

Prediction of Metabolic Stability and Bioavailability with Bioisosteric Replacements

Alison Pui Ki Choy



Clare College

University of Cambridge

Date of submission: September 2017

This dissertation is submitted for the degree of Doctor of Philosophy.

Prediction of Metabolic Stability and Bioavailability with Bioisosteric Replacements

Alison Pui Ki Choy

Abstract

Drug development is a long and expensive process. Potential drug candidates can fail clinical trials due to numerous issues, including metabolic stability and efficacy issues, wasting years of research effort and resource. This thesis detailed the development of *in silico* methods to predict the metabolic stability of structures and their bioavailability.

Coralie Atom-based Statistical SOM Identifier (CASSI) is a site of metabolism (SOM) predictor which provides a SOM prediction based on statistical information gathered about previously seen atoms present in similar environments. CASSI is a real-time SOM predictor accessible via graphical user interface (GUI), allowing users to view the prediction results and likelihood of each atom to undergo different types of metabolic transformation.

Fast Metabolizer (FAME)¹ is a ligand-based SOM predictor developed around the same time by Kirchmair *et al.* In the course of the evaluation of CASSI and FAME performance, the two concepts were combined to produce FamePrint. FamePrint is a tool developed within the Coralie Cheminformatics Platform developed by Lhasa Limited. which can carry out SOM predictions, as well as bioisosteric replacement identification. Same as CASSI, this is available via the Coralie application GUI.

The bioavailability issues caused by the metabolic enzyme, cytochrome P450 3A4, and transporter protein P-glycoprotein are also investigated in this work, along with the potential synergistic relationship between the two systems. *In silico* classifiers to distinguish substrates against non-substrates of the two systems are produced and it was envisaged that these classifiers can be integrated into FamePrint as an additional layer of information available to the user when deciding on bioisosteric replacements to use when optimising a compound.

Preface

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

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text

This dissertation does not exceed the word limit for the Degree Committee.

Acknowledgements

I would like to thank my supervisor Professor Robert Glen for providing me with the opportunity to undertake this study and his support throughout. I would also like to thank Professor Johannes Kirchmair for his guidance and making my time at the Centre for Molecular Informatics a memorable one. I would also like to thank Andrew Howlett for sharing this unforgettable journey with me!

I would also like to thank Lhasa Limited for funding the study, along with all the help and support they have given me. I would particularly like to acknowledge Dr Thierry Hanser, who has been incredibly supportive. Thank you for all the stimulating discussions and making me feel so welcome every time I have visited Lhasa.

Finally, the journey to complete this thesis has not been a smooth or easy one. I am very glad to finally have the chance to thank the friends and family who has been there to support me along the way. Thank you.

TABLE OF CONTENTS

List of Tables	ix
List of Figures	xii
Glossary	xv
List of Abbreviations	xvi
1. INTRODUCTION	17
1.1 DRUG DISCOVERY PROCESS	17
1.2 METABOLISM IN DRUG DISCOVERY	19
1.3 ORAL BIOAVAILABILITY	19
1.4 AIM OF THE STUDY	20
2. IN SILICO TOOLS FOR DRUG DISCOVERY	21
2.1 SITES OF METABOLISM PREDICTION	21
2.1.1 <i>Ligand-based Methods</i>	22
2.1.1.1 Reactivity-based Methods	22
2.1.1.2 Combined Methods	23
2.1.1.3 Machine-learning Methods	25
2.1.1.4 Fingerprint-based Methods	25
2.1.1.5 Summary	26
2.1.2 <i>Metabolite Prediction Methods</i>	27
2.1.3 <i>MetaPrint2D</i>	30
2.1.3.1 Identification of Modified Atoms	30
2.1.3.2 Selection of Transformation Data	31
2.1.3.3 Sites of Metabolism prediction	33
2.1.3.4 MetaPrint2D-React Extension	33
2.1.4 <i>FAst MEtabolizer</i>	38
2.1.4.1 Data Preparation	38
2.1.4.2 Descriptors	38
2.1.4.3 Sites of Metabolism Prediction	39
2.1.4.4 Model Evaluation	40
2.1.5 <i>Summary</i>	41
2.2 BIOISOSTERISM	42
2.2.1 <i>Ligand-based Methods</i>	43
2.2.1.1 Similarity-based Approaches	43
2.2.1.1.1 Physiochemical Property Methods	43
2.2.1.1.2 Pharmacophore Methods	46
2.2.1.2 Knowledge-based Approaches	49
2.2.2 <i>Metabolic Stability & Bioisosteres</i>	51

2.2.3	Summary	55
2.3	CONCLUSION	56
3.	DATA SOURCE AND PREPARATION	57
3.1	DATA SOURCE	57
3.2	IDENTIFICATION OF MODIFIED ATOMS	57
3.3	REACTION TYPES ANNOTATIONS	59
3.4	SELECTION OF STRUCTURES FROM THE DATABASE	61
3.5	CHARGING STRUCTURES	62
3.6	SITES OF METABOLISM ANNOTATION	70
3.7	FRAGMENTATION METHODS	74
3.7.1	<i>Coralie Fragmentation Method</i>	78
3.8	SUMMARY	80
4.	CORALIE ATOM-BASED STATISTICAL SOM IDENTIFIER (CASSI)	81
4.1	INTRODUCTION	81
4.1.1	<i>Coralie Cheminformatics Platform</i>	81
4.2	METHODS	81
4.2.1	<i>Data Source and Preparation</i>	81
4.2.2	<i>Fragmentation of Substrates</i>	82
4.2.3	<i>Model Evaluation</i>	82
4.2.4	<i>Training and Test Dataset Generation</i>	82
4.2.5	<i>Prediction with FAME</i>	82
4.2.6	<i>AUC Calculation</i>	83
4.3	MODEL GENERATION	83
4.3.1	<i>Collection of Metabolic Stability Statistics</i>	83
4.3.2	<i>SOM Prediction</i>	83
4.3.2.1	Reaction Specific Ranking	84
4.3.2.2	Atom Specific Ranking	85
4.4	RESULTS AND DISCUSSION	86
4.4.1	<i>Fragmentation Parameters</i>	86
4.4.2	<i>Model Performance</i>	91
4.5	GRAPHICAL USER INTERFACE	97
4.6	CONCLUSION	99
5.	FAMEPRINT SOM PREDICTOR: FINGERPRINT-BASED SITES OF METABOLISM PREDICTION	100
5.1	INTRODUCTION	100
5.2	METHODS	101
5.2.1	<i>Data Source and Preparation</i>	101

5.2.2	<i>Descriptors</i>	101
5.2.3	<i>Discretisation</i>	102
5.2.4	<i>Fragmentation</i>	102
5.2.5	<i>Topological Atom Pair Fingerprint</i>	102
5.2.6	<i>Training and Test Dataset Generation</i>	102
5.2.7	<i>AUC Calculation</i>	103
5.3	FAMEPRINT DEVELOPMENT	103
5.3.1	<i>FamePrint Workflow</i>	103
5.3.1.1	Model Creation	103
5.3.1.2	Metabolic Stability Prediction	106
5.3.2	<i>Descriptor Calculation</i>	107
5.3.2.1	CDK Version 1.4.18 vs CDK Version 1.5.9	107
5.3.2.2	Handling Invalid Descriptor Values	108
5.3.3	<i>Choice of Discretisation Method</i>	109
5.3.4	<i>Topological Atom Pair Fingerprint</i>	112
5.3.4.1	Fingerprints Versions	113
5.3.4.2	Size of Fingerprints.....	114
5.3.5	<i>Fingerprint Similarity Calculation</i>	115
5.3.5.1	Fingerprint Weighting	117
5.3.6	<i>Metabolic Stability Score Calculation</i>	120
5.4	GRAPHICAL USER INTERFACE	121
5.5	RESULTS AND DISCUSSION	123
5.5.1	<i>Limitation on Fragment Size</i>	123
5.5.2	<i>Performance Measurements</i>	124
5.5.2.1	Coverage Score	124
5.5.2.2	Overlap Score.....	126
5.5.3	<i>Model Evaluation</i>	127
5.5.3.1	Effects of Fragmentation Depths 5 and 6	127
5.5.3.2	Effects of Fragmentation Parameters	129
5.5.3.3	Effects of Fragmentation Depths	132
5.5.3.4	Effects of Fingerprint Versions	134
5.5.3.5	Effects of Number of Discretisation Bins	135
5.5.3.6	Effects of Number of Most Similar Fingerprints Used.....	136
5.5.3.7	Effects of Weighting of Fingerprints	136
5.5.3.8	The Optimised Variables	138
5.5.3.9	Merged Model	139
5.5.3.10	Human-specific Model	140
5.6	CONCLUSION	142
6.	BIOISOSTERIC REPLACEMENTS WITH FAMEPRINT	143

6.1	INTRODUCTION	143
6.2	METHOD	143
6.3	WORKFLOW	143
6.3.1	<i>Identification of Fragment for Replacement</i>	143
6.3.2	<i>Search for Replacement</i>	144
6.3.3	<i>Replacement Compatibility</i>	146
6.3.4	<i>Generation of New Structure</i>	147
6.4	GRAPHICAL USER INTERFACE	147
6.5	MODEL EVALUATION	148
6.5.1	<i>Retrospective Studies</i>	150
6.5.1.1	Case 1: Improvement of Prodrug Oral Bioavailability	150
6.5.1.2	Case 2: Metabolism-driven Optimisation of a Thrombin Inhibitor	153
6.5.1.2.1.	SOM2 Modifications	154
6.5.1.2.2.	SOM3 Modifications – Series 1	155
6.5.1.2.3.	SOM3 Modifications – Series 2	156
6.5.1.3	Case 3: Metabolic Stability Optimisation of a 5-lipoxygenase Inhibitor	158
6.5.1.3.1.	Template Modification	159
6.6	CONCLUSION	160
7.	IMPROVING BIOAVAILABILITY	161
7.1	ORIGIN OF LOW BIOAVAILABILITY	161
7.1.1	<i>Cytochrome P450 3A4</i>	162
7.1.1.1	Atypical Substrate Binding Kinetics	164
7.1.1.2	Existing CYP3A4 Models	165
7.1.2	<i>P-glycoprotein</i>	167
7.1.2.1	Transport Mechanism	168
7.1.2.2	Existing PGP Models	169
7.1.3	<i>CYP3A4 and PGP interplay</i>	170
7.2	STRUCTURE-ACTIVITY RELATIONSHIP MODELS	172
7.2.1	<i>Structure-activity Relationship and Machine-learning Approaches</i>	172
7.2.2	<i>Classification Workflow Design</i>	173
7.3	DATA SOURCES AND PREPARATION	173
7.3.1	<i>Data Sources</i>	173
7.3.1.1	Background Dataset	173
7.3.1.2	CYP3A4 and PGP Substrates Dataset (Stage 1 Classifier)	174
7.3.1.3	CYP3A4 and PGP Substrates/Non-substrate Datasets (Stage 2 Classifier)	175
7.3.2	<i>Data Quality and Limitations</i>	176
7.3.3	<i>Dataset Curation</i>	177
7.4	MATERIALS AND METHODS	179
7.4.1	<i>Descriptor Calculations</i>	179

7.4.1.1	MOE Descriptors	179
7.4.1.2	FamePrint Fingerprints	179
7.4.2	<i>Information Gain Analysis</i>	180
7.4.2.1	On MOE Descriptors	180
7.4.2.2	On FamePrint Fingerprints	180
7.4.3	<i>Machine-learning Methods</i>	181
7.4.4	<i>Multidimensional Scaling</i>	182
7.5	RESULTS AND DISCUSSION	183
7.5.1	<i>Classifiers</i>	183
7.5.1.1	Stage 1 Classifier Results	183
7.5.1.1.1.	CYP3A4	183
7.5.1.1.2.	PGP	184
7.5.1.2	Stage 2 Classifier Results	185
7.5.1.2.1.	CYP3A4	185
7.5.1.2.2.	PGP	186
7.5.1.3	Discussion	187
7.5.2	<i>Multidimensional Scaling</i>	189
7.6	CONCLUSION	190
8.	CONCLUSIONS AND FUTURE WORK.....	191
9.	APPENDICES	195
	Appendix A – Fragmentation Tab in Coralie’s SOM Module.....	195
	Appendix B – Prediction Tab in Coralie’s SOM Module	199
	Appendix C – Analysis Tab in Coralie’s SOM Module.....	204
	Appendix D – Frequency Distribution of CASSI Prediction on Test Dataset	207
	Appendix E – Validation Tab in Coralie’s SOM Module	211
	Appendix F – FamePrint Dataset Creation Wizard	216
	Appendix G – Biostere Tab in Coralie for SOM Prediction	220
	Appendix H – Biostere Tab in Coralie for Bioisosteric Replacement	224
10.	REFERENCES	229

List of Tables

TABLE 2.1	PERFORMANCE OF SMARTCYP AND STARDROP: COMPOUNDS WITHIN THE TOP1, 2 AND 3 RANKED POSITIONS (INCLUSIVE) CONTAINING A TRUE SOM AS IDENTIFIED IN LITERATURE. ADAPTED FROM RYDBERG <i>ET AL.</i> ¹⁷	24
TABLE 2.2	NUMBER OF SUBSTRATES USED IN CROSS VALIDATION ALONG WITH THE TOP 2 ACCURACY SCORES FOR EACH CYP ISOFORM FOR THE RS-WebPREDICTOR MODELS	25
TABLE 2.3	DEFINITION OF ABSOLUTE REASONING LEVELS IN METEOR NEXUS. ³⁰	28
TABLE 2.4	METABOLITE PREDICTION PERFORMANCES OF METEOR NEXUS AND METAPRINT2D-REACT. OVERALL METABOLITE PREDICTION PERFORMANCE ON METABOLITES GENERATED BY CYP3A4 AND/OR CYP2D6 USING A COMBINATION OF THE APPROPRIATE SMARTCYP MODEL (CYP3A4/CYP2D6) WITH METAPRINT2D-REACT. RESULTS REPORTED BASED ON METABOLITE GENERATION BY METAPRINT2D-REACT, GUIDED BY THE TOP3 AND TOP 5 (RESPECTIVELY) SMARTCYP RANKED SOM.	29
TABLE 2.5	PERFORMANCE OF METAPRINT2D WITH MULTI-STEP TRANSFORMATIONS EXCLUDED.	33
TABLE 2.6	SMIRKS PATTERNS USED IN THIS STUDY.	37
TABLE 2.7	PERFORMANCE SOM PREDICTION BY METAPRINT2D-REACT.....	37
TABLE 2.8	METAPRINT2D-REACT'S PERFORMANCE IN PREDICTING THE 5 MOST AND 5 LEAST FREQUENTLY SEEN REACTION TYPES.	37
TABLE 2.9	DEFINITION AND INFORMATION GAIN SCORES OF DESCRIPTORS CHOSEN FOR THE FINAL MODEL. (TABLE ADAPTED FROM FAME ¹).....	39
TABLE 2.10	PERFORMANCE OF FAME MODELS. TOP K SHOWS THE NUMBER OF TOP-RANKED ATOM POSITIONS CONSIDERED FOR PREDICTION SUCCESS. 5-CV SHOWS THE 5-FOLD CROSS-VALIDATION RATES.....	41
TABLE 2.11	EXAMPLES OF WELL-KNOWN BIOISOSTERIC REPLACEMENTS	42
TABLE 2.12	TOP 12 MOST NEUTRAL (BLUE), TOP 10 MOST FAVOURABLE (GREEN) AND TOP 10 MOST UNFAVOURABLE (RED) TRANSFORMATIONS. (%BAD = % OF TRANSFORMATIONS WITH $\Delta P < -25\%$, % NEUTRAL = % OF TRANSFORMATIONS WITH $-25\% \leq \Delta P \leq 25\%$ AND % GOOD = % OF TRANSFORMATION WITH $\Delta P > 25\%$	53
TABLE 3.1	REACTING CENTRE STATUS ⁸⁵	58
TABLE 3.2	TOP 30 MOST FREQUENTLY OCCURRING REACTION LABELS IN THE ACCELRY'S METABOLITE DATABASE (2011.2) 60	
TABLE 3.3	TOP 20 MOST FREQUENTLY SEEN REACTION TYPES IN UNIQUE SET OF SUBSTRATE STRUCTURES EXTRACTED. THE REACTION TYPES AND NUMBERS ARE CONSISTENT WITH ENTRIES FOUND IN TABLE 3.2.	73
TABLE 4.1	FRAGMENTATION PARAMETERS TESTED.....	86
TABLE 4.2	OMITTED DICTIONARIES: FRAGMENTATION PARAMETERS AND PERCENTAGE OF UNKNOWN ATOM ENVIRONMENTS FOUND IN TEST DATASET STRUCTURES.	89
TABLE 4.3	FAME VALIDATION RESULTS. THE PERCENTAGES OF STRUCTURES WHICH CONTAIN AT LEAST ONE SOM ATOM IN THE TOP1, 2 AND 3 POSITIONS AS WELL AS THE MEAN AND MEDIAN OF AUC VALUES PRODUCED BY ALL TEST DATASET STRUCTURES ARE SHOWN.....	92
TABLE 4.4	METAPRINT2D MODELS TRAINED ON ACCELRY'S METABOLITE DATABASE VERSION 2007.1 AND TESTED ON NOVEL COMPOUNDS ADDED TO THE 2008.1 VERSION. ²⁴	92

TABLE 4.5	CASSI VALIDATION RESULTS. THE PERCENTAGES OF STRUCTURES WHICH CONTAIN AT LEAST ONE SOM ATOM IN THE TOP1, 2 AND 3 POSITIONS AS WELL AS THE MEAN AND MEDIAN OF AUC VALUES PRODUCED BY ALL TEST DATASET STRUCTURES ARE SHOWN IN BLACK. THE CORRESPONDING PERCENTAGES OF STRUCTURES WHICH COULD NOT BE EVALUATED AND THUS DID NOT CONTRIBUTE TOWARDS THE RESULTING VALUES ARE SHOWN IN GREY UNDERNEATH THEIR CORRESPONDING VALUES. THE " %?" VALUES REPRESENT THE PERCENTAGE OF ATOMS FOR WHICH NO PREDICTION COULD BE MADE – I.E. UNKNOWN ATOMS (SEE TABLE 4.2 & APPENDIX D – FREQUENCY DISTRIBUTION OF CASSI PREDICTION ON TEST DATASET). 95
TABLE 5.1	THE SEVEN DESCRIPTORS USED AND THEIR DEFINITION. 101
TABLE 5.2	FAME.0, FAME.1 AND FAME.2 VALIDATION RESULTS. THE PERCENTAGES OF STRUCTURES WHICH CONTAIN AT LEAST ONE SOM ATOM IN THE TOP1, 2 AND 3 POSITIONS AS WELL AS THE MEAN AND MEDIAN OF AUC VALUES PRODUCED BY ALL TEST DATASET STRUCTURES WERE SHOWN. THEIR RESPECTIVE P-VALUES (5% SIGNIFICANCE LEVEL), F AND F-CRITICAL VALUES WERE REPORTED. 108
TABLE 5.3	THE NUMBER OF BINS FOR EACH FINGERPRINT VERSION WITH THE RANGE OF DISCRETISATION BINS TESTED. INDIVIDUAL FINGERPRINT SIZES REFER TO THE NUMBER OF BITS PER FINGERPRINT USED (PER DESCRIPTOR) FOR ALL DESCRIPTORS (EXCEPT FOR SYBYLATOMTYPE, SIZE IN BRACKETS). THE TOTAL FINGERPRINT SIZE REFERS TO THE TOTAL NUMBER OF BITS CREATED FOR ALL SEVEN FINGERPRINTS PER FRAGMENT STRUCTURE. 115
TABLE 5.4	INFORMATION GAIN ANALYSIS ON THE SEVEN DESCRIPTORS CHOSEN. FIGURES TAKEN FROM FAME ¹ 117
TABLE 5.5	LIST OF DIFFERENT PARAMETERS TESTED FOR EACH FAMEPRINT WORKFLOW VARIABLE CLASSES. FOR DEFINITION OF FRAGMENTATION PARAMETERS, SEE TABLE 5.6..... 127
TABLE 5.6	KEY FOR FRAGMENTATION PARAMETER USED. 127
TABLE 5.7	NUMBER OF FRAGMENTS AND FINGERPRINTS IN EACH DICTIONARY. 128
TABLE 5.8	AVERAGED PERFORMANCE STATISTICS FOR DICTIONARIES USING EACH COMBINATION OF FRAGMENTATION PARAMETERS. TOP 1, 2 AND 3 SCORES SHOWS THE % OF PREDICTIONS WHERE A FRAGMENT CONTAINS A TRUE SOM WITHIN THE TOP 1, 2 OR 3 POSITIONS. IN BRACKETS, THE COVERAGE SCORE FOR THE FRAGMENT IN THE FIRST, SECOND AND THIRD PLACE. 129
TABLE 5.9	THE AVERAGE RATIO OF SOM ATOM(S) : STRUCTURE ATOMS IN EACH TEST SET PRODUCED. 130
TABLE 5.10	AVERAGED PERFORMANCE STATISTICS FOR DICTIONARIES USING EACH COMBINATION OF FRAGMENTATION PARAMETERS, FRAGMENTATION DEPTHS 0 AND 1 EXCLUDED. TOP 1, 2 AND 3 SCORES SHOWS THE % OF PREDICTIONS WHERE A FRAGMENT CONTAINS A TRUE SOM WITHIN THE TOP 1, 2 OR 3 POSITIONS. IN BRACKETS, THE COVERAGE SCORE FOR THE FRAGMENT IN THE FIRST, SECOND AND THIRD PLACE. 131
TABLE 5.11	AVERAGED PERFORMANCE STATISTICS FOR DICTIONARIES USING EACH TESTED FRAGMENTATION DEPTH. TOP 1, 2 AND 3 SCORES SHOWS THE % OF PREDICTIONS WHERE A FRAGMENT CONTAINS A TRUE SOM WITHIN THE TOP 1, 2 OR 3 POSITIONS. IN BRACKETS, THE COVERAGE SCORE FOR THE FRAGMENT IN THE FIRST, SECOND AND THIRD PLACE..... 132
TABLE 5.12	AVERAGED PERFORMANCE STATISTICS FOR DICTIONARIES USING EACH TESTED VERSION OF FINGERPRINT. TOP 1, 2 AND 3 SCORES SHOWS THE % OF PREDICTIONS WHERE A FRAGMENT CONTAINS A TRUE SOM WITHIN THE TOP 1, 2 OR 3 POSITIONS. IN BRACKETS, THE COVERAGE SCORE FOR THE FRAGMENT IN THE FIRST, SECOND AND THIRD PLACE..... 134
TABLE 5.13	AVERAGED PERFORMANCE STATISTICS FOR DICTIONARIES USING EACH TESTED NUMBER OF DISCRETISATION BIN. TOP 1, 2 AND 3 SCORES SHOWS THE % OF PREDICTIONS WHERE A FRAGMENT CONTAINS A TRUE SOM WITHIN THE TOP 1, 2 OR 3 POSITIONS. IN BRACKETS, THE COVERAGE SCORE FOR THE FRAGMENT IN THE FIRST, SECOND AND THIRD PLACE. 135

TABLE 5.14	AVERAGED PERFORMANCE STATISTICS FOR DICTIONARIES USING EACH TESTED <i>K</i> VALUE. TOP 3 SCORES: AVERAGE CUMULATIVE PERFORMANCE STATISTICS (AVERAGE COVERAGE OF POSITION, NON-CUMULATIVE).....	136
TABLE 5.15	AVERAGED PERFORMANCE STATISTICS FOR DICTIONARIES USING EACH TESTED FINGERPRINT WEIGHTING SCHEME. TOP 1, 2 AND 3 SCORES SHOWS THE % OF PREDICTIONS WHERE A FRAGMENT CONTAINS A TRUE SOM WITHIN THE TOP 1, 2 OR 3 POSITIONS. IN BRACKETS, THE COVERAGE SCORE FOR THE FRAGMENT IN THE FIRST, SECOND AND THIRD PLACE.....	137
TABLE 5.16	OPTIMISED SELECTION OF FAMEPRINT WORKFLOW VARIABLES.	138
TABLE 5.17	EVALUATION OF FAMEPRINT PERFORMANCE WITH OPTIMISED WORKFLOW VARIABLES (GIVEN IN TABLE 5.16)	138
TABLE 5.18	FAME MODEL TRAINED ON TRAINING DATASET USED IN FAMEPRINT STUDY, BOTH USING CDK 1.5.9. RESULTS OF PREDICTIONS CARRIED OUT ON TEST SET 1,2 AND 3 (SAME DATASETS AS TABLE 5.17) ARE SHOWN.	138
TABLE 5.19	EVALUATION OF OPTIMISED FAMEPRINT WORKFLOW PERFORMANCE WITH AN EXTERNAL DATASET.....	139
TABLE 5.20	EVALUATION OF FAMEPRINT PERFORMANCE WITH OPTIMISED WORKFLOW VARIABLES (GIVEN IN TABLE 5.16) (EXCEPT RINGS ARE BROKEN DURING FRAGMENTATION).	139
TABLE 5.21	EVALUATION OF FAMEPRINT PERFORMANCE WITH THE MERGED DICTIONARY.	140
TABLE 5.22	EVALUATION OF FAMEPRINT PERFORMANCE WITH MERGED DICTIONARY OF HUMAN SUBSTRATES.	141
TABLE 6.1	TOP 7 MOST SIMILAR REPLACEMENT SUGGESTIONS RETRIEVED.....	151
TABLE 6.2	MODIFICATIONS TO SOM2 ALONG WITH REPORTED $T_{1/2}$ VALUES (IN HOURS).....	154
TABLE 6.3	SERIES 1 OF MODIFICATIONS TO SOM3 ALONG WITH REPORTED $T_{1/2}$ VALUES (IN HOURS).	155
TABLE 6.4	SERIES 2 OF MODIFICATIONS TO SOM3 ALONG WITH REPORTED $T_{1/2}$ VALUES (IN HOURS).	156
TABLE 6.5	TEMPLATE MODIFICATIONS ALONG WITH REPORTED $T_{1/2}$ VALUES (IN HOURS).....	159
TABLE 7.1	NUMBER OF SUBSTRATE AND BACKGROUND (FOR STAGE 1)/ NON-SUBSTRATE (FOR STAGE 2) STRUCTURES IN THE CYP3A4 AND PGP TRAINING AND TEST DATASETS.	178
TABLE 7.2	LENGTH OF FINGERPRINT FOR EACH DATASET. NUMBER OF BITS PER FINGERPRINT WITH ZERO INFORMATION INCLUDED IN PARENTHESIS.	181
TABLE 7.3	CYP3A4 STAGE 1 CLASSIFIER RESULTS USING MOLECULAR DESCRIPTORS.	183
TABLE 7.4	CYP3A4 STAGE 1 CLASSIFIER RESULTS USING FAMEPRINT FINGERPRINTS.	183
TABLE 7.5	PGP STAGE 1 CLASSIFIER RESULTS USING MOLECULAR DESCRIPTORS.	184
TABLE 7.6	PGP STAGE 1 CLASSIFIER RESULTS USING FAMEPRINT FINGERPRINTS.....	184
TABLE 7.7	CYP3A4 STAGE 2 CLASSIFIER RESULTS USING MOLECULAR DESCRIPTORS.	185
TABLE 7.8	CYP3A4 STAGE 2 CLASSIFIER RESULTS USING FAMEPRINT FINGERPRINTS.	185
TABLE 7.9	PGP STAGE 2 CLASSIFIER RESULTS USING MOLECULAR DESCRIPTORS.	186
TABLE 7.10	PGP STAGE 2 CLASSIFIER RESULTS USING FAMEPRINT FINGERPRINTS.	186
TABLE 7.11	PERCENTAGES OF FINGERPRINT CONTAINING INFORMATION, CALCULATED FROM TABLE 7.2).....	188

