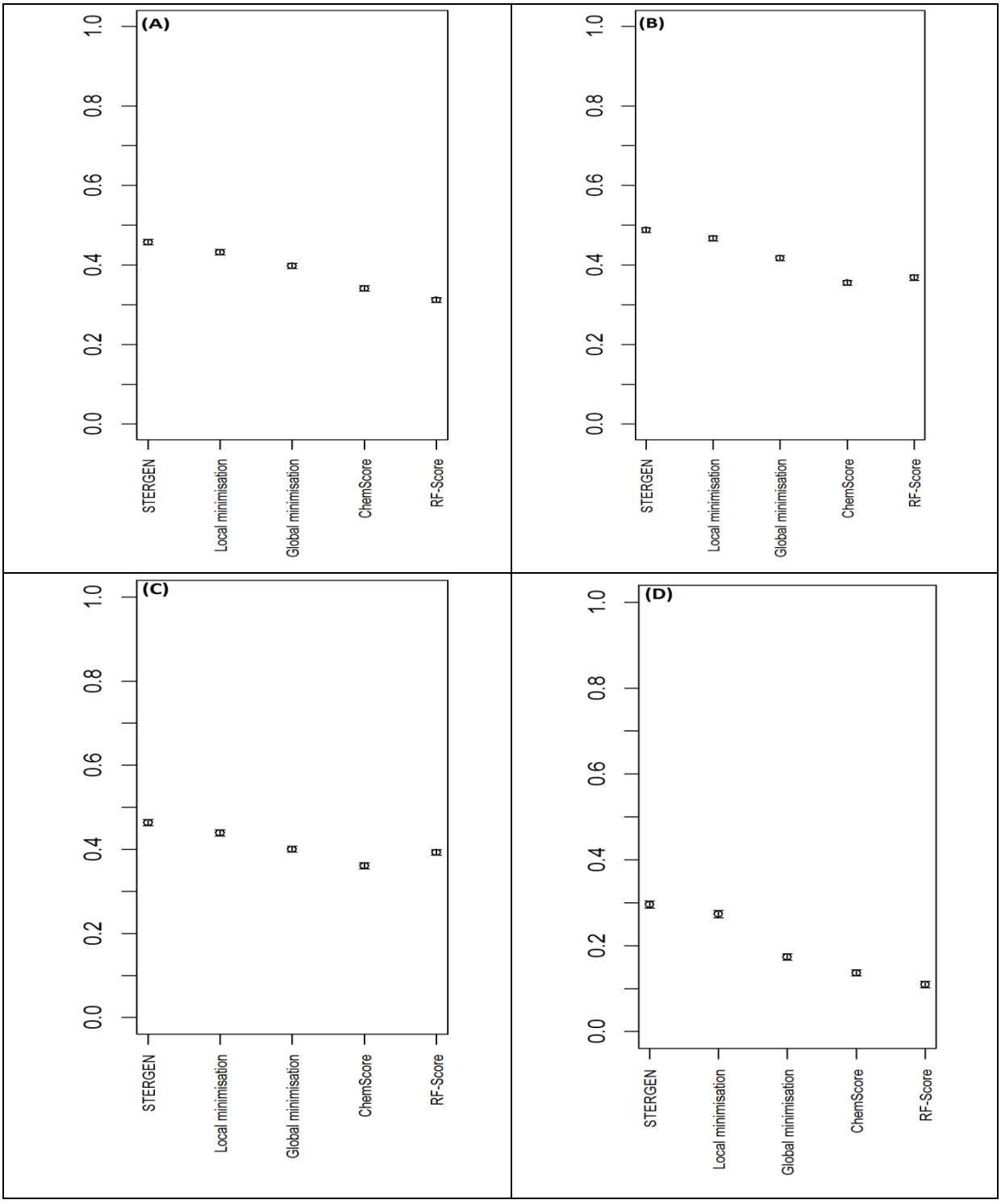


Figure 5.6 Mean R^2 values obtained, for the hERG-196 dataset, using the following descriptors, computed from different 3D structures: (A) ATUSR, (B) USR-ATFP, (C) USR+MACCS, (D) USR.

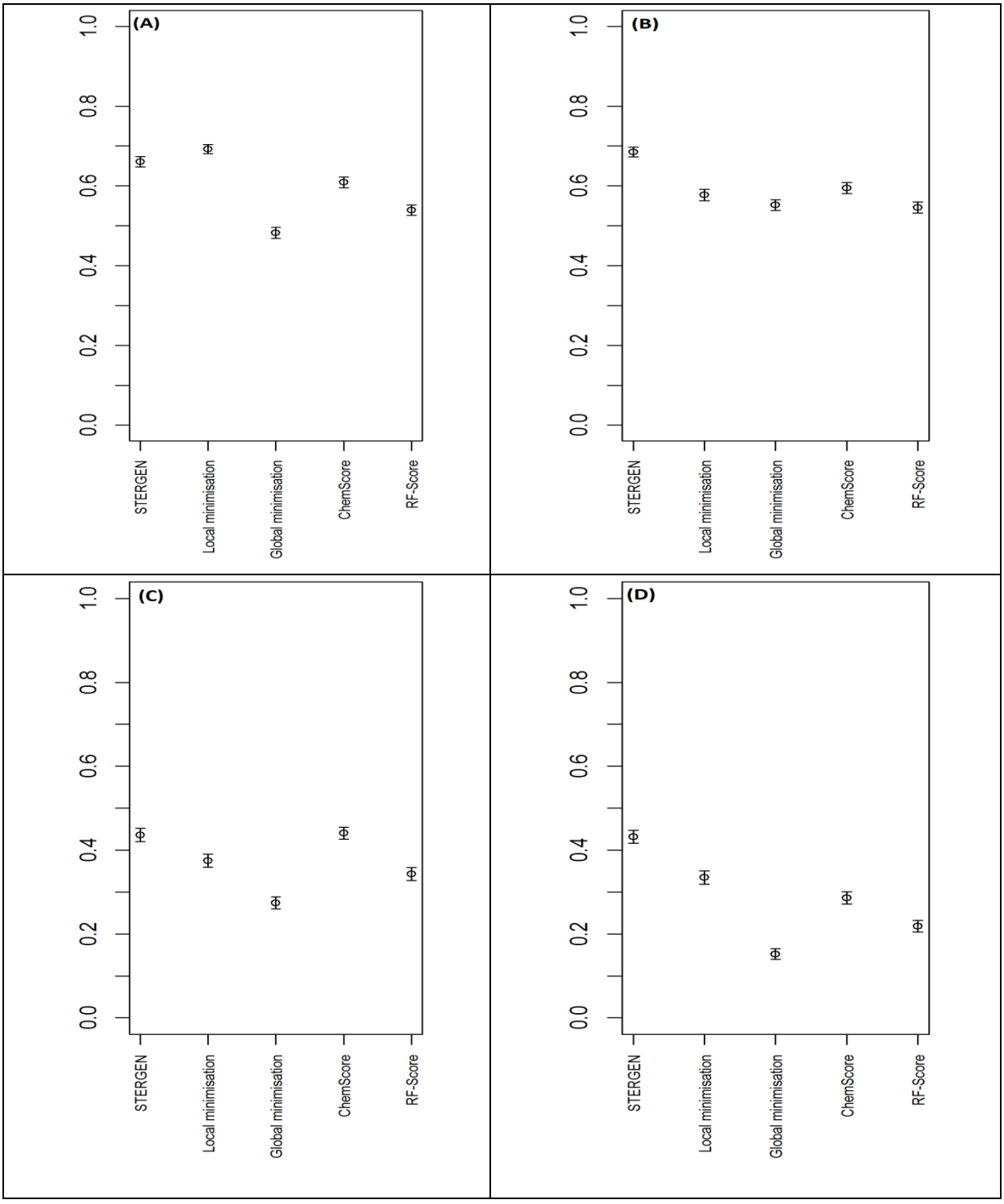


Figure 5.7 Mean MCC values obtained, for the hERG-196: Subset dataset, using the following descriptors, computed from different 3D structures: (A) ATUSR, (B) USR-ATFP, (C) USR+MACCS, (D) USR.

This could be because none of the 'refinement' procedures typically yielded an improved approximation of the bioactive conformer. To gain some partial insight into this, the structures of cisapride and E-4031 from which descriptors were computed for the results presented above were compared to the docking poses presented by Imai *et al.*[305] Since Imai *et al.* presented experimental results in support of these poses, it was supposed that their poses could be taken as good approximations of the actual bioactive conformers. The structures used for descriptor calculations were aligned to these docking poses, extracted from the PDB format files presented by Imai *et al.* using OpenBabel (version 2.3.0a), and the Root Mean Squared Deviation (RMSD) between the heavy atoms computed using a Python script employing OpenEye's OEChem Toolkit (version 1.7.7).[378] USR similarities were also computed. Visual (quantitative) comparisons of the aligned structures are presented in Figure 5.8 (Table 5.2).

It should be noted that the protonation states assigned by Imai *et al.* differed from those assigned for descriptor calculations in the present work and a different stereoisomer of cisapride was used for descriptor calculations (save for the RF-Score selected structure). As USR descriptors are computed from all atoms (i.e. both heavy and hydrogen atoms), the differences in protonation states may *partially* explain the observation that USR similarity did not always increase when the RMSD decreased.

The increase (decrease) in RMSD (USR similarity), relative to the original STERGEN structure, observed upon selecting a structure from a global conformational search or a top scoring docking pose for E-4031 lends credence to the suggestion that the refinement procedures did not typically yield an improved approximation of the bioactive conformer.

It is also worth noting that low Pearson's correlation coefficients (ChemScore: 0.364, RF-Score: 0.287), for the poses selected for descriptor calculation, were obtained between the average pIC_{50} values (used for QSARs) and the docking scores for the hERG-196 dataset. The much lower performance of RF-Score than that previously reported for the PDBbind dataset, comprising experimental protein-ligand complexes,[266,267] further suggests that the docking procedures were (typically) unable to generate improved approximations of the bioactive conformers. Indeed, there are a number of particular difficulties inherent in searching for the bioactive conformer via docking into a hERG model; some of these may be scoring function specific, as discussed by Imai *et al.* - who did not use docking scores to select their experimentally supported docking poses[305] - and further problems may arise due

to the possibility that hERG blockers may bind to multiple states of and/or multiple sites within the ion channel (see Chapter 4, section 4.6.3).

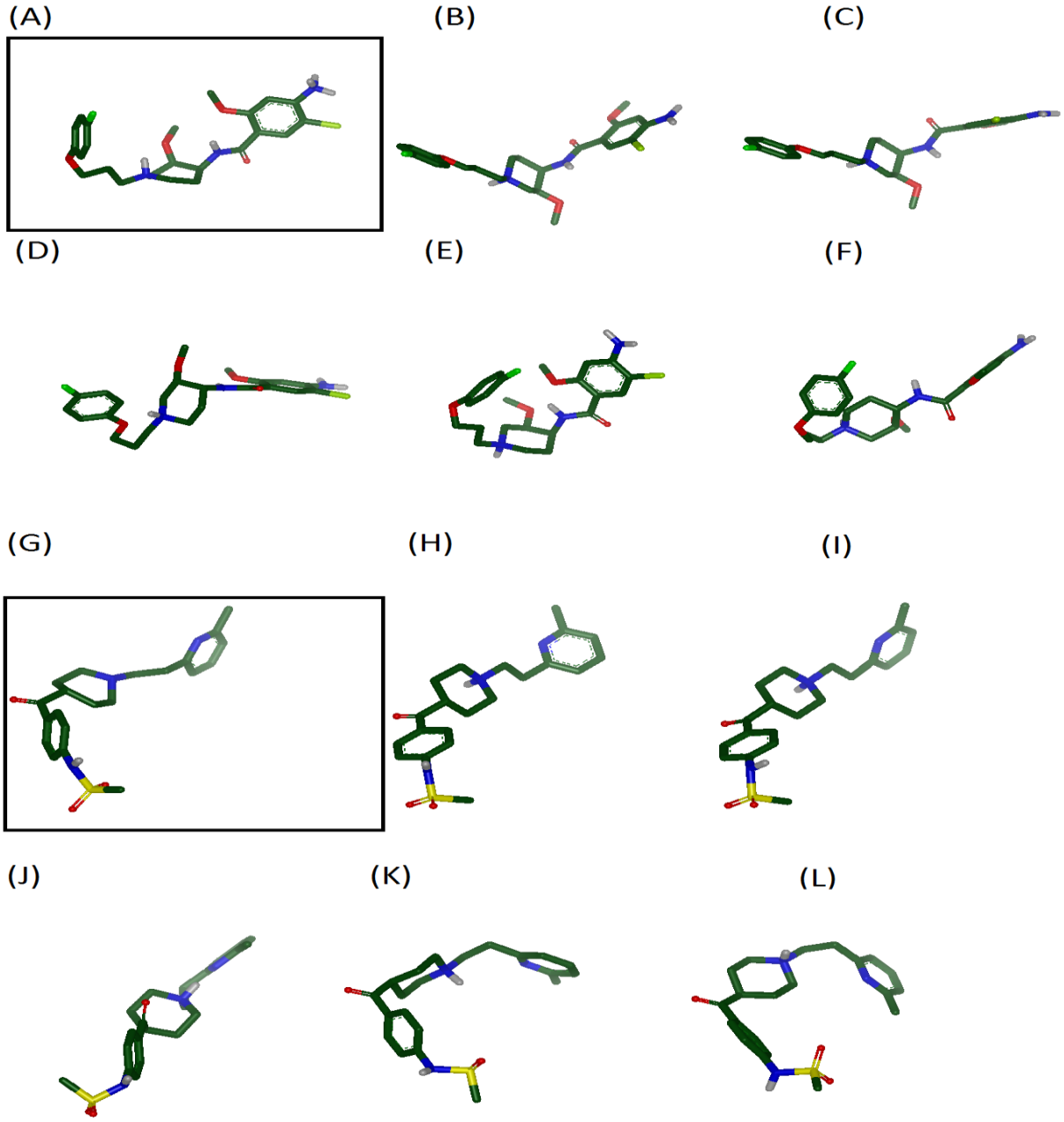


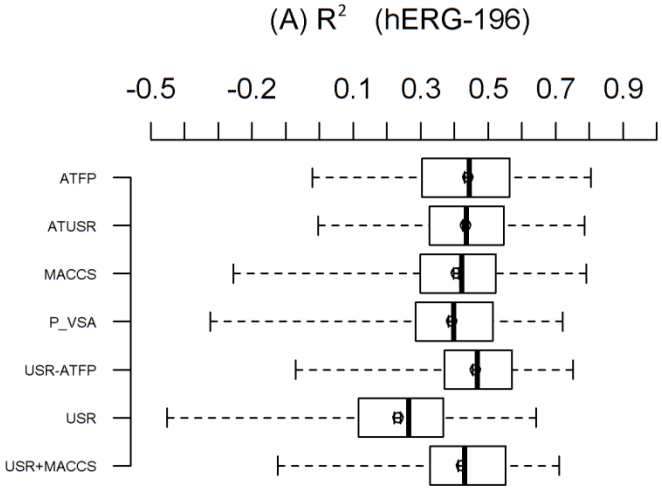
Figure 5.8 Images of Imai *et al.*[305] docked structures for cisapride (A) and E-4031 (G) and corresponding aligned structures obtained via: STERGEN (B,H), local minimisation (C,I), global minimisation (D,J), ChemScore selection (E,K) and RF-Score selection (F,L). All molecular images were generated using VIDA (version 4.1.1).[378]

Compound	Origin of Aligned Structure	RMSD (Å)	USR Similarity
Cisapride	STERGEN	2.44	0.65
	Local minimisation	2.77	0.55
	Global minimisation	2.47	0.69
	ChemScore selection	2.01	0.55
	RF-Score selection	2.24	0.55
E-4031	STERGEN	1.29	0.79
	Local minimisation	1.27	0.77
	Global minimisation	2.14	0.72
	ChemScore selection	2.11	0.54
	RF-Score selection	2.19	0.52

Table 5.2 RMSD values (computed from heavy atoms) and USR similarities (computed from all atoms) upon alignment of the structures obtained here to the docking poses presented by Imai *et al.*[305]

5.4.2 Relative Performance Across Different Datasets

The overall performance of the different descriptor sets for regression (Figure 5.9) and classification (Figure 5.10) modelling of different datasets is summarised below.



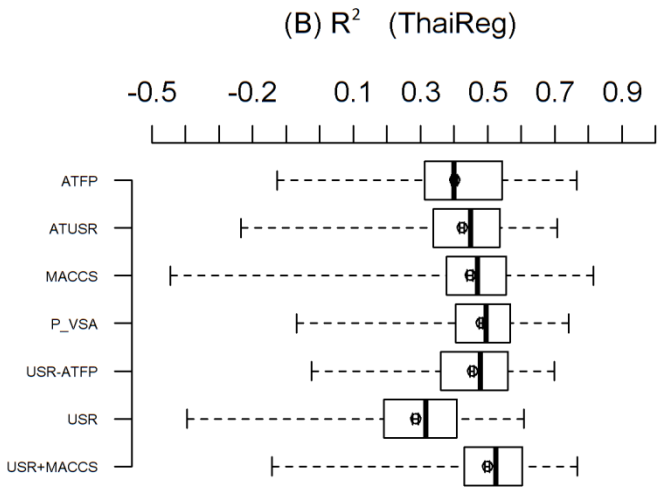
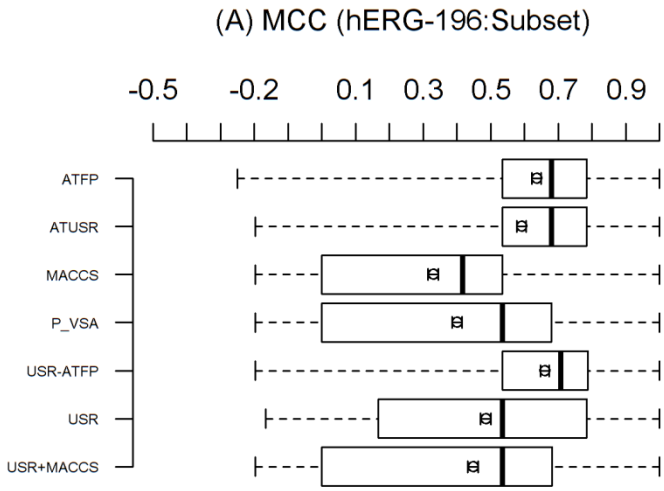
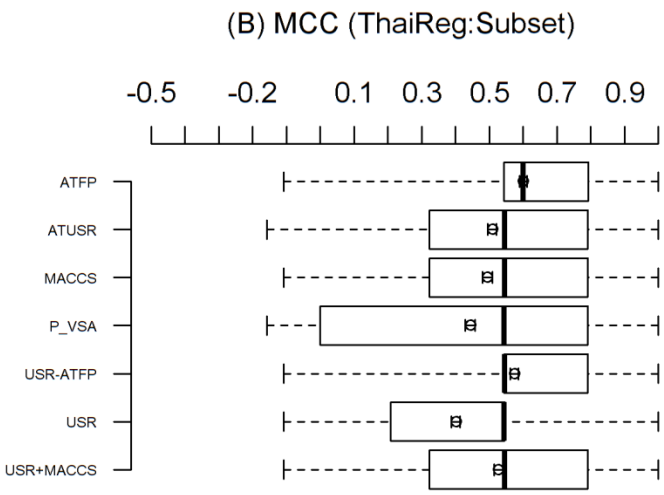
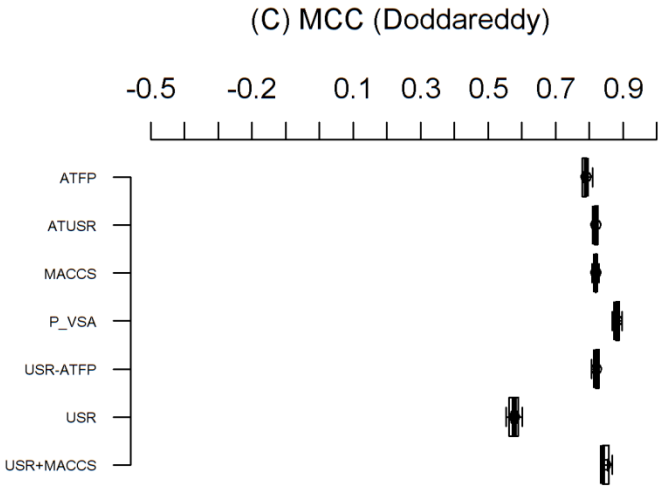


Figure 5.9 R^2 values obtained for: (A) hERG-196 dataset, (B) ThaiReg dataset. In both cases, the structures used for descriptor calculations were obtained from the standard workflow (see 5.3.4.1). The black lines and circles denote the median and mean results respectively.




(B) MCC (ThaiReg:Subset)
-0.5
-0.2
0.1
0.3
0.5
0.7
0.9
ATFP
ATUSR
MACCS
P_VSA
USR-ATFP
USR
USR+MACCS


(C) MCC (Doddareddy)
-0.5
-0.2
0.1
0.3
0.5
0.7
0.9
ATFP
ATUSR
MACCS
P_VSA
USR-ATFP
USR
USR+MACCS

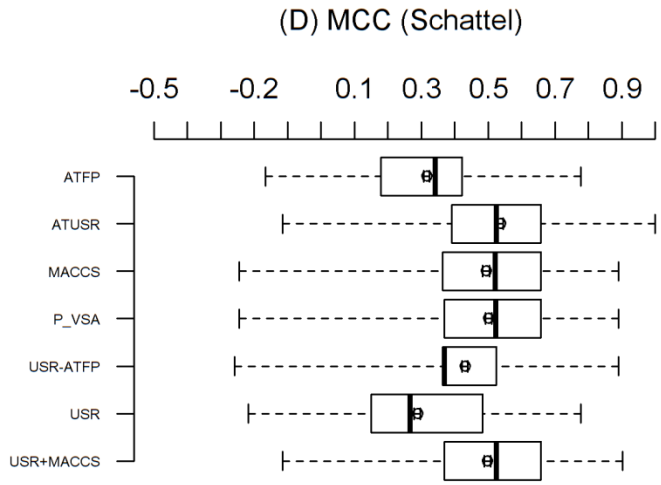


Figure 5.10 MCC values obtained for: (A) hERG-196:Subset dataset, (B) ThaiReg: Subset dataset, (C) Doddareddy dataset and (D) Schattel dataset. In all cases, the structures used for descriptor calculations were obtained from the standard workflow (see 5.3.4.1). The black lines and circles denote the median and mean respectively.

Once more, the regression results obtained for USR were evidently the worst. The overall mean R^2 values (Figure 5.9), Pearson and Spearman correlation coefficients were the lowest, and the overall mean RMSE and MAE values the highest (Appendix A), with USR for both datasets considered. Likewise, the only statistically significant pairwise differences in overall mean R^2 involved USR results (Appendix A).

Whilst, in terms of both the median and mean MCC value (Figure 5.10), the performance of USR was not the worst for the hERG-196: Subset dataset, its performance in terms of the mean MCC *was* the worst for all other classification datasets. Nonetheless, save for the Doddareddy dataset, none of the other descriptor sets was determined (Appendix A) to give significantly different mean MCC values to those obtained with USR.[*] This (given the statistically significant pairwise differences observed for the corresponding regression datasets, noted above) still lends support to the proposal that USR might be intrinsically better at discriminating clearly distinct actives from inactives than in generating regression models.

[*] When the structures used for descriptor calculations were prepared via the standard workflow (section 5.3.4.1).