List of Figures

FIGURE 2.1	TRANSFORMATIONS IN THE ACCELRY'S METABOLITE DATABASE. THE REACTIONS $A \rightarrow B$, $B \rightarrow C$ AND $A \rightarrow C$ WERE ALL PRESENT AS SEPARATE TRANSFORMATIONS.	31
FIGURE 2.2	EXAMPLE METABOLIC SCHEME CONTAINED IN THE ACCELRY'S METABOLITE DATABASE. G1.X REPRESENT THE FIRST GENERATION OF METABOLITES AND G2.X THE SECOND GENERATION OF METABOLITES OF THE PARENT SUBSTRATE.	32
FIGURE 2.3	TYPES OF REDUCED GRAPHS ALONG WITH THE ORDER OF IMPORTANCE OF EACH STRUCTURE TYPE AND FEATURE TYPE. IF MORE THAN ONE STRUCTURE OR FEATURE TYPE WERE PRESENT IN A SINGLE NODE, THE RESULTING NODE WILL BE LABELLED ACCORDING TO THE ORDER OF PRIORITY SHOWN.	45
FIGURE 2.4	CONTEXT DEPENDENT TRANSFORMATIONS.....	54
FIGURE 3.1	ONE OF THE TRANSFORMATIONS FROM THE DATABASE ALONG WITH THE BOND BLOCK (RIGHT) OF THE STRUCTURES FROM THE DATABASE, USING STANDARD CTFILE FORMAT (TABLE 3.1).	57
FIGURE 3.2	A) HYDROLYSIS TRANSFORMATION FOUND IN THE DATABASE. B) HIGHLIGHTED (YELLOW) BONDS = BONDS MARKED AS REACTION CENTRES ACCORDING TO THE DATABASE STRUCTURE. C) HIGHLIGHTED (YELLOW) BONDS = BONDS ALTERED IN HYDROLYSIS REACTION.	58
FIGURE 3.3	TRANSFORMATION WITH BOTH A HYDROLYSIS AND A RING OPENING REACTION CLASS LABEL IN THE DATABASE..	59
FIGURE 3.4	FREQUENCIES OF DIFFERENT REACTION TYPES ACCORDING THEIR LABELS IN THE ACCELRY'S METABOLITE DATABASE.	60
FIGURE 3.5	A TRANSFORMATION WHICH CONTAINS AN R GROUP AND THIS IS THEREFORE EXCLUDED FROM THE STUDY.....	61
FIGURE 3.6	STRONG ACIDS AND BASES. PROTONS TO BE REMOVED ARE HIGHLIGHTED IN GREEN AND ATOMS TO BE PROTONATED HIGHLIGHTED IN AMBER.	63
FIGURE 3.7	ISSUES WITH THE PROTONATION/DEPROTONATION OF STRUCTURES CARRIED OUT BY MOE. THE SAME ISSUE OCCURS IN MOE VERSION 2012.10.....	64
FIGURE 3.8	EXAMPLES OF STRUCTURES CONTAINING TWO DEPROTONATED CARBONYL GROUPS.	64
FIGURE 3.9	ADENOSINE TRIPHOSPHATE AFTER WASHING BY MOE.	65
FIGURE 3.10	ALL SEVEN STRUCTURES WITH POSITIVELY CHARGED NITROGEN ATOMS TWO BONDS AWAY FROM EACH OTHER...	65
FIGURE 3.11	STRUCTURES CONTAINING TWO OR MORE POSITIVELY OR NEGATIVELY CHARGED ATOMS THAT HAVE BEEN CHARGED CORRECTLY.	66
FIGURE 3.12	EXAMPLES OF SUBSTRATE STRUCTURE PAIRS THAT ARE TREATED DIFFERENTLY BY MOE (LEFT) AND OPEN BABEL (RIGHT) CHARGING PROCEDURES.....	68
FIGURE 3.13	SUBSTRUCTURE (HIGHLIGHTED IN GREEN) RECOGNISED AS A LYSINE STRUCTURE AND IS CHARGED BY OPEN BABEL ACCORDING TO SMIRKS PATTERN: <chem>O=C(NCC=O)C(N)CCCC[N:1]</chem> >> <chem>O=C(NCC=O)C(N)CCCC[N+:1]</chem>	69
FIGURE 3.14	ONE OF THE CFP ENTRIES FROM THE METAPRINT2D-REACT FILE (TOP LEFT), THE CORRESPONDING STRUCTURE FROM THE DATABASE (TOP RIGHT) AND THE ATOMS ANNOTATED AS SOM BY THE CFP ENTRY (BOTTOM).....	71
FIGURE 3.15	EXAMPLES OF MOLECULES WHICH CANNOT BE HANDLED BY CDK'S EQUIVALENTCLASSPARTITIONER. BOTH MOLECULES CONTAIN A TERMINAL NITROGEN ATOM WITH A DOUBLE BOND TO EITHER A CARBON OR NITROGEN ATOM.	73
FIGURE 3.16	RECAP CLEAVAGE RULES IMPLEMENTED IN THE CHEMAXON FRAGMENTER.	74

FIGURE 3.17	ALL 25 FRAGMENTS GENERATED FROM STRUCTURE (TOP) BY CDK EXHAUSTIVE FRAGMENTER WITH THE DEFAULT SETTING.	76
FIGURE 3.18	MURCKO FRAMEWORK HIERARCHY AND DEFINITIONS ACCORDING TO BEMIS AND MURCKO. ³⁴	77
FIGURE 3.19	ALL FRAGMENTS GENERATED FROM STRUCTURE (FIGURE 3.17 TOP) BY CDK MURCKO FRAGMENTER WITH THE DEFAULT SETTING.	78
FIGURE 3.20	FRAGMENTATION METHODOLOGY WORKFLOW. ⁹⁴ (IMAGE REPRODUCED WITH AUTHOR'S PERMISSION)	79
FIGURE 3.21	FRAGMENTS PRODUCED FROM THE SAME STRUCTURE USING DIFFERENT FRAGMENTATION PARAMETERS IN CORALIE.	79
FIGURE 4.1	DIFFERENCES BETWEEN RETAINING AND BREAKING SCAFFOLDS.	88
FIGURE 4.2	FRAGMENTS PRODUCED BY ONLY RETAINING FUNCTIONAL GROUPS AT A FRAGMENTATION DEPTH OF 0	90
FIGURE 5.1	A) WORKFLOW FOR CREATING A DICTIONARY OF FRAGMENTS FROM A TRAINING DATASET TO BE USED FOR FRAGMENT STABILITY PREDICTION AND A SOURCE OF POSSIBLE REPLACEMENT FRAGMENTS. B) THE PARAMETERS USED TO CREATE THE DICTIONARY ARE ALSO USED TO PRODUCE FRAGMENT FINERPRINTS FROM STRUCTURES FOUND IN TEST SET 1 – AND SUBSEQUENTLY C) TEST SET 2 AND 3.	105
FIGURE 5.2	WORKFLOW FOR METABOLIC STABILITY PREDICTION	106
FIGURE 5.3	EQUAL INTERVAL (A) AND EQUAL FREQUENCY (B) DISCRETISATION EXAMPLES.	110
FIGURE 5.4	EQUAL INTERVAL (A) AND EQUAL FREQUENCY(B) WITH SKEWED DISTRIBUTION OF DESCRIPTOR VALUES.	111
FIGURE 5.5	OBTAINING THE NON-WEIGHTED SIMILARITY SCORE BETWEEN TWO SETS OF FRAGMENT FINGERPRINTS.	116
FIGURE 5.6	WEIGHTED (FPOONLY) SIMILARITY SCORE CALCULATION	118
FIGURE 5.7	EXAMPLE OF ENTROPY GAIN WEIGHTED (PER FINGERPRINT BIT) SIMILARITY COMPARISON.	119
FIGURE 5.8	EXAMPLE OF COVERAGE CALCULATION AND SCORES.	125
FIGURE 5.9	THE TOP 5 MOST UNSTABLE FRAGMENTS FROM THE QUERY STRUCTURE SUBMITTED IN APPENDIX G – BIOSTERE TAB IN CORALIE FOR SOM PREDICTION.	126
FIGURE 5.10	CHANGES IN PERFORMANCE STATISTICS AS FRAGMENTATIONS DEPTH IS VARIED. ACTUAL PERFORMANCE FIGURES ARE GIVEN IN TABLE 5.11	133
FIGURE 6.1	SUBSTITUTION OF QUERY FRAGMENT WITH A REPLACEMENT FRAGMENT OF IDENTICAL STRUCTURE. AS THE SUBSTITUTION FRAGMENT HAS DIFFERENT CONNECTION POINTS, THIS ALLOWS FOR NOVEL STRUCTURES TO BE GENERATED FROM THE SUBSTITUTION.	145
FIGURE 6.2	REPLACEMENT OF A LARGE FRAGMENT AND THE GENERATION OF A NEW STRUCTURE. THE SUBSTRUCTURE HIGHLIGHTED IN BLUE REMAINS UNCHANGED.	146
FIGURE 6.3	EXAMPLES OF KNOWN ¹⁰¹ BIOISOSTERES IDENTIFIED BY FAMEPRINT.	149
FIGURE 6.4	(R)-N-[4-[2-[[2-HYDROXY-2-(PYRIDIN-3-YL)ETHYL]AMINO]ETHYL]PHENYL]-4-[4-(4-TRIFLUORO-METHYLPHENYL)THIAZOL-2-YL]BENZENESULFONAMIDE (P) AND ITS FIRST GENERATION METABOLITES (M1 – 3).	150
FIGURE 6.5	THE TOP 5 MOST METABOLICALLY UNSTABLE FRAGMENTS PREDICTED FOR STRUCTURE P.	151
FIGURE 6.6	FINAL OPTIMISED STRUCTURE.	152
FIGURE 6.7	TOP 7 MOST UNSTABLE FRAGMENTS OF THE FINAL STRUCTURE PREDICTED BY FAMEPRINT.	152
FIGURE 6.8	3-(2-PHENETHYLAMINO)-6-METHYLPYRAZINONE ACETAMIDE, ITS PRIMARY SOM AND THE FINAL PRODUCT OF METABOLISM-DRIVEN OPTIMISATION CARRIED OUT BY BURGEY <i>ET AL.</i> ¹⁰⁴	153

FIGURE 6.9	ZILEUTON AND THE FINAL OPTIMISED STRUCTURE CARRIED OUT BY BOUSKA <i>ET AL.</i> ¹⁰⁵	158
FIGURE 7.1	CYTOCHROME P450 3A4 STRUCTURE. THE STRUCTURE IS SHOWN IN RIBBONS (COLOURED BLUE AT THE N TERMINUS TO RED AT THE C TERMINUS) AND THE HAEM GROUP SHOWN IN STICKS. THE IMAGE WAS DRAWN USING PYMOL, AND SECONDARY STRUCTURES WERE LABELLED ACCORDING TO THE SCHEME USED IN THE WILLIAMS' STUDY. ¹¹⁵	163
FIGURE 7.2	LIGANDS USED DURING CYP3A4 CRYSTALLISATION. KETOCONAZOLE (TOP LEFT) AND ERYTHROMYCIN (TOP RIGHT), M _R 531 DA AND 734 DA RESPECTIVELY, WERE USED IN THE EKROOS STUDY. ¹¹⁷ METYRAPONE (BOTTOM LEFT) AND PROGESTERONE (BOTTOM RIGHT), M _R 226 DA AND 314 DA RESPECTIVELY, WERE USED IN THE WILLIAMS STUDY. ^{115''}	164
FIGURE 7.3	MULTIPLE BINDING MODES OF CYP3A PROPOSED. THE DIAGRAM SHOWS THE DIFFERENT BINDING MODES SUGGESTED ^{121,125-138} TO ACCOUNT FOR KINETIC DATA OBSERVED. A) SINGLE OCCUPANCY OF LARGE SUBSTRATE A. B) B OCCUPIES THE POCKET WITH DIFFERENT POSSIBLE ORIENTATIONS AND WATER MOLECULES OCCUPY THE REST OF THE POCKET. C) TWO IDENTICAL MOLECULES SIMULTANEOUSLY OCCUPY TWO DISTINCT SITES. D) TWO IDENTICAL MOLECULES SIMULTANEOUSLY OCCUPY THE POCKET WITH MORE THAN ONE POSSIBLE ORIENTATION OF THE BINDING POCKET. E) TWO DIFFERENT MOLECULES SIMULTANEOUSLY OCCUPY TWO DISTINCT SITES. F) THREE DIFFERENT MOLECULES, ONE OF WHICH MAY BE AN EFFECTOR (INDUCER/INHIBITOR), BIND SIMULTANEOUSLY G) TWO IDENTICAL MOLECULES AND AN EFFECTOR BIND SIMULTANEOUSLY, EACH OCCUPYING A UNIQUE SITE. H) TWO EFFECTORS AND A SUBSTRATE BIND SIMULTANEOUSLY, WITH ONE EFFECTOR ACTING AS SUBSTRATE. (THIS DIAGRAM WAS REPRODUCED BASED ON IMAGE FROM EKINS <i>ET AL.</i> ¹³⁹ AND THE SUMMARY WAS FROM THE ORIGINAL DIAGRAM IS USED HERE.).....	165
FIGURE 7.4	CYP3A4 SUBSTRATE PHARMACOPHORE MODEL BY EKINS ¹³⁴ HYDROPHOBIC AREA (BLUE), H-BOND DONOR (RED) AND H-BOND ACCEPTORS (GREEN). DIAGRAM REPRODUCED BASED ON IMAGE FROM EKINS. ¹³⁴	166
FIGURE 7.5	PGP STRUCTURE. THE STRUCTURE IS SHOWN IN RIBBONS (COLOURED BLUE AT THE N TERMINUS TO RED AT THE C TERMINUS). THE IMAGE WAS DRAWN USING PYMOL FROM THE PDB STRUCTURE 3G5U. ¹⁴⁵	167
FIGURE 7.6	CYP3A4 AND PGP INTERPLAY HYPOTHESES. HYPOTHESIS 1 (TOP), HYPOTHESIS 2 (MIDDLE) AND HYPOTHESIS 3 (BOTTOM). 171	
FIGURE 7.7	METRABASE STRUCTURE RECORDS BREAKDOWN BY DATA SOURCE.....	176
FIGURE 7.8	MDS PLOT OF BACKGROUND (BLUE), CYP3A4 SUBSTRATE (RED) AND PGP SUBSTRATE (GREEN) STRUCTURES SHOW OVERLAP OF CYP3A4 AND PGP SUBSTRATE CHEMICAL SPACE.	189

List of Equations

EQUATION 4.1	LIKELIHOOD OF TRANSFORMATION OF A GIVEN TYPE	84
EQUATION 4.2	SUM OF TRANSFORMATION LIKELIHOODS	85
EQUATION 4.3	STABILITY OF A FRAGMENT	85
EQUATION 5.1	INFORMATION GAIN CALCULATION FOR EACH BIT IN A FINGERPRINT.	119
EQUATION 5.2	STABILITY OF A QUERY FRAGMENT.....	120

Glossary

Term	Meaning
EC₅₀	Concentration of drug that produced half the maximum biological response
IC₅₀	Concentration of inhibitor that produced half the maximum biological response
K_i	Equilibrium dissociate constant of a ligand
K_m	Substrate concentration at which the reaction rate is at half the maximum (V_{\max})
Log D	Water-octanol distribution coefficient
Log P	Water-octanol partition coefficient
pK_a	The pH at which a ligand is 50% ionised and 50% unionised
pK_i	Negative logarithm of K _i
V_{max}	Maximum reaction rate achieved at saturating substrate concentration

List of Abbreviations

Abbreviation	Meaning
ADME	Absorption, distribution, metabolism and excretion
ADMET	Absorption, distribution, metabolism, excretion and toxicity
ANOVA	Analysis of variance
ATP	Adenosine triphosphate
AUC	Area under curve
CDK	Chemistry Development Kit
CFP	Circular fingerprint
CYP	Cytochrome P450
CYP3A4	Cytochrome P450 3A4
GUI	Graphical user interface
MCS	Maximum common substructure
MMP	Matched molecular pairs
MOE	Molecular operating environment
PGP	P-glycoprotein
QSAR	Quantitative structure-activity relationship
RAM	Random-access memory
ROC	Receiver operating characteristic
SD file	Structure-data file
SOM	Sites of metabolism
SVM	Support vector machine

1. Introduction

This thesis presents studies in the following areas: the *in silico* prediction of sites of metabolism (SOM) on xenobiotics, the identification of bioisosteric replacements and *in silico* prediction of bioavailability of xenobiotics. Xenobiotics are compounds such as drugs and environmental chemicals which would not normally be expected to be present in an organism.

The current chapter provides an overview of the context of the studies undertaken and a high level description of the issues faced during drug discovery which this thesis aims to address.

1.1 Drug Discovery Process

In recent years, the pharmaceutical industry's productivity has been under threat. The high levels of R&D expenditures combined with the low number of new drug approvals is a cause for concern. Recent figures show that the average development time from project initiation to an approved drug on market is over 13 years and the total amount spend during this process can be as high as 1 billion US dollars.² Preventing the collapse of the pharmaceutical industry in its present form requires new strategies to improve R&D productivity.

The whole process of drug discovery from initialisation to market approval of a drug can roughly be broken down into three main stages:

1. Target Discovery
2. Drug development
3. Clinical Trials

Target discovery involves the identification and validation of the target which causes a disease or unwanted biological activity. Once the target has been confirmed and the clinical need for a drug has been decided, drug development can commence.

Lead identification is the first step of drug development. This involves identifying lead compounds which exhibit some amount of the desired biological activity against the identified target. Lead optimisation follow in order to improve the performance of lead compounds against the target. This is an iterative process involving the fine tuning of the toxicology, pharmacodynamics and pharmacokinetics profiles of lead compounds as well as ensuring the compound can be delivered to the target.³ Pharmacodynamics is the study of drug-target interaction in order to determine the dose-response effect.⁴ This is outside the scope of this study and therefore will not be discussed in

detail. Pharmacokinetics is concerned with the kinetics of absorption, distribution, metabolism and excretion (ADME) of drugs.⁴ This includes the bioavailability of the drug given its method of administration, the metabolism of the drug as well as the route of excretion of the drug. The drug development stage of drug discovery also involves the use of *in vitro* assays and animal tests. Recent trends has also seen a shift towards including *in silico* (computational) methods in addition to tests mentioned above.⁴

Clinical trials following drug development consist of four phases.⁵ Phase I is the first time the drug compound is tested on human subjects, in this case, healthy human volunteers. The aim of this phase is to gather safety information on the drug when applied to the human body, as well as the properties of the drug and its pharmacodynamics. Phase II aims to further examine the drug's safety as well as its effectiveness when given to a small group of humans with the targeted disease. Phase III follows in order to confirm the effectiveness of the drug in a larger test group of patients. Phase IV occurs after the marketing of the new drug and is carried out in order to monitor the efficacy and side effect of the drug.

The use of animal studies in the drug development stage of the drug discovery process provides invaluable information regarding the effect of drug in a living organism. However, these animal models do not provide a sufficient representation of the effects the drug will have on the human body and issues with metabolic stability and the poor oral bioavailability of some drugs are two of the main reasons for phase I and phase II attrition⁶ during drug discovery. Given the cost and time required for drug discovery, earlier failures are preferable in order to allow more time and resources to be made available to other successful drug discovery projects. *In silico* models have been used as an attempt to predict the effects of compounds under study during lead optimisation on the human body. If better *in silico* methods are available that can predict the absorption, distribution, metabolism and excretion (ADME) properties of compounds, this may help reduce the fraction of failed candidate compounds during clinical trials.

If problems with metabolic stability and bioavailability of potential drug candidates can be predicted with higher confidence in order to rule out more unsuitable compounds with undesirable properties earlier on in the drug discovery process, this will help reduce the time and resources spent on drug development (as failures will be ruled out at earlier stages) and thus reduce the delay from project initialisation to market.

The first aim of this thesis is to create *in-silico* models to predict the potential SOM on novel structures which are potential lead compounds, to help reduce the time spent on compounds that

will eventually fail. Once a compound of interest with desirable properties has been identified as having a suitable metabolic stability profile, it should be possible to improve upon the undesirable properties of the compound by selecting suitable bioisosteric replacements. Bioisosterism is a concept well-known to the drug discovery process. A second aim of this study is to identify bioisosteric substructures that could replace an undesirable part of the original structure whilst maintaining its metabolic profile. Finally, development of *in silico* classifiers will also be reported which will help determine the oral bioavailability of a novel compound of interest.

1.2 Metabolism in Drug Discovery

The human body contains a host of enzymes which are capable of detoxifying unwanted and potentially harmful molecules from the immediate environment. A notable class of these enzymes which are capable of catalysing biotransformation are the cytochrome P450 monooxygenases (CYPs), which offer a line of protection against invading aromatic hydrocarbons, and so are of particular relevance to drug design.⁷ CYPs, found in the gut and the liver, are mainly concerned with transforming large amounts of xenobiotic compounds, including drugs, by increasing their hydrophilicity which facilitates their clearance by the kidneys.

There are two main phases of biotransformation in the body, as well as a more recently recognised third phase, Phase I and II and III respectively. Phase I transformations are concerned mainly with CYP oxidative reactions but sometimes also include other non-CYP modifications such as reduction and hydrolysis.⁷ Phase II covers conjugation reactions, typically involves the addition of hydrophilic structure (e.g. sugars, salts, amino acids) to xenobiotic compounds. Phase III describes the extrusion of xenobiotic compounds as well as their metabolites from intracellular space into the intestinal fluid, blood and kidneys via the action of efflux pumps.

This thesis mainly considers Phase I and Phase II metabolism as these are the phases where actual structural modification of a xenobiotic compound occurs. Phase III is more in line with the concerns over bioavailability, which will be discussed below.

1.3 Oral Bioavailability

For an orally administered drug, in order for the drug compound to successfully reach its intended target, it cannot be too lipophilic, otherwise it could remain trapped in cell membrane when attempting to enter the body, likely through the gut intestinal wall. The drug compound also cannot be too hydrophilic, otherwise it would not be able to diffuse through the cell membrane without assistance from transporters. The role of transporters in assisting the movement of xenobiotic

compounds into and out of cells is increasingly being recognised.⁷ Transporter proteins are found in a wide range of tissues and can broadly be classified into two types: influx and efflux transporters. Efflux transporters move xenobiotic compounds out of cells, typically against concentration gradients, by using energy from adenosine triphosphate (ATP) hydrolysis and are of particular interest to drug discovery as they can limit the bioavailability of drug compounds.

A particular efflux transporter of interest is the P-glycoprotein (PGP), which is a major system responsible for moving endogenous compounds and xenobiotics out of cells. PGP is found in the apical cell membrane of the cells lining the gut and work in partnership with high levels of cytochrome P450 3A4 (CYP3A4) in the gut cells (three times the amount compared to in human liver cells) to limit the amount of xenobiotic compounds entering the blood stream.⁸ Both PGP and CYP3A4 have incredibly wide substrate specificity and are both under the control of the same protein expression inducers. This makes PGP and CYP3A4, and the mechanism of their partnership of particular importance to the oral bioavailability of many drug compounds.

1.4 Aim of the Study

The studies reported by this thesis aims to address the difficulties encountered during the drug discovery process by providing *in silico* models to predict metabolic stability and oral bioavailability issues that are likely encountered by the compound of interest.

Chapter two will follow with a presentation of existing SOM prediction methods and methods to identify bioisosteres that have been reported in literature. Chapter three will present the data and methodologies used in the studies of SOM prediction and the identification of bioisosteres. Chapters four and five will report the development and evaluation of CASSI and FamePrint, two separate SOM prediction methodologies. Chapter six will present retrospective case studies of using FamePrint in the context of identifying bioisosteric replacements. Chapter seven will follow with an introduction into the bioavailability faced during drug discovery with particular focus on CYP3A4 and PGP, as they are of particular significance in controlling the oral bioavailability of xenobiotics. Chapter seven will also include the development of *in silico* classifiers for CYP3A4 and PGP substrates. Chapter eight will give a conclusion of the studies reported in this thesis and suggestions of future work.

2. *In Silico* Tools for Drug Discovery

Drug discovery is inherently a multi-objective optimisation problem. During drug discovery, once a lead compound has been identified, multiple cycles of optimisation have to be carried out to fine-tune the compound. Not only does the compound need to bind to the target with adequate affinity and produce the desired physiological effect, other properties such as solubility, toxicity, selectivity, bioavailability and permeability often have to be optimised until all criteria are satisfactory.

In silico methods are increasingly being employed during the drug discovery process as they provides earlier feedback to medicinal chemists, highlighting potential issues faced by the compounds of interest in multiple areas such as metabolic stability, bioavailability, toxicity and pharmacodynamics. The studies reported in this thesis describe the development of *in silico* tools for the prediction of SOM, bioisosteric replacements and oral bioavailability. In this chapter, *in silico* methodologies for SOM prediction and identifying bioisosteric replacements will be reviewed. Methodologies for the *in silico* prediction of bioavailability will be reviewed in chapter 7.

2.1 Sites of Metabolism Prediction

Xenobiotics are compounds that are found within an organism which would not usually be expected to be present in the organism based on a typical diet and metabolism. These include any drugs or environmental chemicals that are introduced into the organism. The metabolism of xenobiotics is a major challenge for drug discovery. Many *in silico* methods to predict SOM on xenobiotic compounds, in particular drug-like compounds, have been produced and reported in the literature; a review of these will be given in this chapter.

SOM prediction methods can broadly be classified into ligand- and structure-based methods. Ligand-based methods obtain knowledge only from the ligands (structures which interact with protein targets, such as drug compounds) and assume that information about the biological target is inherently embedded in properties found in the ligand structures. Structure-based methods require knowledge from biological targets (e.g. enzymes or transporters), usually involving the target's 3D crystal structure, amino acid sequences and binding pocket interaction, especially with respect to the residues that may come into contact with ligands.

Ligand-based methods need to cope with significance uncertainties regarding the ligand's interaction with the binding pocket. The concept that ligand-based methods rely upon is that properties of a ligand which are of significance to its interaction with a target structure are encoded

in in the properties of the ligand itself. As the structural data of a protein target is not always available (and indeed, sometimes the target itself may be unknown), this makes ligand-based methods particularly valuable as no explicit information about the target is required. A drawback of ligand-based methods is that any modification done on the ligand (such as during lead optimisation) may be sterically incompatible with the target's binding pocket, although given the flexibility of metabolic enzyme and transporter binding pockets, this is less likely to be an issue.

Compared to ligand-based methods, structure-based methods tend to require more computational power due to the inclusion of properties of the target protein structure (which are generally much larger than the ligand) when considering protein-ligand interactions. They also require the structure of the target to be known. Some structure-based methods only consider properties of static structures of the target where as others also include the time-dependent conformational fluctuations of both the protein target as well as the ligand (such as molecular dynamic simulations). The studies reported by this thesis are ligand-based methods, therefore structure-based methods are outside of the scope of this literature review.

2.1.1 Ligand-based Methods

Even within ligand-based methods, there are many different classes of approaches, such as reactivity-based methods, fingerprint-based data mining approaches, machine-learning based approaches and combined approaches which consider multiple aspects of metabolic transformations.⁹

As mentioned in 1.2 Metabolism in Drug Discovery, CYP enzymes are of particular importance in the metabolism of xenobiotics, therefore it should come as no surprise that a large number of methodologies reported in the literature focus solely on CYP enzymes (the whole superfamily of enzymes as well as individual members). CYP is a superfamily of haem-thiolate monooxygenases, each member containing a haem cofactor with an iron centre. The active species produced during the catalytic cycle which is responsible for the oxidation of substrate structures is termed Compound I. A number of reactivity-based approaches have been reported in the literature based on the reactivity of ligand structures with regards to Compound I of CYP enzymes.

2.1.1.1 Reactivity-based Methods

QMBO¹⁰ is a quantum mechanical method based on the idea that hydrogen abstraction by Compound I is the rate determining step for CYP catalysis. The method calculates all C-H bond orders in a substrate structure using density functional theory and the deviations from average bond order used to derive the C-H bond strength. Additional corrections are made based on the accessibility of

the hydrogen atom based on the solvent accessible surface area of the atom. QMBO successfully predicted the correct SOM in the top three ranked position in 84% of cases when tested on over 81 structures.

CypScore, like QMBO, is another SOM predictor dedicated to predicting SOM caused by CYP metabolism.¹¹ The CypScore models created are of a non-isoform specific oxidation “P450 super-enzyme”. A unified model instead of separate models for each CYP isoform was created as it is not uncommon for these enzymes to work together – one structure can be a substrate of multiple isoforms of CYP. The advantages of having a unified model is that the resulting reactivity score would already have taken into account competing metabolic transformations carried out by different CYP isoforms. Six different models were created to take different types of generic oxidation reactions into account:

1. Aliphatic hydroxylation, N-dealkylation, O-dealkylation
2. Aromatic hydroxylation
3. Double bond epoxidation/oxidation
4. Amine N-oxidation
5. Imine N-oxidation
6. S-oxidation

These models use a combination of AM1 semi-empirical molecular orbital theory¹² derived atomic reactivity descriptors and molecular surface-based properties calculated using ParaSurf.¹³ When tested against 39 structures, an experimentally observed SOM was identified in the top three ranked positions in 87% of cases.

2.1.1.2 Combined Methods

As well as approaches which focus solely on one aspect of metabolic stability (such as reactivity or accessibility), combined approaches have been reported in the literature which take into account multiple different properties that are significant when considering metabolic stability. MetaSite by Molecular Discovery¹⁴ is one such approach which also targets a selection of the most important human CYP isoforms.¹⁵ MetaSite considers the thermodynamic and kinetic factors during the prediction of the most likely SOM. A molecular interaction field (MIF)-based approach was used to evaluate the thermodynamic aspect of the enzyme-substrate interaction: information regarding the 3D conformation of the enzyme active site pocket was encoded in fingerprints, which would then be compared against fingerprints of the query ligand generated from GRID probe categories (hydrophobic, hydrogen-bond donor, hydrogen-bond acceptor and charge) with their distances binned. The predicted exposure of a site to the catalytic haem group within the CYP active site would account for the site’s accessibility. The accessibility part of the MetaSite methodology is structure

based. MetaSite uses a ligand-based approach when considering the reactivity of the query ligand (relating to the kinetic energy required to reach the transitional state for the catalytic hydrogen atom abstraction), obtained using molecular orbital calculations and fragment recognition. In order for a site to be considered susceptible to CYP-mediated transformation, high scores are required in both the accessibility and reactivity categories. As MetaSite considers the thermodynamic and kinetic factors of the transformation therefore is independent of any training dataset.¹⁶ The authors reported an average accuracy of 85% for identifying a known SOM within the Top 2 positions.

SMARTCyp is another SOM predictor which focused on CYP metabolism.¹⁷ Similar to MetaSite, SMARTCyp also utilises a combined approach and does not rely on training dataset structures but is a purely ligand-based approach. SMARTCyp uses a reactivity model which operates on 2D structures. A reactivity table containing transition state energies derived from density functional theory for substructures with a non-isoform specific CYP haem group was created. This is used during the calculation of reactivity descriptors for each atom of a query structure based on the matching of SMARTS patterns of substructure around the query atom against substructures stored in the reactivity table. Accessibility descriptors are also calculated for each atom based on their relative position to the edges of the structure. 394 CYP3A4 substrates with experimentally identified SOM were used to test the performance of SMARTCyp as well as its performance compared to StarDrop by Optibrium. The results were provided in Table 2.1:

Method	Top 1	Top 2	Top 3
SMARTCyp	65%	76%	81%
StarDrop	59%	75%	84%

Table 2.1 Performance of SMARTCyp and StarDrop: compounds within the top1, 2 and 3 ranked positions (inclusive) containing a true SOM as identified in literature.
Adapted from Rydberg *et al.*¹⁷

The P450 module of StarDrop by Optibrium also focuses on predicting the SOM of CYP metabolism for the following isoforms: 3A4, 2D6, 2C9, 1A2, 2C19, 2C8 and 2E1.¹⁸ These P450 models from StarDrop are based on simulations of the catalytic mechanism carried out by CYP enzymes (with parameters tuned using experimental data). Upon submission of a query structure, first its 3D structure is generated using CORINA¹⁹ and the metabolic vulnerability for each query atom assessed. Site vulnerabilities of atoms are predicted based on the quantum mechanical reaction energies for hydrogen abstraction, calculated using a semi-empirical method. The same CYP model is used for all CYP isoforms. The specificity of the ligand for a given CYP isoform is evaluated separately by the alignment of the query structure to an isoform specific model build based on known substrates of the isoform. Therefore this is an entirely ligand based method.

2.1.1.3 Machine-learning Methods

The Metabolism Module of ADMET (absorption, distribution, metabolism, excretion and toxicity) predictor²⁰ is a SOM predictor produced by Stimulations Plus, specifically tailored to predicting the atoms which are prone to oxidation by different CYP isoforms (1A2, 2A6, 2B6, 2C8, 2C19, 2C9, 2D6, 2E1, and 3A4). The majority of the training data for these SOM prediction models comes from the Accelrys Metabolite Database, Drugbank²¹, as well as curated data from published literature. The SOM prediction models in the Metabolism Module use atom-based descriptors and artificial neural network ensembles to evaluate the metabolic stability of each atom of a query structure. A substrate classification model would then be applied to predict the query structure's likelihood of being a substrate of the CYP isoforms listed above.

Another SOM predictor dedicated to CYP enzymes is the RS-(Web)Predictor, created by Zaretski *et al.*,^{22,23} which includes models for CYP isoforms 1A2, 2A6, 2B6, 2C8, 2C9, 2C19, 2D6, 2E1 and 3A4 (Table 2.2). For each potential SOM, 148 topological descriptors and 392 quantum chemical atom-specific descriptors, some modified to include properties of neighbouring atoms, are used along with a support vector machine (SVM)-based ranking algorithm and a multiple instance learning method to generate a prediction of which atoms are likely to be metabolised by a specified CYP isoform. This outperformed both SMARTCyp and StarDrop when using the top 2 metric.

CYP	1A2	2A6	2B6	2C8	2C9	2C19	2D6	2E1	3A4
substrates	271	105	151	142	226	218	270	145	475
Top2 (%)	83.0	85.7	82.1	83.8	84.5	86.2	85.9	82.8	82.3

Table 2.2 Number of substrates used in cross validation along with the top 2 accuracy scores for each CYP isoform for the RS-WebPredictor models

FAst MEtabolizer (FAME) is another machine-learning based methodology created based on the Accelrys Metabolite Database.¹ The methodology and results of FAME will be reviewed in more detail in section 2.1.4.

2.1.1.4 Fingerprint-based Methods

A notable fingerprint-based SOM study is MetaPrint2D^{24,25} which was created by Adams and is built on data from the Accelrys Metabolite Database²⁶. MetaPrint2D was built upon to create MetaPrint2D-React, which predicts the structure of metabolites after SOM prediction has been carried out. As the dataset used for the SOM prediction studies reported in this thesis have been prepared by the same algorithms used in the data preparation steps in MetaPrint2D, the MetaPrint2D methodology will be discussed in more detail in section 2.1.3.

2.1.1.5 Summary

As almost all SOM prediction models are built upon experimentally obtained biological activity data, which contain intrinsic variance that arises from different experimental conditions, assay types, and human errors to name but a few potential factors. Therefore, a good prediction model should aim to account for around 65 – 85% of the variance in the dataset it is built on; models with better performances are likely to be overfitted.⁹

Given the promiscuity and comprehensiveness of structures affected by CYP metabolism, it is not particularly surprising that a large number of SOM predictors were created to focus on CYP-mediated metabolism. However, a common recurring limitation within these SOM models is that they were created to identify the most likely SOM within a substrate structure, assuming the structure would be a substrate of a CYP enzyme. For example, Metaprint2D gives no result if it is an atom environment not covered by the data it is trained on. Many prediction models always give a prediction result even if the prediction is made on data outside of the region the model has been trained on (i.e. negative predictions are not handled). This particular issue will be addressed in chapter 7: Improving Bioavailability.

Many other structure-based SOM prediction methodologies have been reported, including shape-focused methods, molecular interaction field based methods, and numerous protein-ligand docking studies. These are, however, not discussed here as it is outside the scope of the studies reported.

2.1.2 Metabolite Prediction Methods

In addition to SOM predictors, methodologies created to predict the structure of metabolites are also capable of predicting SOM or regions of metabolic vulnerability on a structure prior to the generation of metabolite structures. One of the first of such metabolite predictors is MetabolExpert from CompuDrug,²⁷ produced in 1985. The tool was designed to predict the structures of metabolites, based on metabolic transformation rules contained within the software. The knowledge database rules consist of substrate and metabolite structures as well as substructures that were prone to promote or inhibit metabolism. The existing rules within MetabolExpert cover the most common metabolic pathways found in mammals, plants and for photodegradation. Users also have the option to add to these rules or modify existing ones in order to improve the prediction performance of the software. This rule-based methodology requires no structural data from the protein targets themselves, it is a purely ligand-based approach.

A similar rule-based metabolite predictor was META, produced by Klopman in 1994.²⁸ META is an expert system, based on a dictionary of transformation rules (each rule consisting of one parent fragment structure and one metabolite fragment structure), gathered from literature by experts. Upon the submission of a new query structure, the relevant rules within the transformation dictionary are consulted and subsequently applied on the query to create a metabolite structure. META was not designed to act directly as a SOM predictor; however, during the prediction of potential metabolite structures, metabolically vulnerable regions of the query structures were identified.

Meteor Nexus (previously Meteor) is a metabolism expert system produced by Lhasa Limited.²⁹ The software can be used to aid the understanding of metabolic vulnerability and outcome of structures under investigation in drug discovery programmes. The knowledge base behind Meteor Nexus contains data gathered from the literature as well as private, confidential data from pharmaceutical companies. Transformation rules were extracted from these sources and stored, adopting descriptors within the representation of structures contained in transformation rules, allowing a more in-depth description of the applicable chemical context of the rule. Appropriate transformation rules are applied on structures to produce metabolites, their likelihood to be generated *in vivo* given as probable, plausible, equivocal, doubted or impossible (Table 2.3).

Absolute Reasoning Level	Definition
Probable	There is at least one strong argument that the proposition is true and there are no arguments against it
Plausible	On balance the weight of evidence supports the proposition.
Equivocal	There is an equal weight of evidence for and against the proposition.
Doubted	On balance the weight of evidence opposes the proposition.
Improbable	There is at least one strong argument that the proposition is false, and there are no arguments that it is true.

Table 2.3 Definition of absolute reasoning levels in Meteor Nexus.³⁰

Systematic Generation of potential Metabolites (SyGMA) is another metabolite predictor, based on rules derived from transformations contained in the Accelrys Metabolite Database.³¹ SyGMA covers a range of human Phase I and Phase II metabolism (70% of observed transformations according to authors in 2008). Unlike other rule-based expert system metabolite predictors, an empirical probability score was added to each rule derived from the Metabolite Database, based on the number of correct predictions made using the rule on training set structures. When a rule is used to generate a metabolite from a query structure, the probability score is applied to the metabolite generated and also used for ranking the resulting predicted structures. This could also be interpreted as the probability of a particular site being metabolised on the query structure, forming a ranking of SOM atoms.

ChemAxon also produced a metabolite predictor, Metabolizer, designed to predict metabolite structures based on a library of transformation rules.³² A number of different biotransformation libraries were available, including human Phase I, human Phase II, mouse, rat, bacteria and plants, which allowed a user to specify the relevant species and metabolic phase for which metabolic stability and metabolite structure predictions should be carried out. The knowledge behind the transformation libraries was composed of manually curated experimental results from the literature. Metabolizer also allowed a user to insert their own rules and libraries either in addition to an existing Metabolizer library or as an independent collection.

Unlike many of the methodologies dedicated to the prediction of SOM, metabolite predictors were not solely dedicated to the prediction of CYP mediated transformations. However, as the reported performance refers to the percentage of successful metabolites predicted, rather than just the identification of correct SOM, this makes a direct comparison of performance statistics impossible.

A recent study by Piechota *et al.* on “Pragmatic approaches to using computational methods to predict xenobiotic metabolism” compared the performance of MetaPrint2D-React, Meteor and

SMARTCyp.³³ It was noted in the study that the prediction of SOM is not equivalent to the prediction of the correct metabolites as the latter would entail the correct ranking of the vulnerabilities of all SOM and the accurate identification of the type of transformation that has occurred. In the study, SMARTCyp was used to identify potential SOM on a structure and MetaPrint2D-React used to predict the metabolites for the top three and top five most likely sites chosen by SMARTCyp. The combination of a reactivity-based SOM predictor to direct the site on which to apply the metabolite prediction and generation proved to be fairly successful in some cases (especially for SMARTCyp's 2D6 model).

Software		Total % of metabolite correctly predicted	
		on homogenous dataset	on diverse dataset
Meteor (EQU3)		73%	85%
MetaPrint2D-React		80%	89%
SMARTCyp	+ CYP 3A4	n/a	44%, 56%
MetaPrint2D-React	CYP 2D6	n/a	73%, 91%

Table 2.4 Metabolite prediction performances of Meteor Nexus and MetaPrint2D-React. Overall metabolite prediction performance on metabolites generated by CYP3A4 and/or CYP2D6 using a combination of the appropriate SMARTCyp model (CYP3A4/CYP2D6) with MetaPrint2D-React. Results reported based on metabolite generation by MetaPrint2D-React, guided by the top3 and top 5 (respectively) SMARTCyp ranked SOM.

2.1.3 MetaPrint2D

MetaPrint2D is a fingerprint-based SOM prediction methodology created by Adams, who also extended the methodology to produce MetaPrint2D-React, a tool for predicting metabolites.^{24,25}

MetaPrint2D gives predictions based on metabolic transformation knowledge mined from the Accelrys Metabolite Database (version 2008.1)²⁶, which is a database consisting of single substrate-single metabolite transformations. Unlike a number of ligand-based SOM predictors reviewed in this chapter, MetaPrint2D does not solely focus on CYP metabolism prediction. Transformation reaction data (such as bonds broken/ formed) is contained in the bond annotations of structures contained in the Accelrys Metabolite Database. However, Adams realised that in some cases these mappings were incorrect. Atom-atom mapping between substrate and metabolite structures also exists in the database and the quality of these mappings were deemed better than the bond annotations but problems identified when using atom mapping to determine SOM for each transformation led to the development of MetaPrint2D's SOM labelling method based on identifying the maximum common substructure (MCS) between the substrate and metabolite, which does not take into account any annotations from the database.

2.1.3.1 Identification of Modified Atoms

When considering a substrate-metabolite pair, MetaPrint2D first attempts to determine whether one structure is entirely contained within the other, which would be the case if an addition or elimination reaction has taken place. This comparison is a simpler and much quicker than the maximum common subgraph-isomorphism problem and is therefore used as a first step to filter out simple transformations. If the transformation cannot be resolved by the simple comparison mentioned above, a constrained MCS search is carried out instead.

MetaPrint2D first compares the substrate-metabolite pair to determine whether the Murcko framework³⁴ (ring atoms and bonds plus any linker atoms and bonds) is conserved between the pair. This includes identifying the scaffolds in both substrate and metabolite structures. A substructure search is used to determine whether one is completely contained within the other when deciding if the scaffold/ ring(s) have been conserved. If a conserved scaffold is found, constraints are then placed on the atoms and bonds of the conserved scaffold. If the scaffold constraint is not met, then MetaPrint2D will attempt to identify conserved ring systems (rings sharing any atoms or bonds or lone rings) by identifying ring systems that are present in both substrate and metabolite structure of the pair under investigation. Once constraints on conserved scaffold, ring systems and rings have been generated, atoms that are not included in the conserved substructures are then allowed to

map onto atoms which are also outside of the conserved substructure when comparing the pair of substrate and metabolite.

After all constraints have been generated, MetaPrint2D then performs a MCS search using an algorithm based on the recursive backtracking algorithm³⁵ developed by Krissinel and Henrick. The algorithm iteratively selects an unmapped atom from a structure and attempts to identify a set of atoms it can map to on the other structure in the pair without violating the constraints already put in place by previously mapped atoms. Once the MCS between the structure pair has been identified, any atoms and/or bonds (including bond order) which are different between the substrate and metabolite structures are identified as reaction centres, i.e. SOM. When there was more than one potential best MCSs, the MCS containing the smallest number of reaction centres with the highest number of unchanged bonds was picked.

Once SOM have been identified on the substrate structures, MetaPrint2D stores information regarding the all atoms and the environments the atoms are present in. This information is stored along with the SOM and total occurrence counts of the atom environments. MetaPrint2D encodes information regarding atom environments in the form of a circular fingerprint (CFP), using SYBYL atom types³⁶ to describe each atom.

2.1.3.2 Selection of Transformation Data

Data in the Accelrys Metabolite Database are presented as transformations. These include single and multi-step transformations. For example, the reaction $A \rightarrow C$ in Figure 2.1 represents the overall transformation which is the result of two single step reactions: $A \rightarrow B$ and $B \rightarrow C$. All three transformations are presented in the Accelrys Metabolite Database as separate transformations. $A \rightarrow B$ and $B \rightarrow C$ are labelled as a “1 Step” transformation and the overall transformation $A \rightarrow C$ is labelled as a “2 Step” transformation.

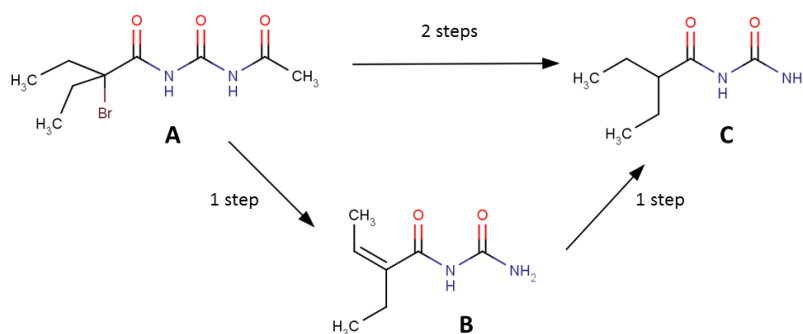


Figure 2.1 Transformations in the Accelrys Metabolite Database. The reactions $A \rightarrow B$, $B \rightarrow C$ and $A \rightarrow C$ were all present as separate transformations.

As all multi-step transformations are only summaries of single step transformations included in the database, only single step transformations are used to avoid data redundancy. However, this is not the only potential source of duplication.

There were transformations which contained the same substrate and the same metabolite. These cases could occur if the substrate of the transformation was an intermediate produced by a different starting parent compound and therefore resulting in the two transformations being present in different reaction schemes. There were also cases where the same compound could be metabolised differently, in the same or separate reaction schemes.

The data in Accelrys Metabolite Database is organised into reaction schemes based on a parent substrate compound. An example of reaction scheme for a parent substrate can be seen below:

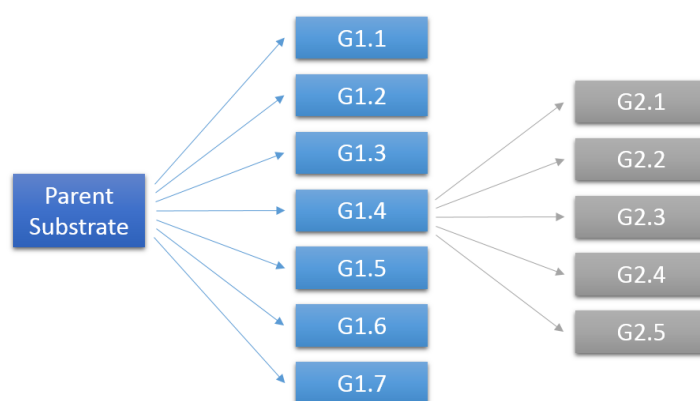


Figure 2.2 Example metabolic scheme contained in the Accelrys Metabolite Database. G1.x represent the first generation of metabolites and G2.x the second generation of metabolites of the parent substrate.

In an overall reaction scheme such as one presented in Figure 2.2, the SOM taking the parent substrate to G1.1 will be marked as a SOM once but also recorded as not being a SOM for the transformations producing the other first generation metabolites. The same issue would occur for any of the first generation metabolites that were further metabolised to produce second generation metabolites. Also, any duplicate transformations involving the same substrate and metabolite occurring in other metabolic schemes would cause the respective SOM to be recorded twice.

Each unique structure in the database can be identified by their ID (unique to each structure). MetaPrint2D has collated all SOM annotations regarding the same substrate structure from all transformations in the database into a single record and associated the SOM annotations with the unique substrate structure. This is termed the merged transformation dataset.

2.1.3.3 Sites of Metabolism prediction

MetaPrint2D is first trained on the Accelrys Metabolite Database. As part of the training, MetaPrint2D stores a list of distinct atom environments (described using CFP based on SYBYL atom types) along with SOM occurrence and substrate occurrence counts of the atom environment (i.e. its stability) computed from information gathered from the database. When carrying out SOM prediction, CFPs are computed for each atom of the query structure and compared to atom environments stored by MetaPrint2D during its training. The number of occurrences of each stored atom environment is compared against the total number of transformation reactions found in the database and the resulting normalised occurrence ratio is used as an indication of the likelihood of metabolism occurring in the given atom environment. The performance is dependent on the literature coverage of the training dataset.

The performance of MetaPrint2D was evaluated on models generated with the different pre-processing options, including the use of all transformations, excluding multi-step transformations, excluding duplicate transformations and merging of all transformation information for each unique structure. It was found that excluding multi-step transformations is the only data pre-processing step which consistently produced statistically significant improvement in MetaPrint2D's performance (Table 2.5).

Top1 %	Top3 %	Mean AUC	Median AUC
59.3	76.5	0.805	0.902

Table 2.5 Performance of MetaPrint2D with multi-step transformations excluded.

2.1.3.4 MetaPrint2D-React Extension

MetaPrint2D was built upon to create MetaPrint2D-React, a tool for the prediction of potential metabolite structures.²⁴ Some transformations in the Accelrys Metabolite Database contain reaction type information. 286 different types of reaction labels were found in the 2008.1 version of the Accelrys Metabolite Database, the most commonly seen reaction labels include C-hydroxylation, hydrolysis, C-oxidation, aromatic hydroxylation and aliphatic hydroxylation. However, not all transformations contain reaction labels and some transformations contain multiple types of reaction labels. There are also inconsistencies between labelling used for the same type of reaction between different releases of the database.

Given the inconsistency and incompleteness of the labels, Adams defined a set of SMIRKS patterns which was used to classify the reaction types separate from the database annotations. SMIRKS patterns were created for the most common reaction classes in the database along with common

reaction types reported in literature. General reaction rules were sought to describe the metabolic transformations in the database. For example, the SMIRKS pattern for a hydroxylation reaction ([*:1]>>[*:1]-[OH]) used a wildcard to represent the atom that was hydroxylated. Therefore, correctly classified instances of Hydroxylation, C-Hydroxylation, Aromatic Hydroxylation and Aliphatic Hydroxylation would all fall under this new hydroxylation category.

The definition of some reaction types used in MetaPrint2D-React were very broad, for example, the SMIRKS pattern for hydrolysis ([*:1]=[*:2]>>[*:1](-[OH])-[*:2]-[OH]) describes a generic hydrolysis but amide and ester hydrolysis would also fall under this SMIRKS pattern. Separate SMIRKS patterns were included for these hydrolysis types. The more specific rules were created to capture specific classes of reactions and the broader rules would capture the remaining instances. Substitution reactions were recognised as a combination of an addition reaction pattern plus an elimination reaction pattern. A list of the SMIRKS patterns used for this study is presented here:

#	Reaction Type	SMIRKS Pattern
0	Unknown	n/a
1	Ester Hydrolysis	<chem>[O:1]-[C\$(*=O):2]>>[*:1].[OH]-[*:2]</chem>
2	Amide Hydrolysis	<chem>[N:1]-[C\$(*=O):2]>>[*:1].[OH]-[*:2]</chem>
3	Thioester Hydrolysis	<chem>[S:1]-[C\$(*=O):2]>>[*:1].[OH]-[*:2]</chem>
4	Phosphorylation	<chem>[*:1]>>[*:1]-P(=O)(-O)-O</chem>
5	Dephosphorylation	<chem>[*:1]-[P\$(P(=O)(-O)-O)]>>[*:1]</chem>
6	Dehalogenation	<chem>[*:1]-[I,Br,Cl,F]>>[*:1]</chem>
7	Dehydrohalogenation	<chem>[*:1]-!:[*:2]-[I,Br,Cl,F]>>[*:1]=[*:2]</chem>
8	Chlorination	<chem>[*:1]>>[*:1]-Cl</chem>
9	Bromination	<chem>[*:1]>>[*:1]-Br</chem>
10	Fluorination	<chem>[*:1]>>[*:1]-F</chem>
11	Epoxidation	<chem>[*:1]=[*:1]>>[*:1]1-[*:1]-O-1</chem>
12	Epoxide Hydrolysis	<chem>[r:1]1-[r:1]-[Or:2]-1>>[*:1](-[OH])-[*:1]-[OH:2]</chem>
13	Epoxide Hydrolysis/Aromatization	<chem>[#6r:1]1-[#6r:2]-[#8r:3]-1>>[#6:1]:[#6:2]-[#8:3]</chem>
14	Epoxide Hydrolysis/Dehydration	<chem>[r:1]1-[r:1]-[Or:2]-1>>[*:1](-[*])=[*:1]-[OH:2]</chem>
15	Epoxide opening (+X)	<chem>[r:1]1-[r:1]-[Or:2]-1>>[*:1](-[*])-[*:1]-[OH:2]</chem>
16	Epoxide opening (3)	<chem>[#6r3:1]@[#8r3:2]>>[#6:1].[#8:2]</chem>
17	Epoxide Dehydration	<chem>[*:1]1-[*:1]-O-1>>[*:1]=[*:1]</chem>
18	Hydroxylation	<chem>[*:1]>>[*:1]-[OH]</chem>
19	Hydroxidation	<chem>[*:1]>>[*:1]-[O-]</chem>
20	Epoxidation/Hydrolysis	<chem>[*:1]=[*:2]>>[*:1](-[OH])-[*:2]-[OH]</chem>
21	Hydroxylation/Tautomerization (=O)	<chem>[*:1]=[*:2]>>[*:1](=O)-[*:2]</chem>

22	Hydroxylation/Tautomerization(=O=O)	[*:1]=[*:2]>>[*:1](=O)-[*:2]=O
23	Oxidation/Dehalogenation	[*:1]=[*:2]-[I,Br,Cl,F]>>[*:1]-[*:2]=O
24	Dehydroxylation	[*:1]-[O;H,-]>>[*:1]
25	Hydration	[*:1]=[*:2]>>[*:1](-[OH])-[*:2]
26	Dehydration	[*:1]-[*:2]-[OH]>>[*:1]=[*:2]
27	Amimation	[*:1]>>[*:1]-[NH2]
28	Nirosation	[*:1]>>[*:1]-N=O
29	Peroxidation	[*:1]>>[*:1]-O-[OH]
30	Sulfation	[*:1]>>[*:1]-S(=O)(=O)-O
31	Sulfuration	[*:1]>>[*:1]-[SH]
32	Sulfonation	[*:1]>>[*:1]-[S](=O)-[CH3]
33	Desulfuration	[*:1]=S>>[*:1]
34	Methoxylation	[*:1]>>[*:1]-O-[CH3]
35	Methiolation	[*:1]>>[*:1]-S-[CH3]
36	Cyanidation	[*:1]>>[*:1]-C#N
37	Oxidation(=O=O)	[*:1]>>[*:1](=O)=O
38	Oxidation(=O-OH)	[*:1]>>[*:1](=O)-[OH]
39	Oxidation(=O-O-)	[*:1]>>[*:1](=O)-[O-]
40	Oxidation(=O)	[*:1]>>[*:1]=O
41	Reduction(=O-O)	[*:1](=O)-[O;H,-]>>[*:1]
42	Reduction(=O)	[*:1]=O>>[*:1]
43	Aromatization	[*:1]-!:[*:2]>>[*:1]=[*:2]
44	Oxidation(-/=)	[*:1]-!:[*:2]>>[*:1]=!:[*:2]
45	Reduction(=/-)	[*:1]=[*:2]>>[*:1]-[*:2]
46	Oxidative Elimination	[*:1](-[*])-[OH:2]>>[*:1]=[O:2]
47	Esterification	[\$([OH]-C=O):1]>>[\$(O-C=O):1]-C
48	Azo cleavage	[N:1]=, #N>>[N:1]
49	Deamination (NH2)	[*:1]-[NH2]>>[*:1]
50	Deamination (NHNH2)	[*:1]-[\$([NH]-[NH2])]>>[*:1]
51	N-dealkylation	[#6:1]-[#7]-[#6]>>[#6:1]
52	Denitration	[*:1]-N(=O)-O>>[*:1]
53	N2-elimination	[*:1]-[\$(N#N)]>>[*:1]
54	N-Dearylation	[N:1]-C>>[N:1]
55	Acetylation	[*:1]>>[*:1]-C(=O)-[CH3]
56	Formylation	[*:1]>>[*:1]-[CH]=O
57	Acylation	[*:1]>>[*:1]-C(=O)-*
58	Demethylation	[*:1]-[CH3]>>[*:1]
59	Demethylation (x2)	[*:1](-[CH3])-[CH3]>>[*:1]

60	Dealkylation(1)	[*:1]-C>>[*:1]
61	Dealkylation(2)	[*:1]-[C:2]>>[*:1].[C:2]-O
62	Dealkylation(3)	[N,O:1]-C-[N,O:2]>>[N,O:1].[N,O:2]
63	Dealkylation(x2)	[*:1]-[#6]-[#6]>>[*:1]
64	Dealkynylation	[*:1]-C#C>>[*:1]
65	Methylation	[*:1]>>[*:1]-[CH3]
66	Alkylation	[*:1]>>[*:1]-[CH2]-[#6]
67	Oxidative deamination (=O-OH)	[*:1]-[N:2]>>[*:1](=O)-[OH]
68	Oxidative deamination (=O)	[*:1]-[N:2]>>[*:1]=O
69	Oxidative deamination (-OH)	[*:1]-[N:2]>>[*:1]-[OH]
70	Elimination (XH)	[#7:1]-[#6]-[#6]-[#7:2]>>[#7:1].[#7:2]
71	Elimination	[#6:1]-[#6:2]-[*]>>[#6:1]=[#6:2]
72	Aromatization/Elimination	[#6:1]-[#6:2]-[*]>>[#6:1]:[#6:2]
73	Elimination (XX)	*-[*:1]-[*:2]-*>>[*:1]=[*:2]
74	Ring opening	[#6:1]~@[#7:2]>>[#6:1]=[#8].[#7:2]
75	Condensation	[#7:1].[#6:2]=[#8]>>[#7:1]=[#6:2]
76	Tautomerization	[*:1]=[*:2]-[*:3]>>[*:1]-[*:2]=[*:3]
77	Rearrangement	[*:1]=[*:2]-[O-]>>[*:1](=O)-[*:2]
78	SS Reduction	[Sv2:1]-[Sv2]>>[SH:1]
79	Acetylcysteination	[*:1]>>[*:1]SCC(C(=O)O)NC(=O)C
80	Cysteamination	[*:1]>>[*:1]-SCCN
81	Protein Binding	[*:1]>>[*:1]SCC(N)C(=O)[#0]
82	CoA Binding	[*:1]>>[*:1]SCCNC(=O)CCNC(=O)C(O)C(C)(C)COP(=O)(O)OP (=O)(O)OCC1C(OP(=O)(O)(O))C(O)C(O1)n1cnc2c1ncnc(N)2
83	Glucosidation (+X)	[*:1]>>[*:1]C1C(O)C(O)C(O)C(CO)O1
84	Glucosidation (+OX)	[*:1]>>[*:1]OC1C(O)C(O)C(O)C(CO)O1
85	Glucuronidation	[*:1]>>[*:1]C1C(O)C(O)C(O)C(C(=O)O)O1
86	Glutathionation (+SX)	[*:1]>>[*:1]SCC(NC(=O)CCC(N)C(=O)O)C(=O)NCC(=O)O
87	Glutathionation (O>SX)	[*:1]-[#8]>>[*:1]SCC(NC(=O)CCC(N)C(=O)O)C(=O)NCC(=O)O
88	Glutathionation (=)	[*:1]=[*:2]>>[*:1]-[*:2]SCC(NC(=O)CCC(N)C(=O)O)C (=O)NCC(=O)O
89	Glycination	[\$(C=O):1]-O>>[C:1]-NCC(=O)-O
90	Glutamation	[C:1](=[O:2])O>>[C:1](=[O:2])NC(CCC(=O)O)C(=O)O
91	Glycosidation (+XP)	[*:1]>>[*:1]-C(C(O)C1(O))OC1COP(-O)(-O)=O
92	Glycosidation (+X)	[*:1]>>[*:1]-C(C(O)C1(O))OC1CO
93	Conjugation (+X)	[*:1]>>[*:1]-[#0]
94	Conjugation (OH>X)	[*:1]-[OH]>>[*:1]-[#0]
95	Conjugation (+SX)	[*:1]>>[*:1]-S-[#0]

96	Conjugation (=)	[*:1]=[*:2]>>[*:1]-[*:2]-[#0]
97	DNA Binding	[*:1]>>[*:1]c1=nc2c(=O)nc(N)=nc=2n1C1CC(O)C(CO)O1

Table 2.6 SMIRKS patterns used in this study.

MetaPrint2D-React performs prediction in a similar manner to MetaPrint2D. During training of MetaPrint2D-React, the same CFP method is used to describe atom environment. Instead of storing the occurrence of SOM, the occurrence of each reaction type observed for each atom environment was scored instead. When making a prediction on a query structure, the normalised occurrence ratio for reaction types are used to determine the stability of each query atom. The stability of the atom is the sum of all reaction type scores for that atom environment in the training data.

After the prediction of SOM along with associated reaction type information, MetaPrint2D-React also generates structures of metabolites by applying the reaction type SMIRKS (Table 2.6) to the query substrate structure. The performance of MetaPrint2D-React's ability to predict SOM (Table 2.7) and its ability to predict different reaction types were evaluated (performance for the 5 most and 5 least frequently seen reaction types given by Adams are presented in Table 2.8).