Nonetheless, for both classification and regression modelling, the performance of USR descriptors was deemed statistically significantly better than random for all datasets when structures were obtained via the standard workflow (see section 5.3.4.1). Indeed, this was the case for all descriptor sets.

There is no indication that the failure of USR(-like) descriptor sets to (significantly) outperform the other descriptor sets, and the typically poor performance of USR in particular, reflects limited diversity of molecular shapes in these datasets.* As indicated by the results in Table 5.3, the hERG-196: Subset dataset, the only dataset for which the median (and mean) MCC for USR was not the worst, appears to be the least diverse in terms of molecular shape of the classification datasets.

Descriptor Set	Dataset Type	Dataset	Number of Clusters
USR	Classification	Doddareddy(Train)	209
		Doddareddy(Test)	36
		Schattel	35
		ThaiReg:Subset	31
		hERG-196:Subset	21
	Regression	ThaiReg	37
		hERG-196	31
logD	Classification	Doddareddy(Train)	51
		Doddareddy(Test)	30
		Schattel	21
		ThaiReg:Subset	21
		hERG-196:Subset	16
	Regression	hERG-196	21
		ThaiReg	19

Table 5.3 Numbers of clusters obtained, upon clustering using predicted logD (or USR descriptors), computed prior to the application of CORINA (or from structures processed via the standard workflow) with a distance cut-off of 0.2 (see section 5.3.6). For both descriptor sets, the classification and regression datasets were ranked in order of decreasing number of clusters - supposed to correspond to decreasing diversity.

For all but the ThaiReg:Subset classification dataset, there would appear to be *some* indication that the actives and inactives are separable in terms of lipophilicity. This is

* For a dataset composed of very similarly shaped molecules, differences in molecular shape could not account for significant differences in bioactivity.

suggested by the p(a) values obtained upon clustering the original datasets using predicted logD values being, commonly, either above or at the top of the range of those obtained when applying the same clustering procedure to the corresponding activity permuted datasets (Figure 5.11).

It is possible that separability of the actives and inactives for the Doddareddy dataset due to differences in lipophilicity could partially account for the good discriminative ability of the P_VSA descriptor set for this dataset (Figure 5.10), since some of its constituent descriptors are based on atomic contributions to logP.[295] Likewise it is possible that the lack of observable separability, in terms of lipophilicity, for the ThaiReg:Subset dataset contributed to the relatively low performance of the P_VSA descriptor set (Figure 5.10). Nonetheless, for this latter dataset, a 2D descriptor set (ATFP) gave the highest mean MCC. This latter observation suggests that failure of the 3D descriptor sets to (significantly) outperform the 2D descriptor sets, for the classification datasets, cannot simply be explained by the actives and inactives being separated in terms of lipophilicity.

Curiously, whilst the results prior to this point on the graph indicate separability of actives and inactives in terms of lipophilicity, there is a noticeable decrease in p(a) values (Figure 5.11) when clustering based on predicted logD is terminated for the Doddareddy training set upon formation of 1290 clusters or more (corresponding to 708 structures or less in the active cluster subset).[*] This result does not appear to be readily explainable.

It is important to remember that the quality of the logD predictions made by the tool employed here (see section 5.3.6),[379] in keeping with other methods, is unlikely to be similar across the whole of chemical space.[338] The (presumably variable) prediction error associated with the logD model employed potentially hampers the discernment of separability, on the basis of lipophilicity, of the actives and inactives in the datasets considered in this work.

[*] Detailed results tables corresponding to, and full size versions of the plots presented in, Figure 5.11 and Figure 5.12 are made available with this thesis (Appendix A).

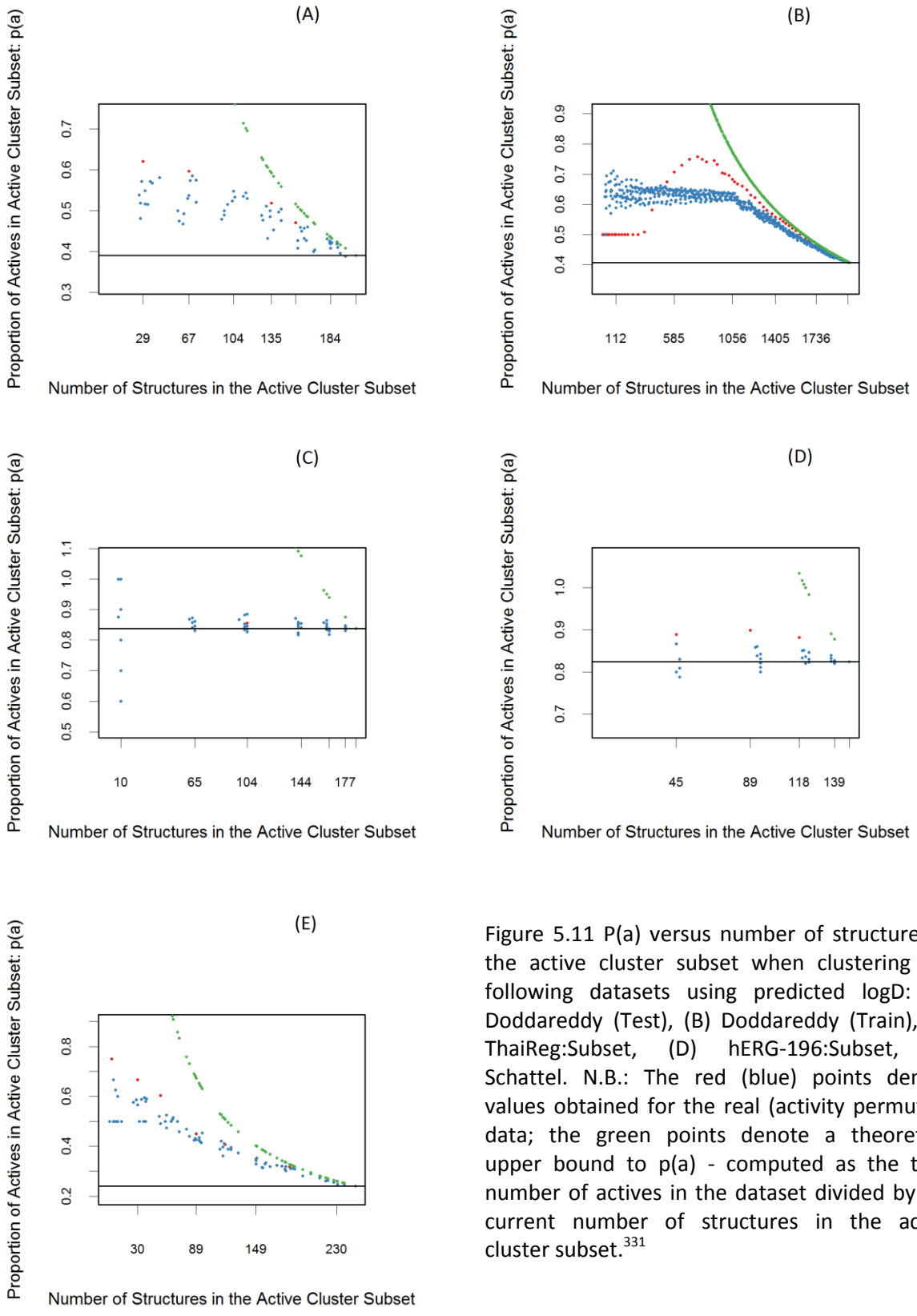


Figure 5.11 P(a) versus number of structures in the active cluster subset when clustering the following datasets using predicted logD: (A) Doddareddy (Test), (B) Doddareddy (Train), (C) ThaiReg:Subset, (D) hERG-196:Subset, (E) Schattel. N.B.: The red (blue) points denote values obtained for the real (activity permuted) data; the green points denote a theoretical upper bound to p(a) - computed as the total number of actives in the dataset divided by the current number of structures in the active cluster subset.[331]

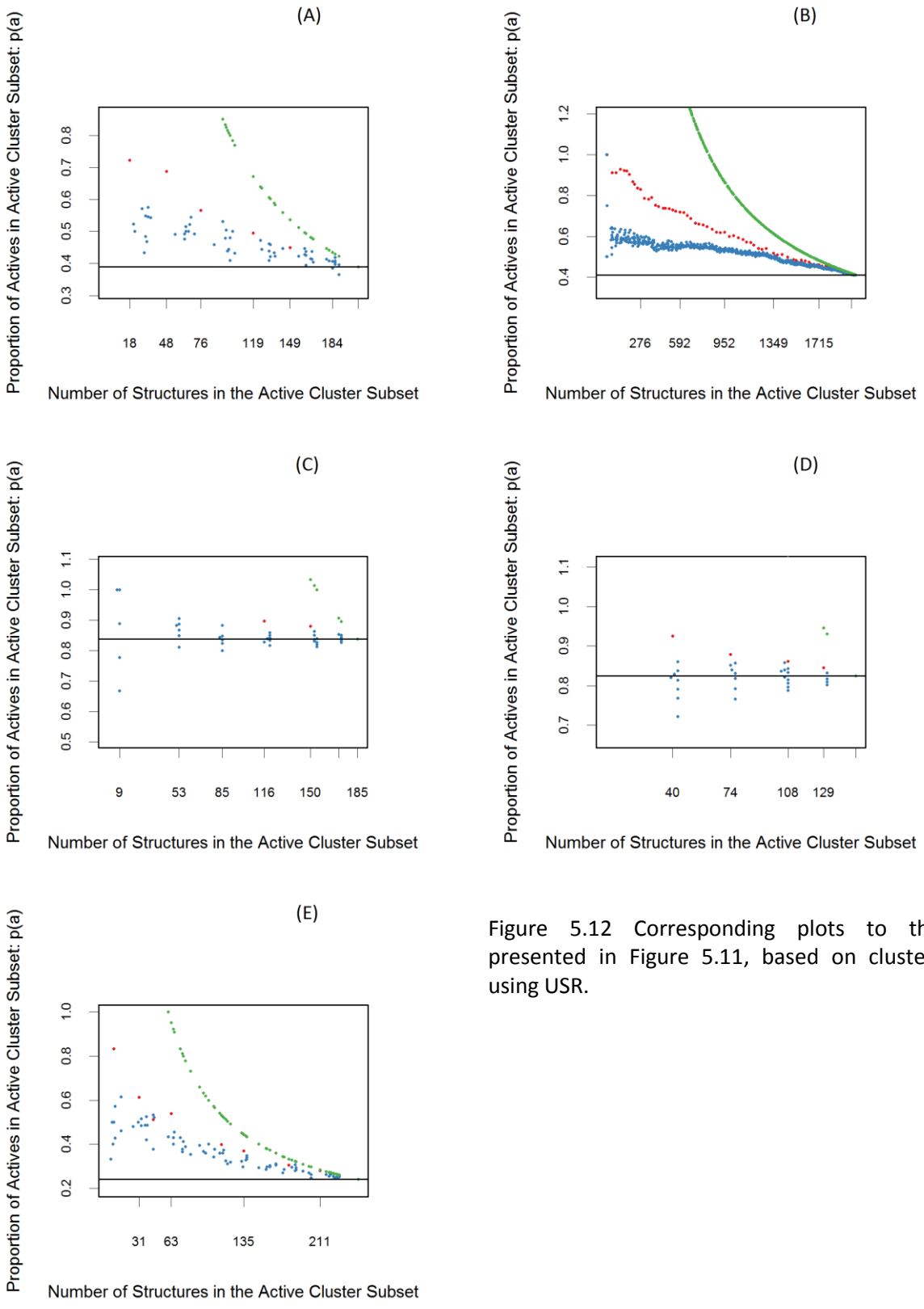


Figure 5.12 Corresponding plots to those presented in Figure 5.11, based on clustering using USR.

5.4.3 Effectiveness of Different Approaches to Combining Chemical Information with USR

An important question here concerns whether or not the ATUSR approach to combining USR with chemical information is (potentially) superior to the "hybrid"[175] approach of combining USR with a 2D chemical fingerprint - as was initially hypothesised. To remove the confounding factor of different atom types, it is therefore appropriate to specifically compare ATUSR to the USR-ATFP, ATFP and USR descriptor sets.

In the following discussion, all references to relative performance should be understood to solely refer to the relative performance of these descriptor sets (albeit no statistical quantities were recalculated) - where performance was measured in terms of the overall mean R^2 or MCC, as applicable. If the original hypothesis was correct, the following ranking would be expected: ATUSR > USR-ATFP > USR/ATFP.

Across all sets of results (obtained for all classification and regression datasets, and all procedures for obtaining 3D structures), the ATUSR descriptor set only yielded the best performance in two cases: for the Schattel and hERG-196:Subset classification datasets, when descriptors were computed from locally minimised structures for the latter. However, in neither case were the differences in descriptor set performance statistically significant. Indeed, the expected ranking (see above) was only observed for the Schattel dataset.

The USR-ATFP "hybrid" descriptor set outperformed all other descriptor sets under consideration for the Doddareddy and ThaiReg datasets as well as for the hERG-196 and hERG-196: Subset datasets under some circumstances.[*] However, considering those sets of results where the best performance was obtained using the USR-ATFP descriptor set, statistically significant differences between the performance obtained using USR-ATFP and any other descriptor set were only observed for the regression datasets and the Doddareddy dataset. Moreover, all such statistically significant differences involved USR descriptors - i.e. the performance of the USR-ATFP or ATUSR descriptor set was not deemed statistically significantly different to the performance of ATFP descriptors.

[*] These circumstances being when the 3D structures used to compute the relevant descriptors were obtained from the standard and STERGEN procedures, as well as the local minimisation procedure for the hERG-196 dataset.

Hence, these results lend inconclusive support to the proposition that the ATUSR descriptor set may yield the best performance under some circumstances. Indeed, these results only weakly support to the proposition that the combination of USR with chemical information leads to improved performance over simply considering chemical information in isolation. This point is reinforced by the observation that statistically significant differences in performance were never observed between the MACCS and USR+MACCS descriptor sets.

Of course, one caveat to note is that the statistical testing protocol employed here (see section 5.3.5) was focussed on controlling Type I error - i.e. erroneously declaring the difference in performance between two descriptors to be genuinely different when the observed difference arose from chance.[226,373] This was deemed appropriate as, arguably, the onus should be on demonstrating that new methods genuinely perform better than those currently in widespread use. The use of adjusted p-values, however, is likely to lead to higher Type II error[303] - i.e. erroneously declaring differences in performance to be statistically insignificant when genuine differences in performance exist.[226,373]

5.5 Conclusions

This chapter described the development of a novel set of 3D descriptors (ATUSR) encoding molecular shape and the three-dimensional distribution of potential macromolecular interaction sites. This was achieved by adding chemical information to the Ultrafast Shape Recognition (USR) descriptor set developed by Ballester and Richards[263,264] in a conceptually distinct fashion to the approach previously proposed by Cannon *et al.*[175] Initial experiments were carried out to evaluate the effectiveness of ATUSR for both regression and classification protein-ligand binding QSAR.

These evaluations included comparisons with standard 2D descriptors as well as comparisons with a combination of USR and a 2D fingerprint (ATFP) based on the same atom types as ATUSR: USR-ATFP. In keeping with the findings of Cannon *et al.*,[377] the results obtained *suggested* that combining USR with 2D chemical information could lead to improved performance, with respect to both USR and the original 2D descriptor set, for certain classification and regression tasks under some circumstances. In one case, the results suggested ATUSR could lead to further improved performance. However, in no cases were these improvements deemed statistically significant.

Indeed, many of the descriptor sets (both 2D and 3D) assessed appeared to perform comparably for both regression and classification QSAR. One notable exception, however,

was USR. Save for the results obtained on one classification dataset, the overall performance of USR was the worst of all descriptors evaluated, emphasising the importance of encoding chemical information. Nonetheless, all descriptor sets almost always yielded statistically significantly better performance than that expected for a random model.

The use of various Molecular Mechanics and docking approaches to yield, ostensibly, improved representations of the bioactive conformer, for one of the datasets employed in this work, did not lead to the expected improvements in performance for the 3D descriptor sets. However, this appeared to reflect the difficulty of employing these methods to find the bioactive conformer for hERG inhibitors.

For the datasets considered in the current work, it appeared that an increased diversity of molecular shapes did not have an effect on the relative performance of USR(-like) descriptor sets.

A particularly interesting issue, however, is that biases in the datasets employed could have favoured 2D descriptor sets with respect to 3D descriptor sets. Qualitative indications were obtained, for some of the classification datasets, that the actives and inactives were partially separable in terms of lipophilicity. This author suggests that this might partially account for the good performance of the 2D descriptor sets relative to the 3D ones on such datasets.

However, further analysis of analogue bias (an issue which there was not time to investigate in the work presented in this chapter)[336,339] would also be warranted, as this may lead to artificially high observed performance for 2D descriptor sets.[330] Some specific proposals are presented in Chapter 7.

The code for the ATUSR descriptor set, and related descriptor sets, is made publicly available to the community (Appendix A), to facilitate further evaluations (or extensions) of these methods.

Chapter 6 Predicting Drug Induced Torsades de Pointes Using Biological Descriptors

This chapter describes the development of QSAR approaches for discriminating between compounds with the potential for inducing Torsades de Pointes (TdP) and those without, denoted TdP+ and TdP- compounds respectively. The novelty introduced in this work is the incorporation of predicted pIC_{50} values for biologically relevant cardiac ion channels as additional/alternative descriptors ('IC-descriptors'). A variation on this idea, namely the incorporation of experimental pIC_{50} values as additional descriptors was also considered. The former types of descriptors are denoted 'predicted IC-descriptors' and the latter 'experimental IC-descriptors'. The former represents an assessment of a purely *in silico* approach, whilst the latter explores how useful *in vitro* data might be in enhancing QSAR approaches to predicting drug induced TdP. Previously presented approaches to predict Torsadogenic potential *in silico*, with a particular emphasis on QSAR studies, are first discussed. The QSAR methods used in the current work to identify TdP causing agents, along with the approaches used to generate IC-descriptors, are then discussed. The datasets used to build and validate TdP models and those used to generate IC-descriptors are subsequently presented - and the challenges associated with modelling this data highlighted. The insights obtained to date into the performance of IC-descriptors are discussed. This chapter concludes by emphasising the need for future work.

6.1 Introduction

The first published study to employ QSAR approaches for directly discriminating between compounds assessed as being TdP causing (TdP+) and supposed non-TdP causing agents (TdP-), was presented by Yap *et al*. in 2004.[27] They developed their models using two sets of 'standard' descriptors, in conjunction with various Machine Learning algorithms.[27]Around the same time, Xue *et al*. employed Support Vector Machines (SVMs), and a filtered collection of commonly used descriptors, to model a similarly derived dataset.[122]

Since then, a number of publications have reported models generated on the datasets made available by these authors. A variety of Machine Learning methods, and feature selection strategies were considered in these studies - although, in most cases, the same descriptors were employed as per the original publications.[124–127,199]

More recently, Frid and Matthews reported the construction of a dataset of 1,632 approved drugs, with data from a variety of sources used to categorise compounds as TdP+ or TdP- (this author's terminology).[33] Models for discriminating between TdP+ and TdP- compounds were developed using various software packages commonly employed in computational toxicology.[29]

Similarly, Clark and Wiseman recently disclosed a dataset of more than 1,000 drugs identified as TdP+ or TdP-, principally using statistical criteria. They developed binary classification models using descriptors based on a dictionary of molecular fragments. [28]

It should be noted that all previously referenced studies sought to relate general purpose descriptors, directly encoding molecular structure, to the Torsadeogenic potential of a compound. In contrast, Gepp and Hutter generated models incorporating (amongst a pool of general purpose descriptors) descriptors based on a pharmacophoric SMARTS pattern for human-ether-a-go-go related gene (hERG) ion channel blockers and similarity to the potent hERG inhibitor astemizole.[128]

Whilst not seeking to directly predict Torsadogenic potential *per se*, studies by Yang *et al.* [107] and Obiol-Pardo *et al.*[104] simulated the effects of drug compounds on the QT interval (see Chapter 1, section 1.3.3). Interestingly, the latter incorporated predicted pIC_{50} values for two cardiac ion channels. Last year (2011), Mirams *et al.* used experimentally measured IC_{50} values for various cardiac ion channels, along with other experimental parameters and simulation derived parameters, as descriptors to build multiclass models for Torsadogenic potential.[159]

The current work builds upon these earlier attempts to incorporate biologically relevant information into the prediction of Torsadogenic potential. This work extends these earlier investigations by seeking to use predicted pIC_{50} values as descriptors in the TdP models. The resulting cardiotoxicity predictions are evaluated on a larger number of compounds than Obiol-Pardo *et al.* [104] and Mirams *et al.* [159] Moreover, the use of conventional (i.e. purely structural) QSAR descriptors, as an alternative/in addition to using (predicted or experimental) pIC_{50} values is assessed, to the best of this author's knowledge, for the first time.

6.2 Modelling Approaches Employed Here

Binary classification models for discriminating between TdP+ and TdP- compounds were built using Random Forest. Models were generated using standard QSAR descriptors and (predicted) pIC_{50} values for (different combinations of) ion channels whose inhibition was deemed to be relevant to Torsadogenic potential. The former set of descriptors are referred to as 'structural-descriptors', with the latter referred to as 'IC-descriptors'. Models were generated using both types of descriptors alone and in combination.

When experimentally obtained pIC_{50} values were used as descriptors, these are denoted 'experimental IC-descriptors'. When predicted pIC_{50} values were employed, these are denoted 'predicted IC-descriptors'. Predicted IC-descriptors were generated using Random Forest regression models, based on the same structural-descriptors used for directly modelling TdP.

6.2.1 Structural-Descriptors

The following combination of structural-descriptors was used to build models for TdP and the ion channel models: Molecular Linear Free Energy Relationship Descriptors (MLFER),[380] McGowan's characteristic volume,[381] as well as a fragment-based bit-string descriptor similar to that used by Clark and Wiseman to develop their models for Torsades de Pointes.[28]

The MLFER descriptors correspond to the 'overall' (or "summation") hydrogen bond acidity (MLFER_A), two descriptors (MLFER_BH, MLFER_BO) representing the 'overall' (or "summation") hydrogen bond basicity, the combined dipolarity/polarizability (MLFER_S), the excess molar refraction (MLFER_E) and the gas-hexadecane partition coefficient (MLFER_L).

These descriptors, along with McGowan's characteristic volume, were calculated using PaDEL-Descriptor (version 2.5).[382,383] The MLFER descriptors represent an implementation of the fragment-based approach, developed by Platts *et al.*,[380] for calculating the experimentally derived solute-descriptors proposed by Abraham[384] to capture different contributions to solute-solvent interactions (personal correspondence with Dr Chun-Wei Yap). As per the corresponding set of descriptors used by Abraham[384] and Yap,[27] the combined set of MLFER descriptors and McGowan's volume is referred to in this thesis as the Linear Solvation Energy Relationship (LSER) descriptors.

Work by Abraham indicated that these LSER descriptors capture properties relevant for protein ligand binding.[384] Since both plasma concentrations (which are related to solubility) and protein interactions are relevant for Torsadogenic potential (see Chapter 1, section 1.3.3), these were appropriate descriptors to use for directly modelling TdP and constructing ion channel models. Indeed, the TdP models developed by Yap *et al.* were based on a subset of these descriptors - albeit, their implementation was not exactly the same as used here (personal correspondence with Dr Chun Wei Yap).[27]

In their 2009 study, Clark and Wiseman used 321 molecular fragments to define a bit-string descriptor encoding the presence or absence of these fragments in a molecule.[28] In earlier work, an occurrence count descriptor based on a subset of these fragments was used to generate regression models for various molecular properties,[385] including hERG inhibition.[222]

In this work, the 319 pseudo-SMARTS patterns presented by Clark and Wiseman,[28] were adapted and a bit-string descriptor implemented based upon the resultant SMARTS patterns. The bit-vector was implemented in Python, using the Pybel module.[177] This Python code, along with a description of the modifications made to the pseudo-SMARTS patterns, is made available with this thesis (Appendix A). There are some discernible differences between this bit-string descriptor and that implemented by Clark and co-workers as illustrated by an example in Figure 6.1.[222]

This bit-string was able to encode the presence of moieties which could engage in specific binding within an ion channel pore. For example, it encoded the possible presence of a tertiary amine, a typical feature of potent hERG inhibitors, potentially facilitating non-classical hydrogen bonds within the ion channel pore,[305] as discussed in Chapter 4. Since the specific interactions allowed for by such fragments would not be encoded by the LSER descriptors, the addition of these bit-string descriptors was deemed appropriate for generating ion channel models as well as directly predicting TdP.

The structural-descriptors presented here do not encode stereochemistry - as evidenced by Platts *et al.*,[380] Abraham and McGowan,[381] and the pseudo-SMARTS patterns obtained from Clark and Wiseman.[28] This informed the processing of the datasets described in section 6.3.

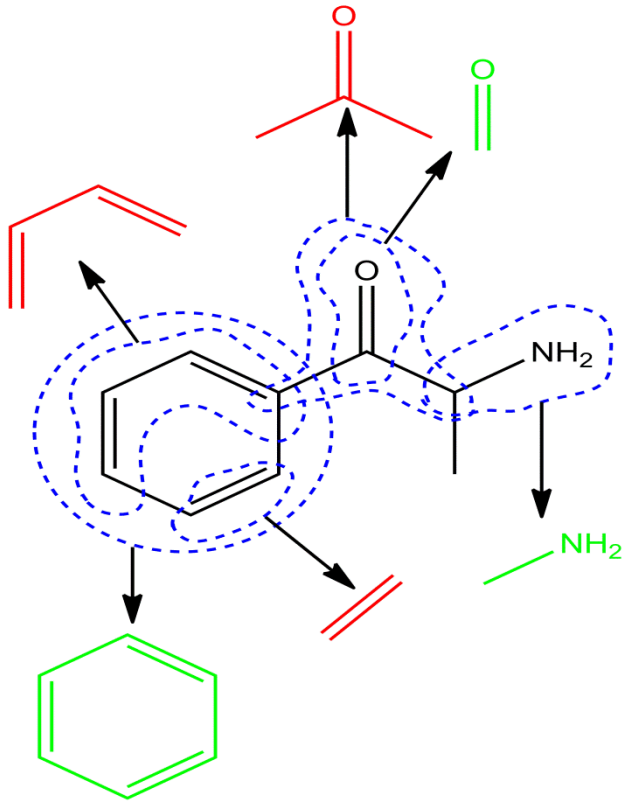


Figure 6.1 Fragments identified in cathinone by Clark and co-workers that were likewise found (green) and not found (red) by this author's implementation of their bit-string descriptor.

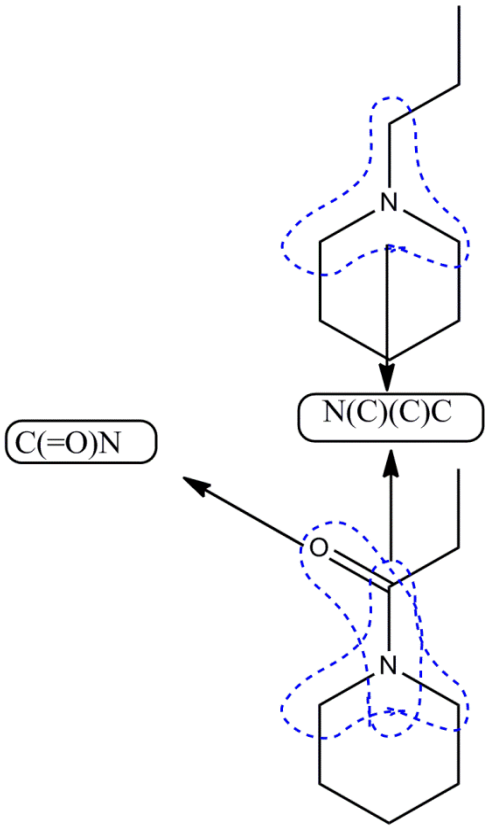


Figure 6.2 SMARTS patterns matching a tertiary amine and the corresponding amide.

6.2.2 Descriptor Pre-processing

Prior to generating the ion channel and TdP models, all descriptors with constant values in the training set were removed. All descriptor values (X) were scaled and centred, using the training set maximum ($X_{max,train}$) and minimum ($X_{min,train}$) values to give new values (X_{scaled}) *designed* to lie between 0 and 1, as per the bit-string fingerprint descriptors (Equation 6.1). This was carried out in anticipation of generating additional models using alternative Machine Learning algorithms, that could be explored in the future, since the Random Forest algorithm used here is unaffected by scaling or centring of descriptors.

$$X_{scaled} = \frac{(X - X_{min,train})}{(X_{max,train} - X_{min,train})} \qquad 6.1$$

6.2.3 TdP Model Generation

Binary classification (TdP+ vs. TdP-) Random Forest models were generated using the randomForest package in R.[225] Models were generated using n_{tree} = 501 (see Chapter 5, section 5.2.2) and the default value for m_{try} (see Chapter 2, section 2.6.3.2). Hyperparameter optimisation was not considered for computational expediency, since Random Forest is noted to typically perform well with its default hyperparameters,[192] and this work was focussed on the effect of using different descriptor combinations to generate TdP models. All models were generated five times, using different random number generator (RNG) seeds.

The scripts employed for TdP model generation are made available with this thesis (Appendix A).

6.2.4 Generation of Predicted IC-descriptors

The Random Forest regression models used to generate predicted IC-descriptors were built on the datasets described in section 6.3.1. Care was taken to ensure that these datasets did not overlap with those used to generate the TdP models, in order to investigate the effect of incorporating genuinely *predicted* IC-descriptors into models for TdP.

The arithmetic mean of all pIC_{50} estimates associated with a given dataset entry was used as the response variable. All structural-descriptors discussed in section 6.2.1 were used, subsequent to the pre-processing described in section 6.2.2.

The Random Forest implementation matched that used to generate the TdP classification models (section 6.2.3). The default number of trees was used. However, in order to better evaluate the effect of using predicted IC-descriptors, it was considered more critical to maximise the predictive power of these models; hence, in contrast to before, efforts were made to optimise m_{try}, by evaluating:

$$m_{try} = max(floor\left(j.\frac{M}{6}\right), 1)$$

for j = 1,2,3,4,5,6. Here M denotes the total number of descriptors, and *floor( )* rounds down to the nearest integer. This sequence was chosen to include the randomForest default (see

Chapter 2, section 2.6.3.2), half and twice this value, as has been recommended, and the maximum possible value of m_{try} which may be appropriate when few descriptors are relevant (a number of the bit-string-descriptors, e.g. the presence of an ethane fragment, were suspected to be largely irrelevant).[386] The model with the highest out-of-bag[192] Pearson's correlation coefficient was selected (see Chapter 2, sections 2.6.3.2 and 2.6.4.2 respectively). The scripts used to build these models are made available with this thesis (Appendix A).

6.2.5 Generation of Experimental IC-descriptors

For a given cardiac ion channel, where experimental estimates for pIC_{50} values were available for a compound in the TdP datasets, their arithmetic mean was used as an experimental IC-descriptor. The origin of these measurements is explained in section 6.3.1. As explained there, pairs of bioactivity files, matching compound IDs to synonyms and (possibly multiple) pIC_{50} estimates, were derived for each cardiac ion channel for which data was available. The compounds present in the TdP datasets were matched to pIC_{50} measurements in these files by comparing the names of the TdP dataset compounds to the synonyms for the ion channel inhibitors presented in these files - ignoring trivial differences such as spaces, capitalisation etc.

6.3 Datasets

6.3.1 Ion Channel Datasets

In order to generate IC-descriptors, inhibition/activation measurements were sought for cardiac ion channels, inhibition or activation of which might be mechanistically related to the induction of Torsades de Pointes. Inhibition measurements were obtained for the following ion channels, denoted by the name of the gene encoding the primary alpha-subunit: SCN5A ($Na_v1.5$), CACNA1C ($Ca_v1.2$), KCNH2 (hERG), KCNQ1 (KVLQT1), which carry the I_{Na}, $I_{Ca,L}$, IK_r and IK_s currents respectively.[387] Inhibition/activation of these currents within ventricular tissue has the potential to increase or reduce the duration of ventricular repolarisation, the prolongation of which may lead to TdP (see Chapter 1, section 1.3.3). Changes in all of these currents have been suggested to play a role in either promoting/counteracting QT interval prolongation and the induction of Torsades de Pointes.[104,106,159,387]

Inhibition measurements were also obtained for the following ion channels (denoted as before): CACNA1H ($Ca_v3.2$),[388] CACNA1G ($Ca_v3.1$), KCNA5 ($K_v1.5$).[387] Due to their (typically) selective expression in non-ventricular - for example, atrial - cells,[387–389] it could be argued that inhibition/activation of their currents is irrelevant to TdP.[389] However, atrial fibrillation has been suggested to be linked with a reduced risk of Torsades de Pointes.[390] Hence, it was deemed reasonable to additionally investigate whether or not generating IC-descriptors based upon inhibition measurements for these ion channels might also enhance models for TdP.

The ChEMBL database (version 11, MySQL format)[269] was searched for inhibition/activation measurements for the cardiac ion channels noted above.[387] The script used to extract data from ChEMBL is made available with this thesis (Appendix A). Measurements, with the maximum ChEMBL confidence score of nine, were extracted for mammalian targets, corresponding to the gene names encoding the primary alpha subunits.[387,388] Measurements which the extracted assay descriptions indicated were derived from competitive binding assays were excluded, since these assays cannot distinguish between channel activation/inhibition,[67] which have distinct effects on the cardiac action potential.[106]

The inhibition measurements derived were pIC_{50} values, IC_{50} values (directly converted into pIC_{50} values) and percent inhibition values. The latter were converted into pIC_{50} estimates, based on the test concentration provided, using the Logit function. Johnson *et al.* found pIC_{50} estimates obtained using the Logit function had a median error of only 0.1 log units for percent inhibition from 20-80 %; hence, only percent inhibition values in this range were considered.[81]

Synonyms and molecular structures (canonical SMILES) were also extracted for the corresponding compounds.

Additional data for KCNH2 was obtained from the Literature-368 dataset described in Chapter 4, as well as the hERG-196 dataset described in Chapter 5. Data for KCNH2 and KCNQ1 was also obtained from Obiol-Pardo *et al.*[104] - kindly provided in electronic format by Dr Manuel Pastor.

The data selected from all three sources was combined for each ion channel. Only a single entry corresponding to a given stereochemically indifferent InChI (see section 6.3.5) was retained in the dataset used to generate the ion channel model; all remaining entries were

separately recorded - i.e. two sets of files, an SDF and a corresponding bioactivity file, were derived per ion channel.

The removal of dataset entries in this fashion was deemed appropriate for the following reasons. Firstly, the stereochemistry of all structures in the Literature-368 dataset had not been checked. Hence, stereochemistry dependent comparisons could have erroneously retained duplicates. Secondly, for some sets of entries in the datasets compiled by Obiol-Pardo *et al.*, the same IC_{50} measurement was assigned to all potential stereoisomers, where this was unknown to Obiol-Pardo and co-workers (private correspondence with Dr Manuel Pastor). Hence, stereochemistry dependent comparisons would have retained duplicate measurements. Thirdly, since stereochemically indifferent structural-descriptors were used to generate IC-descriptors (section 6.2.1), the retention of stereoisomers could have introduced *effective* redundancy when training and validating the corresponding models.

Each dataset entry was assigned an ID - either the name associated with entries in the Obiol-Pardo *et al.* datasets, the ChEMBL molregno, or unique IDs assigned to the combined set of compounds taken from the Literature-368 and hERG-196 datasets. Where pairs of dataset entries, originating from one of these three original sources, were associated with the same stereochemistry dependent InChI, yet their IDs did not indicate they were the same compound/stereoisomers, both dataset entries were removed. Likewise, where the synonyms associated with pairs of entries in the bioactivity files corresponding to the compound sets prepared for the generation of ion channel models, indicated that the entries corresponded to the same compound, both dataset entries were removed. In both cases it was deemed likely that some of the associated structures were topologically incorrect.[*]

This yielded the pairs of files per ion channel which were parsed to generate experimental IC-descriptors as described in section 6.2.5.

Finally, the dataset entries deemed to correspond to compounds in the TdP datasets when generating experimental IC-descriptors (see section 6.2.5) were removed, followed by removal of dataset entries for which their stereochemically indifferent InChIs were matched to those of compounds in the TdP datasets.

---

[*] Alternatively, some of these inhibitors could have been assigned the correct structure but incorrect synonyms - this being the rationale why the bioactivity files available subsequent to filtering these 'suspicious IDs' were parsed to obtain measurements for experimental IC-descriptors (see section 6.2.5).

Ultimately, this yielded datasets of the following size (Table 6.1), which were used to generate the ion channel models employed for calculating predicted IC-descriptors - since the 45 compounds retained in the SCN5A dataset fell below a heuristic minimum of 50 required for model generation, no ion channel model was built on this dataset.

Ion Channel	n.o. Compounds
CACNA1H	73
KCNH2	1,285
CACNA1G	192
CACNA1C	106
KCNQ1	59
KCNA5	300

Table 6.1 Total number of compounds used to generate ion channel models.

All dataset files used to generate these ion channel models are made available with this thesis (Appendix A).

6.3.2 TdP Datasets

TdP models were evaluated on datasets derived from those originally presented by Yap *et al.*[27] (Yap-2004 dataset) and Clark and Wiseman[28] (Clark-2009 dataset). The initial descriptions below refer to these original datasets, whilst the numbers of compounds remaining in the versions of the Yap-2004 and Clark-2009 datasets ultimately used to assess the modelling approaches introduced in the current work, following the removal of compounds described in the intervening sections, are presented in section 6.3.3.4.

Yap-2004 Dataset

Yap and co-workers originally selected TdP+ compounds from agents reported, as of 2003, by the Arizona Center for Education and Research on Therapeutics (ArizonaCERT)[391] to have a risk (or a possible risk) of inducing TdP in humans or otherwise to be avoided by patients with congenital long QT syndrome (LQTS). They aimed to exclude compounds otherwise reported as having a weak association with TdP. Additional TdP+ compounds were elsewhere reported to be associated with TdP/ventricular tachyarrhythmias on the basis of clinical studies/case reports, or were agents with official warnings regarding the occurrence of TdP. All such assignments were based on human studies. Their 243 TdP- compounds

corresponded to an additional set of agents for which they found no reported case of TdP in humans.[27]

Here, structures corresponding to the names presented by Yap *et al.*[27] were obtained from Ligand.Info (ligand_info_ver_1_02.sdf.zip),[392,393] DrugBank,[386–389] PubChem,[169,394] SciFinder,[372] the literature (gentamicin),[395,396] or (anakinra, 1IRP) the PDB (as prompted by DrugBank).[261,262,397] The structures were aggregated into a pair of SDF files, corresponding to the training and test set defined by Yap *et al.*, using the Python Pybel module.[177] These SDF files are made available with this thesis (Appendix A).

Clark-2009 Dataset

Clark and Wiseman derived TdP+ assignments for all but five of the TdP+ compounds in this dataset from the FDA's Adverse Event Reporting System (AERS), with drugs extracted from this database assigned as TdP+ on the basis of reports from the AERS describing compounds as either primary of secondary suspect agents in incidents of TdP. Chi-squared values were computed based upon the number of such reports associating a given drug with TdP, compared to an estimation of the number of such reports expected if there was no causative relationship between taking the drug and TdP. Only if the chi-squared value was statistically significant at the 5% level, was the null hypothesis - that the reports associating the drug with TdP were due to chance co-occurrence - rejected; otherwise, drugs reported in the AERS were assigned to the TdP- class. An additional set of five compounds withdrawn from the market due to QT-associated events were also assigned to the TdP+ class, yielding 71 TdP+ compounds in total.[28]

The 1,307 SMILES presented by Clark and Wiseman were downloaded; non-standard (H) symbols were removed.[179] TdP+ class labels were assigned based on the names of the 71 compounds listed in another of their Supporting Information files.[28]

6.3.3 Resolving Potential Problems Identified with the TdP Datasets

An overview of the identification and removal of problematic entries from the versions of the Yap-2004 and Clark-2009 datasets originally derived in the current work is provided below. The calculations presented here were only carried out on the filtered versions of these datasets. The identities of all compounds removed from both datasets are made available with this thesis (Appendix A).

6.3.3.1Problematic Class Assignments

In their 2006 study, Gepp and Hutter[128] noted that three compounds assigned as TdP- by Yap *et al.* were reported as being weakly associated with TdP by ArizonaCERT. They identified three additional TdP- compounds which were reported to prolong the QT interval and/or might cause TdP - or, in one case, have a "definitive association" with TdP.[129] In order to improve the consistency of class label assignment within the Yap-2004 dataset, all six of these TdP- compounds were removed.[128]

The possibility of unidentified 'false negatives' remains an issue for both the Yap-2004 and Clark-2009 datasets (see Chapter 1, section 1.3.3.2).

6.3.3.2 Problematic Structures

Dataset entries found to correspond to polymers,[*] as well as those identified as corresponding to multiple fragments, were removed. The latter decision was prompted by the detection of compounds (e.g. trimetaphan camsilate) associated with a large organic counterion; the standard approach of discarding minor fragments may be inappropriate for such multifragment compounds, for which the bioactive component is ambiguous.[165] For simplicity, all identified multifragment compounds were removed.

Additional structures in the Clark-2009 dataset were determined to be incorrect via consulting ScifFinder.[372] Some structures with incorrect topology in this dataset were identified via consideration of those cases where the names in the Clark-2009 and Yap-2004 datasets matched, yet not the corresponding stereochemically indifferent InChIs (section 6.3.5). These structures were deemed erroneous based upon their inconsistency with the Yap-2004 structure, which was checked using SciFinder[372] or, in the case of neomycin, a literature reference.[399] Vasopressin was removed from both datasets as no clear structural reference was obtained. Further structures in the Clark-2009 dataset were suspected to be incorrect - for example, dataset entries corresponding to atomic sulfur and lithium - or, in the case of xenon, i.e. a noble gas, to offer no generalisation ability. All these structures were removed from the Clark-2009 dataset.

---

[*] Monomeric structures were originally obtained for the polymeric exchange resins,[128] colestipol and colestyramine (cholestyramine); the need for a distinct approach to QSAR modelling of polymers was emphasised by England.[398]

6.3.3.3 Structural Redundancy

All compounds with non-unique stereochemically indifferent InChIs (see section 6.3.5) were removed.

6.3.3.4 Final Versions of TdP Datasets Employed in this Work

Ultimately, the steps undertaken in the preceding sections yielded filtered versions of the Yap-2004 and Clark-2009 datasets comprising 329 (99 TdP+, 230 TdP-) and 1170 (67 TdP+, 1103 TdP-) compounds respectively. *All references to these datasets presented from here on, unless noted otherwise, should be understood to refer to these filtered versions.*

6.3.4 TdP Dataset Compounds for which Experimental IC-descriptors were Available

The numbers of compounds in the Yap-2004 and Clark-2009 datasets for which experimental pIC_{50} estimates were obtained are presented in Table 6.2.

The original files* presenting these measurements are made available with this thesis (Appendix A).

* N.B.: As these were derived prior to filtering the TdP datasets, the names of some compounds which were ultimately discarded are presented, along with their associated measurements, in these files. These compounds are not included in the counts presented in Table 6.2.

TdP Dataset	Ion Channel	TdP Class	n.o. Compounds with Experimental IC-descriptors
Yap-2004	CACNA1H	TdP+	1
		TdP-	0
	SCN5A	TdP+	1
		TdP-	1
	KCNH2	TdP+	52
		TdP-	8
	CACNA1C	TdP+	0
		TdP-	0
	CACNA1G	TdP+	1
		TdP-	0
	KCNA5	TdP+	1
		TdP-	0
	KCNQ1	TdP+	14
		TdP-	1
Clark-2009	CACNA1H	TdP+	0
		TdP-	0
	SCN5A	TdP+	2
		TdP-	1
	KCNH2	TdP+	37
		TdP-	88
	CACNA1C	TdP+	0
		TdP-	1
	CACNA1G	TdP+	0
		TdP-	0
	KCNA5	TdP+	1
		TdP-	2
	KCNQ1	TdP+	11
		TdP-	10

Table 6.2 Numbers of TdP dataset compounds with experimental IC-descriptors based on comparing names of compounds in TdP and ion channel datasets.

6.3.5 Standardization and Comparison of Structures

All structures in the TdP and ion channel datasets were standardized using the same procedure. All structures in a given dataset, including a couple for which some standardization steps failed, were transferred to a single SDF file using the Python module Pybel,[177] prior to InChI generation, and subsequent calculation of structural descriptors. This

workflow is illustrated in Figure 6.3. The exact options employed for standardization are specified in supplementary files made available with this thesis (Appendix A).

InChIs were computed as per Chapter 5, section 5.3.4.3. All references to stereochemically dependent InChIs in this chapter refer to InChIs generated, via this workflow (Figure 6.3), using default options, whilst stereochemically indifferent InChIs were generated using the SNon flag.

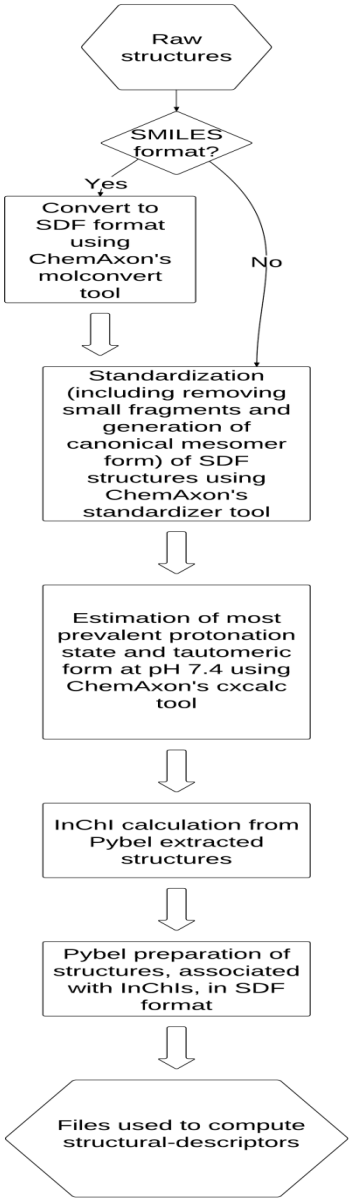


Figure 6.3 Overview of the procedures used to prepare the structures in the ion channel and TdP datasets from which InChIs and structural descriptors were computed.

6.4 Summary of TdP Model Comparisons

All scripts used to run these experiments, along with all raw input files, are made available with this thesis (Appendix A).

6.4.1 Modelling Runs Based on the Complete TdP Datasets

6.4.1.1 Descriptor Combinations Compared

Models were generated based on:

1. The structural-descriptors - i.e. the combination of the fragment-based bit-vector and LSER descriptors described in section 6.2.1.
2. Predicted IC-descriptors.
3. Combinations of both types of descriptor sets.

For all descriptor sets including predicted IC-descriptors, three versions were considered based on the following combinations of ion channel models:

1. All models (see Table 6.1).
2. The hypothetically most biologically relevant (see section 6.3.1) subset (KCNH2, KCNQ1 and CACNA1C).
3. KCNH2 only.