Top1 %	Top3 %	Mean AUC	Median AUC
58.9	78.7	0.812	0.918

Table 2.7 Performance SOM prediction by MetaPrint2D-React

Reaction Type	Count	Top 1 %	Top 3 %	Mean AUC	Median AUC
Hydroxylation	5726	47.20	69.30	0.804	0.891
Dealkylation	4975	71.80	87.50	0.895	0.994
Glucuronidation	4110	73.80	88.10	0.927	1.000
Demethylation	2340	86.60	95.10	0.928	1.000
Oxidation (=O)	2036	55.80	72.20	0.826	0.978
Bromination	7	50.00	70.00	0.810	0.810
Peroxydation	6	10.00	10.00	0.475	0.475
Deamination (NHNH2)	6	0.00	70.00	0.676	0.676
Rearrangement	6	90.00	100.00	0.931	0.931
Dealkynylation	4	80.00	100.00	0.702	0.702

Table 2.8 MetaPrint2D-React's performance in predicting the 5 most and 5 least frequently seen reaction types.

2.1.4 FAsT MEdabolizer

Fast Metabolizer (FAME) is a machine-learning based SOM predictor created by Kirchmair¹ which is built upon knowledge contained within the Accelrys Metabolite Database (version 2011.2)³⁷. Like MetaPrint2D, FAME does not only focus on CYP-mediated metabolism, but covers Phase I and II metabolism as well as providing specie-specific metabolism prediction models (human, rat and dog) as well as a global model containing all species found in the Accelrys Metabolite Database.

2.1.4.1 Data Preparation

FAME uses data exclusively from single-step transformations in the Accelrys Metabolite Database. The recursive backtracking search algorithm in MetaPrint2D^{24,25} were used to identify the MCS of a substrate-metabolite pair in a transformation from the database, giving the SOM of the substrate structure (see section 2.1.3.1). The merge function was also enabled when using MetaPrint2D to extract substrate structures from the database (see section 2.1.3.2), resulting in a dataset of unique substrate structures with all SOM annotations relevant to each substrate structure aggregated onto one structural record. MetaPrint2D-React was also used to compute the reaction types for all transformation and these were used to categorise transformations into Phase I and II. After all structures were extracted, the “Wash” function in MOE³⁸ was used to protonate strong acids and deprotonate strong bases.

2.1.4.2 Descriptors

FAME uses atom-based descriptors to encode properties of substrates and these descriptor values will then be evaluated by a machine learning model in order to generate SOM predictions. Therefore, the selection of descriptors is an important task. Different groups of descriptors investigated were the Chemistry Development Kit (CDK) atomic descriptors (CDK version 1.4.18), span-derived descriptors (Span2End descriptor and components³⁹ with added consideration for hydrogen atoms), SYBYL atom types³⁶ (determined by CDK SybylAtomTypeMatcher) and a revised implementation of the atomic fragment-based descriptors by Long and Rydberg⁴⁰.

All 2D atomic descriptors from CDK were considered and after the removal of descriptors not wholly applicable to the problem, as well as descriptors requiring long calculation time, 10 remained and were investigated. These are termed group A descriptors. The Span2End descriptor and its components³⁹ were chosen as an accessibility descriptor which provides a measure of the steric exposure of an atom to the catalytic components of a target enzyme. These are group B descriptors. SYBYL atom types encodes information regarding an atom’s element type and hybridisation state were also considered due to its effectiveness in MetaPrint2D^{24,25} (see section 2.1.3) and will be called

group C descriptor. Atomic fragment-based descriptors encode the properties (such as number of heavy atoms, rotatable bonds, hydrogen bond donors/ acceptors) of fragments separating an atom from the nearest end of branch of the structure. These are the group D descriptors.

Information gain for all group A, B and C descriptors were calculated using the InfoGainAttributeEval algorithm implemented in Weka version 3.6.9⁴¹, followed by creation and cross-validation of predictive models using different combinations of descriptors involved. The information gain scores calculated correlated well with the performance of predictive models and top seven scoring descriptors from group A, B and C were chosen (their definition and information gain scores are given in Table 2.9). Group D descriptors were appended to the chosen seven descriptors but their inclusion did not yield better performing models and therefore group D was discarded.

Descriptor	Group	Definition	IG
PartialTChargeMMFF94	A	Total partial charges of a heavy atom as derived from the MMFF94 model	0.0741
PartialSigmaCharge	A	Gasteiger–Marsili sigma partial charges in sigma-bonded systems	0.0661
PIElectronegativity	A	Pi electronegativity	0.0608
SigmaElectronegativity	A	Gasteiger–Marsili sigma electronegativity	0.0576
SybylAtomType	C	Sybyl atom type for a specific atom, encoding element type and hybridization state	0.0411
EffectiveAtomPolarizability	A	effective atom polarizability of a heavy atom	0.0180
MaxTopDist	B	maximum topological distance between two atoms of a molecule	0.0149

Table 2.9 Definition and information gain scores of descriptors chosen for the final model.
(Table adapted from FAME¹)

2.1.4.3 Sites of Metabolism Prediction

FAME is a machine-learning based SOM predictor, employing the use of a random forest model when carrying out SOM prediction. Random forest is a collection of decision trees, each trained with a subset of the training data. Classification by random forest is done based on the majority vote by decision trees within the forest. The use of a different number of decision trees was investigated and the SOM prediction performance of FAME increased with an increasing number of trees used, until reaching a performance plateau at 50 trees in the forest (based on the top-k measurement).

Different data balancing techniques were examined (reducing the number of data points from the majority class and oversampling the minority class in order to achieve balance), however, these all

lead to a drop in the model's predictive performance. Therefore, the original dataset was chosen for use instead.

2.1.4.4 Model Evaluation

Several models were created, including a global SOM model as well as specific models for human, rat and dog metabolisms. Metabolic phase specific models (Phase I and Phase II) as well as the combined model were created for all species, plus the global model. The performance of all models were evaluated using a 5-fold cross validation as well as with the use of three test sets:

1. Test set 1: created by a random split of the model's dataset into training and test datasets.
2. Test set 2: a subset of test set 1, containing test set 1 structures with a maximum Tanimoto similarity coefficient of 0.8 when compared to any structure within the training dataset.
3. Test set 3: same as test set 2, with the maximum Tanimoto similarity coefficient set at 0.5.

The evaluation of the prediction model's performance using these test sets should give an indication to the ability of the model to extrapolate into previously unseen chemical space. The results of FAME SOM prediction models created using both metabolic phases are given in Table 2.10.

The performance of the SOM prediction model created using the global dataset was also analysed to determine whether the model is biased toward predicting specific reaction types. Results show that prediction rates are not correlated with the number of instances of particular reaction types in the training dataset. This is possibly due to the fact that some reaction types occur only in very specific chemical environments and therefore are comparatively easy to predict.

The performance of FAME has also been compared to MetaPrint2D^{24,25}, another SOM predictor not limited to a particular family of metabolising enzyme, was carried out for the global dataset as well as the human specific models. MetaPrint2D was retrained with the same training dataset used by FAME and testing was performed with test set 1, 2 and 3 for both FAME and MetaPrint2D. In all models evaluated, FAME showed 10-20% higher top-*k* scores for all datasets evaluated. FAME also showed stronger extrapolating ability compared to MetaPrint2D. Comparison of FAME with other CYP specific SOM predictors was also carried out, despite the fact that FAME offers SOM predictions for a wide range of metabolic enzymes and both Phase I and II transformations. It was found that FAME offers competitive accuracy when compared to CYP specific models.

Metabolic phase	Species	Top <i>k</i>	5-CV	Test set		
				1	2	3
Phase I + II	all	1	0.70	0.71	0.361	0.57
		2	0.81	0.81	0.76	0.78
		3	0.87	0.87	0.83	0.85
Phase I + II	human	1	0.69	0.80	0.58	0.55
		2	0.80	0.80	0.74	0.73
		3	0.86	0.87	0.82	0.83
Phase I + II	rat	1	0.68	0.70	0.61	0.53
		2	0.80	0.81	0.75	0.72
		3	0.86	0.87	0.83	0.79
Phase I + II	dog	1	0.60	0.64	0.56	0.56
		2	0.74	0.75	0.69	0.72
		3	0.82	0.84	0.80	0.82

Table 2.10 Performance of FAME models. Top *k* shows the number of top-ranked atom positions considered for prediction success. 5-CV shows the 5-fold cross-validation rates.

2.1.5 Summary

A number of ligand-based SOM prediction as well as metabolite prediction methodologies have been reviewed here. A significant number of SOM prediction methodologies concentrated on providing prediction only for transformations catalysed by the CYP family. The metabolite prediction methods discussed do not have this limitation, but instead are knowledge-based methods providing transformation predictions based on rules.

Of the SOM prediction tools reviewed, FAME and MetaPrint2D cover a more diverse chemical space than a number of SOM predictors that have been reviewed in this chapter. By expanding the ability to carry out SOM prediction to cover multiple metabolic phases as well as being non-enzyme family specific, SOM predictors such as FAME and MetaPrint2D can offer medicinal chemists a more extensive, and perhaps more useful insight into the metabolic vulnerability of their compound during the drug discovery process. Once metabolic lability or other undesirable properties have been identified in a compound, methodologies such as bioisosteric replacements can be used to improve the compound. A review of *in silico* bioisosteric replacement methodologies is provided the following section.

2.2 Bioisosterism

Drug discovery and lead optimisation is an iterative process and often requires the simultaneous improvement of multiple parameters that are problematic whilst keeping the desired characteristics that the original compound already processed. The usage of bioisosteric replacement techniques are particularly powerful in this case and have been extensively used by medicinal chemists to identify potential sub-structural replacements which would reduce undesirable properties in a compound whilst retaining other desirable characteristics. There are a number of classical, textbook examples of bioisosteric analogues (for example, Table 2.11), however, these are not always applicable. The replacements of more complex or unusual groups would also require iterations of trial and error, even with experienced medicinal chemistry expertise, before an acceptable substitution could be found, particularly in cases where activity cliffs, receptor flexibility and multiple binding sites may come into play. One of the key advantages of *in silico* methods is the ability to calculate and consider (estimate) many important properties in ligand-target interaction which have to be mimicked by any replacement group. This makes *in silico* methods particularly suitable for dealing with complex group replacements.

Group	Examples
Carbonyl Group	
Carboxylic Acid	

Table 2.11 Examples of well-known bioisosteric replacements

Similar to SOM prediction methodologies, the tools used to identify potential bioisosteric groups can generally be categorised into ligand-based and structure-based, where ligand-based methods extract information only from the ligands themselves and structure-based methods obtain knowledge from the targets. As the focus of this study is on ligand-based methods, structure-based methods will not be discussed here.

The term scaffold hopping was first introduced by Schneider in his attempt to identify “isofunctional molecular structures with significantly different molecular backbones.”⁴² Nowadays, scaffold hopping is often considered in conjunction with bioisosteric replacement methods and aims to replace the core (scaffold) of the molecule with another structurally distinct core whilst retaining the

biological effect produced by the original core. There are methodologies that have been developed with the aim of scaffold hopping, however, scaffold hopping itself is not a separate technique, but rather a subset of the bioisosteric replacement problem, which different bioisosteric replacement methods address to a greater or lesser degree. It will be discussed here along with other ligand-based approaches. Scaffold hopping is used for the same purposes as ‘conventional’ bioisosteric replacement methods, to improve potency, absorption, selectivity etc. but also has the added benefit of potentially discovering new compounds with analogous biological effects to existing compounds, but which are novel and are not covered by existing patents.^{43–48}

2.2.1 Ligand-based Methods

Ligand-based approaches have generally be categorised into similarity-based approaches and knowledge-based approaches (data mining)^{47,49,50}. The idea behind similarity-based approaches is that similar structures will have similar biological effects. In the context of bioisosteric replacement, an exact chemical structural match is undesirable. Instead, a query is often characterised by descriptor values, which are used to query other compounds containing similar descriptor values in the database, resulting in a list of structures that are similar within the descriptor space but that do not have the same chemical structure as the query. The aim of all similarity-based approaches is to choose a description of the structure/fragment that correlates well with biological activity and to adopt an appropriate similarity measure.

While similarity-based approaches attempt to determine bioisosteric pairs from first principles, knowledge-based approaches attempt to extract relevant information contained in a collection of chemical or biological data repositories. Knowledge-based approaches are retrospective analyses that aim to find chemical transformations in large structural repositories and associate them with the induced change in biological properties.⁵¹ However, as these methods were built on databases of historical data, knowledge-based approaches will never be able to return truly novel bioisosteric replacement suggestions, unlike similarity-based approaches. As the studies reported in this thesis are similarity-based methods, knowledge-based approaches will only be mentioned briefly here.

2.2.1.1 Similarity-based Approaches

2.2.1.1.1. Physiochemical Property Methods

Many methods were developed over the years that utilise physiochemical properties calculated by *in silico* methods to identify potential bioisosteric replacement groups using a ligand-based approach. These were developed based on numerous existing *in silico* methods created to calculate

physiochemical properties and the availability of large amounts of bioactivity data (public,^{52,53} proprietary^{37,54} and private databases). The WWW-based molecular modelling system at Novartis was one of the first methods to make use of *in silico* methods to calculate substituent parameters on a large scale. Properties relevant to ligand-target interactions (logP, molar refractivity, electron-donating power, electron-withdrawing power, heavy atom count and maximum topological length) were identified and calculated for a large number of substituents (80,000,⁵⁵ later extended to 850,000⁵⁶). These values are stored in a database, and a user can submit a substituent that has the same descriptor calculation as a query which will be used to identify potential bioisosteric substituents.

A 2D R-group descriptor⁵⁷ with certain analogies to the WWW-based molecular modelling system described previously was developed by Holliday *et al.* to calculate the similarities between substituents. The descriptors selected were based on the work done by Martin *et al.*,⁵⁸ originally designed to measure diversity in combinatorial libraries. The R-group descriptors were calculated for substituents with a maximum topological distance of six bonds away from a fixed point, such as the attachment point of the substituent to the molecule's core. Seven R-group descriptors, encoding atomic weight, hydrophobicity, molar refractivity, atomic charge, polar surface area, hydrogen bond donor and hydrogen bond acceptor information, were calculated for each substituent. These descriptor results could be combined or used separately, and together they describe the distribution of the selected atomic properties at up to six bonds away from the attachment point. This method was originally intended for combinatorial library designs which was then found suitable for identifying bioisosteric replacements.

The different studies and methods outlined above all attempted to identify bioisosteric substitutions by finding groups with similar combinations of volume/topology, electronic and steric properties. Kier and Hall presented general guidance for selecting bioisosteric groups which can be employed when modifying a compound by adding or replacing a group, modifying or replacing rings and linking fragments within the molecule.⁵⁹ The 2D method analysed compounds based on three characteristics:

1. Volume: the size, shape and steric properties of the group, estimated from the number of σ , π and lone pair electrons
2. Electrotopological property: the electron accessibility of the atom, taking into account the atom's topological position
3. Hydrophobic effect: the influence of the group on surrounding water molecules

These properties were plotted against each other in 2D plots for ease of visualisation and to aid identification of potential bioisosteric replacement groups. The advantages of this approach are the

rapidly available descriptor values, the interpretability of the results and similarly to the methods described above, the reduction of dimensionality which makes interpreting the descriptor values easier.

Devereux *et al.* took a similar but different approach by incorporating quantum mechanical descriptors in the description of potential bioisosteric groups. The web-based tool Quantum Isostere Database is designed to find bioisosteric replacement groups using pre-calculated *ab initio* descriptor values.⁶⁰ Conformers are generated for all fragments before the following descriptors are calculated:

1. Shape: conformation, size, shape, charge and spatial distribution of charges and properties
2. Electronic: atomic and fragment multipole moments and polarity
3. Surface properties: electrostatic potential and local ionisation energy
4. H-bonding: donor and acceptor strengths
5. Others: atom and bond properties, electron delocalisation and bond order

Fragments have to be aligned before similarity assessment can be made. A significant amount of computational time was used while creating the database of *ab initio* descriptor values which were then stored and accessed via a web interface to allow for bioisosteric replacement predictions.

Birchall *et al.* implemented a method which utilised reduced graphs (RGs) to identify bioisosteric replacements.⁶¹ A structure was first fragmented by recursively cutting along non-terminal, acyclic single bonds, with the exception of acyclic sp³ carbon to acyclic sp³ carbon bond, acyclic heteroatom to acyclic heteroatom bond and acyclic heteroatom to acyclic sp² carbon bond. Each fragment produced was represented by one reduced graph node (Figure 2.3). When calculating the similarity between two RGs, the MCS of the two RGs was identified first and the Dice coefficient obtained.⁶¹

Reduced Graph Nodes			
Aromatic negatively ionisable	Aliphatic negatively ionisable	Acyclic negatively ionisable	<div>High priority</div> <div>Features</div> <div>Low priority</div>
Aromatic positively ionisable	Aliphatic positively ionisable	Acyclic positively ionisable	
Aromatic joint donor-acceptor	Aliphatic joint donor-acceptor	Acyclic joint donor-acceptor	
Aromatic donor	Aliphatic donor	Acyclic donor	
Aromatic acceptor	Aliphatic acceptor	Acyclic acceptor	
Aromatic featureless	Aliphatic featureless	Acyclic	
<div>High priority</div> <div>Structures</div> <div>Low priority</div>			

Figure 2.3 Types of reduced graphs along with the order of importance of each structure type and feature type. If more than one structure or feature type were present in a single node, the resulting node will be labelled according to the order of priority shown.

The authors found that a significant portion of bioisosteric fragments obtained from BIOSTER⁶², which contains literature examples of bioisosteric replacements, were encoded in a single RG node.

On the other hand, this discovery meant the approach did not (and would not be expected to) perform well as a scaffold hopping method.

2.2.1.1.2. Pharmacophore Methods

Aside from methodologies concentrated on physiochemical properties, pharmacophores also play a role in identifying bioisosteric replacement. A pharmacophore is defined as “an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response.”⁶³ The relevance of pharmacophores in ligand-target interaction lead to the development of numerous methods aimed to identify bioisosteric replacement groups based on the similarities between their pharmacophores.

Gridding and partitioning (GaP) was developed to characterise monomers present in combinatorial libraries by identifying the similarities based on positions occupied by pharmacophores in 3D space.⁶⁴ The method aimed to identify gaps in libraries in order to aid library design and monomer acquisition. This same method, however, has also been used in the context of identifying bioisosteric pairs.⁶⁰ 2D monomer structures were converted into 3D low-energy conformers, the connection point of the monomer was placed at the origin of a 3D Cartesian coordinate space and the monomer rotated about the X-axis. The cells in the Cartesian coordinate space occupied by a pharmacophoric group during the rotation were recorded and the ‘hits’ converted into a fingerprint. Monomers with a higher number of flexible bonds were penalised. GaP collected information regarding the 3D occupancy of pharmacophoric properties contained in a monomer. The authors argued that this is more directly related to the properties relevant to molecular recognition than information gathered based on 2D molecular graphs.

Whilst some methods, such as the R-group descriptor and GaP, concentrated on substituent similarities, work was also carried out to assess similarities between the core/scaffold part of the molecule, often with the intention of identifying potential scaffold hopping candidates. Lewell *et al.* compiled information on rings and scaffolds from structures contained in corporate and commercial databases to produce the Drug Rings Database which could be accessed via a web-interface.⁶⁵ A total of 5.5 million compounds (191,000 unique rings) were investigated, which included proprietary compounds, commercially available compounds, late-stage development compounds and natural products. Structures were first fragmented to obtain the rings substructures with attachment points marked. The 3D structure of two- and three-connection rings were then generated using CONCORD^{66,67} and CORINA¹⁹, and then descriptors for the rings were calculated. These included counts of properties (such as hydrogen bond donor/acceptor counts, acids and bases counts, number of connections and number of rings), logical parameters (if structure contained fused rings

and spiro structure) and numerical parameters (such as molecular weight and frequency of ring presence in databases). The database search capability included descriptor value range search, structural similarity (using the Tanimoto coefficient), SMARTS structural search, exact match structure search and data look up. This application could be used for idea generation and scaffold hopping for medicinal chemists during lead optimisation.

Stiefl *et al.* presented a 2D method which combined reduced graph and pharmacophore property pairs to produce a pharmacophore-type node descriptor (ErG).⁶⁸ The ability to identify scaffold hopping pairs using this method compared to DAYLIGHT fingerprints, which encodes the presence and absence of a pre-defined list of sub-structural elements, was highlighted. The reduced graph was generated using the following protocol:

1. Hydrogen bond donor and acceptor labels were assigned (structures charged at physiological pH)
2. Terminal hydrophobic features with three heavy atoms (as well as thiol groups) were encoded as 'endcap' groups
3. Ring systems were abstractified and each ring's centroid was assigned either an aromatic or hydrophobic flag

The features on the reduced graphs were then each converted into property points (e.g. H-bonding, charge, endcap) then into a binary fingerprint with each bit position representing the presence or absence of: [Property Point 1] – [topological distance] – [Property point 2]. Compounds spanning 11 activity classes from the MDL Drug Data Report database (MDDR) were used and retrieval rates for active compounds for the top 1% using ErG was compared to DAYLIGHT fingerprints; ErG was seen to outperform DAYLIGHT fingerprints in 10 out of 11 classes.

Wagener *et al.* created another 2D, interactive method for identifying bioisosteric replacements.⁶⁹ Topological pharmacophore fingerprints were used to describe a database of 700,000 fragments, with a maximum of 12 heavy atoms per fragment. An atom pair description in the form of [pharmacophore] – [topological distance] – [pharmacophore] was used and eight pharmacophore properties were chosen for the fingerprint:

1. Attachment point
2. Hydrogen bond donor
3. Hydrogen bond acceptor
4. Hydrophobe
5. Conjugated atom
6. Aromatic atom
7. Positively charged atom
8. Non-hydrogen atom

These were transformed into a topological fingerprint by enumerating all possible pairs of atoms and their pharmacophore properties. The final version of the fingerprint, which was implemented as an intranet tool, allows for a maximum of 3 occurrences of any atom pair containing an attachment point to be accounted for. The user of the tool had the option to ‘fuzzify’ the fingerprint, which allowed for the same atom pairs on two fragments that differed by one bond distance to have one bit in common rather than none. If the atom pairs and distances matched exactly, the fragments share three bits in common. Both Euclidean and Soergel distances ($= 1 - \text{Tanimoto}$) were investigated as a similarity measure in this study. Euclidean proved to be a better method in this context and the authors speculated that this was due to the significant importance of the absence of pharmacophore groups as well as their presence. Certain analogies can be seen between this method and Stiefl’s ErG method; both methods produced a fingerprint which described the topological distance between a pair of pharmacophores.

Much like the methodologies reported from SOM prediction, there are also a number of molecular shape based methods and methods focused on fields and molecular potentials. These methods all require the 3D geometry of a molecule and are outside of the scope of studies reported in this thesis, therefore will not be discussed in detail here. It is worth noting that if the active, bound conformation of a ligand is not available (no crystal structures with bound ligand), low energy conformers are often generated using methods such as CORINA¹⁹ before fields and potentials are calculated. This inherently introduces error into the starting structures entered for calculation. It has been noted that for scaffold hopping (and therefore likely for general bioisosteric replacement) in the absence of the active, bound conformation of ligands, there are no large discrepancies between the effectiveness of 3D methods compared to 2D methods. However, when used in retrospective studies where the active conformation of a ligand is known (such as from crystal structures), 3D methods are reported to perform better than their 2D counterpart.⁴⁵

An interesting study was carried out by Schuffenhauer *et al.*, when they compared the effectiveness of 2D molecular substructure fingerprints and 3D field-based similarity searching methods (FBSS) in identifying bioisosteric replacements.⁷⁰ The UNITY 2D fingerprint from TRIPOS was used in this study. The fingerprint encoded the presence or absence of a set of pre-defined structural patterns using 992-bit binary vectors. FBSS utilised a genetic algorithm to identify the best alignment between two structures. The Carbò similarity index was used to assess how well the electrostatic, steric and hydrophobic fields overlap individually and in combination. Structures from the BIOSTER database were used to compare the effectiveness of the two methods. The 2D UNITY fingerprint returned more bioactive molecules, however the 3D and computationally more demanding FBSS method

returned more structurally diverse molecules. It was demonstrated that both methods were capable of identifying similarities between bioisosteric pairs but each have different priorities. UNITY fingerprints proved to be particularly sensitive to heteroatom substitutions. FBSS was less concerned with the atom types but was sensitive to fields projected by the atoms and therefore heavily dependent on the accuracy of the 3D structure presented for similarity calculations. Similarity values produced by UNITY fingerprints were combined with FBSS similarity results using data fusion. The authors concluded that as both methods were capable of identifying some aspects of the bioisosteric relationship, the effectiveness of a similarity search could be improved by combining the different similarity measures.

A number of similarity-based approaches that have been discussed here attempted to identify bioisosteric replacements based on similar physiochemical properties, pharmacophores, molecular topology or a combination of these properties. These bioisosteric replacement methods utilised the similarity principle and assumed that structures which are similar must also have similar biological effects. Some of these methods were developed with the intention of aiding virtual screening or library design but have been found to be equally useful as a method for the identification of bioisosteric pairs.

2.2.1.2 Knowledge-based Approaches

The expert system Example-Mediated Innovation for Lead Evolution (EMIL) was one of the first automated procedures built to identify bioisosteric replacements from existing knowledge.⁷¹ EMIL extracted observed structural modifications to form empirical bioisosteric transformation rules and apply these rules to new query structures. These modifications were mined from sequences of structural modification patterns for drugs and pesticides in several classes.

BIOSTER (a database of bioisosteres) was produced around the same time as EMIL, and whilst the two approaches were developed independently of each other, they are also fairly similar. The first version of BIOSTER (1992) contained 479 bioisosteric pairs obtained from literature.⁷² The database was updated regularly and is currently available from Digital Chemistry. Primary and secondary literature on medicinal chemistry, pesticide chemistry and bioorganic chemistry continued to be added to BIOSTER and the 1997 version of the database contained more than 1500 transformations.⁶² The literature coverage of BIOSTER has now expanded to contain general biochemistry, pest management, prodrugs, propesticides, fragrances, natural products and their synthetic analogous, resulting in over 30,000 bioisosteric transformations in BIOSTER version 15.1.⁵⁴ The database contains bioisosteric pairs, using a broader definition of bioisosteres which include bioanalogs, defined as “molecules or groups that, in the context of a given biological parameter,

elicit analogous responses".⁷³ The database was also developed using 2D structures as authors pointed out that most searchable databases used in drug discovery utilise 2D structures as they are computationally less demanding than their 3D counterparts as well as being visually informative. Unlike EMIL, which presents rules for transformations, BIOSTER presents structurally analogous pairs with interchangeable fragments as a bioisosteric transformation, where the exchanges of the bioisosteric groups were depicted in a style analogous to traditional chemical reactions. A large number of different types of bioisosteric transformations can be found in the database and its comprehensiveness of the database is one of the reasons that BIOSTER have also been used in multiple studies as a validation dataset during the development of other bioisosteric replacement methods.^{57,69,70,74,75}



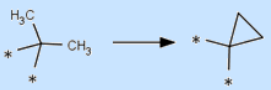
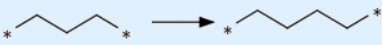
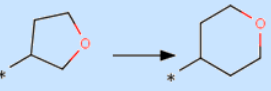
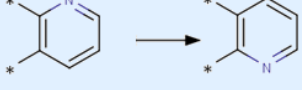


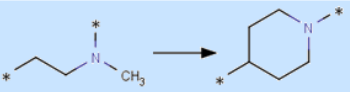
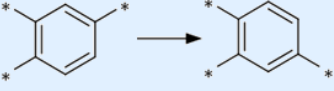

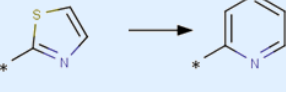
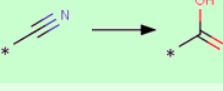
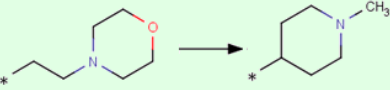
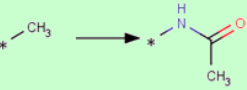
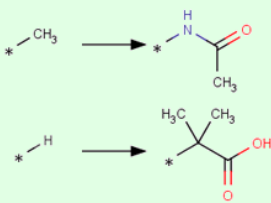
Drug Guru (drug generation using rules) is a web-based software program produced by Stewart *et al.* which contains a collection of 'rule-of-thumb' transformation strategies that were employed by medicinal chemists in previous projects, and which could be suggested for application on new structures.⁷⁶ The collection of bioisosteric transformations in Drug Guru was extracted from publications and stored in SMIRKS format. The original version of Drug Guru contained 186 rules, derived from a traditional medicinal chemistry knowledge base, which were either functional group exchanges or molecular framework modifications. This was later extended to 530 transformations in the 2011 version of Drug Guru through an on-going manual curation of data obtained from the primary literature and proprietary studies. Transformations suggested by Drug Guru may also contain supporting examples of the transformation, such as the (proprietary) historical use of the transformation and unpublished knowledge regarding the transformation of interest. Automatic iterative application of transformations on newly generated structures is included as an option in the program, as well as ranking of new molecules based on calculated physiochemical properties.

Drug Guru is similar to BIOSTER, the main difference between the two being that Drug Guru compiles a list of general rules rather than explicit examples in BIOSTER. EMIL is similar to BIOSTER in that each entry in the database serves as a 'rule' instead of having changes categorised into transformation types.