The KCNH2 only descriptor set was considered for the following reasons. Firstly, given the larger quantities of data available for building these models (Table 6.1), it was suspected the KCNH2 dataset might have greater coverage of the chemical space of the TdP datasets, hence yield higher quality predictions for these datasets and, consequently, more useful predicted IC-descriptors. Secondly, given the current focus on assessing hERG (KCNH2) inhibition as a surrogate for Torsadogenic potential in the pharmaceutical industry (see Chapter 1, section 1.3.2),[62,400] it was considered valuable to assess the extent to which the inclusion of hERG inhibitory information alone might improve predictive models for TdP.

6.4.1.2 Partitioning into Training and Test Sets

For both TdP datasets, models were generated and validated using (multiple repetitions of) 10-fold stratified cross-validation (10CV). Whilst 10 repetitions of 10CV were carried out on

the Yap-2004 dataset, 10CV was only carried out once on the larger Clark-2009 dataset for computational expediency. For each train/test partition, Random Forest models were generated five times, using five different random number generator (RNG) seeds.

6.4.2 Modelling Runs Based on Compounds with Experimental IC-descriptors

6.4.2.1Subsets of TdP Datasets Employed

Since a sizeable number of compounds with experimental IC-descriptors was only obtained for the KCNH2 channel (see Table 6.2), TdP models were built on the subsets of the TdP datasets for which KCNH2 experimental IC-descriptors were available. It was considered appropriate to generate these based on a subset of pIC_{50} estimates obtained under more experimentally consistent conditions than those originally obtained. Hence, KCNH2 experimental IC-descriptors were recalculated via only averaging estimates obtained from electrophysiological measurements in mammalian, heterologous expression systems.[67] Measurements for terfenadine which were deemed to be derived from a suboptimal IonWorks™ assay protocol,[78,401] or had been erroneously annotated in ChEMBL[402] were also excluded. The files containing these reduced sets of measurements are also made available with this thesis (Appendix A).

However, recalculating KCNH2 experimental IC-descriptors based upon excluding measurements, as described above, led to a reduction in the size of the TdP datasets. The subsets of the Yap-2004 and Clark-2009 datasets based on these recalculated descriptors contained 51 (7) and 36 (85) TdP+ (TdP-) compounds respectively. Hence, TdP models were also generated and validated using the datasets corresponding to the original KCNH2 experimental IC-descriptors (see Table 6.2).

6.4.2.2 Descriptor Combinations Compared

Models were generated based on:

1. The structural-descriptors (section 6.2.1).
2. KCNH2 experimental IC-descriptors.
3. Combinations of both types of descriptor sets.

6.4.2.3 Partitioning into Training and Test Sets

Due to their reduced size, 10 repetitions of stratified cross-validation were carried out using 10, 8 and 7 folds for the subsets of the Clark-2009, Yap-2004 (original experimental IC-descriptors) and Yap-2004 (recalculated experimental IC-descriptors) datasets respectively.

6.4.3 Statistical Comparisons

Overall TdP model performance was assessed in terms of the overall mean MCC. Differences in these mean MCC values obtained were determined to be statistically significant using the procedure described for the assessment of repeated cross-validation results in Chapter 5, section 5.3.5. Whilst the exact constitution of these descriptor sets varied, depending upon which ion channels were considered, all results obtained with either the 'IC-descriptors' or 'combination' descriptor sets, corresponding to either predicted or experimental IC-descriptors, were treated as a single family for the purposes of adjusting the p-values obtained for comparisons to the random predictor baseline. Pairwise comparisons were only carried out between the 'structural-descriptors', 'IC-descriptors' and 'combination' descriptor sets for a given dataset and a given combination of ion channels.[*] However, all corresponding sets of comparisons involving predicted IC-descriptors and experimental IC-descriptors were considered to be distinct families for the purpose of computing adjusted p-values.

6.5 Results and Discussion

The scripts used to compute performance statistics for both the ion channel and TdP models are made available with this thesis (Appendix A).

6.5.1 Results Obtained for Ion Channel Models

In order to estimate the expected performance of these models, for compounds inside their applicability domains, leave-one-out cross-validation (LOOCV) was employed.[†] For each

[*] This was because the primary focus of these investigations was to draw general conclusions, if possible, regarding the predictivity conferred by the use of IC-descriptors. Increasing the number of pairwise comparisons could have led to less powerful statistical tests of these pairwise comparisons upon generating adjusted p-values.[303]

[†] Save where this would have been computationally prohibitive. In practice, this lead to the performance of the KCNH2 model, based on a dataset of more than 1,000 compounds (see Table 6.1), being assessed using two repetitions of 10-fold cross-validation (10CV).

fold, m_{try} was re-optimised using the cross-validation training set, in order to avoid biasing the cross-validated performance estimates. All cross-validated performance estimations presented in these sections were obtained from pooling the results across all folds.

Finally, the performance of these selected ion channel models on those TdP dataset compounds for which experimental pIC_{50} values were obtained is presented.

Model performance is quantified in terms of the Root Mean Square Error (RMSE), the coefficient of determination (R^2), the Pearson (r) and Spearman's rank (ρ) correlation coefficients and their associated one-tail p-values (see Chapter 2, section 2.6.4.2.).

6.5.1.1 Cross-validated Performance

The cross-validated results obtained for the ion channel models are presented in Table 6.3. A number of the models, notably for the key ion channels KCNQ1 and CACNA1C identified in section 6.3.1, are clearly not as predictive as might be hoped.

Nonetheless, all correlation coefficient p-values obtained are statistically significant at the 5% level, suggesting that the positive correlation observed between the predicted and experimental pIC_{50} values was not due to chance.

Ion Channel	CV Repetition	RMSE	R^2	r	p-value(r)	ρ	p-value(ρ)
KCNH2	1	0.71	0.52	0.72	0.0E+00	0.69	1.3E-182
	2	0.70	0.53	0.73	0.0E+00	0.70	3.6E-190
CACNA1H	N/A	0.55	0.81	0.91	0.0E+00	0.78	1.7E-16
CACNA1C		0.73	0.25	0.50	2.0E-08	0.51	1.4E-08
KCNQ1		0.91	0.20	0.46	1.2E-04	0.49	3.7E-05
KCNA5		0.46	0.29	0.55	0.0E+00	0.55	1.5E-25
CACNA1G		0.42	0.69	0.83	0.0E+00	0.83	9.2E-51

Table 6.3 Cross-validated results obtained on the entirety of the ion channel datasets derived for the generation of predicted IC-descriptors.

6.5.1.2 Performance of Ion Channel Models for TdP Dataset Compounds

The predicted IC-descriptors were compared to the experimental IC-descriptors - i.e. predicted and (mean) experimental pIC_{50} values were compared - for the compounds in the TdP datasets for which experimental IC-descriptors were available. For the KCNH2 model,

this external validation was repeated after filtering the pIC_{50} measurements obtained for the TdP dataset compounds as described in section 6.4.2.1.

The results obtained for the former and latter validations are presented in Table 6.4 and Table 6.5 respectively. Supposing the statistically significant (at the 5% level) positive correlation between predicted and experimentally obtained pIC_{50} values obtained for the KCNH2 model held across the entirety of the TdP datasets, these results suggest these predictions are appropriate alternatives, for these datasets, to the inclusion of experimental pIC_{50} values as descriptors for predicting TdP.

TdP Dataset	Ion Channel	RMSE	R^2	r	p-value (r)	ρ	p-value (ρ)	n.o. Compounds
Yap-2004	KCNH2	0.90	0.50	0.76	7.1E-13	0.80	0.0E+00	60
	KCNQ1	1.47	-1.35	-0.16	7.1E-01	0.00	5.0E-01	15
Clark-2009	KCNH2	0.99	0.37	0.67	0.0E+00	0.69	4.6E-19	125
	KCNA5	2.21	-3.36	-0.84	8.1E-01	-1.00	1.0E+00	3
	KCNQ1	1.51	-0.80	-0.09	6.6E-01	-0.08	6.4E-01	21

Table 6.4 Performance of ion channel models on the TdP datasets; the performance of those models for which only a single TdP dataset compound with an experimental IC-descriptor was obtained is omitted.

TdP Dataset	Ion Channel	RMSE	R^2	r	p-value (r)	ρ	p-value (ρ)	n.o. Compounds
Yap-2004	KCNH2	0.87	0.52	0.78	1.7E-13	0.81	0.0E+00	58
Clark-2009	KCNH2	0.99	0.37	0.68	0.0E+00	0.70	2.7E-19	121

Table 6.5 Performance of KCNH2 model on TdP datasets, after removing problematic measurements and increasing the experimental consistency of the retained measurements.

6.5.2 Results Obtained for TdP Models

In addition to the performance summaries presented below, detailed results - including the outcomes of the statistical tests noted below - are made available with this thesis (Appendix A). Unless stated otherwise, all figures of merit presented are the mean across all five calls to randomForest( ), all cross-validation folds and, where applicable, all repetitions of cross-validation. The results obtained across all repetitions of cross-validation, all folds and all

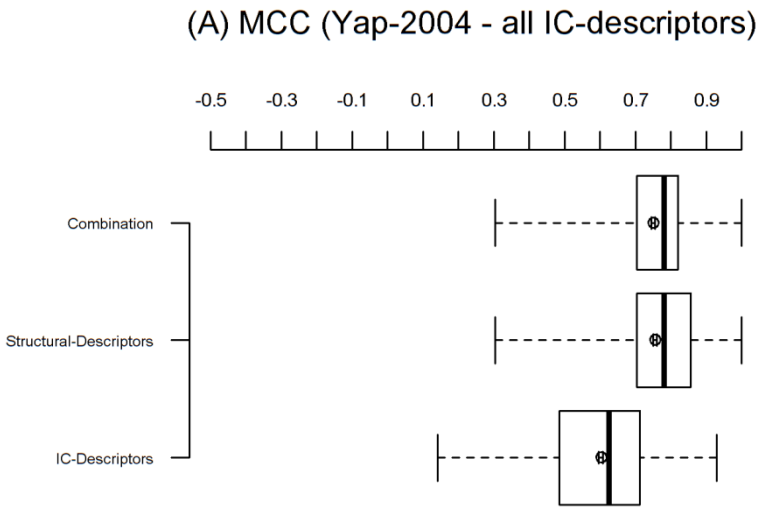
RNG seeds (i.e. calls to randomForest( )) are graphically summarised as per Chapter 5, section 5.4.

Unless noted otherwise, all following references to model performance refer to model performance in terms of the overall mean MCC.

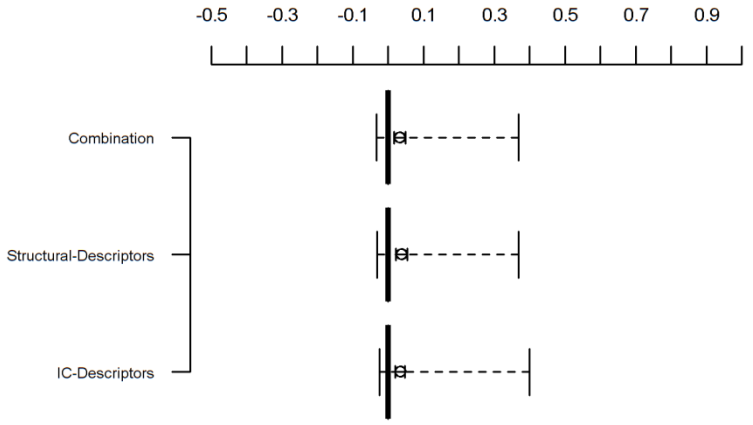
6.5.2.1 Effect of IC-descriptors on Performance

6.5.2.1.1 Effect of Predicted IC-descriptors on Performance

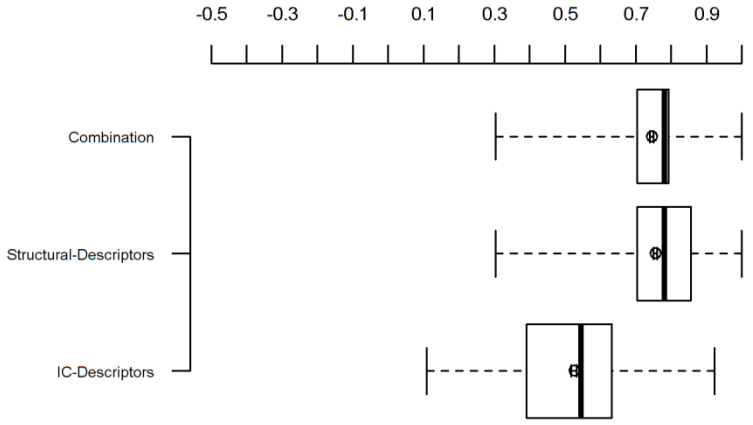
Figure 6.4 summarises the performance of all TdP models generated using predicted IC-descriptors, and the corresponding models generated using structural-descriptors alone.



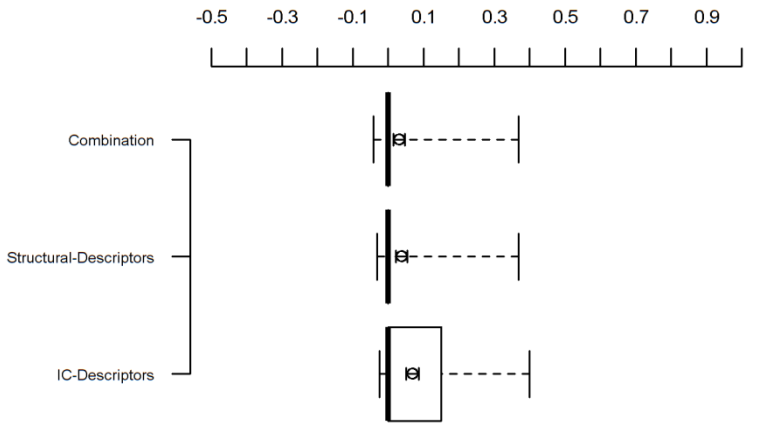
(B) MCC (Clark-2009 - all IC-descriptors)



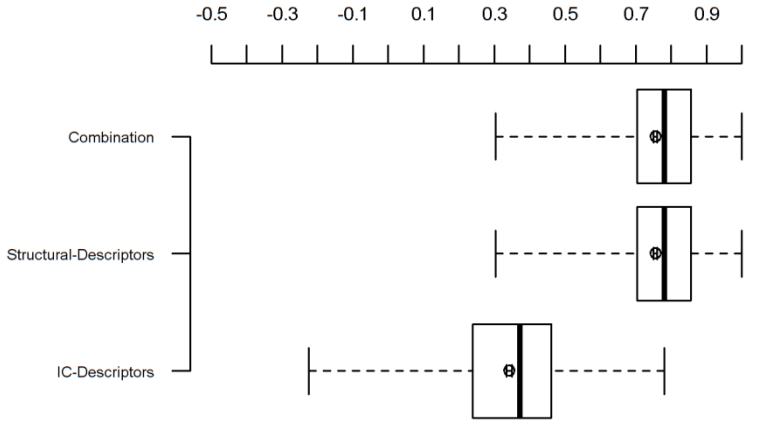
(C) MCC (Yap-2004 - KCNH2,KCNQ1,CACNA1C IC-descriptors)



(D) MCC (Clark-2009 - KCNH2,KCNQ1,CACNA1C IC-descriptors)



(E) MCC (Yap-2004 - KCNH2 IC-descriptors)



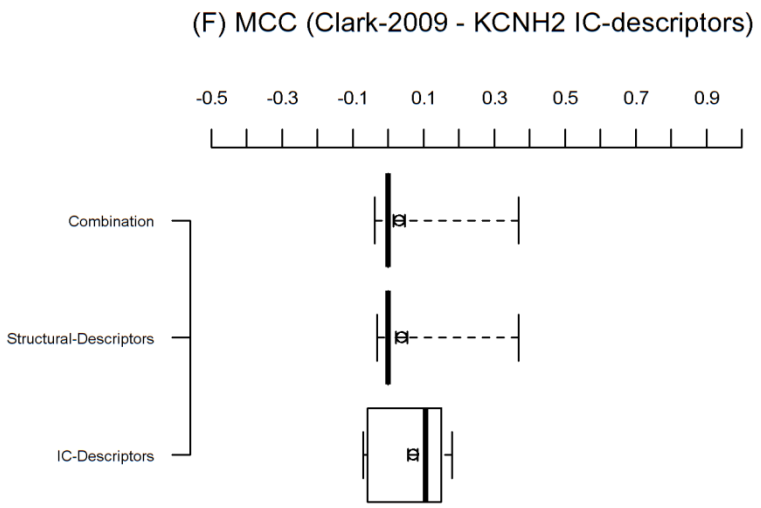


Figure 6.4 MCC values (black line: median, circle: mean) obtained using predicted IC-descriptors, and the corresponding results obtained using structural-descriptors alone, across all cross-validation folds, repetitions and RNG seeds. Results obtained on the Yap-2004 (A,C,E) and Clark-2009 (B,D,F) datasets, using predicted IC-descriptors generated from all models (A,B), the putative most relevant set - i.e. KCNH2, KNCQ1, CACNA1C (C,D) and just the KCNH2 model (E,F).

The results obtained on the Yap-2004 dataset indicate that the inclusion of predicted IC-descriptors has no appreciable effect on the performance of models generated using structural-descriptors alone, whilst the performance of models generated using predicted IC-descriptors alone is appreciably worse. Irrespective of the combination of ion channel models, the mean MCC values obtained using IC-descriptors alone were lower, and statistically significantly different (Appendix A), to those obtained with the other descriptor sets; no statistically significant differences between the mean MCC values obtained with the other two descriptor sets were observed.

Whilst yielding worse performance than the other descriptor sets, the overall mean MCC obtained with, any combination of, predicted IC-descriptors alone was nonetheless, in keeping with both other descriptor sets, statistically significantly higher than the value expected for a random predictor (zero) for the Yap-2004 dataset.

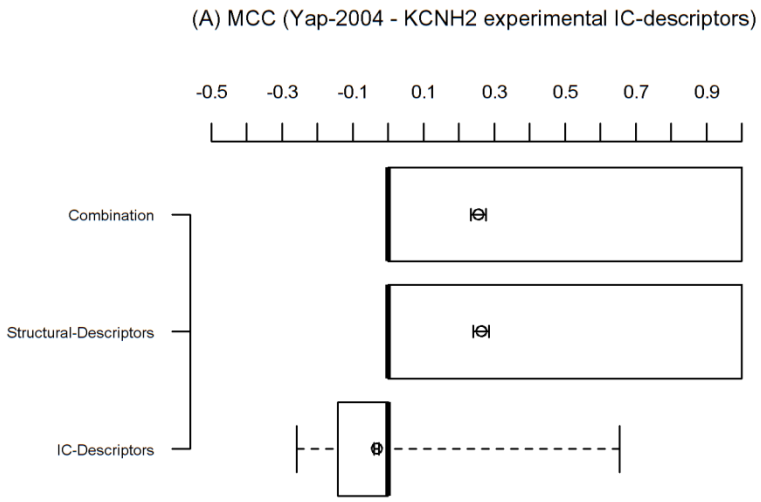
Interestingly, the overall mean MCC obtained using only these descriptors consistently fell, for this dataset, upon reducing the number of ion channel models used to generate them. Given the supposedly greater biological relevance of inhibiting CACNA1C, KCNQ1 and KCNH2 (see section 6.3.1) - as opposed to the superset of all ion channels - and the indicated

poorer predictivity of the CACNA1C and KCNQ1 models, with respect to the KCNH2 model, this is somewhat surprising.

The low quality of all models generated on the Clark-2009 dataset, which may well reflect the considerable imbalance in classes (see section 6.3.3.4),* offers little basis for discerning the relative predictivity expected with the different descriptor sets. No models performed statistically significantly better than random. No pairwise differences in overall mean MCC were statistically significant.

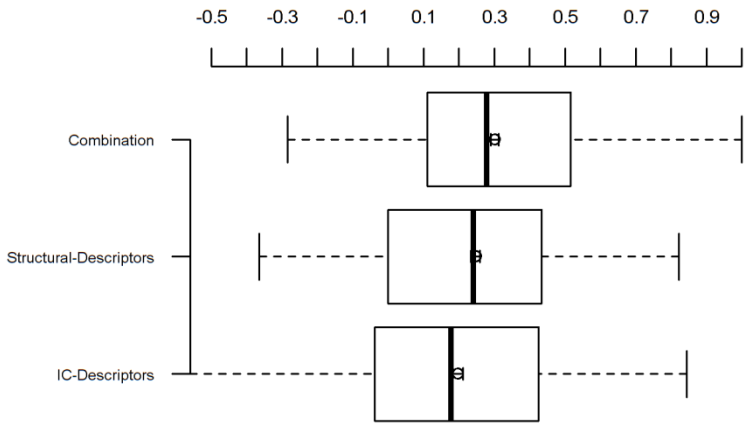
6.5.2.1.2 Effect of Experimental IC-descriptors on Performance

Figure 6.5 summarises the performance of all models generated using experimental KCNH2 IC-descriptors, and the corresponding models based on structural-descriptors alone.

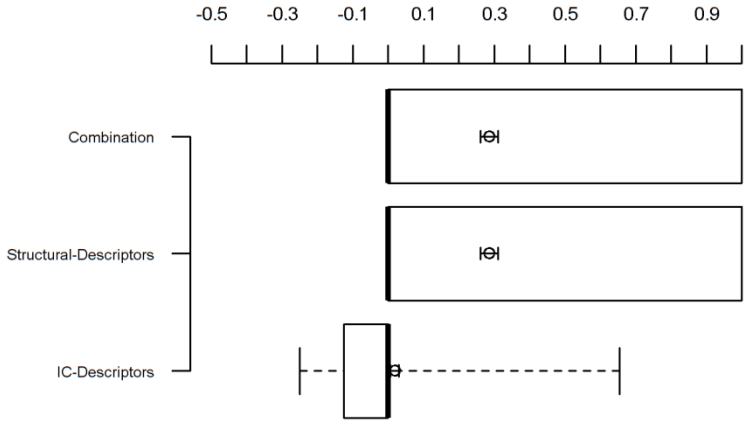


* This imbalance appearing to bias the models towards simply becoming majority class predictors, as suggested by the observation that, for all descriptor sets, the overall mean recall was less than 0.15 and greater than 0.94 for the TdP+ and TdP- (majority class) compounds respectively.

(B) MCC (Clark-2009 - KCNH2 experimental IC-descriptors)



(C) MCC (Yap-2004 - filtered KCNH2 experimental IC-descriptors)



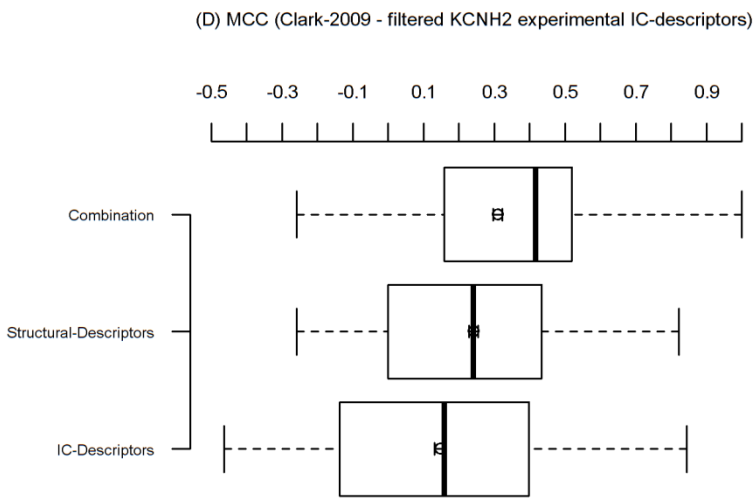


Figure 6.5 MCC values (black line: median, circle: mean) obtained using experimental KCNH2 IC-descriptors, and the corresponding results obtained using structural-descriptors alone, across all cross-validation folds, repetitions and RNG seeds. Results obtained on subsets of the Yap-2004 (A,C) and Clark-2009 (B,D) datasets corresponding to all compounds for which KCNH2 experimental IC-descriptors were obtained (A,B) and subsequent to reducing the inconsistency in the experimental conditions used to obtain the underlying measurements (C,D).

No statistically significant differences (Appendix A) in overall mean MCC values were observed between the three sets of descriptors for any of these sets of compounds. However, this could well reflect, at least in part, the small size of these datasets (see section 6.4.2.1). The reduced size of the corresponding validation set sizes would be expected to lead to increased instability of the results obtained on a given validation set and the smaller size of the training sets would be expected to lead to an increase in the 'signal to noise ratio' - making it harder for any modelling approach to learn to discriminate between TdP+ and TdP- compounds. This latter contention is supported by the observation that, for the subsets of the Clark-2009 dataset, only the use of the ‘combination’ descriptor set led to statistically significant improvements in performance with respect to a random predictor and that this was true for none of the modelling approaches assessed on subsets of the Yap-2004 dataset.

The considerable class imbalance (see section 6.4.2.1) in the subsets of the Yap-2004 dataset employed here appears to have biased the corresponding models towards becoming majority class predictors.*

6.5.2.2 Contributions of Different Descriptors to the Models

The contributions of different descriptors to the Random Forest TdP models were assessed using the Gini importance measure.[223] However, given the particular bias of this measure towards continuous descriptors,[227] the importance of the binary substructural descriptors was expected to be artificially lowered with respect to the LSER descriptors and IC-descriptors. Nonetheless, the relative importance of the IC-descriptors with respect to each other and the LSER descriptors could yield valuable insights into the contributions of these novel descriptors to the predictions made by the TdP models. The focus below is on the contributions made by IC-descriptors; hence, all the discussion below refers to those models generated using one of the considered combinations of IC-descriptors on its own, or in combination with structural-descriptors.

The final measure of descriptor importance employed here was the median (across all RNG seeds, and cross-validation partitions for a given dataset) Gini importance based rank,† lower values indicating more important features. All calculated importance values are presented in CSV files made available with this thesis (Appendix A).

6.5.2.2.1 Contributions of Predicted IC-descriptors

Across every single model (i.e. for every cross-validation split, every RNG seed , every predicted IC-descriptor combination and both datasets), the KCNH2 IC-descriptor was the most important descriptor. For all other predicted IC-descriptors, the Gini measure based ranks were less stable across different models.

This could be taken as affirmation of the particular biological significance of KCNH2 inhibition for Torsadogenic potential. However, it is clearly necessary to also take into

---

* As is presumably reflected in the median recall for the TdP+ (majority) class and TdP- class, in all cases, being one and zero respectively (see Appendix A).

† The ranks were generated for *all* descriptors, including the binary descriptors corresponding to molecular fragments.

account the predictivity of the ion channel models for the compounds in the TdP datasets. The higher predicitivity of the KCNH2 model (see section 6.5.1) compared to the models for the other two ion channels widely understood to be important for the induction of TdP (i.e. CACNA1C and KCNQ1, see section 6.3.1) may also have skewed the results. Also, the non-linearity of Random Forest means that the importance of KCNH2 inhibition suggested here should *not* be interpreted as indicating that KCNH2 inhibition can be considered in isolation when evaluating Torsadogenic potential.

Whilst the observation that the CACNA1C descriptor was assigned the second most important median rank for all relevant sets of models built on the Yap-2004 dataset might be supposed to affirm the significance of inhibiting this channel for the induction of TdP (see section 6.3.1), its median rank for the Clark-2009 dataset was not greater than the ranks for the KCNA5 (supposedly mechanistically irrelevant) and KCNQ1 IC-descriptors in both sets of models including all predicted IC-descriptors. These latter results, however, could simply be an artefact of the difficulty in generating discriminative models for the Clark-2009 dataset (see section 6.5.2).

6.5.2.2.2 Contributions of Experimental IC-descriptors

For the models built on the subset of the Clark-2009 dataset, the KCNH2 IC-descriptor was once more consistently assigned the most important rank. That this was not the case for the models built on the subset of the Yap-2004 dataset seems likely to be an artefact of the difficulty in generating discriminative models for this dataset (see section 6.5.2).

6.5.3 The Value Added by IC-Descriptors

The results obtained offer no clear evidence that IC-descriptors, as an alternative or complement to structural-descriptors, improve the performance of models for Torsadogenic potential. That no clear evidence of improved performance was observed with both predicted and experimental IC-descriptors suggests that this finding is not simply an artefact of the poor quality of some of the ion channel models used to generate predicted IC-descriptors, nor the possibility that fractions of the TdP datasets might lie outside the applicability domain of these models.

To put this in context, it should be noted that attempts, reported in the recent literature, to build models, for *in vivo* endpoints, based on experimentally derived biological descriptors have usually, though not always, suggested that the combination of biological and traditional

(i.e. purely structural) descriptors leads to improved bioactivity predictions over those obtained using structural-descriptors alone.[187–189]

However, it has been indicated that the manner in which biological measurements are turned into descriptors may determine whether or not an improvement in predictivity is observed upon their incorporation into the modelling procedure.[188] This point is explored further, with respect to designing new types of IC-descriptors, in Chapter 7.

Specifically regarding the development of predicted biological descriptors, this author is unaware of any other attempts to generate models for an *in vivo*, or clinical, endpoint based on the paradigm proposed here (i.e. the generation of predicted bioactivities based on the same structural-descriptors directly used for modelling the ultimate endpoint of interest). However, work by Cunningham *et al.* indicated that carcinogenicity predictions based upon fragment descriptors and docking predicted protein targets were improved with respect to models built on either type of information in isolation.[191] Perhaps of greater relevance for the paradigm proposed in this work, Zhu *et al.* found that *in vivo* toxicity predictions based on purely structural-descriptors were improved when predictions were separately generated for subclasses of the data predicted, using the same descriptors, to show different associations between *in vitro* and *in vivo* biological information.[190]

Even if they do not improve predictivity, IC-descriptors could be valuable by yielding more mechanistically interpretable models for TdP. For example, suppose a model were to predict a compound to be Torsadogenic. It would be valuable to be able to rationalise the prediction in mechanistic terms; for example, 'compound X is anticipated to induce TdP, because it strongly inhibits the KCNH2 channel and only weakly inhibits the CACNA1C channel'. This could suggest strategies to a medicinal chemist for reducing the anticipated Torsadogenic liability; for example, might the KCNH2/CACNA1C inhibition of 'compound X' be reduced/ increased, given the synthetic accessibility of derivatives etc.?

Such an analysis was, unfortunately, not possible with the implementation of Random Forest employed in this work. However, Kuz'min *et al.* recently presented a means of obtaining the signed contributions of descriptors towards the predictions made by Random Forest models for individual compounds. For example, in the current context, 'this descriptor contributes positively towards the predicted Torsadogenic potential of compound X'.[229]

6.6 Conclusions

The key objective of the work presented in this chapter was to assess the viability of discriminating between TdP inducing and non-inducing compounds using predicted pIC_{50} values for (potentially) biologically relevant ion channels as well as the potential for enhancing conventional QSAR approaches (based on structural-descriptors) to this task by including predicted/experimental pIC_{50} values as additional descriptors (IC-descriptors).

It was originally hypothesised that the addition of such novel biological descriptors would, on average, improve performance over models generated using structural-descriptors alone. However, no clear evidence to support this hypothesis was obtained.

Nonetheless, the potential value of (predicted) IC-descriptors may lie in their ability to yield mechanistically interpretable predictions of Torsadogenic potential.

Further work is required to realise the potential mechanistic interpretability of models generated using IC-descriptors; this could entail using the approach proposed by Kuz'min *et al.* to interpret the contributions of different descriptors to the predictions made by Random Forest models for specific compounds.[229] Possible improvements to the manner in which IC-descriptors are generated are discussed in Chapter 7.

Chapter 7 Conclusions and Future Work

The aim of the work presented in this thesis was the development of novel models/modelling approaches which could be used to identify, *in silico*, potential drug compounds inducing important toxicity endpoints. Chapter 1 explained why *in silico* predictive toxicology is of immense value in pharmaceutical research. This chapter proceeded to explain the importance of anticipating, and overviewed current scientific understanding of, the key toxicity endpoints which were the focus of the work presented in this thesis: mutagenicity, carcinogenicity, hERG inhibition and Torsades de Pointes (TdP). Chapter 2 presented an overview of the approaches available to predict toxicity *in silico*, with particular reference to those methods (notably Quantitative Structure-Activity Relationships, or QSARs) upon which the novel work presented in this thesis was founded.

7.1 Conclusions

In Chapter 3, a consensus model for mutagenicity/carcinogenicity was developed by combining the output generated by two commonly employed predictive toxicology programs: Toxtree[160] and Derek for Windows™.[271] Various options for combining the output of these programs or using one of these programs in isolation, to generate predictions for both endpoints were evaluated - notably, on a large, QSAR ready database (ISSCAN).[51] In terms of the Matthews Correlation Coefficient (MCC), the top performing models - for both endpoints - were combined models. Moreover, combining the outputs from both programs allowed the model to be tuned to minimise false predictions of toxicity, as was desirable given the model's intended application as a screening tool in an early stage discovery project.[195] The selected model was used to remove compounds predicted to exhibit both endpoints during a virtual screening workflow which successfully identified possible starting points for structurally novel anti-tuberculosis drugs. Subsequent analysis of the wealth of toxicity predictions generated for a range of endpoints suggests the bioactive compounds identified would not necessarily be lost to attrition for toxicity reasons and pointed to some potential toxicities that should be experimentally tested.

In Chapter 4, a novel modelling approach based on (an extension of) Nigsch's version of the Winnow algorithm[184,213] was presented - incorporating for the first time, to the best of this author's knowledge, information encoded by numeric descriptors into QSAR models generated using this technique and investigating the use of multiple training cycles. The additional features based upon numeric descriptors were found to make important

contributions to the models, whilst the use of multiple training cycles was found to potentially yield improved results.

This approach was used to develop models for the rapid identification of potent hERG inhibitors, the development of which would usually be abandoned in the pharmaceutical industry.[62,69] These binary classification models were rigorously externally validated using carefully selected datasets and directly compared to approaches previously proposed for this task by Thai and Ecker[69] and Dubus *et al.*[91].

The Winnow based models were found to perform competitively, or better, to those previously presented in the literature, and the Winnow algorithm was found (to this author's knowledge, for the first time in QSAR research) to perform comparably or better to models generated from the same descriptors using Random Forest and SVM - the memory efficient manner in which Winnow learns from the training set (as explained in Chapter 2, section 2.6.3.2) making this a notable finding. This work also highlighted the considerable variability in the estimated performance of a given QSAR approach which may be observed when training and testing the resultant model on different data.

Chapter 5 introduced a novel approach to combining Ballester and Richard's Ultrafast Shape Recognition (USR) descriptors[263] with chemical information: Atom Type USR (ATUSR). The performance of this 3D descriptor set for the generation of both classification and regression QSAR models for protein-ligand binding, primarily for hERG inhibition, was benchmarked on various datasets and compared to (extensions of) USR (analogous to those previously presented by Cannon *et al.*),[175] as well as commonly used 2D descriptor sets. USR was almost always found to yield the worst performance - yet statistically significantly better performance than expected for a random predictor - for both classification and regression modelling.

More generally, the ordering of descriptor performance was different across different datasets. In only some cases were ATUSR descriptors (weakly) suggested to offer improved performance over a conceptually analogous approach (USR-ATFP) to that proposed by Cannon *et al.*[175] for extending USR.

Moreover, the use of various Molecular Mechanics and docking approaches to yield (ostensibly) improved approximations (compared to those generated by CORINA)[333] of the bioactive conformer, employed for descriptor calculations, did not lead to the expected increases in predictive performance for ATUSR. However, this could partly reflect the

particular difficulties associated with employing docking to find improved representations of the bioactive conformer for the (hERG) datasets considered. Moreover, some limited evidence was obtained that the actives and inactives in some of the classification datasets were partially separable based on lipophilicity, which may have biased the evaluations in favour of 2D descriptors.

In Chapter 6, a novel approach to directly discriminating between compounds with TdP causing potential, and non-TdP causing compounds, was introduced. The proposed approach incorporated predicted (or experimental) pIC_{50} values for cardiac ion channels (including hERG) previously indicated to play a role in the induction of TdP as additional descriptors (IC-descriptors). This approach was evaluated on datasets (subsets of those previously used for QSAR modelling of TdP) derived from assessments for TdP causing potential in humans.

In contrast to many of the earlier studies which employed biological measurements as additional descriptors for toxicity endpoints,[187–189] the results obtained did not (clearly) indicate that the addition of predicted (experimental) IC-descriptors led to improved performance with respect to the use of traditional (structural) descriptors in isolation. However, the real value of such 'hybrid' models for TdP may lie in improved mechanistic interpretability of the predictions.

7.2 Future Work

Novel Models for TdP Based on Biological Descriptors

Given the novelty of the approach and the potential for improved interpretability of predictions for medicinal chemists, additional work is warranted to assess whether alternative approaches to incorporating IC-descriptors into models for TdP would be more successful (in terms of predictivity) than those described in Chapter 6.

Improved versions of IC-descriptors might correspond to binary descriptors denoting levels of inhibition above or below appropriate thresholds. These thresholds might best be selected via internal validation of the resultant TdP models. The use of such descriptors, particularly if compounds close to the 'active vs. inactive' boundary were excluded, could reduce the significance of experimental noise, and more effectively enable the combination of data from different assays, leading to improved predictivity of the ion channel models as well as more robust experimental IC-descriptors.[90,93] Were this to facilitate the inclusion of data from a wider variety of assays, and more compounds, this could both extend the applicability

domain of the ion channel models, which may not currently cover the entirety of the TdP datasets, and yield more TdP dataset compounds with experimental IC-descriptors; this could improve the value of the predicted IC-descriptors when applied to the entirety of the TdP datasets and strengthen the basis for drawing conclusions regarding the value of the experimental IC-descriptors.

Sedykh *et al.* recently found that experimental biological descriptors derived from multiple points on a dose response curve yielded better toxicity predictions than simple binary encoding of the corresponding assay data.[188] Hence, it would be appropriate to see whether IC-descriptors based upon multiple dose-response parameters and other parameters which can be derived from electrophysiological assays, would yield improved IC-descriptors. Such detailed information, from the same electrophysiological hERG assay, has recently been made available for more than 300,000 compounds.[97]

Moreover, as noted in Chapter 6, the use of the approach proposed by Kuz'min *et al.* for the interpretation of Random Forest models,[229] in order to realise the potential mechanistic interpretability of models generated with these kinds of descriptors, would also be warranted.

Unbiased Assessment of ATUSR Descriptors

As briefly discussed above, as well as in Chapter 5, dataset biases may result in the observed performance of 2D descriptors exceeding that of 3D descriptors, even if the 3D descriptors capture useful information that the 2D descriptors do not. Examples of dataset biases which may have this effect are (some forms of) “analogue bias” and “property bias”.[336] In general, analogue bias refers to the actives in a dataset being disproportionately comprised of ('trivially') structurally related analogues. Where these analogues are similar in terms of (properties which are directly related to) their 2D structural characteristics, artificially improved performance of 2D descriptors may be observed.[330] Property bias refers to separation of the actives and inactives in terms of simple molecular properties,[336] such as logP,[339,403] which may be highly related to 2D descriptors.[337] There is a particular need in the pharmaceutical industry for methods which can identify compounds with, say, lower hERG inhibition, yet similar lipophilicity.[338]

Recently, Rohrer and Baumann developed a framework for quantifying both kinds of dataset bias, and for generating "unbiased" versions of an original dataset.[339,403] They applied this framework to generate the Maximum Unbiased Validation (MUV) binary classification

datasets based on bioactivity data for 30 actives and 15,000 inactives for 17 protein target interaction based bioactivities.

It would be particularly interesting to evaluate the relative performance of ATUSR descriptors, using a similar set of experiments to those presented in Chapter 5, on these unbiased, benchmark datasets. Moreover, an initial inspection of the PDB[261] suggests that relevant (antagonist/agonist bound) crystal structures are available for the protein targets against which inhibition/activation is measured in the relevant PubChem assays.[169] For example, the 3NMQ and 1GWR complexes correspond to antagonist and agonist bound complexes of the HSP90 and ER-alpha targets respectively. This would allow further investigation of how structure-based approaches to approximating the bioactive conformer might improve the relative performance of ATUSR descriptors for 3D QSAR - Celik *et al.* having previously successfully docked small molecules into the 1GWR structure.[404] Indeed, given that ER-alpha agonists may act as endocrine disrupting chemicals (EDCs), inducing EDC toxicity,[405] investigating the performance of ATUSR descriptors for modelling activation of the ER-alpha receptor could serve as a test of their usefulness in computational toxicology.

Moreover, the MUV protocol[339] could be applied to the (hERG) datasets previously modelled in this thesis, and the change (if any) in the relative performance of ATUSR descriptors investigated.

As implementations of the descriptor sets presented in Chapter 5 are made available to the community with this thesis (Appendix A), this should facilitate further studies of their effectiveness. Additionally, careful evaluations of their potential computational efficiency would be merited if further investigations demonstrated their usefulness - which might entail re-implementing the ATUSR descriptor set.

Improved Screening for Potent hERG Inhibitors

The results presented in Chapter 4 suggest that Winnow models generated using features obtained from ECFP_4 fingerprints and discretized simple molecular properties, such as logP, pK_a values and the Wiener Index, could perform well compared to models previously presented in the literature for discriminating potent ($IC_{50} < 1$ μM) hERG inhibitors from moderate ($IC_{50:}$ 1-10 μM) and weak ($IC_{50} \geq 10$ μM) inhibitors. Additional work is warranted to:

1. Build/evaluate such models on a larger set of compounds. This data might be obtained from the ChEMBL database;[269] to obtain appreciably larger quantities of data though, it might be necessary to compromise (see Chapter 6) the rigorous specifications regarding experimental conditions employed when compiling the Literature-368 and hERG-196 datasets discussed in Chapter 4 and Chapter 5 respectively. Alternatively, experimentally consistent hERG inhibition measurements for more than 300,000 compounds are currently available via HERGCentral.[97,105]
2. More rigorously define the applicability domain of these models. As extensively discussed in Chapter 2, section 2.6.6, clearly delimiting the applicability domain of QSAR models is of considerable importance, yet remains a highly active area of research. It is possible that an approach based upon missing features, or the variation in class scores across different training set orders might serve as useful "distance to model" metrics that could be used to define an applicability domain as per Sushko *et al.*[47]
3. Implement the models as open source, freely available tools that would be readily available to non-experts. This could, in part, entail using the freely available implementation of extended connectivity fingerprints in the software program jCompoundMapper.[406]

In summary, the work presented in this thesis has investigated a variety of novel approaches to computationally predicting drug induced toxicity and offers starting points for avenues of future research as well as readily available predictive tools.

Bibliography

(1) Combs, A. B.; Acosta, D. Jr. In *Computational Toxicology: Risk Assessment for Pharmaceutical and Environmental Chemicals*; Ekins, S., Ed.; Wiley Series on Technologies for the Pharmaceutical Industry; John Wiley and Sons: Hoboken, New Jersey, 2007; pp. 3–20.

(2) Nigsch, F. Computational Prediction of Molecular Properties for Drug Discovery. PhD Thesis, University of Cambridge, United Kingdom, 2008.

(3) Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. *Nat. Rev. Drug Discovery* **2006**, *5*, 993–996.

(4) Toxicity Testing Overview. http://alttox.org/ttrc/tox-test-overview/ (accessed January 11 2012).

(5) Xu, J. J. In *Computational Toxicology: Risk Assessment for Pharmaceutical and Environmental Chemicals*; Ekins, S., Ed.; Wiley Series on Technologies for the Pharmaceutical Industry; John Wiley and Sons: Hoboken, New Jersey, 2007*;* pp. 21–32.

(6) Müller, L.; Breidenbach, A.; Funk, C.; Muster, W.; Pähler, A. In *Computational Toxicology: Risk Assessment for Pharmaceutical and Environmental Chemicals*; Ekins, S., Ed.; Wiley Series on Technologies for the Pharmaceutical Industry; John Wiley and Sons: Hoboken, New Jersey, 2007; pp. 545–580.

(7) Kramer, J. A.; Sagartz, J. E.; Morris, D. L. *Nat. Rev. Drug Discovery* **2007**, *6*, 636–649.

(8) Kola, I.; Landis, J. *Nat. Rev. Drug Discovery* **2004**, *3*, 711–715.

(9) U S Food and Drug Administration Home Page. http://www.fda.gov/ (accessed January 11 2012).

(10) Qureshi, Z. P.; Seoane-Vazquez, E.; Rodriguez-Monguio, R.; Stevenson, K. B.; Szeinbach, S. L. *Pharmacoepidemiol. Drug Saf.* **2011**, *20*, 772–777.

(11) Lazarou, J.; Pomeranz, B. H.; Corey, P. N. *JAMA, J. Am. Med. Assoc.***1998**, *279*, 1200–1205.

(12) Vitola, J.; Vukanovic, J.; Roden, D. M. *J. Cardiovasc. Electrophysiol.* **1998**, *9*, 1109–1113.

(13) Wysowski, D. K.; Bacsanyi, J. *N. Engl. J. Med.* **1996**, *335*, 290–291.

(14) Nigsch, F.; Lounkine, E.; McCarren, P.; Cornett, B.; Glick, M.; Azzaoui, K.; Urban, L.; Marc, P.; Müller, A.; Hahne, F.; Heard, D. J.; Jenkins, J. L. *Expert Opin. Drug Metab. Toxicol.* **2011**, *7*, 1497–1511.

(15) Gavaghan, C. L.; Arnby, C. H.; Blomberg, N.; Strandlund, G.; Boyer, S. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 189–206.

(16) Johnson, D.E.; Rodgers, A.D. *Curr. Opin. Drug Discovery Dev.* **2006**, *9*, 29–37.

(17) Olson, H. M.; Davies, T. S. In *Predictive Toxicology in Drug Safety*; Xu, J. J.; Laszlo, U., Eds.; Cambridge University Press: New York, 2011; pp. 1–17.

(18) Bleicher, K. H.; Böhm, H.-J.; Müller, K.; Alanine, A. I. *Nat. Rev. Drug Discovery* **2003**, *2*, 369–378.

(19) Mitchell, J. B. O. *Future Med. Chem.* **2011**, *3*, 451–467.

(20) Gleeson, M. P.; Modi, S.; Bender, A.; Marchese Robinson, R. L.; Kirchmair, J.; Promkatkaew, M.; Hannongbua, S.; Glen, R. C. *Curr. Pharm. Des.* **2012**, *18*, 1266–1291.

(21) Davenport, A. J.; Möller, C.; Heifetz, A.; Mazanetz, M. P.; Law, R. J.; Ebneth, A.; Gemkow, M. J. *Assay Drug Dev. Technol.* **2010**, *8*, 781–789.

(22) Valerio, L. G. Jr. *Toxicol. Appl. Pharmacol.* **2009**, *241*, 356–370.

(23) Egan, W. J.; Zlokarnik, G.; Grootenhuis, P. D. J. *Drug Discovery Today: Technol.* **2004**, *1*, 381–387.

(24) Valerio, L. G. Jr. *Hum. Exp. Toxicol.* **2008**, *27*, 757–760.