2.2.2 Metabolic Stability & Bioisosteres

There have been few bioisosteric replacement methodologies which are specifically focused on metabolic stability. However, a notable data mining study was carried out by Papadatos *et al.*⁷⁷ on the Lilly metabolic stability assay with experimental values from human microsomal metabolic stability measurements. The database held 43,340 structures and their SMILES strings (2D structures) were used in this study. The authors employed the popular method developed by Hussain and Rea⁷⁸ to exhaustively generate all possible MMPs without supervision from the database structures, with a maximum of 14 heavy atoms in any MMP substructure. The smallest transformation was kept from each pair of structures. These transformations would have to contain fewer atoms than the MCS, i.e. the rest of the structure. The chemical contexts of the resulting MMPs were described on a local and global level. The global context descriptor involved a whole molecule description using Murcko frameworks³⁴, which gave an abstract description of a molecule's rings and linkers along with their atom and bond types. A local context descriptor based on SYBYL atom types³⁶ of neighbouring atoms up to a distance of three bonds away was then used to describe the local environment around the attachment points. This local context descriptor provided an increasingly more detailed description of the transformation compared to the Murcko frameworks. Aside from the descriptors, each pair of MMPs also had their pairwise property differences (ΔP) assigned: favourable ($\Delta P > \text{threshold}$), unfavourable ($\Delta P < \text{threshold}$) and no effect ($-\text{threshold} \leq \Delta P \leq \text{threshold}$). The threshold for this study was set at 25% (of metabolised compounds) but could be altered for other assays and properties. The inclusion of the context descriptors by the authors was intended to improve upon the MMP approach where it was assumed that differences in properties must be a result of the transformation, regardless of the surrounding context of the exchange.

832,037 distinct transformations were identified and 424 of these occurred more than 30 times. The skewness of transformation occurrences found was not unexpected as this appeared to be a common trend found in MMP studies^{79–81}. The most frequently observed replacements were small transformations. The top 20 most common transformations involved a maximum of four heavy atoms, with the exception of two regio-specific phenyl replacements. 370 of the 424 transformation with over 30 occurrences were terminal side chain transformations. The authors examined transformations that had minimal effect on metabolic stability, the transformations with a maximum of 25% ΔP in either direction (which could therefore be considered as bioisosteres), as well as transformations that brought about the largest increase or decrease in metabolic stability (shown in Table 2.12). If the direction of transformations which brought about a decrease in metabolic stability was reversed, they should instead bring about an increase in metabolic stability, an example being the removal of metabolically labile *n*-butyl groups and *ortho*-fluorophenyl.

Transformation	Count	Mean ΔP	% neutral	% bad	% good
	31	0.9	96.8	0.0	3.2
	44	0.6	95.5	2.3	2.3
	65	-0.9	96.8	0.0	4.6
	58	2.2	95.5	5.2	0.0
	54	0.2	95.4	1.9	3.7
	30	7.2	94.8	6.7	0.0
	72	-1.5	91.7	4.2	4.2
	33	3.8	90.9	9.1	0.0
	33	-3.5	90.9	0.0	9.1
	31	0.4	90.3	3.2	6.5
	41	0.0	90.2	2.4	7.3
	71	0.5	90.1	5.6	4.2
	38	-41.9	36.8	0.0	63.2
	32	-34.7	37.5	0.0	62.5
	30	-23.5	50.0	0.0	50.0
	35	-19.9	71.4	0.0	28.6

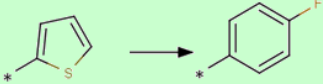

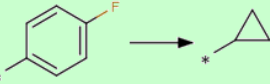
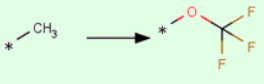
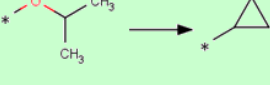
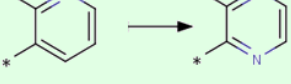



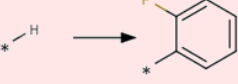
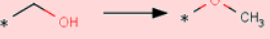
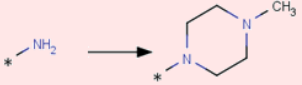



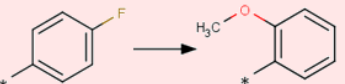
	35	-12.0	74.3	0.0	25.7
	38	-14.3	76.3	0.0	23.7
	36	-12.5	77.8	0.0	22.2
	42	-9.4	83.3	0.0	16.7
	31	-7.1	83.9	0.0	16.1
	43	-8.6	86.1	0.0	14.0
	32	32.8	50.0	50.0	0.0
	36	24.8	58.3	41.7	0.0
	57	19.4	59.7	40.4	0.0
	34	19.8	61.8	38.2	0.0
	45	17.6	62.2	37.8	0.0
	31	21.1	64.5	35.5	0.0
	36	25.3	66.7	33.3	0.0
	35	7.0	80.0	20.0	0.0
	31	7.7	80.7	19.4	0.0
	31	10.3	80.7	19.4	0.0

Table 2.12 Top 12 most neutral (blue), top 10 most favourable (green) and top 10 most unfavourable (red) transformations. (%bad = % of transformations with $\Delta P < -25\%$, % neutral = % of transformations with $-25\% \leq \Delta P \leq 25\%$ and % good = % of transformation with $\Delta P > 25\%$)

When contextual information of transformations was considered, it was discovered that some well-known examples of bioisosteric replacement pairs (e.g. $H \leftrightarrow F$) exhibit a metabolic stability profile that was highly dependent on the chemical context the substitution occurred in. This result was also discovered in Wassermann's large-scale MMP study on ChEMBL structures.^{80,81}

The $NH_2 \rightarrow OH$ substitution (Figure 2.4.a) was observed 155 times, with around 25% examples bringing about an improvement in metabolic stability. However, if the substitution was found in the context shown in Figure 2.4b, 60% of the 42 contributing examples brought about an increase in metabolic stability. The $H \rightarrow$ mesyl substitution (Figure 2.4c) was seen to mostly bring about minimal change in metabolic stability in 147 examples. However, when found in the context seen in Figure 2.4d, the majority of the 27 examples where this was observed lead to a decrease in stability instead. This was also found to be true for the $H \rightarrow F$ transformation (Figure 2.4e) where the majority of examples saw very little change in the stability but when found in the context of Figure 2.4f, most examples saw a larger than 25% decrease in metabolic stability after the transformation was applied. Figure 2.4 contained the few examples the authors provided of transformations where their effects on metabolic stability were heavily context dependent. The authors emphasised that results provided by this study were statistical in nature but could be used to generate new ideas for medicinal chemists, which might otherwise have been missed.

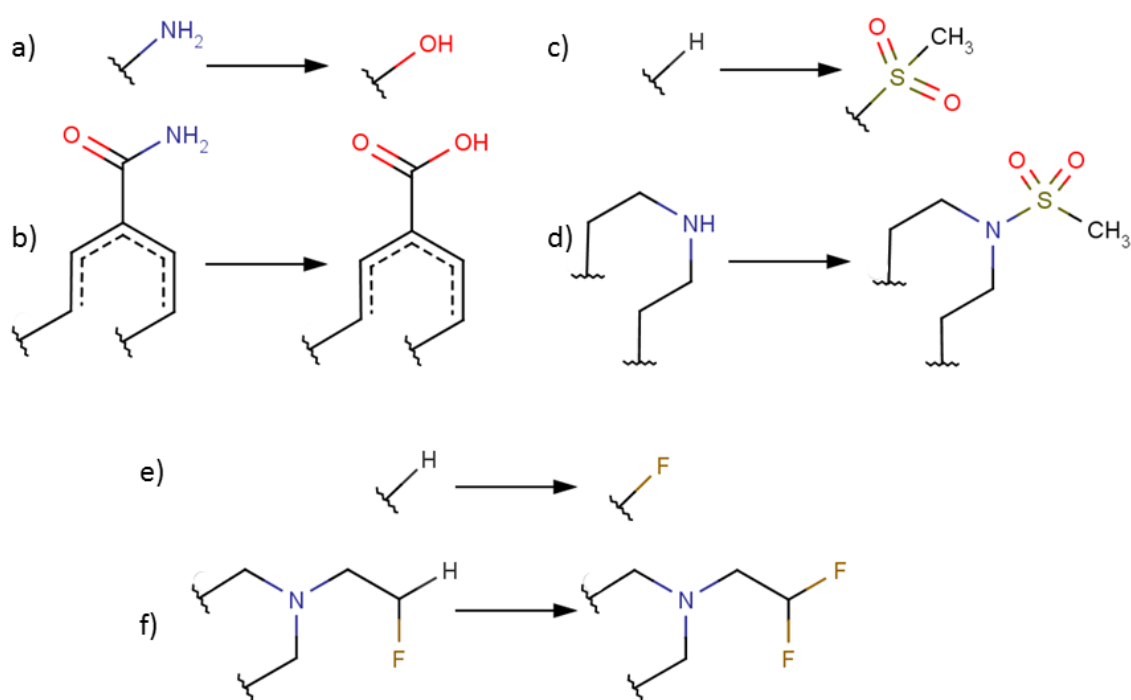


Figure 2.4 Context dependent transformations

The methodology used in the study described here was originally developed by Papadatos *et al.* for use in analysing MMPs extracted from the GlaxoSmithKline ADME datasets containing hERG, solubility and lipophilicity data.⁷⁹ In this original study, a MCS approach (Daylight Toolkit) in identifying MMPs was compared to the Hussain and Rea fragment indexing algorithm. The results produced by both were very similar but the MCS approach took 4.5 days on a pre-filtered dataset whereas the fragment indexing algorithm required only 8 hours on the same machine without the pre-filtering requirement. Different types of global context descriptors (RGs, Murcko frameworks and Daylight fingerprints) and local context descriptors (localised RGs and SYBYL atom types based atom environments) were also considered. This study demonstrated that the additional information could provide more refined information about the suitability of a transformation for a particular problem. This study, along with its application on the metabolic stability dataset, further demonstrated that some methodologies and approaches are transferable between different projects and end points when used correctly.

2.2.3 Summary

A number of ligand-based methods to identify bioisosteric replacements have been reviewed here. The studies described are by no means an exhaustive list of methods available in the literature, rather a sample of different methodologies that have been employed to locate bioisosteres. Similarity-based approaches have the advantage of being able to identify novel replacements that were previously unseen in other studies. However, this also means that the replacements suggested may have undesirable properties, such as O-O bonds and synthetically inaccessible structures. It is possible that after the generation of new structures based on replacement suggested, filters such as the synthetic accessibility filter⁸² can be applied to remove undesirable compounds from the results. Knowledge-based methods, on the other hand, do not usually have this issue as results returned come from synthetic projects where compounds have been previously synthesised. Whilst being resilient to returning nonsensical results, these methods will not yield any truly novel suggestions, unlike similarity based methods. A small selection of knowledge-based methods have been reviewed here to give an idea of the datasets that can be used for validation of a new bioisosteric replacement methodology.

2.3 Conclusion

This chapter provided a review of different SOM prediction and bioisosteric replacement methodologies that have been reported in the literature. A summary of 2D and 3D methods have been presented here. Using either representation of a ligand has its advantages and disadvantages. 2D methods are usually faster, compared to their 3D counterparts. The rapid speed of 2D methods means it is possible to produce interactive methods where real-time exploration of results was possible (where 3D methods failed). Avoiding 3D structures also circumvented the error introduced when no active bound conformation of the ligand was available and conformations had to be generated instead. The utilisation of incorrect conformers and inappropriate conformation generation can result in poor performance of 3D methods.⁸³

Validation statistics of methods based on 2D structures have shown that they can perform as well as methods utilising 3D structures, although if a binding mode or bioactive conformation of a ligand was available, 3D methods were often found to perform better.⁴⁵ This was not surprising as 3D methods arguably capture the interaction between ligands and receptors more accurately. However, given that the active conformations of ligands are not always known (sometimes not even the structure or identity of the target), fast interactive 2D methods could be used as a starting point for idea generation and 3D methods for refinement once the active conformations of targets and ligands are known. As was previously noted, 2D and 3D studies could provide complementary results⁸⁴ and results from both should be considered together. The performance gap between 2D and 3D methods suggests that current methodologies are still incapable of identifying the active conformation of ligands in the absence of experimental data. There are still also gaps in determining the relative importance of different features in the ligand-target interaction.

Given the importance of optimising all desirable properties of a structure during lead optimisation, the following chapters will describe development of methodologies to identify metabolically vulnerable regions of a molecule and attempts to find suitable bioisosteric replacements to improve other properties of the compound (such as bioavailability) whilst maintaining the compound's metabolic stability profile. The SOM prediction methods aim to provide medicinal chemists with a broad overview of the metabolic vulnerability of compounds and therefore will not be restricted to any single enzyme family or metabolic phase. A similarity-based approach will be used to identify bioisosteric replacements in an attempt to generate novel replacement ideas. The construction of a straightforward graphical user interface through which the method could be easily accessed will also be described.

3. Data Source and Preparation

This chapter provides details of the data source and preparation steps that has been carried out to create the dataset which will be used for the studies being presented in chapters 4, 5 and 6.

3.1 Data Source

The Accelrys Metabolite Database (version 2011.2)³⁷ is used as a source of data in this study. The Accelrys Metabolite Database contains over 100,000 metabolic transformations, compiled from the primary literature, conference proceedings and non-proprietary metabolism studies from FDA drug applications. These metabolic transformations provide information on the metabolic fate of xenobiotics and are annotated with the literature source(s) of the transformation, experimental techniques used for detection, assay type, animal species and the reaction type of the transformation involved. Though not all transformations contain all the details listed above. The database contains both Phase I and Phase II biotransformations from species such as human, rat, rabbit and dog. This is the same database used in the FAME study (section 2.1.4) and is a newer version of the database used in MetaPrint2D (section 2.1.3).

3.2 Identification of Modified Atoms

The substrates and metabolite structures present in the transformations in the Accelrys Metabolite Database also contain reaction centre annotations. These are marked in the bond block, using the standard CTfile annotation (Figure 3.1).

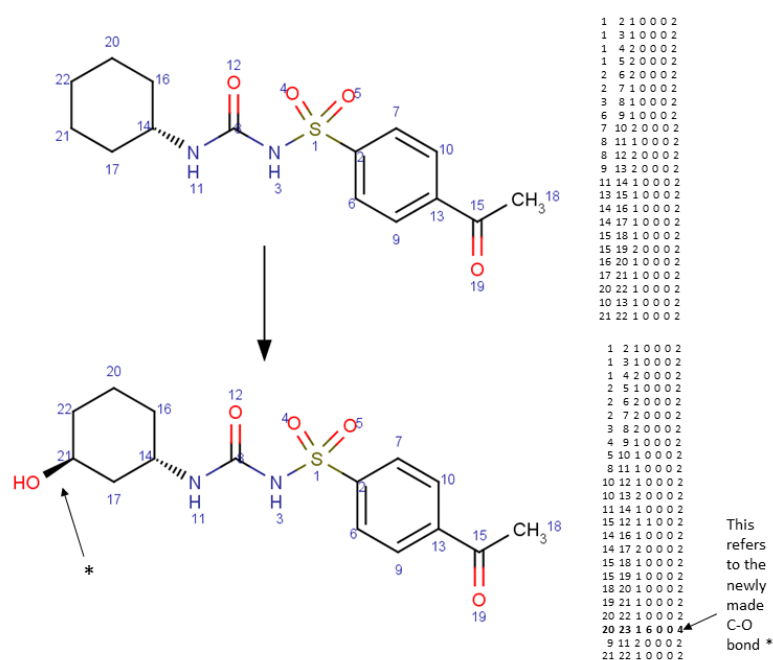


Figure 3.1 One of the transformations from the database along with the bond block (right) of the structures from the database, using standard CTfile format (Table 3.1).

Value	Meaning
0	Unmarked
-1	Not a reacting centre
1	Generic reacting centre
2	No change
4	Bond made/broken
8	Bond order changes
5 = (4 + 1)	Bond made/broken and changes
9 = (8 + 1)	
12 = (4 + 8)	
13 = (12 + 1)	

Table 3.1 Reacting centre status⁸⁵

However, although the annotations are available from the database, they are not always sensible or referring to the correct bonds. The hydrolysis reaction shown in Figure 3.2a is found in the database, and according to the annotation in the bond block of the substrate structure, the reacting C-N bonds are highlighted in Figure 3.2b. However not all of the annotations are correct, as the hydrolysis reaction should have altered the terminal carbonyl and caused the breakage of one of the annotated C-N bonds (Figure 3.2c), leaving the second C-N unchanged.

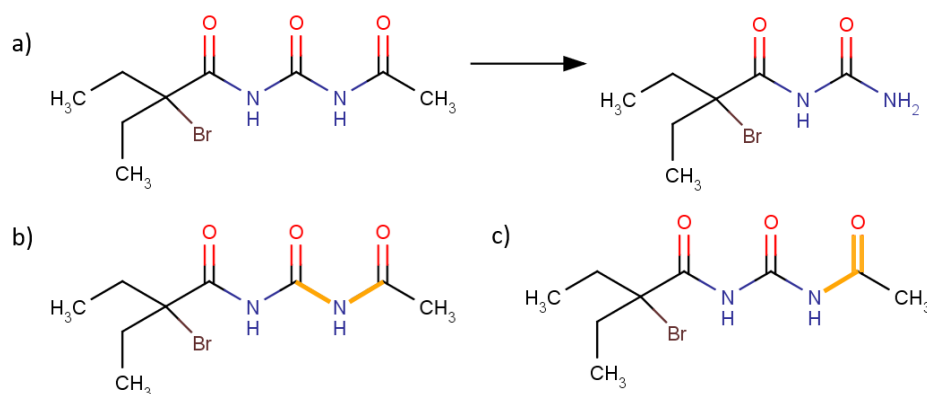


Figure 3.2 a) Hydrolysis transformation found in the database. b) Highlighted (yellow) bonds = bonds marked as reaction centres according to the database structure. c) Highlighted (yellow) bonds = bonds altered in hydrolysis reaction.

Aside from incorrectly labelling bonds during some transformations, mapping is also incomplete within the database. This, along with atom-atom mapping errors in the database, means identifying SOM purely based on these annotations will lead to erroneous annotations. Adams, who created MetaPrint2D^{24,25} (section 2.1.3) based on the 2008.1 version of the Accelrys Metabolite Database, found the same problems and consequently created an MCS-based algorithm (as described in in

section 2.1.3.1.) to identify the correct SOM. During the creation of FAME¹ (section 2.1.4), Dr. Kirchmair used the same algorithm on the Accelrys Metabolite Database (version 2011.2) to extract SOM annotations and this dataset is used for the studies outlined in chapters 4 and 5.

3.3 Reaction Types Annotations

As mentioned in section 2.1.3.4, some transformation entries in the Accelrys Metabolite Database carry reaction type information. However, some transformations in the database contain no reaction class labels and some transformations contain multiple reaction class labels (Figure 3.3). This was also discovered by Adams when using an older version of the database to create MetaPrint2D-React.

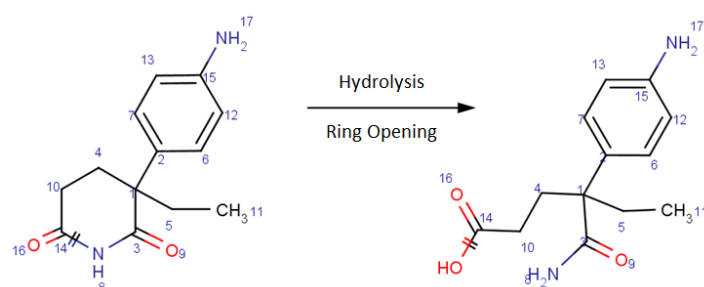


Figure 3.3 Transformation with both a Hydrolysis and a Ring Opening reaction class label in the database.

There are 290 different reaction class labels found in the 2011.2 version of the Accelrys Metabolite Database, the most common are shown in Table 3.2. There are a total of 132,601 instances of reaction class labels in the entire database, but due to there being multiple labels per transformation in some cases, only 81,274 out of the total 103,908 transformations in the database have associated reaction class label(s). Also, according to the reaction class type labels in the database, 23 different reaction types are only observed once throughout the entire database. Over 100 reaction types are seen 20 times or fewer, and the frequency distribution is extremely skewed (Figure 3.4). As well as being incomplete, the labelling system is also inconsistent. For example, hydroxylation (addition of an OH group) of a carbon atom has been found under the labels Hydroxylation, C-Hydroxylation, Aromatic Hydroxylation and Aliphatic Hydroxylation (and incorrectly found under the N-Deglucosidation label).

As the same issues are present in the newer Accelrys Metabolite Database (version 2011.2) (compared to the version used by MetaPrint2D-Reaction version 2008.1), the same list of SMIRKS patterns defined by Adams (Table 2.6) and the annotation method used by MetaPrint2D-React (section 2.1.3.4) are used when preparing the dataset. The extraction of reaction type information from the Accelrys Metabolite Database (version 2011.2) along with the extraction of SOM

annotations (section 3.2) was carried out by Dr. Kirchmair using Adams' algorithm. This was done when preparing the dataset for FAME¹ (section 2.1.4). This dataset is used for the studies outlined in chapters 4 and 5.

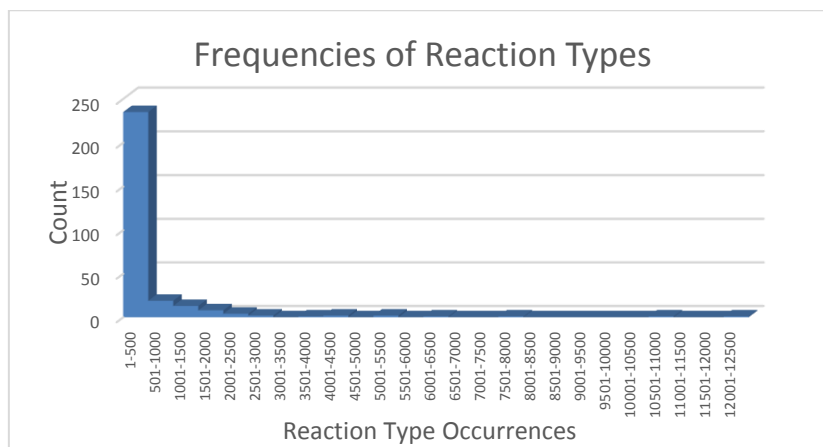


Figure 3.4 Frequencies of different reaction types according their labels in the Accelrys Metabolite Database.

Reaction Type	Count	Reaction Type	Count
C-Hydroxylation	12276	Dehydrogenation	1851
Hydrolysis	10860	Aromatization	1787
C-Oxidation	7980	O-Demethylation	1786
O-Glucuronidation	6078	Conjugation	1586
Aromatic Hydroxylation	5428	Epoxidation	1579
Aliphatic Hydroxylation	5189	Dehalogenation	1576
N-Dealkylation	4392	Dearomatization	1571
Reduction	4195	DNA Binding	1544
Ring Opening	3809	Protein Binding	1418
O-Dealkylation	2690	N-Acetylation	1340
Oxidation	2634	Optical Resolution	1315
Glutathionation	2326	S-Oxidation	1293
O-Sulfation	2299	O-Conjugation	1199
Hydrogenation	2217	Oxidative Deamination	1179
N-Demethylation	2014	Covalent Binding	1132

Table 3.2 Top 30 most frequently occurring reaction labels in the Accelrys Metabolite Database (2011.2)

3.4 Selection of Structures from the Database

There are 103,908 transformations in the 2011.2 version of the Accelrys Metabolite Database. Each transformation only contains a single substrate and a single metabolite. There are also entries in the database which contain R groups in the substrate and/or metabolite structures (e.g. Figure 3.5). As these transformations contain incomplete structures, these are excluded from the study. After the application of these selection criteria, 79,238 transformations remains eligible.

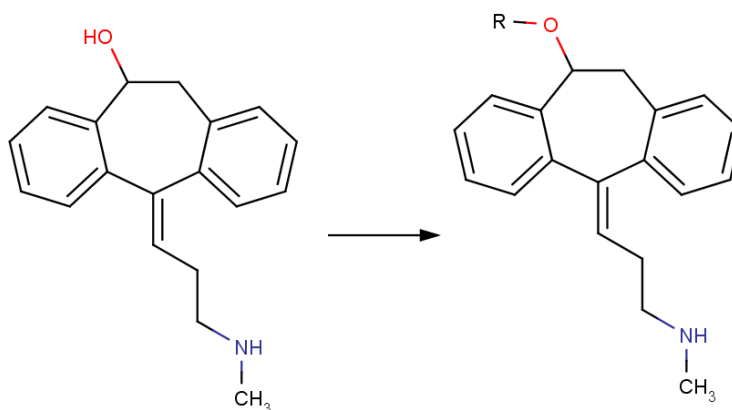


Figure 3.5 A transformation which contains an R group and this is therefore excluded from the study.

Not all 103,908 transformations are single step reactions. Some of these transformations are multi-step transformations, which are summaries of two or more single step transformations. All these transformations are individually listed in the database. As Adams found during the evaluation of MetaPrint2D, excluding multi-step transformations brought about consistent improvement to the SOM prediction performance of MetaPrint2D (see section 2.1.3.2 and 2.1.3.3). Due to this reason, and to avoid data redundancy, only single step transformations are used in the dataset for the studies that will be presented in the following chapters.

The Accelrys Metabolite Database can contain multiple transformations for a single substrate structure if the substrate is present in different reaction schemes (see 2.1.3.2). MetaPrint2D allows for the extraction of unique substrate structures from the database by aggregating all transformation information regarding the substrate into a single record to produce a merged dataset. Adams found the merging of transformation data produced no statistically significant difference to the SOM prediction performance of MetaPrint2D. In order to minimise the computational resources required to store and handle the dataset, a merged dataset is extracted and used.

When preparing the data for FAME (sections 3.2 and 3.3), Dr. Kirchmair extracted single step only transformations from the 2011.2 version of the Accelrys Metabolite Database, using the data preparation algorithms created by Adams for MetaPrint2D and MetaPrint2D-React (section 2.1.3). The merging option was enabled and the resulting dataset will be used for the studies that will be presented in chapters 4, 5 and 6. A text file containing one CFP record for each substrate structure and a SD file containing all the unique substrate structures from the Accelrys Metabolite Database were provided by Dr. Kirchmair.

Each CFP entry in the MetaPrint2D-React CFP file corresponds to a unique substrate structure in the Accelrys Metabolite Database. Each CFP entry contains all reaction identifiers corresponding to all single step transformations involving the substrate structure, as well as the collective reaction type information gathered from all of these transformations. Each line of the CFP block corresponds to one atom line in the atom block of the substrate structure in the extracted SD file and any SOM records (with reaction type) associated with the atom of that substrate structure are added to the end of the line.

The set of unique substrate identifiers present in the MetaPrint2D-react CFP file is used to extract the unique set of substrate structures from the SD file containing all substrate structures from the database. The substrate structure, if it contains no invalid R groups, corresponding to the first reaction identifier in each CFP entry is extracted and kept for use. Inorganic and organometallic compounds are also removed from the dataset as they are usually not handled correctly by cheminformatics descriptors. A total of 30,467 unique substrate structures are extracted and will undergo charging and SOM annotations as outlined in sections 3.5 and 3.6.

3.5 Charging Structures

The substrate structures are first prepared in MOE version 2011.10³⁸. The “Wash” function is applied with all options kept at their default settings (with the exception of disabling the “Add Explicit Hydrogens” option). If a terminal electropositive atom (lithium, sodium, potassium, rubidium or caesium) is covalently attached via a single bond to one of carbon, nitrogen, oxygen, fluorine, phosphorus, sulphur, chlorine, selenium, bromine or iodine, and they have an overall charge of 0, the electropositive element is removed as an ion with a +1 charge, resulting in a -1 charge on the atom it was attached to. This washing step also removes smaller molecular fragments (determined by the total sum of heavy atom in each fragment), keeping only the largest fragment. This removes any solvent and non-bonded counter-ions which may be associated with the database structures.

Protonation states of functional groups that are usually always protonated or deprotonated at physiological pH (i.e. functional groups with a pK_a significantly larger or smaller than 7) are adjusted (examples given in Figure 3.6). However, functional groups with a pK_a close to 7 are assumed to be neutral.

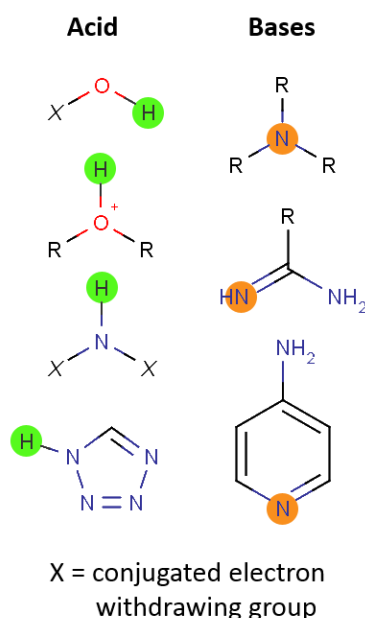


Figure 3.6 Strong acids and bases. Protons to be removed are highlighted in green and atoms to be protonated highlighted in amber.

As the protonation and deprotonation of structures is carried out according to the presence or absence of certain substructures, rather than the overall pK_a value of the structure after charging, there are certain issues with the results of the charging step. For example, the substrate structure in Figure 3.7 (left) has two non-conjugated nitrogen atoms in the piperazine substructure, both are recognised by MOE as basic nitrogen atoms and subsequently protonated, resulting in a doubly protonated piperazine substructure. This is very unlikely to be the correct protonation state of the compound under physiological conditions, as studies on the pK_a values of different piperazine structures suggests that one of the nitrogen atoms should be protonated (as all pK_a values of the first amine are higher than 7.4), however, the second nitrogen atom should be neutral (as all pK_a values of the second amine are lower than 5.4).⁸⁶ There are 216 structures which contain a doubly protonated piperazine substructure.

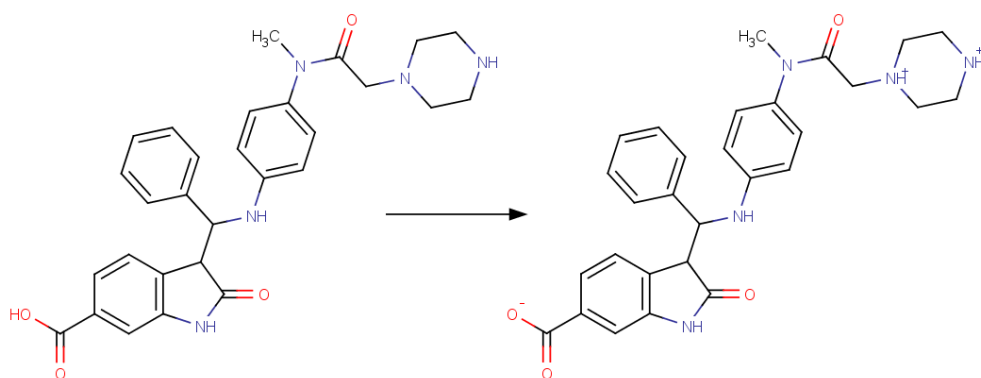


Figure 3.7 Issues with the protonation/deprotonation of structures carried out by MOE. The same issue occurs in MOE version 2012.10.

There are 3,399 unique substrate structures in total with two or more positively or negatively charged atoms within their structures. 1,227 structures contain two or more carboxylic acid groups (both deprotonated by MOE). These structures have an average heavy atom count of 38, which is noticeably higher than the overall average of 25 heavy atoms among the list of unique substrate structures. 1,022 out of the 1,227 structures contain only two carbonyl groups (Figure 3.8), most of them are not (topologically speaking) next to each other, therefore it is more plausible that they are both deprotonated at physiological pH.

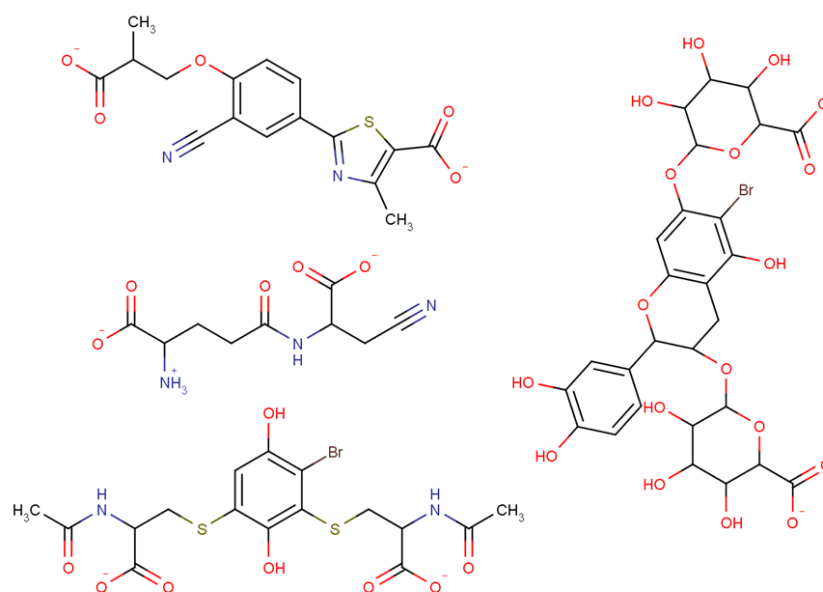


Figure 3.8 Examples of structures containing two deprotonated carbonyl groups.

622 structures contain phosphate groups where the negatively charged oxygen atoms (all oxygen attached to phosphorus via a single bond) are responsible for the multiple negative charges on the structure. 310 structures contain a diphosphate substructure and 46 of those are triphosphates. In all cases, all hydroxyl groups connected to a phosphorus atom are deprotonated. For a single phosphate group, a study carried out on various phosphoric acid⁸⁷ (such as monomethyl, monoethyl,

mono-*n*-propyl phosphoric acids) shows that the first pK_a value ranged from 1.5 to 1.9, with the pK_a value increasing as the alkyl group length increased. The second pK_a value shows the same trend and ranges from 6.3 to 6.9 (these experiments were conducted at 298K). However, for di- and tri-phosphates, this may not be the case. Adenosine triphosphate (ATP) is a substrate in the list of unique substrates extracted and all four available acidic oxygen atoms in the triphosphate group are deprotonated. However, according to protonation states study carried out by Storer and Cornish-Bowden⁸⁸, the dominant species at pH 7.4 is $(Mg^{2+})ATP^{2-}$, not ATP^{4-} (Figure 3.9).

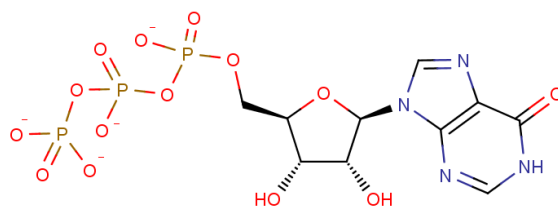


Figure 3.9 Adenosine triphosphate after washing by MOE.

There are 7 instances where two positively charged nitrogen atoms are two bonds away from each other (Figure 3.10). These protonation states are unlikely to be correct at physiological pH, although no experimental information can be found apart from the first pK_a values of similar compounds.

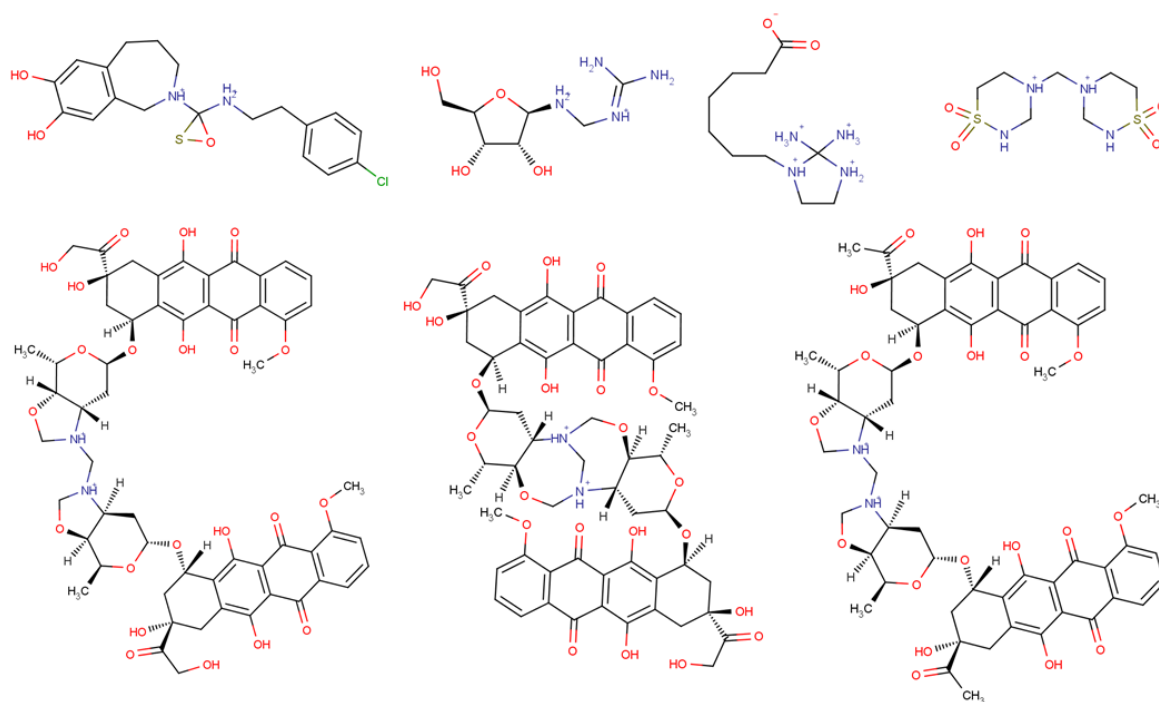


Figure 3.10 All seven structures with positively charged nitrogen atoms two bonds away from each other.

However, not all of these 3,399 compounds are incorrectly charged. Over 300 of them have a nitro group in combination with another charged carboxylate or amine. This means the structure is

included in the list of (at least) doubly charged compounds, even though the nitro group is not formally charged.

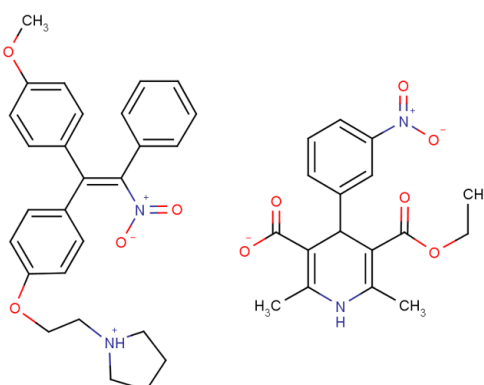


Figure 3.11 Structures containing two or more positively or negatively charged atoms that have been charged correctly.

Open Babel version 2.3.0 utilises a similar but more detailed approach to identify atoms which require protonation/deprotonation. SMIRKS patterns of functional groups (with associated pK_a values of each transformation) are used to identify and charge substructures. Although the charging protocol requires all implicit hydrogen atoms to be made explicit, these are appended to the end of the atom block and therefore does not affect the subsequent SOM labelling which require consistent ordering of lines in the atom block. Open Babel is free to use and easily accessible. Its charging method is transparent and produces output that is compatible with the data structure required for the next steps, Open Babel is therefore investigated as an alternative charging method.

All 30,467 unique structures are charged separately by MOE and Open Babel to produce two sets of structures for comparison. First the 30,467 structures are charged with MOE then have all implicit hydrogen atoms made explicit by Open Babel without changing the protonation state of any atoms. The pre-washed unique substrate structures are also processed by Open Babel with the “Add hydrogens (make explicit)” option enabled and “Add hydrogens appropriate for this pH” set to 7.4. The corresponding substrate structures that have been charged by the two methods are then compared using canonical SMILES strings generated for each structure by the SmilesGenerator from CDK (version 1.5.9).

A total of 2,314 pairs of structures are identified as being different (having different canonical SMILES strings) after being processed by the two different washing methods. Of these, 880 structure pairs have a positive charge incorrectly introduced onto carbon atoms by Open Babel (their counterpart charged by MOE have no such issue), resulting in structures with an incorrect number of electrons and bonds (Figure 3.12a). 310 structures contain an inappropriate negative charge on one

or more carbon atoms (Figure 3.12b). All except two are errors introduced by Open Babel. The two exceptions (Accelrys Metabolite Database IDs: RMTB00026391, RMTB00028921) each contain a negatively charged carbon atom, both are introduced by the Accelrys Metabolite Database. These atoms are not altered by either MOE or Open Babel during the charging process. There are 6 cases where Open Babel introduced negative charge(s) and positive charge(s) onto different carbon atoms within one structure, an example of which could be seen in Figure 3.12c.

146 structure pairs contain azide groups, which are correctly handled by MOE but incorrectly disconnected as a salt by Open Babel (Figure 3.12d). The diazomethane groups are also incorrectly handled by Open Babel (Figure 3.12e). There are 84 structure pairs of structures containing sulanilamide as a substructure, including sulfisoxazole (Figure 3.12f), where the nitrogen atoms bonded to the sulphur atoms are all deprotonated by MOE but kept neutral by Open Babel. The deprotonation, in the case of sulfisoxazole at least, is appropriate as the nitrogen atom has a pK_a value of ca. 5⁸⁹. This is likely to be the case in similar structures, although experimental pK_a values are required for verification.

There are 59 structure pairs containing the thiazolidine-2,4-dione substructure. MOE has deprotonated the nitrogen atom in the thiazolidine-2,4-dione substructure in all 59 cases. Structures charged by Open Babel have kept them all neutral (Figure 3.12g). Several of these structures are known drugs and experimental pK_a values that are available in some cases ranged from 5.8 – 6.3^{90,91}, suggesting that at least in those cases (and likely in others with similar structures), the deprotonation of the nitrogen atom by MOE is appropriate for physiological pH conditions.

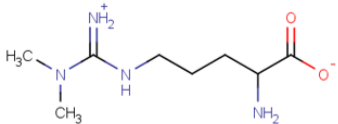
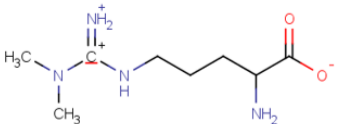
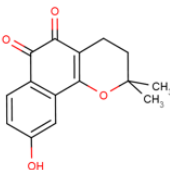
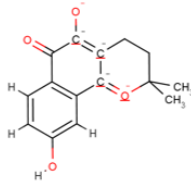
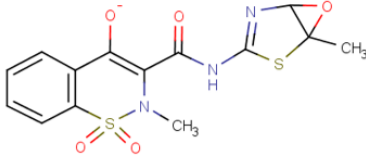
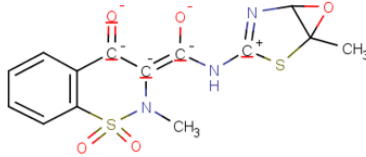
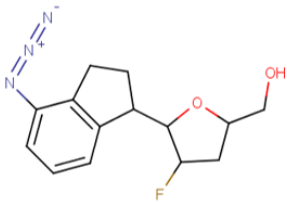
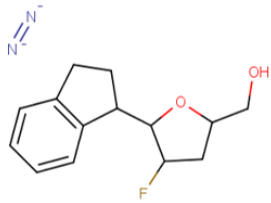
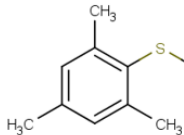
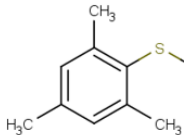
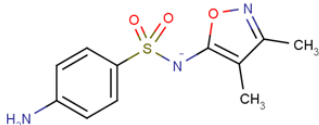
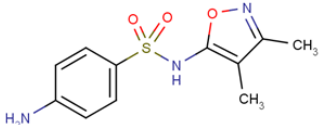
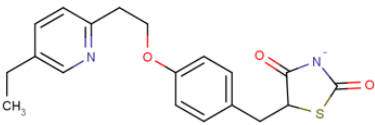
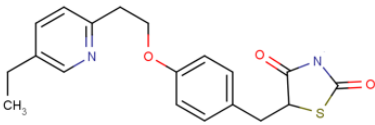
	MOE	Open Babel	Occurrence
a) Positively charged carbon atom(s)			810
b) Negatively charged carbon atom(s)			310
c) Positively + negatively charged carbon atoms			6
d) Azide groups			146
e) Diazo-methane groups			3
f) Sulanilamide group			84
g) Thiazolidine-2,4-dione group			59

Figure 3.12 Examples of substrate structure pairs that are treated differently by MOE (left) and Open Babel (right) charging procedures.

The discrepancies between the remaining structure pairs include differences between the number of deprotonated oxygen and/or sulphur atoms in phosphate or phosphorothioate groups in 68 structure pairs and various heterocycles with an amine group attached that are protonated (positive charge on one nitrogen atom in the ring where delocalisation is possible) by MOE but not Open Babel. There are also 15 pairs of structures, all with 96 or more heavy atoms, where the nitrogen atom of one of the carbamate substructures in the structure is charged by Open Babel but not MOE. This is due to substructures being recognised as a Lysine structure by the Open Babel SMIRKS patterns (Figure 3.13).

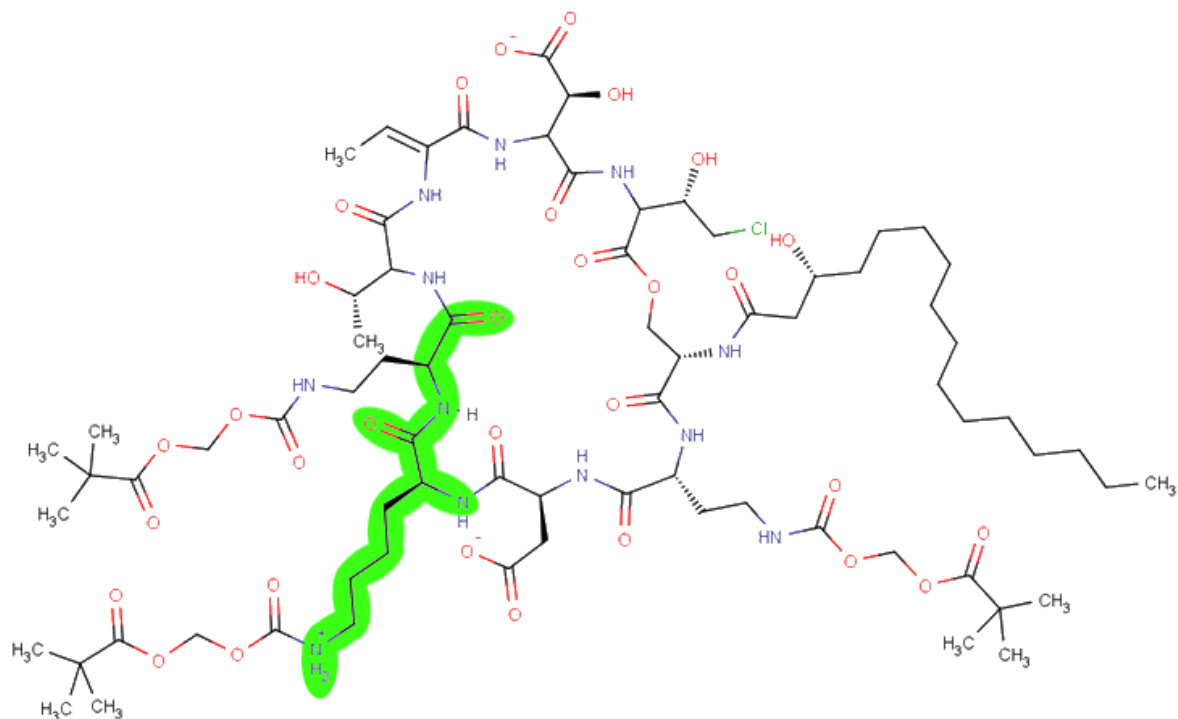


Figure 3.13 Substructure (highlighted in green) recognised as a lysine structure and is charged by Open Babel according to SMIRKS pattern: O=C(NCC=O)C(N)CCCC[N:1] >> O=C(NCC=O)C(N)CCCC[N+:1]

Both methods produce the same incorrectly charged structures, including the protonation of both nitrogen atoms of piperazine substructures.

From the list of structures that are handled differently by the two methods, it is clear that the MOE washing method produces more accurate results overall. The other option is to neutralise all structures to avoid errors introduced by washing. However, the neutralising procedure may introduce unexpected mistakes to the structures and structures will be in inappropriate protonation states for physiological pH.

There are a total of 2,135 structures with charged atoms from the set of unique substrate structures taken directly from the Accelrys Metabolite Database. A total of 16,012 structures are charged by the MOE washing process. 10,272 of these structures contain only one charged atom. 5,740 structures with two or more charged atoms, a portion of these are overall neutral (e.g. structure that contains one negatively charged carboxylic acid and one positively charged amine). 10,185 structures contain a single charged atom has one of the following groups (occurrences in brackets):

1. Positively charged nitrogen atom (5673)
 - Tertiary amine (3807)
 - Secondary amine (1642)
 - Pyridine/quinolone/isoquinoline (224)
2. Negatively charged oxygen (4185)
 - Carboxylic acid (3852)
 - Phosphate (177)
 - Sulphate (92)
 - α , β - unsaturated carbonyl (64)
3. Negatively charged nitrogen atom (287)
 - 2,4-Thiazolidinedione (58)
 - Sulphanilamide (177)
 - Tetrazole (52)
4. Negatively charged sulphur atom (40)
 - Carbamothioic acid (17)
 - Phosphorothioate groups (23)

From inspection, the majority of charged structures look chemically reasonable, therefore the MOE washing protocol is employed to charge all structures to approximate their physiologically relevant states prior to SOM annotation and descriptor calculation. There are some erroneously charged structures but on the whole, from manual inspection and the number of examples given above where the functional groups should be charged at physiological pH, erroneously charged structures is estimated at fewer than 10%. There are more structures in the correctly charged form than structures that are charged incorrectly compared to if the entire dataset is neutralised.

3.6 Sites of Metabolism Annotation

After the unique set of substrate structures are charge, CDK version 1.5.9 is used to carry out SOM annotations on all structures. CDK version 1.4.18 was used in the FAME (section 2.1.4) for SOM Annotation. However, some structures in the database contains selenium atoms, which are not supported in the atom types provided in CDK 1.4.18 (atom type Se.2). This is no longer an issue in CDK version 1.5.9, which is used in this study.

The MetaPrint2D-React CFP file contains only one entry for every unique substrate structure in the Accelrys Metabolite Database which are involved in single step transformations. Each of these entries contained all single step transformation identifiers (which the substrate structure is involved in) and these are used to extract the 30,467 unique substrate structures.

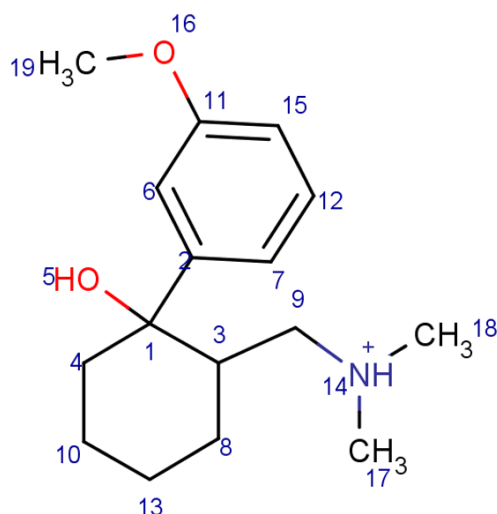
After the washing procedure is carried out on the extracted substrate structures, SOM annotations are extracted from the CFP file. The CFP block of each structure contains one line for each atom in the structure and each line describes one atom in the structure. The order of these lines matches

those in the atom block of the SD file. When producing the CFP file, MetaPrint2D-React uses each transformation involving the same substrate structure and attempts to identify the type of reaction undergone in the transformation using the SMIRKS patterns (Table 2.6). Each atom is annotated with a reaction type label, followed by the instance of that atom in that reaction. An example is given here:

CFP entry:

601_022171_0322_012241_022171_01	26:1
621_0122_022271_032171_0241_02	
601_03_02214171_0322_22_2171	
601_02_022171_0222_2241_022171	26:0
671_01_0221_0322_012241_022171	26:2,30:0,85:0
621_22_012271_032171_03_0141	
621_22_0122_022171_0371_0241	
601_02_03_01214171_0222_22	68:0
601_0141_04_022171_0122_22	
601_02_02_012171_0122_2241	
621_2271_0122_0121_0271_03	
621_22_22_012171_0371_03	18:0
601_02_02_02_214171_0222	18:0
641_03_01_02_022171_0122	19:0,58:0,68:1
621_22_2271_0121_01_0271	
671_0121_22_22_0121_0271	58:0
601_41_02_01_02_022171	58:1
601_41_02_01_02_022171	58:1
601_71_21_22_22_0121	58:1

Structure from SD file:



Sites of Metabolism extracted from the CFP entry:

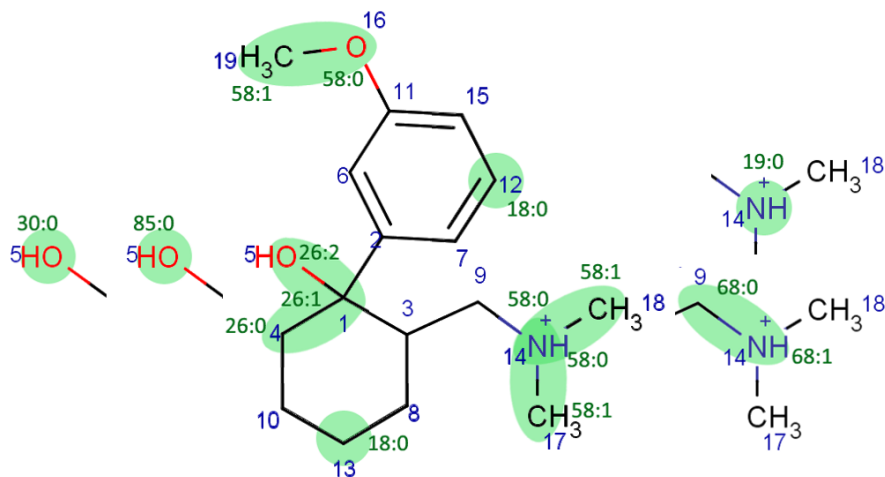


Figure 3.14 One of the CFP entries from the MetaPrint2D-React file (top left), the corresponding structure from the database (top right) and the atoms annotated as SOM by the CFP entry (bottom).

The order of lines in the CFP entry corresponds to the ordering of atoms (labelled in blue in both structures) in the structure. Each line in the CFP entry block contains the atom's CFP (in grey). If an atom is modified in a transformation, SOM information is added at the end of the line. Take the example of the first atom (and first line of the CFP block) – atom 1 has been labelled as 26:1. The

number 26 refers to the type of reaction undergone by atom 1. Reaction type 26 refers to a dehydration transformation with the SMIRKS pattern [:1]-[:2]-[OH]>>[:1]=[:2] (see Table 2.6). The number 1 in 26:1 means atom 1 is the second atom in this reaction as specified by the SMIRKS pattern, as the numbering scheme begins at 0 (these will be referred to as atom levels from now on). Atom 4 and atom 5 (26:0 and 26:2 respectively) completes this transformation, which requires 3 atoms as specified by the SMIRKS pattern (highlighted in Figure 3.14 bottom).

During SOM annotation, each structure is processed and its SOM annotations retrieved from the CFP entry. Each reaction type has a required number of atoms in order to fulfil the SMIRKS pattern. The annotation algorithm checks that the SOM label contains the correct number of atoms before associating the SOM label with the molecule. Problems arise in the case where there are two transformations of the same type occurring where both transformations point to the same atom at the same atom level. In these specific cases, only one instance of the SOM annotation will appear in the CFP block, an example can be seen for reaction type 58 in Figure 3.14. Reaction 58 refers to the demethylation reaction [:1]-[CH3]>>[:1]. In the example above, there are two demethylation transformations involving atom 14 at atom level 0, one involving atoms 14 and 18, the other atoms 14 and 17. However, atom 14 is only labelled as 58:0 once in the CFP block. These occurrences are identified and resolved to create two separate SOM annotations.

After joint SOM are resolved, symmetrical atom environments are identified using the CDK `EquivalentClassPartitioner`. If existing SOM annotations are found on any atom with a symmetrical counterpart in the structure, SOM labels are propagated to the symmetrical atoms if none of that reaction type exists on the symmetrical atoms. This is necessary to ensure symmetrical atoms are not recorded as labile once and stable once, when in fact, the unlabelled atoms are as likely to be metabolised as the labelled atoms.

The `EquivalentClassPartitioner` fails to handle 648 molecules correctly. These molecules all contain a terminal nitrogen atom with a double bond to either a carbon or nitrogen atom (Figure 3.15). SOM annotation but not symmetry checks have been carried out on these structures, therefore there may be a number of SOM labels missing from the resulting dataset. These only make up 2% of the entire dataset and from visual inspection, not many are symmetrical molecules. The symmetrical structure from Figure 3.15 (left) is one exception, however none of the SOM labels involved atoms in the benzene ring, therefore no extra annotation is required regardless of the inability of `EquivalentClassPartitioner` to handle the structure.

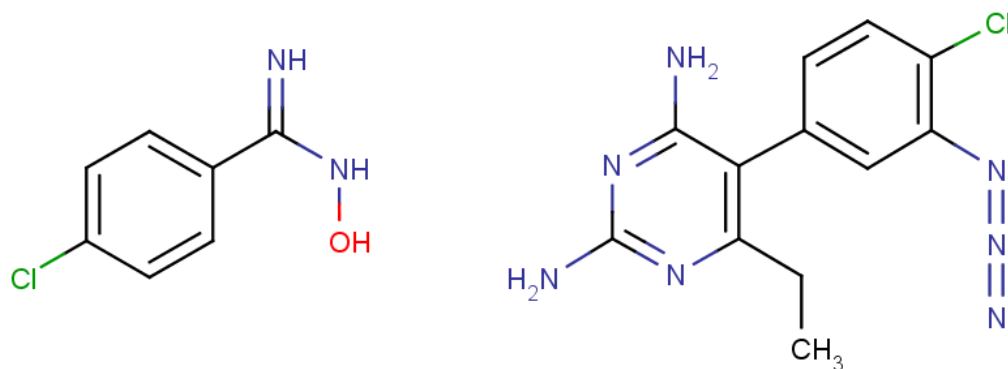


Figure 3.15 Examples of molecules which cannot be handled by CDK's EquivalentClassPartitioner. Both molecules contain a terminal nitrogen atom with a double bond to either a carbon or nitrogen atom.

After SOM annotations are carried out on the 30,391 structures as outlined above, the entire set of structures contain 86,574 reactions, 25,564 (30%) of which are of an unknown reaction type (not recognised by the SMIRKS patterns in Table 2.6). The most common reaction types are listed here:

#	Reaction Type	Count	#	Reaction Type	Count
0	Unknown	25564	65	Methylation	1007
18	Hydroxylation	13543	6	Dehalogenation	991
60	Dealkylation(1)	10275	11	Epoxidation	979
85	Glucuronidation	5952	24	Dehydroxylation	948
58	Demethylation	3715	38	Oxidation (=O-OH)	907
40	Oxidation (=O)	2665	86	Glutathionation (+SX)	876
30	Sulfation	1971	55	Acetylation	815
44	Oxidation (-/=)	1967	4	Phosphorylation	685
45	Reduction (=/-)	1866	19	Hydroxylation	673
69	Oxidative deamination (-OH)	1466	68	Oxidative deamination (=O)	644

Table 3.3 Top 20 most frequently seen reaction types in unique set of substrate structures extracted. The reaction types and numbers are consistent with entries found in Table 3.2.

3.7 Fragmentation Methods

In order to predict the potential SOM on new query molecules, a list of chemical environments that are typically stable or labile need to be compiled, along with the likelihood of metabolism occurring in these environments. The Accelrys Metabolite Database is used as a source of metabolic stability information. It is the aim to produce a SOM predictor which can be extended upon to suggest bioisosteric replacements for structures which has a desired metabolic stability profile but contains other less desirable properties. This requires the substitution of one substructure for another. Therefore, it is decided that a dataset of fragments with associated metabolic stability information will allow for the prediction of SOM and subsequent identification of potential fragments for substitution. A number of fragmentation methods are identified and their suitability for producing descriptive fragments of sensible sizes for use in is evaluated.