(25) Worth, A.; Lapenna, S.; Piparo, E. L.; Mostrag-Szlichtyng, A.; Serafimova, R. *A Framework for Assessing In Silico Toxicity Predictions: Case Studies with Selected Pesticides*; JRC Scientific and Technical Reports; European Commission, Joint Research Centre, 2011.

(26) Johnson, D. E.; Rodgers, A. D.; Sudarsanam, S. In *Computational Toxicology: Risk Assessment for Pharmaceutical and Environmental Chemicals*; Ekins, S., Ed.; Wiley Series on Technologies for the Pharmaceutical Industry; John Wiley and Sons: Hoboken, New Jersey, 2007; pp. 725–749.

(27) Yap, C. W.; Cai, C. Z.; Xue, Y.; Chen, Y. Z. *Toxicol. Sci.* **2004**, *79*, 170–177.

(28) Clark, M.; Wiseman, J. S. *J. Chem. Inf. Model.* **2009**, *49*, 2617–2626.

(29) Frid, A. A.; Matthews, E. J. *Regul. Toxicol. Pharmacol.* **2010**, *56*, 276–289.

(30) Bender, A.; Scheiber, J.; Glick, M.; Davies, J. W.; Azzaoui, K.; Hamon, J.; Urban, L.; Whitebread, S.; Jenkins, J. L. *ChemMedChem* **2007**, *2*, 861–873.

(31) Pharmatrope Press Release. http://pharmatrope.com/?q=node/9 (accessed January 12 2012).

(32) Nigsch, F.; Mitchell, J. B. O. *Toxicol. Appl. Pharmacol.* **2008**, *231*, 225–234.

(33) Matthews, E. J.; Frid, A. A. *Regul. Toxicol. Pharmacol.* **2010**, *56*, 247–275.

(34) Foster, J.R. In *Carcinogenesis*; Comprehensive Toxicology, Vol. 14, 2nd ed.; Roberts, R., Ed.; Elsevier Limited.: Kidlington, United Kingdom, 2010; pp. 1–10.

(35) Boyle, P.; Ferlay, J. *Ann. Oncol.* **2005**, *16*, 481–488.

(36) Jansen, J. D. *Mutat. Res.* **1988**, *205*, 3–12.

(37) Benigni, R.; Bossa, C.; Jeliazkova, N.; Netzeva, T.; Worth, A. *The Benigni / Bossa Rulebase for Mutagenicity and Carcinogenicity - A Module of Toxtree*; JRC Scientific and Technical Reports; European Commission, Joint Research Centre, 2008.

(38) Benigni, R.; Bossa, C. *Chem. Rev.* **2011**, *111*, 2507–2536.

(39) *Guidance on Genotoxicity Testing and Data Interpretation for Pharmaceuticals Intended for Human Use, S2(R1), Step 4 Version*; ICH Harmonised Tripartite Guideline; International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use, 2011.

(40) Zeiger, E. *Mutat. Res., Genet. Toxicol. Environ. Mutagen.* **2001**, *492*, 29–38.

(41) Shelby, M. D.; Bishop, J. B.; Mason, J. M.; Tindall, K. R. *Environ. Health Perspect.* **1993**, *100*, 283–291.

(42) Custer, L. L.; Kreatsoulas, C.; Durham, S. K. In *Computational Toxicology: Risk Assessment for Pharmaceutical and Environmental Chemicals*; Ekins, S., Ed.; Wiley Series on

Technologies for the Pharmaceutical Industry; John Wiley and Sons: Hoboken, New Jersey, 2007; pp. 391–401.

(43) *S1B Testing for Carcinogenicity of Pharmaceuticals*; Guidance for Industry; International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use, 1997.

(44) Jacobs, A. *Toxicol. Sci.* **2005**, *88*, 18–23.

(45) Cohen, S. M. *Toxicol. Sci.* **2004**, *80*, 225–229.

(46) Ames, B. N. *Cancer* **1984**, *53*, 2034–2040.

(47) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Cherkasov, A.; Li, J.; Gramatica, P.; Hansen, K.; Schroeter, T.; Müller, K.-R.; Xi, L.; Liu, H.; Yao, X.; Öberg, T.; Hormozdiari, F.; Dao, P.; Sahinalp, C.; Todeschini, R.; Polishchuk, P.; Artemenko, A.; Kuz'min, V.; Martin, T. M.; Young, D. M.; Fourches, D.; Muratov, E.; Tropsha, A.; Baskin, I.; Horvath, D.; Marcou, G.; Muller, C.; Varnek, A.; Prokopenko, V. V.; Tetko, I. V. *J. Chem. Inf. Model*. **2010**, *50*, 2094–2111.

(48) Kamber, M.; Flückiger-Isler, S.; Engelhardt, G.; Jaeckh, R.; Zeiger, E. *Mutagenesis* **2009**, *24*, 359–366.

(49) Zeiger, E. *Cancer Res.* **1987**, *47*, 1287–1296.

(50) Cohen S.M. *Exp. Toxicol. Pathol.* **2010**, *62*, 497–502.

(51) ISSCAN: Istituto Superiore di Sanita, Chemical Carcinogens: Structures and Experimental Data. http://www.epa.gov/ncct/dsstox/sdf_isscan_external.html (accessed January 16 2012).

(52) Backus, G. S.; Wolf, M. A.; Burch, J.; Richard, A. M. DSSTox EPA Integrated Risk Information System (IRIS) Toxicity Review Data: SDF File and Documentation.

http://www.epa.gov/ncct/dsstox/sdf_iristr.html (accessed January 16 2012).

(53) *Guidelines for Carcinogen Risk Assessment*; FRL-2984-1; Risk Assessment Forum, US Environmental Protection Agency: Washington, DC, 1986.

(54) *Guidelines for Carcinogen Risk Assessment*; EPA/630/P-03/001F; Risk Assessment Forum, US Environmental Protection Agency: Washington, DC, 2005.

(55) Hansen, K.; Mika, S.; Schroeter, T.; Sutter, A.; ter Laak, A.; Steger-Hartmann, T.; Heinrich, N.; Müller, K.-R. *J. Chem. Inf. Model.* **2009**, *49*, 2077–2081.

(56) Leadscope Toxicity Databases. http://www.leadscope.com/toxicity_databases/ (accessed January 16 2012).

(57) Hillebrecht, A.; Muster, W.; Brigo, A.; Kansy, M.; Weiser, T.; Singer, T. *Chem. Res. Toxicol.* **2011**, *24*, 843–854.

(58) Sanguinetti, M. C.; Tristani-Firouzi, M. *Nature* **2006**, *440*, 463–469.

(59) Crumb, W.; Cavero, I. *Pharm. Sci. Technol. Today* **1999**, *2*, 270–280.

(60) MedTerms Medical Dictionary. http://www.medterms.com/script/main/hp.asp (accessed January 19 2012).

(61) Hoffmann, P.; Warner, B. *J. Pharmacol. Toxicol. Methods* **2006**, *53*, 87–105.

(62) Yao, X.; Anderson, D. L.; Ross, S. A.; Lang, D. G.; Desai, B. Z.; Cooper, D. C.; Wheelan, P.; McIntyre, M. S.; Bergquist, M. L.; MacKenzie, K. I.; Becherer, J. D.; Hashim, M. A. *Br. J. Pharmacol.* **2008**, *154*, 1446–1456.

(63) Scuderi, P. E. *Anesthesiology* **2010**, *113*, 772–775.

(64) Hondeghem, L. M. *Acta Cardiol.* **2011**, *66*, 685–689.

(65) *S7B The Non-Clinical Evaluation of the Potential for Delayed Ventricular Repolarization (QT Interval Prolongation) By Human Pharmaceuticals*, *Step 4 Version*; ICH Harmonised Tripartite Guideline; International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use, 2005.

(66) Friedrichs, G. S.; Patmore, L.; Bass, A. *J. Pharmacol. Toxicol. Methods* **2005**, *52*, 6–11.

(67) Polak, S.; Wiśniowska, B.; Brandys, J. *J. Appl. Toxicol.* **2009**, *29*, 183–206.

(68) Redfern, W. S.; Carlsson, L.; Davis, A. S.; Lynch, W. G.; MacKenzie, I.; Palethorpe, S.; Siegl, P. K. S.; Strang, I.; Sullivan, A. T.; Wallis, R.; Camm, A. J.; Hammond, T. G. *Cardiovasc. Res.* **2003**, *58*, 32–45.

(69) Thai, K.-M.; Ecker, G. F. *Bioorg. Med. Chem.* **2008**, *16*, 4107–4119.

(70) Murphy, S. M.; Palmer, M.; Poole, M. F.; Padegimas, L.; Hunady, K.; Danzig, J.; Gill, S.; Gill, R.; Ting, A.; Sherf, B.; Brunden, K.; Stricker-Krongrad, A. *J. Pharmacol. Toxicol. Methods* **2006**, *54*, 42–55.

(71) Williams, M.; Jarvis, M. F. In *Drug Discovery Technologies*; Clark, C. R.; Moos, W. H., Eds.; Ellis Horwood Series in Pharmaceutical Technology; Ellis Horwood Limited: Chichester, England, 1990; pp. 129–166.

(72) Chiu, P. J. S.; Marcoe, K. F.; Bounds, S. E.; Lin, C.-H.; Feng, J.-J.; Lin, A.; Cheng, F.-C.; Crumb, W. J.; Mitchell, R. *J. Pharmacol. Sci.* **2004**, *95*, 311–319.

(73) Diaz, G. J.; Daniell, K.; Leitza, S. T.; Martin, R. L.; Su, Z.; McDermott, J. S.; Cox, B. F.; Gintant, G. A. *J. Pharmacol. Toxicol. Methods* **2004**, *50*, 187–199.

(74) Price, G. W.; Riley, G. J.; Middlemiss, D. N. In *Medicinal Chemistry: Principles and Practice*, 2nd ed.; King, F. D., Ed.; The Royal Society of Chemistry: Cambridge, United Kingdom, 2005; pp. 91–117.

(75) Cheng, Y.-C..; Prusoff, W. H. *Biochem. Pharmacol.* **1973**, *22*, 3099–3108.

(76) Wood, C.; Williams, C.; Waldron, G. J. *Drug Discovery Today* **2004**, *9*, 434–441.

(77) Dunlop, J.; Bowlby, M.; Peri, R.; Vasilyev, D.; Arias, R. *Nat. Rev. Drug Discovery* **2008**, *7*, 358–368.

(78) Bridgland-Taylor, M. H.; Hargreaves, A. C.; Easter, A.; Orme, A.; Henthorn, D. C.; Ding, M.; Davis, A. M.; Small, B. G.; Heapy, C. G.; Abi-Gerges, N.; Persson, F.; Jacobson, I.; Sullivan, M.; Albertson, N.; Hammond, T. G.; Sullivan, E.; Valentin, J.-P.; Pollard, C. E. *J. Pharmacol. Toxicol. Methods* **2006**, *54*, 189–199.

(79) Guo, L.; Guthrie, H. *J. Pharmacol. Toxicol. Methods* **2005**, *52*, 123–135.

(80) Tang, Q.; Jin, M.-W.; Xiang, J.-Z.; Dong, M.-Q.; Sun, H.-Y.; Lau, C.-P.; Li, G.-R. *Biochem. Pharmacol.* **2007**, *74*, 1596–1607.

(81) Johnson, S. R.; Yue, H.; Conder, M. L.; Shi, H.; Doweyko, A. M.; Lloyd, J.; Levesque, P. *Bioorg. Med. Chem.* **2007**, *15*, 6182–6192.

(82) Kirsch, G. E.; Trepakova, E. S.; Brimecombe, J. C.; Sidach, S. S.; Erickson, H. D.; Kochan, M. C.; Shyjka, L. M.; Lacerda, A. E.; Brown, A. M. *J. Pharmacol. Toxicol. Methods* **2004**, *50*, 93–101.

(83) Tie, H. H. Cellular Mechanism of QT Prolongation and Proarrhythmia Induced by Non-antiarrhythmic Drugs. MD Thesis, University of New South Wales, Australia, 2002.

(84) Wible, B. A.; Hawryluk, P.; Ficker, E.; Kuryshev, Y. A.; Kirsch, G.; Brown, A. M. *J. Pharmacol. Toxicol. Methods* **2005**, *52*, 136–145.

(85) Ko, C. M.; Ducic, I.; Fan, J.; Shuba, Y. M.; Morad, M. *J. Pharmacol. Exp. Ther.* **1997**, *281*, 233–244.

(86) Zhou, Z.; Vorperian, V. R.; Gong, Q.; Zhang, S.; January, C. T. *J. Cardiovasc. Electrophysiol.* **1999**, *10*, 836–843.

(87) Fenichel, R. R.; Malik, M.; Antzelevitch, C.; Sanguinetti, M.; Roden, D. M.; Priori, S. G.; Ruskin, J. N.; Lipicky, R. J.; Cantilena, L. R. *J. Cardiovasc. Electrophysiol.* **2004**, *15*, 475–495.

(88) Witchel, H. J.; Milnes, J. T.; Mitcheson, J. S.; Hancox, J. C. *J. Pharmacol. Toxicol. Methods* **2002**, *48*, 65–80.

(89) Su, B.-H.; Shen, M.; Esposito, E. X.; Hopfinger, A. J.; Tseng, Y. J. *J. Chem. Inf. Model.* **2010**, *50*, 1304–1318.

(90) Li, Q.; Jørgensen, F. S.; Oprea, T.; Brunak, S.; Taboureau, O. *Mol. Pharmaceutics* **2008**, *5*, 117–127.

(91) Dubus, E.; Ijjaali, I.; Petitet, F.; Michel, A. *ChemMedChem* **2006**, *1*, 622–630.

(92) Marchese Robinson, R. L.; Glen, R. C.; Mitchell, J. B. O. *Mol. Inf.* **2011**, *30*, 443–458.

(93) Doddareddy, M. R.; Klaasse, E. C.; Shagufta; IJzerman, A. P.; Bender, A. *ChemMedChem* **2010**, *5*, 716–729.

(94) O'Brien, S. E.; de Groot, M. J. *J. Med. Chem.* **2005**, *48*, 1287–1291.

(95) AID 376 - PubChem BioAssay Summary.

http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=376 (accessed January 19 2012).

(96) CHEMBL240 Target Report Card.

https://www.ebi.ac.uk/chembldb/target/inspect/CHEMBL240 (accessed January 19 2012).

(97) Du, F.; Yu, H.; Zou, B.; Babcock, J.; Long, S.; Li, M. *Assay Drug Dev. Technol.* **2011**, *9*, 580–588.

(98) QSAR World Home Page. http://www.qsarworld.com/ (accessed January 4 2012).

(99) Tox-Portal Home Page (English Version). http://www.tox-portal.net/index.html (accessed January 19 2012).

(100) Hishigaki, H.; Kuhara, S. *Database* **2011**, *2011*, bar017.

(101) Fenichel, R.R. Receptor Binding Database.

http://www.fenichel.net/pages/Professional/subpages/QT/Tables/pbydrug.htm (accessed June 21 2012).

(102) Aureus Sciences Home Page. http://www.aureus-sciences.com/aureus/web/guest (accessed Jan 19 2012).

(103) Sunset Molecular Home Page. http://www.sunsetmolecular.com/ (accessed January 19 2012).

(104) Obiol-Pardo, C.; Gomis-Tena, J.; Sanz, F.; Saiz, J.; Pastor, M. *J. Chem. Inf. Model.* **2011**, *51*, 483–492.

(105) hERG Central Home Page. http://www.hergcentral.org/ (accessed January 19 2012).

(106) Varró, A.; Baczkó, I. *Br. J. Pharmacol.* **2011**, *164*, 14–36.

(107) Yang, P.-C.; Kurokawa, J.; Furukawa, T.; Clancy, C. E. *PLoS Comput. Biol.* **2010**, *6*, e1000658.

(108) Gupta, A.; Lawrence, A. T.; Krishnan, K.; Kavinsky, C. J.; Trohman, R. G. *Am. Heart J.* **2007**, *153*, 891–899.

(109) Rosen, M. R.; Janse, M. J. *J. Cardiovasc. Pharmacol.* **2010**, *55*, 428-437.

(110) Sibbald, B. *Can. Med. Assoc. J.* **2003**, *168*, 1035.

(111) Recanatini, M.; Cavalli, A.; Masetti, M. *ChemMedChem* **2008**, *3*, 523–535.

(112) Viskin, S.; Rosso, R. *J. Am. Coll. Cardiol.* **2010**, *56*, 1585–1588.

(113) *A Guide to Drug Safety Terms at FDA*; FDA Consumer Health Information; US Food and Drug Administration, 2008.

(114) Smalley, W.; Shatin, D.; Wysowski, D. K.; Gurwitz, J.; Andrade, S. E.; Goodman, M.; Chan, K. A.; Platt, R.; Schech, S. D.; Ray, W. A. *JAMA, J. Am. Med. Assoc.* **2000**, *284*, 3036–3039.

(115) Lee, N.; Authier, S.; Pugsley, M. K.; Curtis, M. J. *Toxicol. Appl. Pharmacol.* **2010**, *243*, 146–153.

(116) Yan, G.-X.; Antzelevitch, C. *Circulation* **1998**, *98*, 1928–1936.

(117) Curtis, M. J. *Cardiovasc. Res.* **2012**, *93*, 10–11.

(118) De Ponti, F.; Poluzzi, E.; Montanaro, N. *Eur. J. Clin. Pharmacol.* **2001**, *57*, 185–209.

(119) *E14 The Clinical Evaluation of QT/QTc Interval Prolongation and Proarrhythmic Potential for Non-Antiarrhythmic Drugs, Step 4 Version*; ICH Harmonised Tripartite Guideline; International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use, 2005.

(120) Ursem, C. J.; Kruhlak, N. L.; Contrera, J. F.; MacLaughlin, P. M.; Benz, R. D.; Matthews, E. J. *Regul. Toxicol. Pharmacol.* **2009**, *54*, 1–22.

(121) Arizona Center for Education and Research on Therapeutics. Process for Assigning Risk. http://www.azcert.org/medical-pros/drug-lists/why-lists.cfm (accessed January 20 2012).

(122) Xue, Y.; Li, Z. R.; Yap, C. W.; Sun, L. Z.; Chen, X.; Chen, Y. Z. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1630–1638.

(123) Ivanciuc, O. *Internet Electron. J. Mol. Des.* **2006**, *5*, 488–502.

(124) Yang, S.-Y.; Huang, Q.; Li, L.-L.; Ma, C.-Y.; Zhang, H.; Bai, R.; Teng, Q.-Z.; Xiang, M.-L.; Wei, Y.-Q. *Artif. Intell. Med.* **2009**, *46*, 155–163.

(125) Fu, G.-H.; Cao, D.-S.; Xu, Q.-S.; Li, H.-D.; Liang, Y.-Z. *J. Chemom.* **2011**, *25*, 92–99.

(126) Cao, D.-S.; Xu, Q.-S.; Liang, Y.-Z.; Chen, X.; Li, H.-D. *Chemom. Intell. Lab. Syst.* **2010**, *103*, 129–136.

(127) Bhavani, S.; Nagargadde, A.; Thawani, A.; Sridhar, V.; Chandra, N. *J. Chem. Inf. Model.* **2006**, *46*, 2478–2486.

(128) Gepp, M. M.; Hutter, M. C. *Bioorg. Med. Chem.* **2006**, *14*, 5325–5332.

(129) Fermini, B.; Fossa, A. A. *Nat. Rev. Drug Discovery* **2003**, *2*, 439–447.

(130) Arizona Center for Education and Research on Therapeutics. QT Drug List by Risk Groups. http://www.azcert.org/medical-pros/drug-lists/drug-lists.cfm (accessed November 25 2011).

(131) Micromedex Home Page. http://www.micromedex.com/ (accessed January 22 2012).

(132) *Meyler's Side Effects of Drugs: The International Encyclopedia of Adverse Drug Reactions and Interactions*, 15th ed.; Online Version.

http://www.sciencedirect.com/science/referenceworks/9780444510051 (accessed January 22 2012).

(133) Adverse Event Reporting System (AERS).

http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm (accessed January 4 2012).

(134) Lindquist, M. *Drug Inf. J.* **2008**, *42*, 409–419.

(135) Uppsala Monitoring Centre Home Page. http://who-umc2010.phosdev.se/ (accessed January 21 2012).

(136) Kuhn, M.; Campillos, M.; Letunic, I.; Jensen, L. J.; Bork, P. *Mol. Syst. Biol.* **2010**, *6*, 343.

(137) MedWatch Home Page. http://www.fda.gov/Safety/MedWatch/default.htm (accessed January 22 2012).

(138) *IUPAC Glossary of Terms Used in Toxicology*, 2nd ed.; Online Version, Terms Starting with C. http://sis.nlm.nih.gov/enviro/iupacglossary/glossaryc.html (accessed December 19 2011).

(139) Matthews, E. J.; Daniel Benz, R.; Contrera, J. F. *J. Mol. Graphics Modell.* **2000**, *18*, 605–615.

(140) Gombar, V. K.; Mattioni, B. E.; Zwicki, C.; Deahl, J. T. In *Computational Toxicology: Risk Assessment for Pharmaceutical and Environmental Chemicals*; Ekins, S., Ed.; Wiley Series on Technologies for the Pharmaceutical Industry; John Wiley and Sons: Hoboken, New Jersey, 2007; pp. 183–195.

(141) Brown, A. C.; Fraser, T. R. *J. Anat. Physiol.* **1868**, *2*, 224–242.

(142) Kubinyi, H. *Quant. Struct.-Act. Relat.* **2002**, *21*, 348–356.

(143) Lipnick, R. L.; Filov, V. A. *Trends Pharmacol. Sci.* **1992**, *13*, 56–60.

(144) Hansch, C.; Maloney, P. P.; Fujita, T.; Muir, R. M. *Nature* **1962**, *194*, 178–180.

(145) Hansch, C.; Muir, R. M.; Fujita, T.; Maloney, P. P.; Geiger, F.; Streich, M. *J. Am. Chem. Soc.* **1963**, *85*, 2817–2824.

(146) Hansch, C.; Fujita, T. *J. Am. Chem. Soc.* **1964**, *86*, 1616–1626.

(147) Fujita, T.; Iwasa, J.; Hansch, C. *J. Am. Chem. Soc.* **1964**, *86*, 5175–5180.

(148) Free, S. M.; Wilson, J. W. *J. Med. Chem.* **1964**, *7*, 395–399.

(149) Maran, U.; Sild, S. *Artif. Intell. Rev.* **2003**, *20*, 13–38.

(150) *Computational Toxicology: Risk Assessment for Pharmaceutical and Environmental Chemicals*; Ekins, S., Ed.; Wiley Series on Technologies for the Pharmaceutical Industry; John Wiley and Sons: Hoboken, New Jersey, 2007.

(151) Enoch, S. J.; Cronin, M. T. D.; Schultz, T. W.; Madden, J. C. *Chem. Res. Toxicol.* **2008**, *21*, 513–520.

(152) Bassan, A.; Worth, A. P. In *Computational Toxicology: Risk Assessment for Pharmaceutical and Environmental Chemicals*; Ekins, S., Ed.; Wiley Series on Technologies for the Pharmaceutical Industry; John Wiley and Sons: Hoboken, New Jersey, 2007; pp. 751–775.

(153) Tropsha, A. *Mol. Inf.* **2010**, *29*, 476–488.

(154) Judson, P. N. In *Computational Toxicology: Risk Assessment for Pharmaceutical and Environmental Chemicals*; Ekins, S., Ed.; Wiley Series on Technologies for the Pharmaceutical Industry; John Wiley and Sons: Hoboken, New Jersey, 2007; pp. 521–543.

(155) Kontijevskis, A.; Komorowski, J.; Wikberg, J. E. S. *J. Chem. Inf. Model.* **2008**, *48*, 1840–1850.

(156) Du, L.; Li, M.; You, Q.; Xia, L. *Biochem. Biophys. Res. Commun.* **2007**, *355*, 889–894.

(157) Du-Cuny, L.; Chen, L.; Zhang, S. *J. Chem. Inf. Model.* **2011**, *51*, 2948–2960.

(158) Reisfeld, B.; Mayeno, A. N.; Lyons, M. A.; Yang, R. S. H. In *Computational Toxicology: Risk Assessment for Pharmaceutical and Environmental Chemicals*; Ekins, S., Ed.; Wiley Series on Technologies for the Pharmaceutical Industry; John Wiley and Sons: Hoboken, New Jersey, 2007; pp. 33–70.

(159) Mirams, G. R.; Cui, Y.; Sher, A.; Fink, M.; Cooper, J.; Heath, B. M.; McMahon, N. C.; Gavaghan, D. J.; Noble, D. *Cardiovasc. Res.* **2011**, *91*, 53–61.

(160) Judson, P. N.; Marchant, C. A.; Vessey, J. D. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1364–1370.

(161) Judson, P. N.; Vessey, J. D. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1356–1363.

(162) Jeliazkova, N. *Toxtree*; Ideaconsult Limited: Sofia, Bulgaria. http://toxtree.sourceforge.net/ (accessed June 22 2012).

(163) Patlewicz, G.; Jeliazkova, N.; Safford, R. J.; Worth, A. P.; Aleksiev, B. *SAR QSAR Environ. Res.* **2008**, *19*, 495–524.

(164) *OncoLogic*; US Environmental Protection Agency: Washington, DC.

http://www.epa.gov/oppt/sf/pubs/oncologic.htm (accessed December 22 2011).

(165) Fourches, D.; Muratov, E.; Tropsha, A. *J. Chem. Inf. Model.* **2010**, *50*, 1189–1204.

(166) Schattel, V.; Hinselmann, G.; Jahn, A.; Zell, A.; Laufer, S. *J. Chem. Inf. Model.* **2011**, *51*, 670–679.

(167) The Cheminformatics and QSAR Society: Data Sets.

http://www.qsar.org/resource/datasets.htm (accessed May 16 2012).

(168) ChEMBLdb Home Page. https://www.ebi.ac.uk/chembldb/ (accessed May 16 2012).

(169) The PubChem Project Home Page. http://pubchem.ncbi.nlm.nih.gov/ (accessed November 24 2011).

(170) DrugBank Home Page. http://drugbank.ca/ (accessed March 16 2011).

(171) ChemSpider Home Page. http://www.chemspider.com/ (accessed May 16 2012).

(172) Young, D.; Martin, T.; Venkatapathy, R.; Harten, P. *QSAR Comb. Sci.* **2008**, *27*, 1337–1345.

(173) Williams, A. J.; Ekins, S. *Drug Discovery Today* **2011**, *16*, 747–750.

(174) Muresan, S.; Petrov, P.; Southan, C.; Kjellberg, M. J.; Kogej, T.; Tyrchan, C.; Varkonyi, P.; Xie, P. H. *Drug Discovery Today* **2011**, *16*, 1019–1030.

(175) Cannon, E. O.; Nigsch, F.; Mitchell, J. B.O. *Chem. Cent. J.* **2008** *2*, 3.

(176) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.

(177) O'Boyle, N. M.; Morley, C.; Hutchison, G.R. *Chem. Cent. J.* **2008**, *2*, 5.

(178) *The IUPAC International Chemical Identifier*; International Union of Pure and Applied Chemistry: North Carolina, USA. http://www.iupac.org/home/publications/e-resources/inchi.html (accessed June 22 2012).

(179) Weininger, D. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

(180) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244–255.

(181) *Tripos Mol2 File Format*; Tripos: St Louis, USA. http://tripos.com/data/support/mol2.pdf (accessed June 22 2012).

(182) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Methods and Principles in Medicinal Chemistry, Vol. 11; Wiley-VCH: Weinheim, Germany, 2000.

(183) Bender, A. *Expert Opin. Drug Discovery* **2010**, *5*, 1141–1151.

(184) Nigsch, F.; Mitchell, J. B. O. *J. Chem. Inf. Model.* **2008**, *48*, 306–318.

(185) Rogers, D.; Hahn, M. *J. Chem. Inf. Model* **2010**, *50*, 742–754.

(186) Eposito, E. X.; Hopfinger, A. J.; Madura, J. D. In *Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery*; Bajorath, J., Ed.; Methods in Molecular Biology, Vol. 275; Humana Press: New Jersey, USA, 2004; pp. 131–213.

(187) Zhu, H.; Rusyn, I.; Richard, A.; Tropsha, A. *Environ. Health Perspect.* **2008**, *116*, 506-513.

(188) Sedykh, A.; Zhu, H.; Tang, H.; Zhang, L.; Richard, A.; Rusyn, I.; Tropsha, A. *Environ. Health Perspect.* **2011**, *119*, 364–370.

(189) Low, Y.; Uehara, T.; Minowa, Y.; Yamada, H.; Ohno, Y.; Urushidani, T.; Sedykh, A.; Muratov, E.; Kuz'min, V.; Fourches, D.; Zhu, H.; Rusyn, I.; Tropsha, A. *Chem. Res. Toxicol.* **2011**, *24*, 1251–1262.

(190) Zhu, H.; Ye, L.; Richard, A.; Golbraikh, A.; Wright, F. A.; Rusyn, I.; Tropsha, A. *Environ. Health Perspect.* **2009**, *117*, 1257–1264.

(191) Cunningham, A. R.; Qamar, S.; Carrasquer, C. A.; Holt, P. A.; Maguire, J. M.; Cunningham, S. L.; Trent, J. O. *SAR QSAR Environ. Res.* **2010**, *21*, 463–479.

(192) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958.

(193) Bishop, C. M. *Pattern Recognition and Machine Learning*; Information Science and Statistics; Springer Science+Business Media,LLC: New York, USA, 2006.

(194) Lowe, R.; Mussa, H. Y.; Mitchell, J. B. O.; Glen, R. C. *J. Chem. Inf. Model.* **2011**, *51*, 1539–1544.

(195) Simon-Hettich, B.; Rothfuss, A.; Steger-Hartmann, T. *Toxicology* **2006**, *224*, 156–162.

(196) Peterson, K. L. In *Reviews in Computational Chemistry,* Vol. 16; Lipkowitz, K. B.; Boyd, D. B., Eds.; John Wiley and Sons, Incorporated: Hoboken, New Jersey, USA, 2000; pp. 53–140.

(197) Jensen, F. *Introduction to Computational Chemistry*, 2nd ed.; John Wiley and Sons Limited: Chichester, England, 2007.

(198) Hinselmann, G.; Rosenbaum, L.; Jahn, A.; Fechner, N.; Ostermann, C.; Zell, A. *J. Chem. Inf. Model.* **2011**, *51*, 203–213.

(199) Arodź, T.; Yuen, D. A.; Dudek, A. Z. *J. Chem. Inf. Model.* **2006**, *46*, 416–423.

(200) Benigni, R.; Bossa, C. *J. Chem. Inf. Model.* **2008**, *48*, 971–980.

(201) Gleeson, M.P.; Hersey, A.; Hannongbua, S. *Curr. Top. Med. Chem.* **2011**, *11*, 358–381.

(202) Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912–1928.

(203) Keerthi, S. S.; Lin, C.-J. *Neural Comput* **2003**, *15*, 1667–1689.

(204) Strobl, C.; Malley, J.; Tutz, G. *An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests*; Technical Report Number 55; Department of Statistics, University of Munich, 2009.

(205) Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. *Comput. Chem.* **2001**, *26*, 5–14.

(206) Cartwright, H. In *Reviews in Computational Chemistry,* Vol. 25; Lipkowitz, K. B.; Cundari, T.R., Eds.; John Wiley and Sons , Incorporated: Hoboken, New Jersey, USA, 2007; pp. 349–389.

(207) Golbraikh, A.; Tropsha, A. *J. Mol. Graphics Modell.* **2002**, *20*, 269–276.

(208) Zheng, W.; Tropsha, A. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185–194.

(209) Ivanciuc, O. In *Reviews in Computational Chemistry,* Vol. 23; Lipkowitz, K. B.; Cundari, T.R., Eds.; John Wiley and Sons , Incorporated: Hoboken, New Jersey, USA 2007; pp. 291–400.

(210) Müller, K.-R.; Mika, S.; Rätsch, G.; Tsuda, K.; Schölkopf, B. *IEEE T. Neural Networ.* **2001**, *12*, 181–201.

(211) Breiman, L. *Mach. Learn.* **2001**, *45*, 5–32.

(212) Xia, X.; Maliski, E. G.; Gallant, P.; Rogers, D. *J. Med. Chem.* **2004**, *47*, 4463–4470.

(213) Nigsch, F.; Bender, A.; Jenkins, J. L.; Mitchell, J. B. O. *J. Chem. Inf. Model.* **2008**, *48*, 2313–2325.

(214) Labute, P. In *Proceedings of the Pacific Symposium on Biocomputing '99*; Altman, R. B.; Dunker, A. K.; Hunter, L.; Klein, T. E.; Lauderdale, K., Eds.; World Scientific: New Jersey, USA, 1999; pp. 444−455.

(215) Mussa, H. Y.; Hawizy, L..; Nigsch, F.; Glen, R. C. *J. Chem. Inf. Model.* **2011**, *51*, 4–14.

(216) Littlestone, N. *Mach. Learn.* **1988**, *2*, 285–318.

(217) Åberg, K. M.; Jacobsson, S. P. *J. Chemom.* **2010**, *24*, 650–654.

(218) Fisher, R. A. *Annals of Eugenics* **1936**, *7*, 179–188.

(219) Mitchell, J. M. O. In *Machine Learning, Neural and Statistical Classification*; Michie, D.; Spiegelhalter, D. J.; Taylor, C. C., Eds.; Ellis Horwood Series in Artificial Intelligence; Ellis Horwood Limited: Hemel Hempstead, England, 1994; pp. 17–28.

(220) Henery, R. J. In *Machine Learning, Neural and Statistical Classification*; Michie, D.; Spiegelhalter, D. J.; Taylor, C. C., Eds.; Ellis Horwood Series in Artificial Intelligence; Ellis Horwood Limited: Hemel Hempstead, England, 1994; pp. 107–124.

(221) Hsu, C.-W.; Chang, C.-C.; Lin, C.-J. *A Practical Guide to Support Vector Classification*; Department of Computer Science, National Taiwan University, 2010.

(222) Song, M.; Clark, M. *J. Chem. Inf. Model* . **2006**, *46*, 392–400.

(223) Menze, B. H.; Kelm, B. M.; Masuch, R.; Himmelreich, U.; Bachert, P.; Petrich, W.; Hamprecht, F. A. *BMC Bioinf.* **2009**, *10*, 213.

(224) Rusinko, A.; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017–1026.

(225) *R: A Language and Environment for Statistical Computing*; R Development Core Team, R Foundation for Statistical Computing: Vienna, Austria. http://www.r-project.org/ (accessed June 23 2012).

(226) Hugill, M. *Advanced Statistics*; Bell and Hyman Limited: London, England, 1985.

(227) Strobl, C.; Boulesteix, A.-L.; Zeileis, A.; Hothorn, T. *BMC Bioinf.* **2007**, *8*, 25.

(228) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, 2nd ed.; Springer Series in Statistics; Springer Science+Business Media,LLC: New York, USA, 2009.

(229) Kuz'min, V. E.; Polishchuk, P. G.; Artemenko, A. G.; Andronati, S. A. *Mol. Inf.* **2011**, *30*, 593–603.

(230) Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A. F.; Nielsen, H. *Bioinformatics.* **2000**, *16*, 412–424.

(231) Thai, K.-M.; Ecker, G. *Mol. Diversity* **2009**, *13*, 321–336.

(232) Gorodkin, J. *Comput. Biol. Chem.* **2004**, *28*, 367–374.

(233) DeGroot, M.H. *Probability and Statistics*, 2nd ed.; Addison-Wesley Series in Statistics; Addison-Wesley Publishing Company, Incorporated: USA, 1989.

(234) Cohen, J. *Psychol. Bull.* **1968**, *70*, 213–220.

(235) Demel, M. A.; Janecek, A. G. K.; Thai, K.-M.; Ecker, G. F.; Gansterer, W. N. *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 91–110.

(236) Konovalov, D. A.; Llewellyn, L. E.; Vander Heyden, Y.; Coomans, D. *J. Chem. Inf. Model.* **2008**, *48*, 2081–2094.

(237) Best, D. J.; Roberts, D. E. *J. Roy. Stat. Soc. C-App.* **1975**, *24*, 377–379.

(238) Hughes, L. D.; Palmer, D. S.; Nigsch, F.; Mitchell, J. B. O. *J. Chem. Inf. Model.* **2008**, *48*, 220–232.

(239) Palmer, D. S.; O'Boyle, N. M.; Glen, R. C.; Mitchell, J. B. O. *J. Chem. Inf. Model.* **2007**, *47*, 150–158.

(240) Martin, J. K.; Hirschberg, D. S. *Small Sample Statistics for Classification Error Rates I: Error Rate Measurements*; Technical Report No. 96-21; Department of Information and Computer Science, University of California, Irvine, USA, 1996.

(241) Cawley, G.C.; Talbot N.L.C. *J. Mach. Learn. Res.* **2010**, *11*, 2079-2107.

(242) Zhu, J. X.; McLachlan, G. J.; Ben-Tovim Jones, L.; Wood, I. A. *J. Stat. Plan. Infer.* **2008**, *138*, 374–386.

(243) Hawkins, D. M. *J. Chem. Inf. Comput. Sci.***2004**, *44*, 1–12.

(244) Hansen, K.; Rathke, F.; Schroeter, T.; Rast, G.; Fox, T.; Kriegl, J. M.; Mika, S. *J. Chem. Inf. Model.* **2009**, *49*, 1486–1496.

(245) Molinaro, A. M.; Simon, R.; Pfeiffer, R. M. *Bioinformatics* **2005**, *21*, 3301–3307.

(246) Parker, B. J.; Günter, S.; Bedo, J. *BMC Bioinf.* **2007**, *8*, 326.

(247) Rücker, C.; Rücker, G.; Meringer, M. *J. Chem. Inf. Model.* **2007**, *47*, 2345–2357.

(248) Dragos, H.; Gilles, M.; Alexandre, V. *J. Chem. Inf. Model.* **2009**, *49*, 1762–1776.

(249) Netzeva, T. I.; Worth, A. P.; Aldenberg, T.; Benigni, R.; Cronin, M. T. D.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A.; Myatt, G.; Nikolova-Jeliazkova, N.; Patlewicz, G. Y.; Perkins, R.; Roberts, D. W.; Schultz, T. W.; Stanton, D. T.; Van De Sandt, J. J. M.; Tong, W.; Veith, G.; Yang, C. *ATLA, Altern. Lab. Anim.* **2005**, *33*, 155–173.

(250) Kühne, R.; Ebert, R.-U.; Schüürmann, G. *J. Chem. Inf. Model.* **2009**, *49*, 2660–2669.

(251) Tetko, I. V.; Bruneau, P.; Mewes, H.-W.; Rohrer, D. C.; Poda, G. I. *Drug Discovery Today* **2006**, *11*, 700–707.

(252) Roberts, D. W.; Patlewicz, G.; Kern, P. S.; Gerberick, F.; Kimber, I.; Dearman, R. J.; Ryan, C. A.; Basketter, D. A.; Aptula, A. O. *Chem. Res. Toxicol.* **2007**, *20*, 1019–1030.

(253) Gourley, D. G.; Shrive, A. K.; Polikarpov, I.; Krell, T.; Coggins, J. R.; Hawkins, A. R.; Isaacs, N. W.; Sawyer, L. *Nat. Struct. Biol.* **1999**, *6*, 521–525.

(254) Blomberg, L. M.; Mangold, M.; Mitchell, J. B. O.; Blumberger, J. *J. Chem. Theory Comput.* **2009**, *5*, 1284–1294.

(255) Tizón, L.; Otero, J. M.; Prazeres, V. F. V.; Llamas-Saiz, A. L.; Fox, G. C.; van Raaij, M. J.; Lamb, H.; Hawkins, A. R.; Ainsa, J. A.; Castedo, L.; González-Bello, C. *J. Med. Chem.* **2011**, *54*, 6063–6084.

(256) Hong, H.-J.; Hutchings, M. I.; Hill, L. M.; Buttner, M. J. *J. Biol. Chem.* **2005**, *280*, 13055–13061.

(257) Koul, A.; Arnoult, E.; Lounis, N.; Guillemont, J.; Andries, K. *Nature* **2011**, *469*, 483–490.

(258) Lipinski, C. A. *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 235–249.

(259) ZINC Subsets. http://zincdocking.org/browse/subsets/ (accessed January 24 2012).

(260) Irwin, J. J.; Shoichet, B. K. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.

(261) RCSB PDB Home Page. http://www.pdb.org/ (accessed November 24 2011).

(262) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(263) Ballester, P. J.; Richards, W. G. *J. Comput. Chem.* **2007**, *28*, 1711–1723.

(264) Ballester, P. J. *Future Med. Chem.* **2011**, *3*, 65–78.

(265) *GOLD*; The Cambridge Crystallographic Data Centre: Cambridge, United Kingdom. http://www.ccdc.cam.ac.uk/products/life_sciences/gold/ (accessed June 23 2012).

(266) Ballester, P. J.; Mitchell, J. B. O. *Bioinformatics* **2010**, *26*, 1169–1175.

(267) Ballester, P. J.; Mitchell, J. B. O. *J. Chem. Inf. Model.* **2011**, *51*, 1739–1741.

(268) Robinson, D. A.; Stewart, K. A.; Price, N. C.; Chalk, P. A.; Coggins, J. R.; Lapthorn, A. J. *J. Med. Chem.* **2006**, *49*, 1282–1290.

(269) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. *Nucleic Acids Res.* **2012***, 40*, D1100-D1107.

(270) Everitt, B. S. In *An R and S-Plus Companion to Multivariate Analysis*; Springer-Verlag Limited: London, United Kingdom, 2005; pp. 115–136.

(271) *Derek for Windows*; Lhasa Limited: 22-23 Blenheim Terrace, Woodhouse Lane, Leeds, LS2 9HD. https://www.lhasalimited.org/ (accessed June 22 2012).

(272) Hayashi, M.; Kamata, E.; Hirose, A.; Takahashi, M.; Morita, T.; Ema, M. *Mutat. Res., Genet. Toxicol. Environ. Mutagen.* **2005**, *588*, 129–135.

(273) Contrera, J. F.; Kruhlak, N. L.; Matthews, E. J.; Benz, R. D. *Regul. Toxicol. Pharmacol.* **2007**, *49*, 172–182.

(274) Matthews, E. J.; Ursem, C. J.; Kruhlak, N. L.; Daniel Benz, R. ; Sabaté, D. A.; Yang, C.; Klopman, G.; Contrera, J. F. *Regul. Toxicol. Pharmacol.* **2009**, *54*, 23–42.

(275) *Python Programming Language*. http://www.python.org/ (accessed June 22 2012).

(276) Ellison, C. M.; Sherhod, R.; Cronin, M. T. D.; Enoch, S. J.; Madden, J. C.; Judson, P. N. *J. Chem. Inf. Model.* **2011**, *51*, 975–985.

(277) Provost, F.; Fawcett, T. *Mach. Learn.* **2001**, *42*, 203–231.

(278) Karlberg, A.-T.; Bergström, M. A.; Börje, A.; Luthman, K.; Nilsson, J. L. G. *Chem. Res. Toxicol.* **2008**, *21*, 53–69.

(279) Pichler, W. J. *Ann. Intern. Med.* **2003**, *139*, 683-693.

(280) Burman, W. J. *Clin. Infect. Dis.* **2010**, *50*, S165–S172.

(281) Lubasch, A.; Erbes, R.; Mauch, H.; Lode, H. *Eur. Respir. J.* **2001**, *17*, 641-646.

(282) Chekmarev, D. S.; Kholodovych, V.; Balakin, K. V.; Ivanenkov, Y.; Ekins, S.; Welsh, W. J. *Chem. Res. Toxicol.* **2008**, *21*, 1304–1314.

(283) Roche, O.; Trube, G.; Zuegge, J.; Pflimlin, P.; Alanine, A.; Schneider, G. *ChemBioChem* **2002**, *3*, 455-459.

(284) Tobita, M.; Nishikawa, T.; Nagashima, R. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 2886–2890.

(285) *MOE (Molecular Operating Environment)*; Chemical Computing Group Incorporated: Montreal, Canada. http://www.chemcomp.com/ (accessed September 23 2012)

(286) Nisius, B.; Göller, A. H. *J. Chem. Inf. Model.* **2009**, *49*, 247–256.

(287) *SciFinder Scholar*; Chemical Abstracts Service: Columbus, Ohio, USA. http://www.cas.org/ (accessed September 23 2012).

(288) Fanoe, S.; Jensen, G. B.; Sjøgren, P.; Korsgaard, M. P. G.; Grunnet, M. *Br. J. Clin. Pharmacol.* **2009**, *67*, 172–179.

(289) Aronov, A. M. *J. Med. Chem.* **2006**, *49*, 6917–6921.

(290) Morik, K.; Brockhausen, P.; Joachims T. In *Proceedings of the 16th International Conference on Machine Learning*; Bratko, I.; Dzeroski, S., Eds.; Morgan Kaufmann Publishers Incorporated: San Fransisco, USA, 1999; pp. 268–277.

(291) Joachims, T. In *Advances in Kernel Methods: Support Vector Learning*; Scholkopf, B.; Burges, C.; Smola, A., Eds.; MIT Press, 1999; pp. 169–184.

(292) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708–1718.

(293) Fayyad, U.M. Irani, K.B. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*; 1993; pp. 1022–1029.

(294) Demsar, J. Zupan, B. *Orange: From Experimental Machine Learning to Interactive Data Mining*; White *Paper*; Faculty of Computer and Information Science, University of Ljubljana, 2004. http://orange.biolab.si/wp/orange.pdf (accessed June 23 2012).

(295) Labute, P. *J. Mol. Graphics Modell.* **2000**, *18*, 464–477.

(296) Thai, K.-M.; Ecker, G. F. *Chem. Biol. Drug Des.* **2008**, *72*, 279–289.

(297) Wiener, H. *J. Am. Chem. Soc.* **1947**, *69*, 17–20.

(298) ChemAxon Home Page. http://www.chemaxon.com/ (accessed November 26 2011).

(299) Jamieson, C.; Moir, E. M.; Rankovic, Z.; Wishart, G. *J. Med. Chem.* **2006**, *49*, 5029–5046.

(300) Zachariae, U.; Giordanetto, F.; Leach, A. G. *J. Med. Chem.* **2009**, *52*, 4266–4276.

(301) *Pipeline Pilot Student Edition*; Accelrys Incorporated: San Diego, California, USA. http://accelrys.com (accessed June 23 2012).

(302) Weininger, D.; Weininger, A.; Weininger, J. L. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101.

(303) Dudoit, S.; Popper Shaffer, J.; Boldrick, J. C. *Statist. Sci.* **2003**, *18*, 71–103.

(304) Kubat, M.; Matwin, S. In *Proceedings of the 14th International Conference on Machine Learning*; Fisher, D.H., Ed.; Morgan Kaufmann Publishers: San Francisco, USA, 1997.

(305) Imai, Y. N.; Ryu, S.; Oiki, S. *J. Med. Chem.* **2009**, *52*, 1630–1638.