The Fragmenter API and command line tool are produced and distributed by ChemAxon.⁹² There are several fragmentation methods available within the tool. The RECAP method implemented cleaves single bonds according to the list of retrosynthetic analysis compatible rules:

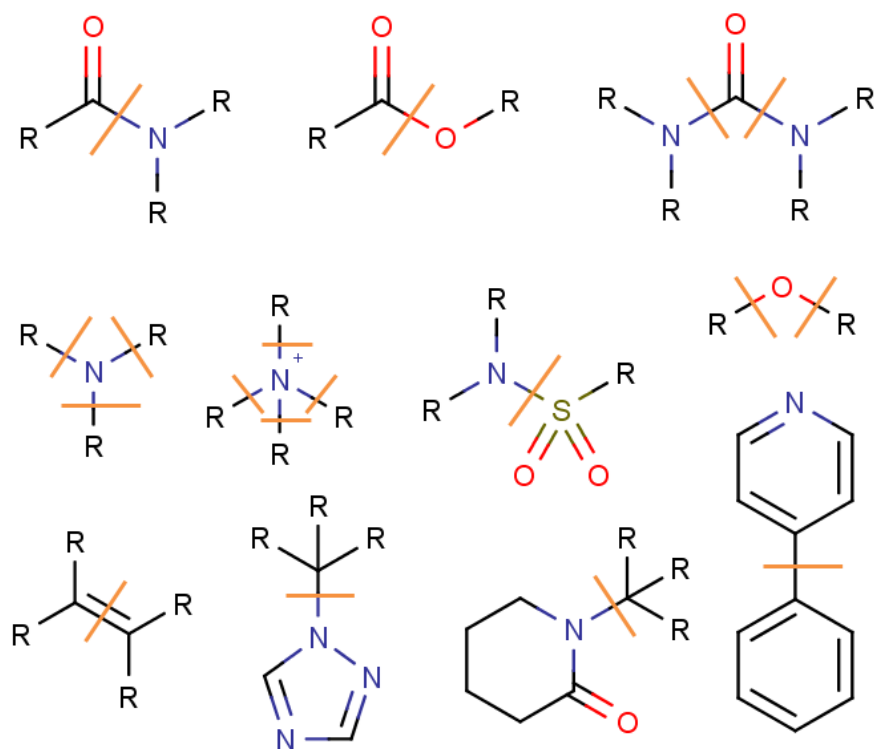


Figure 3.16 RECAP cleavage rules implemented in the ChemAxon Fragmenter.

These rules are not appropriate for the purposes of this study as they are very restrictive and may only produce one or two cuts per structure or none at all, if none of the substructures listed above

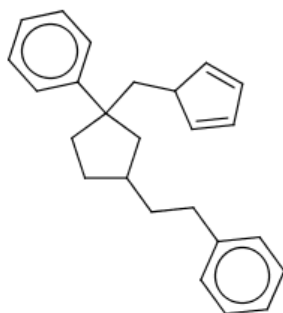
are found. There is the option to customise or add to the list of existing rules, however, more appropriate options can be found elsewhere.

Another fragmentation method implemented in the ChemAxon Fragmenter is the CCQ method. This is a fast, simple fragmentation where a cut is only made between two carbon atoms where at least one is adjacent to a heteroatom. This rule is put in place so cuts are made adjacent to functional groups which are kept intact. Heteroatoms connected to ring(s) are not affected. Aliphatic rings are cleaved and aromatic ring substructures are not affected. The CCQ method does not require any rules and is quick, as no substructure search is required. The method claims this reduces the risk of combinatorial explosion (as no hydrocarbon chains are cut without an adjacent heteroatom). This method can generate fragments which are relatively large in size compared to the original structure but can be considered if applied with a filter to remove larger structures after fragmentation has occurred.

The makefraglib command line tool available from OpenEye⁹³ allows a user to supply a list of structures to be fragmented. However, aside from the fragmentation process, makefraglib also generates the low energy 3D conformers of fragments as part of the fragmentation process. A constraint of 5 kcal above the global minimum conformer is used to filter out conformers for flexible ring systems. As 3D structures are not required in this study, the generation of conformers for all fragments is a time consuming and unnecessary process and this fragmentation method is therefore considered inappropriate.

There are two notable fragmentation implementations from CDK: the exhaustive fragmenter and Murcko fragmenter. The exhaustive fragmenter identifies all non-ring single bonds and creates a cut along all possible cut sites recursively. This has the potential to create a large number of fragments per structure (Figure 3.17), even with a minimum fragment size limit. The default minimum fragment size is set at 6 atoms, although this can be specified by the user.

Structure:



Fragments (minimum allowed size = 6):

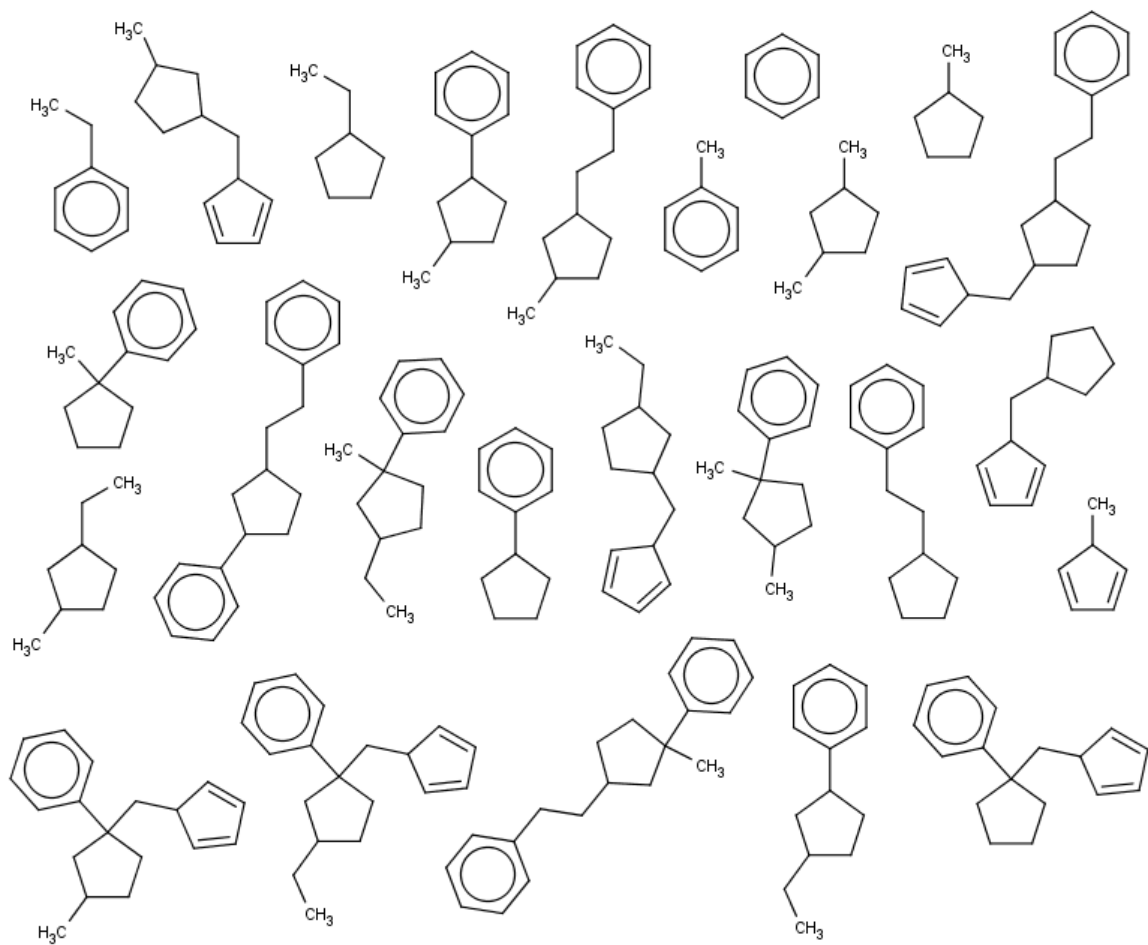


Figure 3.17 All 25 fragments generated from structure (top) by CDK exhaustive fragmenter with the default setting.

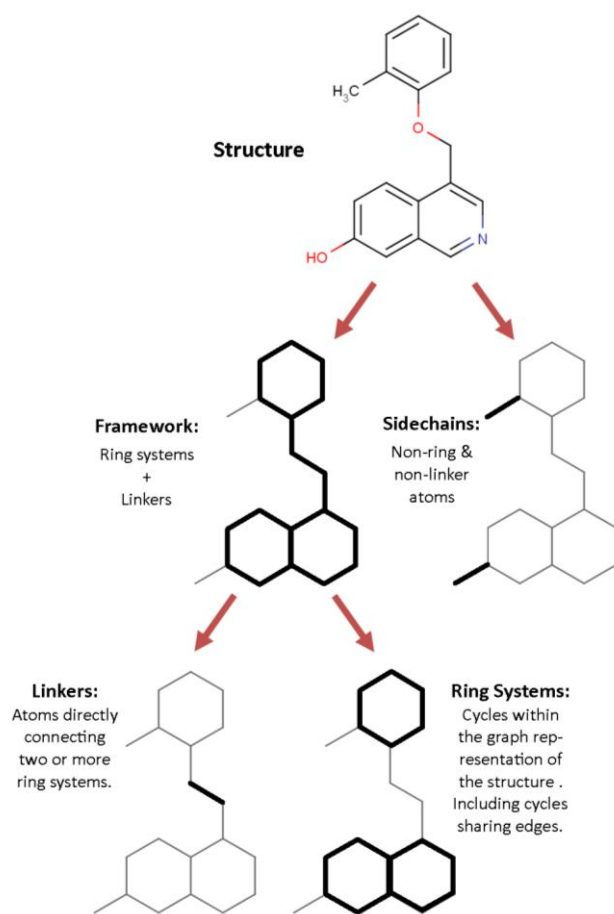


Figure 3.18 Murcko framework hierarchy and definitions according to Bemis and Murcko.³⁴

The Murcko fragmenter produces fragments in line with the definitions of ring, framework, linker and side chains as specified by Bemis and Murcko.³⁴ A structure can be separated into sidechain(s) and framework. The framework can be further broken down into ring system(s) and linker(s) (Figure 3.18). CDK's Murcko framework fragmenter also include the option to produce only a single framework with a specified minimum fragment size. The structure shown in the exhaustive fragmenter example (Figure 3.17) is fragmented using the Murcko fragmenter with default settings. The resulting fragments are shown in Figure 3.19.

Fragments:

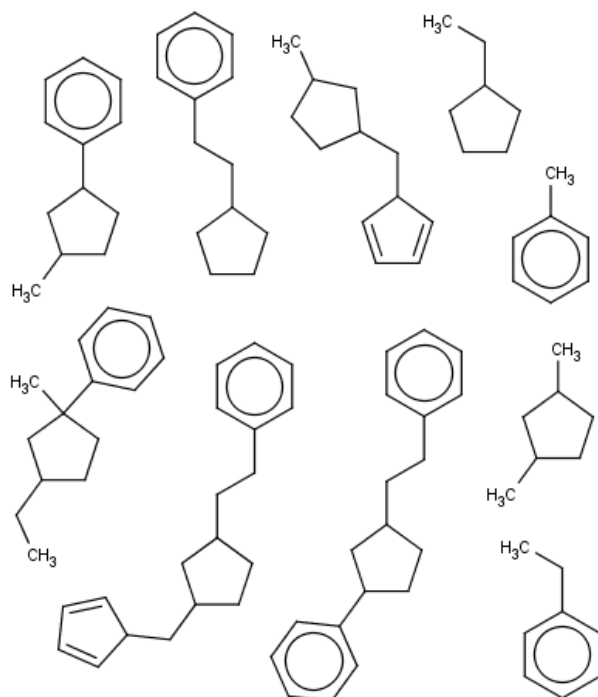


Figure 3.19 All fragments generated from structure (Figure 3.17 top) by CDK Murcko fragmenter with the default setting.

3.7.1 Coralie Fragmentation Method

A fragmentation algorithm developed at Lhasa (outlined in Figure 3.20)⁹⁴ is available for use within Lhasa's Coralie Cheminformatics Platform. The algorithm first identifies the different structural units making up the 2D structure. These structural units are determined by several parameters which can be adjusted by the user via the options available in the Coralie application GUI. These include the option to retain ring systems, retain scaffold (if retain ring systems is enabled), retain functional group and fragmentation depth (which controls the maximum size of growth in any direction of the resulting fragment). When the retain functional group option is enabled, double bonds, triple bonds, bonds to a heteroatom and aromatic ring bonds are retained during fragmentation.

Different combinations of these fragmentation parameters can be used to generate structural units with different properties; these can be anything from large scaffolds, fused rings, individual rings or single atoms. During fragmentation, a set of intermediate reduced feature graphs is extracted from the original structure, with each node in the graph representing a structural unit (Figure 3.20). Each node (fragment) can then be extended upon to include atoms and bonds up to the distance specified by the fragmentation depth (adjusted by the user). A simple example in Figure 3.21 shows the different fragments produced from the same structure using different combinations of fragmentation parameters.

Several knowledge-based bioisosteric studies have shown that the contextual information surrounding a group is important,^{77,80,81} therefore more descriptive fragments than those produced by simple cuts between rings and linkers are desirable.

A fragmentation method which allows linkers and part of a ring system, including aromatic rings, to be included within the same fragment may produce a better description of the chemical context of sidechains and linkers attached to aromatic ring systems (which is not uncommon), whilst keeping fragment sizes small and manageable. Due to its adaptability and ability to create small fragments containing a partial rings as well as linker/sidechain atoms, the fragmentation method implemented in Coralie is chosen.

3.8 Summary

This chapter has provided detail on the source of the dataset (Accelrys Metabolite Database version 2011.2), the selection and extraction of substrate structures, the SOM annotations and preparation steps carried out on the extracted substrate structures when preparing the dataset. This dataset will be used in the studies that will be reported in chapters 4, 5 and 6. The choice and detail of the chosen fragmentation methodology which will be used in the studies mentioned above are also given in this chapter.

4. Coralie Atom-based Statistical SOM Identifier (CASSI)

4.1 Introduction

This chapter reports the development of the Coralie Atom-based Statistical SOM Identifier (CASSI), a SOM predictor. CASSI predicts metabolically labile atoms of a query structure based on statistical metabolic stability information collected from fragments which make up the query structure.

CASSI is similar to MetaPrint2D as both methodologies are created based on data obtained from the Accelrys Metabolite Database and both approaches predict SOM based on statistical information. However, where MetaPrint2D uses CFP and SYBYL atom types³⁶ to represent atom environments, CASSI uses fragment structures to represent the local context of an atom.

4.1.1 Coralie Cheminformatics Platform

The Coralie Cheminformatics Platform (version 2.0) is an application created at Lhasa Limited. It is an application used by in-house cheminformatics research. The platform contains the capability to read and interpret common file formats for representing chemical structures, such as SMILES, SD file and Molfile. The application supports substructure search as well as exploring properties of structures contained within a specified dataset. The fragmentation algorithm developed by Lhasa (section 3.7.1) that was chosen for this study is also available via the Coralie Cheminformatics Platform. For this reason, the CASSI methodology has been developed inside the Coralie Cheminformatics Platform. A graphical user interface (GUI) has also been implemented in the application to allow easier access to the CASSI methodology.

4.2 Methods

4.2.1 Data Source and Preparation

To create the SOM predictor, a data source with metabolic stability information is required. The Accelrys Metabolite Database (version 2011.2) is used as the data source (section 3.1). The determination of SOM and reaction type based on metabolic transformation data available from the Accelrys Metabolite Database is carried out as detailed in sections 3.2 and 3.3. This results in a dataset of substrate structures and a matching CFP text file containing information regarding the atoms that are modified for each structure, along with reaction type information where the reaction types are determined by the same SMIRKS patterns used by MetaPrint2D-React (Table 2.6).

The substrate structures from Accelrys Metabolite Database has undergone the same selection process as detailed in section 3.4, producing a unique set of 30,467 substrate structures. These structures, each containing SOM annotations, are charged by MOE³⁸ according to the procedure described in section 3.5. Once the unique set of substrate structures have been washed, the 30467 substrate structures were processed along with the CFP text file (details given in section 3.6). The dataset contains 30,467 unique substrate structures in SD file format where each structure contains SOM annotations (plus reaction type information) in their respective SD tags.

4.2.2 Fragmentation of Substrates

The fragmentation algorithm within the Coralie Cheminformatics Platform (detailed in section 3.7.1) is used to fragment all structures in this study.

4.2.3 Model Evaluation

All validations on CASSI are carried out by comparing the SOM prediction results generated by CASSI against the SOM annotations extracted from the Accelrys Metabolite Database (detailed in section 4.2.1).

4.2.4 Training and Test Dataset Generation

The development of CASSI was carried out at around the same time that FAME¹ (section 2.1.4) was developed; as CASSI and FAME both utilise the same data source and both aim to predict metabolic stability, the performance of all CASSI models will also be compared against the performance of FAME. As FAME is used as part of the performance comparison and time required for calculation of descriptors used in FAME increases exponentially with increasing number of heavy atoms, structures with more than 100 heavy atoms are removed from the dataset. The remaining 30,391 substrates are randomly split into training and test datasets using the Coralie Cheminformatics Platform.

The Split Dataset method within Coralie utilises the `java.util.Random` class in order to select a subset of structures randomly until the user specified number of training dataset structures are picked; the rest of the unpicked structures are used to create the test dataset. A 70:30 split is carried out on the whole dataset, resulting in a training and test dataset of 21,270 and 9,116 structures respectively (five structures which could not be handled by the CDK descriptors were discarded).

4.2.5 Prediction with FAME

SOM predictions are carried out on the training dataset using FAME¹ version 1.0.

4.2.6 AUC Calculation

A method to calculate the AUC values for each test structure given the SOM prediction scores for all atoms has been implemented based on the approaches outlined by Mason *et al.*⁹⁵. This approach also takes tied prediction values into account.

4.3 Model Generation

CASSI carries out SOM predictions based on statistical information gathered during the model training phase. When training the model, each substrate structure in the training dataset is fragmented and each fragment created is stored.

4.3.1 Collection of Metabolic Stability Statistics

All SOM annotations (including reaction types) concerning the atoms contained in the fragment will be used to annotate the fragment. Each transformation type identified within the same fragment structure has its own record (SOM entry) stored within the properties of the fragment. During the fragmentation of the training dataset, if a fragment created by a training dataset structure has already been generated and stored in the dictionary of fragments, the SOM annotations from this new training dataset structure will be appended to the existing record instead. Information on the number of times a particular fragment has been observed during the fragmentation of the training dataset, along with the number of times each atom of the fragment has been found with a SOM annotation and the number of times each reaction type has been found associated with each atom in the fragment, are all collected concurrently during the fragmentation procedure for CASSI. These metabolic stability statistics are stored along with the SOM annotations under each fragment structure as SOM transformation records.

This results in a dictionary of unique fragment structures, each containing SOM annotations obtained from the training dataset structures (originally collected from transformations contained in the Accelrys Metabolite Database) as well as the fragment's metabolic stability statistic figures collected during fragmentation. Although all substrate structures in the training dataset contain at least one SOM atom, it is possible that there are fragments without any SOM transformation records. This can occur if the fragment is produced from a region of the substrate structure where metabolic transformation has not been observed.

4.3.2 SOM Prediction

CASSI generates a metabolic stability score for each atom within a query structure. In order for CASSI to produce the predicted score, first the query structure has to be fragmented using the same

fragmentation algorithm used to create the dictionary of fragments. For each query fragment generated, a search for the same fragment structure in the dictionary of fragments will be carried out. If a match is found, the SOM transformation records for each atom in the dictionary fragment will be propagated to the relevant atoms in the query structure. If there are two SOM transformation records of the same transformation type from different dictionary fragment structures, both pointing to the same atom(s) of the query structure, the transformation record from the larger dictionary fragment is kept as it provides a more accurate description of the chemical context. After all transformation records from matching fragments are propagated to the relevant atoms of the query structure, any symmetrical atoms within the query structure have any missing transformations added from their equivalent counterparts. When all query fragments have been processed and transformation records have been added to all atoms of the query structure, each query atom is ranked according to one of two implemented ranking algorithms: reaction specific and atom specific ranking.

4.3.2.1 Reaction Specific Ranking

In the reaction specific ranking algorithm, the transformation types listed under each atom are treated as being in competition with each other. For each fragment produced, there is a record of the types of transformations the fragment undergoes. For each transformation type, the likelihood of that transformation occurring is obtained using:

$$T_i \text{ likelihood} = \frac{\text{number of } T_i \text{ observations associated with } F}{\text{total number of } F \text{ observations}}$$

T_i = transformation of type *i*

Equation 4.1 Likelihood of transformation of a given type

The transformation with the highest likelihood (obtained using Equation 4.1) is selected to represent the metabolic vulnerability of the atom – the rationale being that these are competing reactions that could occur on the same atom, and therefore only the most likely reaction type should be used to inform on the atom's metabolic stability. This distinction between different reaction types is not taken into account in MetaPrint2D (section 2.1.3) or FAME (section 2.1.4). If there are two equally likely transformations that can occur at a single atom, the transformation type whose record contained a higher number of supporting examples (number of different parent structures that produced the fragment) is used. If the two competing reaction types contain the same number of supporting examples then the first one found (arbitrarily) is used.

The reaction specific ranking algorithm does not take into account the collected statistics associated with the unknown reaction type (type 0) as the records could refer to the same or different types of

transformations. Any information regarding unknown reaction types will play no part in the calculation of atoms' stability scores. However for information purposes, this information is accessible within the GUI (section 4.5).

4.3.2.2 Atom Specific Ranking

Unlike reaction specific ranking, the atom specific ranking approach ignores transformation types. Instead of each transformation reaction being in competition with each other at each atom site, all their recorded likelihoods are combined and all records contribute to the final metabolic stability score for the atom.

This ranking method was created to examine whether comparing the total likelihood of an atom being metabolised (rather than treating different reaction types as competitors) would produce any improvement on the performance of the method. The atom specific ranking approach takes into account the metabolic stability statistics collected from reaction type 0 (unknown reaction type), which is disregarded by the reaction specific ranking.

Note that as all SOM annotations regarding the same substrate structure are merged to form one structure entry during the extraction of substrate structures from the Accelrys Metabolite Database (section 3.4), there could be multiple transformation annotations, of the same or different types, associated with one atom of the fragment (e.g. type 58 on atom 14 and types 30, 85 and 26 on atom 5 in Figure 3.14). Therefore, the sum of all transformation likelihoods (each calculated using Equation 4.1) for one atom could exceed 1 and directly using the sum of these likelihoods (Equation 4.2) could result in a negative stability score.

$$Stability\ of\ F_{(likelihoods)} = 1 - \sum_{i=0}^n (T_i\ likelihood\ associated\ with\ F)$$

Equation 4.2 Sum of transformation likelihoods

Instead of utilising transformation likelihoods, the number of times a fragment containing the atom of interest was generated by a unique parent structure with no SOM annotation associated with the atom (T_{none}) was used in the calculation. In the atom specific ranking algorithm, the overall metabolic stability of the atom is obtained using:

$$Stability\ of\ F = \frac{number\ of\ T_{none}}{total\ number\ of\ F\ observations}$$

Equation 4.3 Stability of a fragment

4.4 Results and Discussion

4.4.1 Fragmentation Parameters

As CASSI relies on directly matching the structure of a query fragment with a fragment generated by the training dataset, the effects of using of different fragmentation parameters requires examination. The fragmentation algorithm implemented in the Coralie Cheminformatics Platform (section 3.7.1) caters for the following parameterisation:

- Retain scaffold (if enabled, retain ring must also be enabled),
- Retain ring,
- Retain functional group, and
- Fragmentation depth

The training dataset is fragmented 12 times using different combinations of the fragmentation parameters with different fragmentation depths to produce 12 different fragment dictionaries (Table 4.1). Each of these dictionaries is then used to process structures contained in the test dataset and the distribution of all stability scores predicted will be used for boundaries selection.

Retain Scaffolds	Retain Rings	Retain Functional Groups	Fragmentation Depths
Yes	Yes	Yes	0, 1, 2
No	Yes	Yes	1, 2, 3
No	No	Yes	1, 2, 3
No	No	No	1, 2, 3

Table 4.1 Fragmentation parameters tested.

The predicted metabolic stability values of all atoms contained in the test dataset structures were recorded and analysed – their frequency distributions are provided in Appendix D – Frequency Distribution of CASSI Prediction on Test Dataset. Out of all structures contained in the test dataset, a total of 28625 atoms were marked as SOM according to the information available from the Accelrys Metabolite Database. These made up 13% of all atoms contained in the test dataset. The predicted metabolic stability value above which 13% of the test dataset values were found was identified in for each of the tested fragmentation parameters and indicated in Appendix D – Frequency Distribution of CASSI Prediction on Test Dataset by orange lines (with the exception of 4 sets of prediction results).

When the reaction specific ranking algorithm was applied to the dictionaries of fragments produced by the retention of all scaffolds, rings and functional groups at various fragmentation depths (0, 1 and 2) during analysis of the test set, over 25% of the test structure atoms were in an environment not covered by the dictionary. As a result, less than three quarters of the atoms had prediction results. More than 22% of atoms were also marked as being in an unseen environment for the atom specific ranking algorithm when using the dictionary as created with a fragmentation depth of 0, although the coverage improved with increasing fragmentation depth. These results are not surprising because by keeping the scaffolds, larger ring systems were retained and not broken up, resulting in larger fragments with more chemical context included. However, the information stored within those larger ring systems will not be available.

An example can be seen in Figure 4.1. The same structure in the dataset was used to produce two different fragment dictionaries, one with the retention of scaffolds (left) and one without (right). When these two dictionaries were used for the prediction of metabolic stability of a structurally similar query structure, no information from the scaffold would be passed on from the dictionary structure shown in the case where scaffolds were retained. However, when the retain scaffolds option was disabled, the information on two of the matching rings between the dictionary and query structure can now be carried across and taken into account during the prediction.

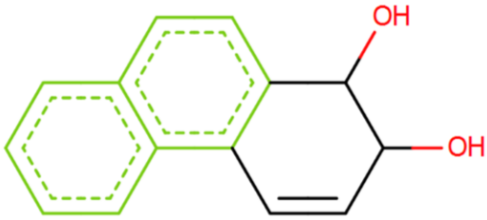
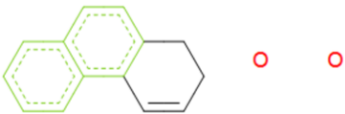

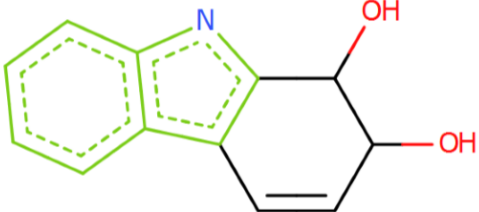
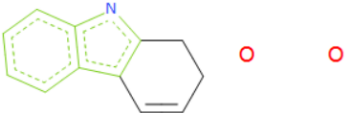
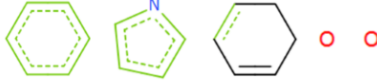
Dictionary Structure		
Fragmentation Parameters	<ul style="list-style-type: none"> • Retain Scaffolds • Retain Ring systems • Retain Functional Groups • Fragmentation Depth = 0 	<ul style="list-style-type: none"> • Break Scaffolds • Retain Ring systems • Retain Functional Groups • Fragmentation Depth = 0
Dictionary Fragments		
Query Structure		
Query Fragments		

Figure 4.1 Differences between retaining and breaking scaffolds.

When using the atom specific ranking algorithm on the same fragment dictionaries, reductions in unknown atom environments were observed (a reduction of over 20% in some cases). This is expected because the atom specific algorithm only require an atom to be observed once in order for a prediction to be made, regardless of whether or not there are SOM annotations associated with that atom.

Three fragment dictionaries were created and tested with both ranking algorithms during the search for appropriate highlighting boundaries but were excluded from the frequency distribution graphs in Table 4.1 and Appendix D – Frequency Distribution of CASSI Prediction on Test Dataset. These fragment dictionaries all had scaffold retention disabled and all three dictionaries had a fragmentation depth set to 0 (Table 4.2). These dictionaries were considered inappropriate for the purpose of generating metabolic stability predictions as a large amount of information was lost.

#	Retain Scaffolds	Retain Ring Systems	Retain Functional	Fragmentation Depth	Unknown %	
					Reaction	Atom

Groups					Specific	Specific
A	No	No	No	0	73.6	72.3
B	No	No	Yes	0	68.4	63.2
C	No	Yes	Yes	0	26.3	19.2

Table 4.2 Omitted dictionaries: fragmentation parameters and percentage of unknown atom environments found in test dataset structures.

All fragments generated for the first omitted dictionary (#A) with no retention options enabled were single atoms. When used to carry out prediction on the test dataset structures, 74% and 73% of atoms were unknown to the dictionary of fragments when the reaction and atom specific ranking algorithms were (respectively) employed. High proportions of atoms were considered unknown as a single carbon atom on its own was not considered a valid fragment by the algorithm.

The second omitted dictionary (#B) was produced by disabling all retention parameters except for the retention of functional groups and using a fragmentation depth of 0. The fragments produced by these parameters included single atoms and non-ring functional groups (Figure 4.2). No ring systems were found and all information contained in non-functionalised ring atoms was lost. When this dictionary of fragments was used with the reaction or atom specific ranking algorithm, 68% and 63% of atoms (respectively) in the test dataset structures were treated as unseen as a significant portion of atoms in the test dataset structures were found in rings and scaffolds.