(306) Mitcheson, J. S. *Br J Pharmacol* **2003**, *139*, 883–884.

(307) Thai, K.-M.; Windisch, A.; Stork, D.; Weinzinger, A.; Schiesaro, A.; Guy, R.H.; Timin, E.N.; Hering, S.; Ecker G.F. *ChemMedChem* **2010**, *5*, 436–442.

(308) Daylight Theory: SMILES.

http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html (accessed March 25 2012).

(309) Kortagere, S.; Krasowski, M. D.; Ekins, S. *Trends in Pharmacological Sciences* **2009**, *30*, 138–147.

(310) Nicholls, A.; McGaughey, G. B.; Sheridan, R. P.; Good, A. C.; Warren, G.; Mathieu, M.; Muchmore, S. W.; Brown, S. P.; Grant, J. A.; Haigh, J. A.; Nevins, N.; Jain, A. N.; Kelley, B. *J. Med. Chem.* **2010**, *53*, 3862–3886.

(311) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. *J. Comput. Chem.* **1996**, *17*, 1653–1666.

(312) Good, A. C.; Richards, W. G. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 112–116.

(313) Grant, J. A.; Pickup, B. T. *J. Phys. Chem.* **1995**, *99*, 3503–3510.

(314) Nicholls, A.; MacCuish, N. E.; MacCuish, J. D. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 451–474.

(315) *ROCS*; OpenEye Scientific Software: Santa Fe, New Mexico, USA.

http://www.eyesopen.com/rocs/ (accessed June 23 2012).

(316) Haque, I. S.; Pande, V. S. *J. Comput. Chem.* **2010**, *31*, 117–132.

(317) Kalliokoski, T.; Ronkko, T. P.; Poso, A. *Mol. Inf.* **2010**, *29*, 293–296.

(318) Kubinyi, H. *Drug Discovery Today* **1997**, *2*, 457-467.

(319) Zauhar, R. J.; Moyna, G.; Tian, L.; Li, Z.; Welsh, W. J. *J. Med. Chem.* **2003**, *46*, 5674–5690.

(320) Mak, L.; Grandison, S.; Morris, R. J. *J. Mol. Graph.Model.* **2008**, *26*, 1035–1045.

(321) Wilson, J. A.; Bender, A.; Kaya, T.; Clemons, P. A. *J. Chem. Inf. Model.* **2009**, *49*, 2231–2241.

(322) Armstrong, M. S.; Morris, G. M.; Finn, P. W.; Sharma, R.; Richards, W. G. *J. Mol. Graph.Model.* **2009**, *28*, 368–370.

(323) Zhou, T.; Lafleur, K.; Caflisch, A. *J. Mol. Graphics Modell.* **2010**, *29*, 443-449.

(324) Venkatraman, V.; Pérez-Nueno, V. I.; Mavridis, L.; Ritchie, D. W. *J. Chem. Inf. Model.* **2010**, *50*, 2079–2093.

(325) Ballester, P. J.; Finn, P. W.; Richards, W. G. *J. Mol. Graph.Model.* **2009**, *27*, 836–845.

(326) Ballester, P. J.; Westwood, I.; Laurieri, N.; Sim, E.; Richards, W. G. *J. R. Soc. Interface* **2010**, *7*, 335–342.

(327) Bissantz, C.; Kuhn, B.; Stahl, M. *J. Med. Chem.* **2010**, *53*, 5061–5084.

(328) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. *J. Med. Chem.* **2007**, *50*, 74–82.

(329) Kirchmair, J.; Distinto, S.; Markt, P.; Schuster, D.; Spitzer, G. M.; Liedl, K. R.; Wolber, G. *J. Chem. Inf. Model.* **2009**, *49*, 678–692.

(330) Sheridan, R. P.; Kearsley, S. K. *Drug Discov. Today* **2002**, *7*, 903–911.

(331) Brown, R. D.; Martin, Y. C. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572–584.

(332) Foloppe, N.; Chen, I.-J. *Curr. Med. Chem.* **2009**, *16*, 3381–3413.

(333) *CORINA*; Molecular Networks GmbH – Computerchemie: Erlangen Germany.

http://www.molecular-networks.com/products/corina (accessed February 8 2012).

(334) Jones, G.; Willett, P.; Glen, R. C. *J. Mol. Biol.* **1995**, *245*, 43–53.

(335) Chen, I.-J.; Foloppe, N. *Drug Dev. Res.* **2011**, *72*, 85–94.

(336) Rupp, M. Kernel Methods for Virtual Screening. PhD Thesis, Goethe Universität, Frankfurt am Main, 2009.

(337) Brown, R. D.; Martin, Y. C. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1–9.

(338) Shamovsky, I.; Connolly, S.; David, L.; Ivanova, S.; Nordén, B.; Springthorpe, B.; Urbahns, K. *J. Med. Chem.* **2008**, *51*, 1162–1178.

(339) Rohrer, S. G.; Baumann, K. *J. Chem. Inf. Model.* **2009**, *49*, 169–184.

(340) Daylight Theory: SMARTS.

http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed June 23 2012).

(341) Open Babel SMARTS. http://openbabel.org/wiki/SMARTS (accessed March 1 2012).

(342) Clayden, J.P.; Greeves, N.; Warren, S.; Wothers, P. D. In *Organic Chemistry*; Oxford University Press Incorporated: New York, 2001; pp. 1136–1137.

(343) Wojciechowski, M.; Grycuk, T.; Antosiewicz, J. M.; Lesyng, B. *Biophys. J.* **2003**, *84*, 750–756.

(344) Herr, R. J. *Bioorg. Med. Chem.* **2002**, *10*, 3379–3393.

(345) Babić, S.; Horvat, A. J. M.; Mutavdžić Pavlović, D.; Kaštelan-Macan, M. *TrAC, Trends Anal. Chem.* **2007**, *26*, 1043–1061.

(346) Mills, J. E. J.; Dean, P. M. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 607–622.

(347) Böhm, M.; Klebe, G. *J. Med. Chem.* **2002**, *45*, 1585–1597.

(348) Goldgur, Y.; Craigie, R.; Cohen, G. H.; Fujiwara, T.; Yoshinaga, T.; Fujishita, T.; Sugimoto, H.; Endo, T.; Murai, H.; Davies, D. R. *PNAS* **1999**, *96*, 13040–13043.

(349) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. *J. Med. Chem.* **1999**, *42*, 3251–3264.

(350) Steiner, T. *Angew. Chem., Int. Ed.* **2002**, *41*, 48–76.

(351) Marcou, G.; Rognan, D. *J. Chem. Inf. Model.* **2007**, *47*, 195–207.

(352) Wildman, S. A.; Crippen, G. M. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868–873.

(353) Böhm, H.; Banner, D.; Bendels, S.; Kansy, M.; Kuhn, B.; Müller, K.; Obst-Sander, U.; Stahl, M. *ChemBioChem* **2004**, *5*, 637–643.

(354) Landrum, G. *RDKit* http://www.rdkit.org/ (accessed June 23 2012).

(355) Politzer, P.; Lane, P.; Concha, M. C.; Ma, Y.; Murray, J. S. *J. Mol. Model.* **2007**, *13*, 305–311.

(356) Auffinger, P.; Hays, F. A.; Westhof, E.; Ho, P. S. *PNAS* **2004**, *101*, 16789–16794.

(357) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33–38.

(358) *VMD (Visual Molecular Dynamics)*; Theoretical and Computational Biophysics Group, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign: Illinois, USA. http://www.ks.uiuc.edu/Research/vmd/ (accessed March 5 2012).

(359) Armstrong, M. S.; Morris, G. M.; Finn, P. W.; Sharma, R.; Moretti, L.; Cooper, R. I.; Richards, W. G. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 789–801.

(360) Gedeck, P.; Rohde, B.; Bartels, C. *J. Chem. Inf. Model* **2006**, *46*, 1924–1936.

(361) *Maestro and MacroModel*; Schrödinger, LLC: New York, USA.

http://www.schrodinger.com/productsguide/ (accessed June 23 2012).

(362) *GOLDSuite*; Cambridge Crystallographic Data Centre: Cambridge, United Kingdom.

http://www.ccdc.cam.ac.uk/products/gold_suite/ (accessed June 23 2012).

(363) Halgren, T. A. *J. Comput. Chem.* **1999**, *20*, 720–729.

(364) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.

(365) Ponder, J. W.; Richards, F. M. *J. Comput. Chem.* **1987**, *8*, 1016–1024.

(366) Chang, G.; Guida, W. C.; Still, W. C. *J. Am. Chem. Soc.* **1989**, *111*, 4379–4386.

(367) Zambre, V. P.; Murumkar, P. R.; Giridhar, R.; Yadav, M. R. *J. Mol. Graph. Model.* **2010**, *29*, 229–239.

(368) Schreyer, A.; Blundell, T. *Chem. Biol. Drug Des.* **2009**, *73*, 157–167.

(369) Clark, R. D.; Norinder, U. *WIREs Comput. Mol. Sci.* **2012**, *2*, 108–113.

(370) Elridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425.

(371) Baxter, C. A.; Murray, C. W.; Clark, D. E.; Westhead, D. R.; Eldridge, M. D. *Proteins* **1998**, *33*, 367–382.

(372) *SciFinder*; Chemical Abstracts Service: Columbus, Ohio, USA. https://scifinder.cas.org (accessed November 24 2011).

(373) Bouckaert, R. R.; Frank, E. In *Proceedings of the 8th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining '04*; Dai, H.; Srikant, R.; Zhang, C., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2004; pp. 3–12.

(374) Benjamini, Y.; Hochberg, Y. *J.R. Statist. Soc.* **1995**, *57*, 289–300.

(375) Benjamini, Y.; Yekutieli, D. *Ann. Statist.* **2001**, *29*, 1165–1188.

(376) Benjamini, Y. *Am. Stat.* **1988**, *42*, 257–262.

(377) Cannon, E.O. Chemical Informatics of Prohibited Substances. PhD Thesis, University of Cambridge, United Kingdom, 2008.

(378) OpenEye Home Page. http://www.eyesopen.com/ (accessed March 18 2012).

(379) Martin, Y. C.; Porter, W.; Wu, J.; Metz, J.; Fu, W. *$Log_{10}D^{octanol/buffer}$ Computation: Which Combination of Computational Predictors Gives the Best Prediction of Experimental Results?* Oral Presentation, ChemAxon US User Group Meeting, 2010.

http://www.chemaxon.com/library/evaluation-of-software-for-property-predictions/ (accessed April 26 2012).

(380) Platts, J. A.; Butina, D.; Abraham, M. H.; Hersey, A. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 835–845.

(381) Abraham, M. H.; McGowan, J. C. *Chromatographia* **1987**, *23*, 243–246.

(382) Yap, C. W. *J. Comput. Chem.* **2011**, *32*, 1466–1474.

(383) Yap, C. W. *PaDEL-Descriptor*; National University of Singapore: Singapore.

http://padel.nus.edu.sg/software/padeldescriptor/ (accessed June 23 2012).

(384) Abraham, M. H. *Chem. Soc. Rev.* **1993**, *22*, 73–83.

(385) Clark, M. *J. Chem. Inf. Model*. **2005**, *45*, 30–38.

(386) Liaw, A. *R News* **2002**, *2*, 3:18-22.

(387) Grant, A. O. *Circ.: Arrhythmia Electrophysiol.* **2009**, *2*, 185–194.

(388) Markandeya, Y. S.; Fahey, J. M.; Pluteanu, F.; Cribbs, L. L.; Balijepalli, R. C. *J. Biol. Chem.* **2011**, *286*, 2433–2444.

(389) Jackson, C. M.; Blass, B.; Coburn, K.; Djandjighian, L.; Fadayel, G.; Fluxe, A. J.; Hodson, S. J.; Janusz, J. M.; Murawsky, M.; Ridgeway, J. M.; White, R. E.; Wu, S. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 282–284.

(390) Darbar, D.; Kimbrough, J.; Jawaid, A.; McCray, R.; Ritchie, M. D.; Roden, D. M. *J. Am. Coll. Cardiol.* **2008**, *51*, 836–842.

(391) Arizona Center for Education and Research on Therapeutics. http://www.azcert.org/ (accessed November 19 2011).

(392) Ligand.Info Home Page. http://ligand.info/ (accessed March 12 2011).

(393) von Grotthuss, M.; Koczyk, G.; Pas, J.; Wyrwicz, L. S.; Rychlewski, L. *Comb. Chem. High Throughput Screening* **2004**, *7*, 757–761.

(394) Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. In *Annual Reports in Computational Chemistry*, Vol. 4; Cornell, W., Ed.; Elsevier BV., 2007; Chapter 12.

(395) Chu, J.; Zhang, S.; Zhuang, Y.; Chen, J.; Li, Y. *Process Biochem.* **2002**, *38*, 815–820.

(396) Yoshizawa, S.; Fourmy, D.; Puglisi, J. D. *EMBO J.* **1998**, *17*, 6437–6448.

(397) Stockman, B. J.; Scahill, T. A.; Strakalaitis, N. A.; Brunner, D. P.; Yem, A. W.; Deibel Jr., M. R. *FEBS Lett.* **1994**, *349*, 79–83.

(398) England, N. W. Automatic Analysis and Validation of Open Polymer Data. PhD Thesis, University of Cambridge, United Kingdom, 2011.

(399) Kaul, M.; Barbieri, C. M.; Srinivasan, A. R.; Pilch, D. S. *J. Mol. Biol.* **2007**, *369*, 142–156.

(400) Bass, A. S.; Vargas, H. M.; Valentin, J.-P.; Kinter, L. B.; Hammond, T.; Wallis, R.; Siegl, P. K. S.; Yamamoto, K. *J. Pharmacol. Toxicol. Methods* **2011**, *64*, 7–15.

(401) Aslanian, R.; Piwinski, J. J.; Zhu, X.; Priestley, T.; Sorota, S.; Du, X.-Y.; Zhang, X.-S.; McLeod, R. L.; West, R. E.; Williams, S. M.; Hey, J. A. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 5043–5047.

(402) Zhu, B.-Y.; Jia, Z. J.; Zhang, P.; Su, T.; Huang, W.; Goldman, E.; Tumas, D.; Kadambi, V.; Eddy, P.; Sinha, U.; Scarborough, R. M.; Song, Y. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 5507–5512.

(403) Rohrer, S. G.; Baumann, K. *J. Chem. Inf. Model.* **2008**, *48*, 704–718.

(404) Celik, L.; Lund, J. D. D.; Schiøtt, B. *Chem. Res. Toxicol.* **2008**, *21*, 2195–2206.

(405) Shanle, E. K.; Xu, W. *Chem. Res. Toxicol.* **2010**, *24*, 6–19.

(406) Hinselmann, G.; Rosenbaum, L.; Jahn, A.; Fechner, N.; Zell, A. *J Cheminf.* **2011**, *3*, 3.

(407) Psyco Home Page. http://psyco.sourceforge.net/ (accessed March 3 2012).

(408) Python-Statlib Home Page. http://code.google.com/p/python-statlib/ (accessed June 23 2012).

(409) MySQL for Python Home Page. http://sourceforge.net/projects/mysql-python/ (accessed June 23 2012).

(410) Numpy Home Page. http://numpy.scipy.org/ (accessed June 23 2012).

(411) Cygwin Home Page. http://www.cygwin.com/ (accessed September 23 2012).

Appendix A. Supplementary Files

All code, additional results and data files referred to in this thesis are to be found, along with explanatory README files, in chapter specific subdirectories of the ZIP file (Final_PhD_Thesis_SI_RLMarcheseRobinson.zip) saved onto the DVD attached to the inside cover.

Appendix B. Performance of Toxicity Models Previously Reported in the Literature

Study	Descriptors	Algorithm	Training (n.o. compounds, details)	Validation (n.o. compounds, details)	MCC	P-Value (MCC)	Recall (A)	Recall (I)
Li *et al.*[90].	GRIND	SVM (WEKA), linear	476 (54 Active, 422 Inactive)	LOOCV (internal validation)*	0.28	1.0E-09	0.33	0.92
				WOMBAT-PK: 66 (19 Active, 47 Inactive), external test	0.40	1.2E-03	0.68	0.74
		SVM (WEKA), non-linear		LOOCV (internal validation)	0.18	8.6E-05	0.16	0.96
				WOMBAT-PK	0.29	1.8E-02	0.83	0.47
Dubus *et al.*[91]	P_VSA	QuaSAR-Classify	160 (80 Active, 80 Inactive)	43 (16 Active, 27 Inactive), external test of P_VSA models	0.66	1.5E-05	0.94	0.74
	Dubus-Rel				0.56	2.4E-04	0.94	0.63
	P_VSA			Training set	0.95	2.9E-33	0.98	0.98
	Dubus-Rel				0.91	1.2E-30	0.98	0.94
Tobita *et al.*[284].	MOE descriptors (2D) + MACCS (feature selection)	SVM (WEKA), RBF	73 (28 Active,45 Inactive)	10-fold CV (internal?)	0.80	8.2E-12	0.86	0.93

* Private correspondence with Dr Olivier Taboureau.

Study	Descriptors	Algorithm	Training (n.o. compounds, details)	Validation (n.o. compounds, details)	MCC	P-Value (MCC)	Recall (A)	Recall (I)
Thai and Ecker 2008 [69]	P_VSA	Binary QSAR	240 (81 Active,159 Inactive)	LOOCV	0.57	1.0E-18	0.58	0.94
	Thai-Rel				0.56	4.1E-18	0.61	0.91
	P_VSA			73 (19 Active, 54 Inactive)	0.70	2.2E-09	0.68	0.96
	Thai-Rel				0.82	2.5E-12	0.79	0.98
	P_VSA		223 (81 Active,142 Inactive), R-COOH excluded	LOOCV	0.51	2.6E-14	0.56	0.91
	Thai-Rel				0.58	4.78E-18	0.62	0.92
	P_VSA			64 (19 Active, 45 Inactive), R-COOH Excluded	0.78	4.4E-10	0.68	1.00
	Thai-Rel				0.85	1.0E-11	0.84	0.98
Thai and Ecker 2009 [231]	SIBAR (based on Thai-Rel)	Binary QSAR	194	58 (5 Active, 53 Inactive), external test	0.56	2.0E-05	0.60	0.96
Gava-ghan *et al.*[15]	DRONE (2D+3D)+ SELMA (2D)	Hier -archical PLS	436	7,520 (605 Active, 6915 Inactive), external test	0.55	0.0E+00	0.63	0.96

Table B.1 Performance of hERG blocker binary classification models for which the two classes were defined as per Chapter 4 (IC50 threshold: 1 µM). N.B.: Gavaghan *et al.*[15] specifically applied their regression model to binary classification. All values highlighted in bold were calculated by this author. Where not explicitly provided, TP, TN, FP, FN values required to compute the MCC were estimated, to the nearest integer, using the reported recall for each class, in conjunction with the number of validation compounds in each class. All p-values corresponding to the MCC values were computed from the corresponding chi-squared statistic, as per Baldi *et al.,*[230] supposing one degree of freedom (see Chapter 2, section 2.6.4.1), using the CHIDIST( ) function in Excel 2007 (32-bit).

Appendix C. Additional Computational Details

Unless stated otherwise in the main text,* the following software and hardware was employed.

Software Tools

Python

Save for control scripts run on 'non-local' machines (see below), this author primarily used Python[275] version 2.5.2 (32-bit), and, if applicable, the following non-standard modules.

- Pybel.[177] Version 1.4 (Chapter 3 and Chapter 4). Version 1.6 (Chapter 5 and Chapter 6). See below for corresponding Open Babel versions.
- Psyco.[407] Version 1.6.
- Statlib.[408] Version 1.1.0.
- MySQLdb.[409] Version 1.2.2. See below for corresponding MySQL versions.
- NumPy.[410] Version 1.2.1 (Chapter 3 and Chapter 4). Version 1.6.0 (Chapter 5 and Chapter 6).

Python version 2.7.3 (32-bit) was used to run all scripts employed for calling, and parsing the output of, the R scripts used for generation and analysis of the FOM values obtained in Chapter 5 and Chapter 6.

R

The versions of R[225] and the corresponding non-standard packages used for the following calculations and analyses are presented in Table C.1.

Calculations/Analyses	Version of R	Non-Standard R Packages
• Chapter 4: Random Forest model generation and PCA plot generation • Chapter 5: Training RF-Score	2.7.0, 32-bit	• randomForest (version 4.5-30)
• Chapter 3: Plot generation • Chapter 5 and Chapter 6: Machine Learning.	2.12.1, 64-bit	• randomForest (version 4.6-2) • caret (version 4.98) , with its dependencies: iterators, itertools, plyr, reshape (versions 1.0.4, 0.1-1, 1.5.2 and 0.8.4

* Details explicitly specified in the main text are not repeated here.

		respectively). • gtools (version 2.6.2). • gplots (version 2.10.1), with its dependencies: caTools, gdata and bitops (versions 1.12, 2.8.2 and 1.0-4.1 respectively).
• Chapter 5 and Chapter 6: Computing, plotting and statistical analysis of FOM values.	2.15.1, 64-bit	• gplots (version 2.11.0), with its dependencies: caTools, gdata, bitops and gtools (versions 1.13, 2.11.0, 1.0-4.1 and 2.7.0 respectively).

Table C.1 R setup for all relevant calculations/analyses presented in this thesis.

Molecular Operating Environment (MOE)

MOE[285] versions 2008.10 (Chapter 4) and 2011.10 (Chapter 5) were employed, unless noted otherwise, using their default settings.

Pipeline Pilot

Pipeline Pilot Student Edition[301] (version 6.1.5) was used throughout.

Open Babel

Open Babel versions 2.2.1 (Chapter 3 and Chapter 4) and 2.3.0a (Chapter 5 and Chapter 6) were employed.

ChemAxon

All references to ChemAxon[298] tools in Chapter 4 refer to version 5.1.3. Otherwise, all references to ChemAxon tools refer to those installed via the jchem-5.5.0.1-windows_with_jre installer.

MySQL

MySQL version 5.5.13 was used to load the ChEMBL database as required in Chapter 6.

Cygwin

The unix-like utilities (e.g. *rm*, *cp*) made available with Cygwin[411] were employed in various scripting workflows ran on machines running Windows operating systems (see below).

Computational Resources

Description	Calculations	Specifications
Standalone machine (local workstation)	• Chapter 3 (save for plot generation in R). • Chapter 4 (save for Winnow, Winnow derivatives and MOE calculations). • RF-Score training and validation on the PDBbind database, plus Pipeline Pilot calculations, in Chapter 5.	• Windows XP Professional, Service Pack 3, 32-bit
Standalone machine (local workstation)	• Chapter 5 (save where otherwise noted in this table). • Chapter 6 (save where otherwise noted in this table).	• Windows 7 Professional, Service Pack 1, 64-bit • Intel® Core™2 Duo CPU E7500 @ 2.93 GHz
Standalone machine (local workstation)	• Chapter 5 and Chapter 6: Computing, plotting and statistical analysis of FOM values.	• Windows 7 Enterprise, Service Pack 1, 64-bit • Intel® Core™ 2 Duo CPU P8600 @ 2.40 GHz
Various standalone machines	• Winnow feature selection, external validation, OSB generation and raw feature importance computation (Chapter 4) • CORINA (Chapter 5). • MOE (Chapter 4).	• Linux
University of Cambridge's distributed computing facility, CAMGRID	• SVM^{light} (Chapter 4). • Winnow training cycles selection	• Linux

	(Chapter 4).	
Cluster	• GOLD docking runs (Chapter 5).	• Linux • Intel Xeon X5650

Table C.2 Relevant details for all computational resources used.