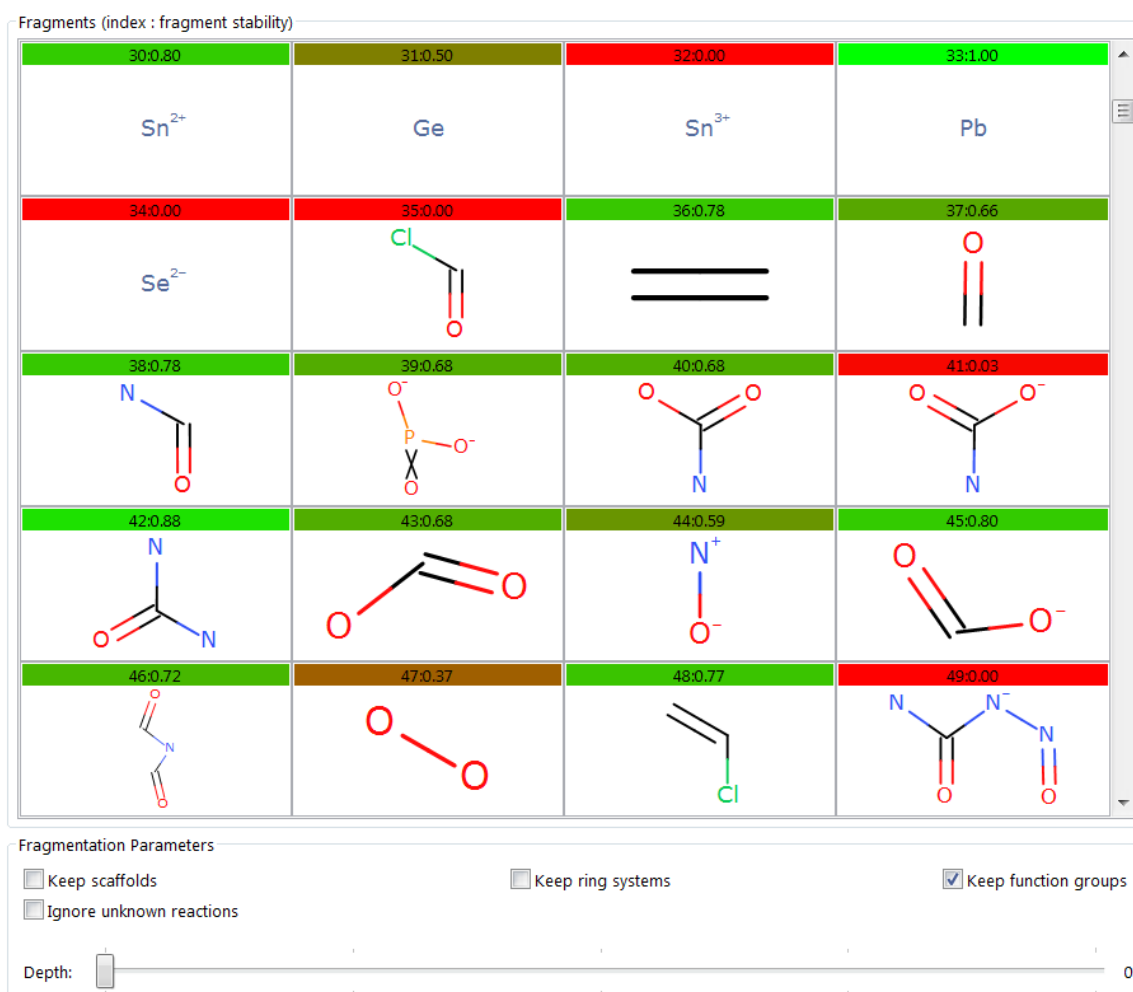


Figure 4.2 Fragments produced by only retaining functional groups at a fragmentation depth of 0.

The third omitted dictionary (#C) was the most successful of the three. However, 26% and 19% unknown atoms is still unacceptably high. These values, along with the results mentioned so far, showed that setting the fragmentation depth to 0 and decoupling all rings and linkers in the fragments produced leads to considerable loss of information.

Aside from the instances mentioned above, no other dictionary/ranking combinations tested (Table 4.1) produced greater than 6% of atoms in an unknown environment. The percentages of unknown atom environments produced by the atom specific ranking algorithm in all of the tested dictionaries were lower than their reaction specific counterparts. This is in line with expectations as more detailed information is required in order to produce reaction specific ranking results. These results together suggest that the atom specific ranking has better extrapolation power compared to the reaction specific algorithm.

4.4.2 Model Performance

The performance of CASSI models produced using the different fragmentation parameters (Table 4.1) are evaluated on the test dataset. Performance of both reaction and atom specific ranking algorithms are also assessed.

Several statistics have been considered as measures quantifying the performance of CASSI. The top- k metric measures the percentages of instances out of all tested structures where at least one known SOM is found in the highest ranked k positions. This is chosen as it is a standard method and has been used in the evaluation of FAME (section 2.1.4) and MetaPrint2D (section 2.1.3). However, although using the top- k metric allows for a direct comparison between methods, it is sensitive to both the size of the query structure as well as the number of positive examples present in each instance – in this case, the number of SOM atoms per structure.

If there are fewer atoms in the structure, the chances of a real SOM atom being picked randomly in the top 1, 2 or 3 positions is a lot higher than in a structure that is much larger – if they both contain the same number of SOM atoms. For example, given two structures, A and B, both with one SOM atom but A having 10 atoms in total and B having 20 atoms, a prediction methodology offered no discriminating power picks one atom at random as the most unstable within the structure, yielding a 10% chance that the correct atom is picked from structure A but only a 5% chance from structure B.

The other problem occurs when there are a different number of positive examples in each structure. If there are two structures, each containing 25 atoms but structure C only has one atom annotated as a SOM and structure D has five, then if one atom is picked randomly, there is a 4% chance of the correct atom being picked from structure C but a 20% chance for structure D. However, as both FAME and MetaPrint2D, which use the same data source, also use the top three metric, the top- k metric provides a good direct comparison between the three different methodologies.

Another performance measure considered is the area under the ROC curve.⁹⁶ A ROC curve is a graphical visualisation of the performance of a classifier which has operated on binary input data and provides a continuous value as its output. The changes in the classifier's true positive and false positive rates as the discrimination threshold is varied produce the ROC curve. Although ROC curves were originally developed during World War II for RADAR development, in more recent years they have been widely used as a performance measure in machine-learning and data mining, among other fields. Unlike the top- k metric and other common measures of performance (such as accuracy, sensitivity, specificity), ROC curves are insensitive to the number of atoms within each structure and to the ratio of positive to negative instances per structure. They also require no threshold calibration

unlike other frequently used classifier performance measures such as (weighted and unweighted) accuracy, precision, sensitivity and specificity. The area under a ROC curve (AUC) has been shown to measure the ability of a classifier to correctly assign a randomly chosen instance to one of two classes or, in other words, rank the positive instances with greater confidence over negative instances. An algorithm has been implemented to calculate the AUC score for each test structure (4.2.6).

An AUC value is produced for each structure contained in the test dataset. This is done instead of an overall AUC value for the entire test dataset as individual AUC values can give an indication of CASSI's ability to predict the most unstable site in each structure rather than the most unstable atoms out of all atoms contained in the structures found in the test dataset. The mean and median of AUC scores from all test structures are used in addition to the top-*k* metrics when evaluating the performance of CASSI models. This is also carried out on FAME prediction values for comparison (Table 4.3). The results from Adams' evaluation of MetaPrint2D used the same AUC measurement and are included in Table 4.4 for comparison.

Performance of 28 dictionaries produced by different combinations of fragmentation parameters as well as results for both ranking algorithms are given in Table 4.5.

Top 1	Top 2	Top 3	Mean AUC	Median AUC
66.517	77.817	84.893	0.856	0.941

Table 4.3 FAME validation results. The percentages of structures which contain at least one SOM atom in the top1, 2 and 3 positions as well as the mean and median of AUC values produced by all test dataset structures are shown.

Model		Top 1	Top 3	Mean AUC	Median AUC
MetaPrint2D	All	59.6	77.2	0.804	0.900
	No Multistep Transformations	59.3	76.5	0.805	0.902
	No Duplicate Transformations	60.3	77.2	0.803	0.900
	Symmetry Mapping Added	59.6	76.7	0.803	0.900
	Transformations of same compound merged	60.0	75.7	0.803	0.913
	At least one Phase I SOM	60.5	76.2	0.799	0.892
MetaPrint2D-React	All transformation types	58.9	78.7	0.812	0.918

Table 4.4 MetaPrint2D models trained on Accelrys Metabolite Database version 2007.1 and tested on novel compounds added to the 2008.1 version.²⁴

Retain					Reaction Specific Ranking						Atom Specific Ranking					
#	Scaffolds	Rings	FGs	Depth	Top 1	Top 2	Top 3	AUC		%?	Top 1	Top 2	Top 3	AUC		%?
								Mean	Median					Mean	Median	
1	Yes	Yes	Yes	0	44.671	56.959	67.413	0.667	0.804	45.552	25.747	42.915	50.662	0.555	0.622	22.515
					3.039	4.553	7.175	3.039			2.293	3.324	4.487	2.293		
2	Yes	Yes	Yes	1	49.450	61.205	70.278	0.734	0.833	25.209	27.714	43.969	52.119	0.583	0.613	3.920
					2.183	2.644	3.884	2.183			1.865	2.139	2.666	1.865		
3	Yes	Yes	Yes	2	53.006	63.010	71.759	0.744	0.852	25.148	27.736	44.069	51.839	0.580	0.608	3.920
					2.183	2.644	3.862	2.183			1.865	2.139	2.666	1.865		
4	Yes	Yes	Yes	3	54.183	63.448	72.196	0.751	0.866	25.145	27.602	43.443	51.302	0.578	0.607	3.919
					2.183	2.644	3.862	2.183			1.865	2.139	2.666	1.865		
5	Yes	Yes	Yes	4	55.832	65.096	73.777	0.762	0.882	25.143	27.457	43.175	51.124	0.575	0.604	3.919
					2.183	2.644	3.862	2.183			1.865	2.139	2.666	1.865		
6	Yes	Yes	Yes	5	56.797	66.005	74.170	0.767	0.889	25.143	27.334	42.851	50.855	0.573	0.600	3.919
					2.183	2.644	3.862	2.183			1.865	2.139	2.666	1.865		
7	Yes	Yes	Yes	6	57.257	66.375	74.450	0.770	0.895	25.143	27.121	42.515	50.576	0.571	0.600	3.919
					2.183	2.644	3.862	2.183			1.865	2.139	2.666	1.865		
8	No	Yes	Yes	0	46.537	59.204	69.334	0.727	0.836	26.336	29.617	48.504	57.07	0.655	0.750	19.164
					1.009	1.657	2.688	1.009			0.614	1.02	1.624	0.614		
9	No	Yes	Yes	1	48.152	60.201	70.529	0.779	0.849	5.493	31.57	48.419	58.424	0.677	0.722	0.376
					0.570	0.614	0.834	0.570			0.439	0.461	0.560	0.439		
10	No	Yes	Yes	2	52.367	62.639	71.875	0.791	0.864	5.432	32.077	49.355	58.567	0.679	0.727	0.375

					0.570	0.614	0.823	0.570			0.439	0.461	0.560	0.439		
11	No	Yes	Yes	3	54.154	64.151	73.629	0.802	0.882	5.429	32.176	49.278	58.545	0.677	0.727	0.374
					0.570	0.614	0.823	0.570			0.439	0.461	0.560	0.439		
12	No	Yes	Yes	4	55.887	65.718	75.141	0.813	0.897	5.427	32.176	49.069	58.479	0.674	0.725	0.374
					0.570	0.614	0.823	0.570			0.439	0.461	0.560	0.439		
13	No	Yes	Yes	5	56.824	66.468	75.483	0.818	0.903	5.428	32.11	48.584	58.193	0.673	0.725	0.374
					0.570	0.614	0.823	0.570			0.439	0.461	0.560	0.439		
14	No	Yes	Yes	6	57.575	67.174	75.979	0.821	0.907	5.428	31.945	48.231	58.017	0.672	0.722	0.374
					0.570	0.614	0.823	0.570			0.439	0.461	0.560	0.439		
15	No	No	Yes	0	41.817	54.683	66.719	0.524	0.600	68.345	28.869	45.825	56.268	0.446	0.465	63.165
					2.194	6.363	12.792	2.194			1.459	4.191	9.183	1.459		
16	No	No	Yes	1	40.316	52.147	64.209	0.714	0.750	3.042	29.239	46.185	56.615	0.722	0.788	0.394
					0.592	0.636	0.845	0.592			0.494	0.516	0.625	0.494		
17	No	No	Yes	2	46.606	58.459	69.738	0.750	0.804	2.983	29.89	47.641	58.423	0.731	0.795	0.394
					0.592	0.636	0.834	0.592			0.494	0.516	0.625	0.494		
18	No	No	Yes	3	48.096	59.828	70.235	0.767	0.833	2.982	30.331	48.6	59.261	0.731	0.800	0.393
					0.592	0.636	0.834	0.592			0.494	0.516	0.625	0.494		
19	No	No	Yes	4	48.714	59.110	69.142	0.777	0.842	2.980	31.235	49.923	60.209	0.731	0.800	0.393
					0.592	0.636	0.834	0.592			0.494	0.516	0.625	0.494		
20	No	No	Yes	5	50.811	61.881	71.681	0.787	0.861	2.980	32.051	50.75	60.981	0.731	0.800	0.393
					0.592	0.636	0.834	0.592			0.494	0.516	0.625	0.494		
21	No	No	Yes	6	51.959	62.101	71.063	0.797	0.870	2.979	32.624	51.014	61.367	0.734	0.800	0.393

					0.592	0.636	0.834	0.592			0.494	0.516	0.625	0.494		
22	No	No	No	0	32.023	47.669	58.875	0.323	0.000	73.644	32.203	47.929	59.326	0.345	0.000	72.296
					0.922	5.485	15.206	0.922			0.658	4.191	13.033	0.658		
23	No	No	No	1	36.066	47.975	59.796	0.690	0.719	0.067	38.576	56.025	67.351	0.745	0.827	0.005
					0.044	0.088	0.187	0.044			0.033	0.066	0.154	0.033		
24	No	No	No	2	39.546	51.454	62.397	0.709	0.748	0.028	38.74	56.223	67.746	0.754	0.828	0.005
					0.044	0.088	0.176	0.044			0.033	0.066	0.143	0.033		
25	No	No	No	3	40.007	52.464	62.562	0.719	0.759	0.028	39.102	56.64	67.921	0.750	0.827	0.005
					0.044	0.088	0.176	0.044			0.033	0.066	0.143	0.033		
26	No	No	No	4	41.17	53.496	63.572	0.731	0.778	0.028	39.618	57.21	68.218	0.748	0.8214	0.005
					0.044	0.088	0.176	0.044			0.033	0.066	0.143	0.033		
27	No	No	No	5	42.465	55.142	65.503	0.737	0.782	0.028	40.244	57.638	68.382	0.746	0.817	0.005
					0.044	0.088	0.176	0.044			0.033	0.066	0.143	0.033		
28	No	No	No	6	42.564	55.285	65.251	0.750	0.804	0.028	40.727	58.461	68.898	0.748	0.818	0.005
					0.044	0.088	0.176	0.044			0.033	0.066	0.143	0.033		

Table 4.5 CASSI Validation Results. The percentages of structures which contain at least one SOM atom in the top1, 2 and 3 positions as well as the mean and median of AUC values produced by all test dataset structures are shown in black. The corresponding percentages of structures which could not be evaluated and thus did not contribute towards the resulting values are shown in grey underneath their corresponding values. The “%?” values represent the percentage of atoms for which no prediction could be made – i.e. unknown atoms (see Table 4.2 & Appendix D – Frequency Distribution of CASSI Prediction on Test Dataset).

None of the fragment dictionaries performed better than FAME in any of the five performance measures. With the exception of dictionaries #7, 13 and 14, the performance statistics of most dictionaries, especially when combined with the atom specific ranking algorithm, shows that CASSI performed poorly compared to both MetaPrint2D and FAME. Fragment dictionaries #7, #13 and #14 in combination with the reaction specific ranking algorithm produced results that are comparable to the performance obtained by MetaPrint2D and MetaPrint2D-React.

It is interesting to note that with the exception of dictionaries #22 - 28, where both ranking algorithms produced fairly similar results, the reaction specific ranking algorithm consistently performed better than its atom specific counterpart in all performance measures (except for a small difference in AUC mean and median in dictionary #16). This shows that the inclusion of transformation type information (and the exclusive usage of likelihood of the most frequently observed transformation) brought about an improvement in the performance of CASSI. Given that around 30% of all transformations contained within the database are of an unknown reaction type, reaction specific ranking algorithm produced equivalent or better performing models compared to atom specific ranking algorithm, despite only utilising a subset of the knowledge included in the database. Also, the performance of #7 is one of the best out of all dictionaries tested, which is surprising as 25% of atoms cannot be processed as they cannot be matched to any fragments within the dictionary of fragments. It is possible that because of the more stringent selection criteria of the reaction specific ranking algorithm (usage of transformation statistics rather than the lack of them), more appropriate, relevant information was passed on and used during the prediction.

By breaking up the scaffolds, then rings and lastly functional groups during fragmentation at the same fragmentation depth (dictionaries #14, #21 then #28), the coverage of atom environments by these dictionaries of fragments shows a steady improvement – an initial drop from 25.143% to 5.428% when scaffolds are broken then a further decrease to 2.979% when ring retention is disabled and finally to 0.028% when functional groups are also broken down. However, once the rings are broken, the performance of CASSI dropped when the reaction specific ranking algorithm is used. These findings are expected since even though more atoms are covered, it is also easier for fragments and atoms that are not in the appropriate environment (within their parent structures) to be recognised and thus erroneously contribute towards the atom's metabolic stability score.

The same trend is not seen with the use of the atom specific ranking algorithm. The performances of the reaction specific ranking algorithms are best when combined with the fragmentation parameters: break scaffolds, retain rings and retain functional groups (dictionaries #8 - #14), the performance of the atom specific ranking algorithm, although not as good as their reaction specific

counterparts, are best with both ring and scaffold retention options disabled during fragmentation. This is possibly because the breaking of rings and scaffolds allowed more of the statistics collected in the dictionary of fragments to become available for all query fragments.

In almost all cases (with three small exceptions), an increase in fragmentation depth consistently brought about an improvement in performance of CASSI, even in cases where the increase in fragmentation depth brought no discernible improvement in the coverage of atom environments. This suggests that the usage of larger fragments over smaller fragments for stability scoring lead to the matching of query structure fragments with more appropriate chemical environments contained within the dictionaries.

4.5 Graphical User Interface

The Coralie Cheminformatics Platform is a Java application created based on the Eclipse framework. Functionalities within the application are typically contained within individual modules. A SOM module has been created for CASSI within the application. Within the SOM module, there are four separate tabs: Fragmentation, Prediction, Analysis and Validation.

The “Fragmentation” tab (Appendix A – Fragmentation Tab in Coralie’s SOM Module) within the SOM module allows exploration of the dictionary dataset structures, which are displayed in a matrix, and also shows a panel where different combinations of fragmentation parameters can be specified by the user before initialising fragmentation of the dictionary dataset. There is an option to filter out unknown transformation types (not recognised by SMIRKS pattern in Table 2.6) in the fragmentation panel. When this is selected, the collection of transformation statistics will ignore all transformations with a type 0, which indicates that the reaction type was not recognised by MetaPrint2D-React. A dictionary of fragments can be created by fragmenting the training dataset within the “Fragmentation” tab.

Once a dictionary of fragments has been created, the “Prediction” tab (Appendix B – Prediction Tab in Coralie’s SOM Module) within the SOM module allows individual query structures to be submitted via the structure editor. The query structure submitted via the structure editor will be fragmented using the same fragmentation parameter used to create the dictionary and these are used to produce the predicted metabolic stability score on each atom. SOM prediction using atom or reaction specific ranking algorithms, as well as the option to view prediction results on all atoms or only the top three most unstable atoms in both cases, are available within the tab. When an option to only highlight the top three most unstable sites is selected, the most unstable atom is highlighted in red, the second most unstable atom in amber and the third most unstable atom in yellow.

Selecting the option to highlight all atoms applies the appropriate highlight colour (red, amber and yellow) to those atoms with predicted metabolic stability scores based on a set of arbitrary, discrete cut off values. If an atom has a predicted stability score of 1, it will be highlighted in green. If an atom is unknown with regards to the training dataset, it will be highlighted in grey. This colour scheme also applies for the “Analysis” and “Validation” tab.

Upon the selection of an atom in the query structure, all reaction types which contributed to its final calculated stability score will be displayed along with the reaction type’s likelihood (Equation 4.1) and its occurrence counts (when creating the dictionary of fragments). If a particular reaction type is selected, the supporting examples (relevant substrate structure from the training dataset) will be displayed in the same tab.

If a query structure has been submitted for SOM prediction in the “Prediction” tab, the fragments produced by the query structure along with each fragment’s metabolic stability statistics (from the dictionary of fragments) can be accessed in the “Analysis” tab (Appendix C – Analysis Tab in Coralie’s SOM Module). When a fragment is generated by the submitted query structure, examples of training dataset structure which contain the fragment will be displayed along with all collected metabolic stability statistics of the fragment.

A “Validation” tab (Appendix E – Validation Tab in Coralie’s SOM Module) has also been created in the SOM Module in Coralie to allow for visual inspection of the prediction results carried out on a test dataset containing structures annotated with SOM transformation records (in this case, gathered from the Accelrys Metabolite Database). Atoms that are annotated as SOM are highlighted in purple. Prediction results produced by CASSI are highlighted in red, amber, yellow, green and grey (according to the same rules used in the “Prediction” and “Analysis” tab). These prediction results highlights are applied on top of the purple highlights marking SOM retrieved from the Accelrys Metabolite Database, allowing for a visual comparison. Both reaction and atom specific ranking approaches are supported.

If the input test dataset has also been annotated with pre-calculated top three SOM prediction values of another SOM predictor (in this case, FAME¹ was used), these prediction results are displayed alongside prediction scores from CASSI, allowing for a visual comparison between different methods.

4.6 Conclusion

This chapter reported the development of CASSI, a SOM prediction method based on statistical metabolic stability information collected from fragments. The overall poor performance of CASSI suggests that whilst the method carries out predictions rapidly, merely using the structure of atoms (which includes the elemental information of atoms as well as the bond order, including aromatic bonds) is not an adequate description of the chemical context of the atoms involved.

A method that can better gauge the appropriateness of the fragment being used in metabolic stability prediction, possibly by use of a similarity measurement between dictionary and query fragment structure, should be able to predict SOM on query structures with greater accuracy. The development of FamePrint is an attempt to utilise a similarity measurement in order to produce a better SOM prediction model. This is reported in the next chapter.

5. FamePrint SOM Predictor: Fingerprint-based Sites of Metabolism Prediction

5.1 Introduction

With the encouraging performance of FAME in predicting SOM with only seven atom-based descriptors, the expansion of the atom-based descriptions used in FAME to a fragment/substructure level description in order to capture a broader chemical environment may prove to be an appropriate description of a fragment. This may allow a fragment-based SOM prediction methodology using a similarity-based approach to be created. The fragment/substructure level description based on the seven descriptors employed in FAME (Table 2.9) can be used to gauge the similarity between a query fragment and a dictionary fragment (containing metabolic stability information). The similarity score can then be used to weight the dictionary fragment's contribution towards the final predicted metabolic stability score of the query fragment.

This chapter reports the development of FamePrint, a fingerprint-based SOM prediction method developed based on an atom pair description of the seven atom-based descriptors (Table 2.9) used by the FAME¹ models. If proven to be successful in identifying metabolically vulnerable sites, the fragment-based nature of FamePrint means that the method can readily be expanded to incorporate the identification of bioisosteric replacements and the generation of new structures whilst maintaining a compound's metabolic stability. As seen in other literature examples outlined in 2.2, similarity based methodologies developed for other purposes, such as to aid combinatorial library design,^{57,58} can also be adapted to identify bioisosteric replacement groups.

FamePrint, like CASSI, has also been implemented in the Coralie Cheminformatics Platform (see section 4.1.1) in order to leverage the fragmentation algorithm within the application.

5.2 Methods

5.2.1 Data Source and Preparation

FamePrint is a SOM predictor and like CASSI, a data source with metabolic stability information is required. Version 2011.2 of the Accelrys Metabolite Database (section 3.1) is used as a source of data. SOM annotations are generated from transformations contained within the database (sections 3.2). Reaction types are determined by SMIRKS pattern used by MetaPrint2D-React (Table 2.6) and propagated to structures in the dataset as detailed in section 3.3. Structures are selected from the Accelrys Metabolite Database according to procedure outlined in section 3.4, producing a unique set of 30467 substrate structures. These structures are then washed by MOE³⁸ (section 3.5) before being processed with the corresponding CFP text file containing SOM annotations and reaction type records (section 3.6). After the propagation of SOM and reaction type information to the appropriate atoms, the final dataset contains 30,467 unique substrate structures in SD file format where each structure contains SOM annotations (plus reaction type information) in their respective SD tags.

5.2.2 Descriptors

The same seven atom-based descriptors used in the FAME (section 2.1.4) study is calculated for all atoms in the training and test datasets using CDK version 1.5.9:

Descriptor	Descriptor definition
PartialTChargeMMFF94 (CDK)	Total partial charges of a heavy atom as derived from the MMFF94 model
PartialSigmaCharge (CDK)	Gasteiger–Marsili sigma partial charges in sigma-bonded systems
PiElectronegativity (CDK)	pi electronegativity
SigmaElectronegativity (CDK)	Gasteiger–Marsili sigma electronegativity
SybylAtomType (CDK)	Sybyl atom type for a specific atom, encoding element type and hybridization state
EffectiveAtomPolarizability (CDK)	Effective atom polarizability of a heavy atom
MaxTopDist (Span2End ³⁹)	Maximum topological distance between two atoms of a molecule (including explicit hydrogen atoms)

Table 5.1 The seven descriptors used and their definition.

The maximum topological distance (MaxTopoDist) describes the longest topological distance of the molecule which gives an idea of the size of the molecule. Sybyl atom types encode implicit information regarding the atom's valence and hybridisation state. The remaining CDK descriptors are all responsible for encoding some aspect of the atom's electronic state; together they give an idea of the size, density and softness of the electron cloud around the atom. The selection of these descriptors aimed to capture the properties that are most significant in the interaction between a ligand and target pocket in a metabolising enzyme.

5.2.3 Discretisation

An equal frequency binning⁹⁷ algorithm has been implemented based on the method found in Weka⁴¹ under the list of unsupervised attribute filters. The adapted implementation is used to identify the cut off descriptor values used in this study.

5.2.4 Fragmentation

The fragmentation algorithm within the Coralie Cheminformatics Platform (detailed in section 3.7.1) is used to fragment all structures in this study.

5.2.5 Topological Atom Pair Fingerprint

A topological atom pair (TAP) fingerprint has been implemented in the following format: [discretised descriptor value of atom 1] – [topological distance between the two atoms] – [discretised descriptor value of atom 2]. The descriptors used are given in Table 5.1. A single fingerprint is created for one descriptor, resulting in a set of seven fingerprints generated for each structure. The descriptor values are discretised using an implementation of an equal frequency discretisation method (section 5.2.3). For each descriptor, each bit of the TAP fingerprint refers to a unique combination of [discretised descriptor value of atom 1] – [topological distance between the two atoms] – [discretised descriptor value of atom 2].

5.2.6 Training and Test Dataset Generation

As is the case in FAME¹, due to the length of time required to calculate the required descriptors for structures with more than 100 heavy atoms, these structures have been removed from the dataset. The remaining 30,391 substrate are randomly split (70:30) into training and test datasets using the Coralie Cheminformatics Platform (as detailed in section 4.2.4), resulting in a training and test dataset of 21,270 and 9,116 structures respectively (five structures which could not be handled by the CDK descriptors were discarded). Out of the 9,116 test dataset structures, three different test datasets are generated as follows:

- Test set 1: 30% of the whole dataset, consisting of all 9116 test structures.
- Test set 2: subset of test set 1, consisting of the top 25% of structures that are most dissimilar to any structure found in the training dataset where similarity is determined by the average Tanimoto similarity score among all seven TAP fingerprints for the structure.
- Test set 3: created in the same manner as test set 2, but consisting of the top 5% most dissimilar structures to the training dataset.

5.2.7 AUC Calculation

This uses the same AUC calculation algorithm implemented for CASSI (see section 4.2.6).

5.3 FamePrint Development

5.3.1 FamePrint Workflow

5.3.1.1 Model Creation

The training dataset containing substrate structures with SOM annotations obtained from the Accelrys Metabolite Database (section 5.2.1) is first subject to descriptor calculation (section 5.2.2), producing 7 descriptor values for each atom in each substrate structure. The descriptor values generated are then discretised (section 5.2.3) by an equal frequency discretisation method. All substrate structures in the training dataset are then fragmented (section 5.2.4) to produce a dictionary of fragments.

The previously discretised descriptor values are then used to generate a set of seven TAP fingerprints (detailed in section 5.3.4) for the fragment, one fingerprint for each descriptor in Table 5.1. As these fingerprints are generated based on the descriptor values calculated from atoms in their original chemical context within the parent structure, they contain information regarding the atom's environment before fragmentation has occurred. Metabolic stability information associated with the fragment (obtained from the parent structure's SOM annotations) are gathered and stored along with the TAP fingerprints, therefore each set of TAP fingerprint has its own stability record.

For each unique fragment structure, there can be (and are in majority of cases) more than one pair of TAP fingerprints and stability score. As fragments are generated from different parent structures where the atoms of the fragment are in different chemical contexts, this gives rise to different descriptor values. Pairs of TAP fingerprints and stability score referring to the same fragment structure are stored together under the same structure, producing a dictionary of unique fragments

with varying number of associated fingerprints and stability scores. This dictionary of fragments is used to generate prediction of the metabolic stability of query structures. The workflow for producing the dictionary of fragments is provided in Figure 5.1a.

As shown by the performance statistics of CASSI, the fragmentation parameters are expected to play an important role in prediction performance. A similar combination of fragmentation parameters to the set used for CASSI evaluation are also tested for their suitability with the FamePrint approach, except the dictionaries that produced only single atoms fragments (e.g. #22 in Table 4.5), as two or more atoms are required to generate the TAP fingerprints. Unlike CASSI, reaction types have not been included in the first instance. The performance, speed and size of the resulting FamePrint dictionaries were evaluated before the feasibility of including the reaction types within an interactive environment was determined.

There are other points in the workflow where decisions have to be made and performance needs to be evaluated to identify the optimum choice. Choices such as which CDK library version should be used for descriptor calculation, the discretisation method used, implementation of TAP fingerprint, choice of similarity measure and data points to use for stability score prediction are given in the following sections.

The workflow diagram in Figure 5.3 also detailed the creation of the three test sets (section 5.2.6). These test sets are used to investigate the performance of FamePrint as well as testing its ability to extrapolate into more distant/unseen chemical space.

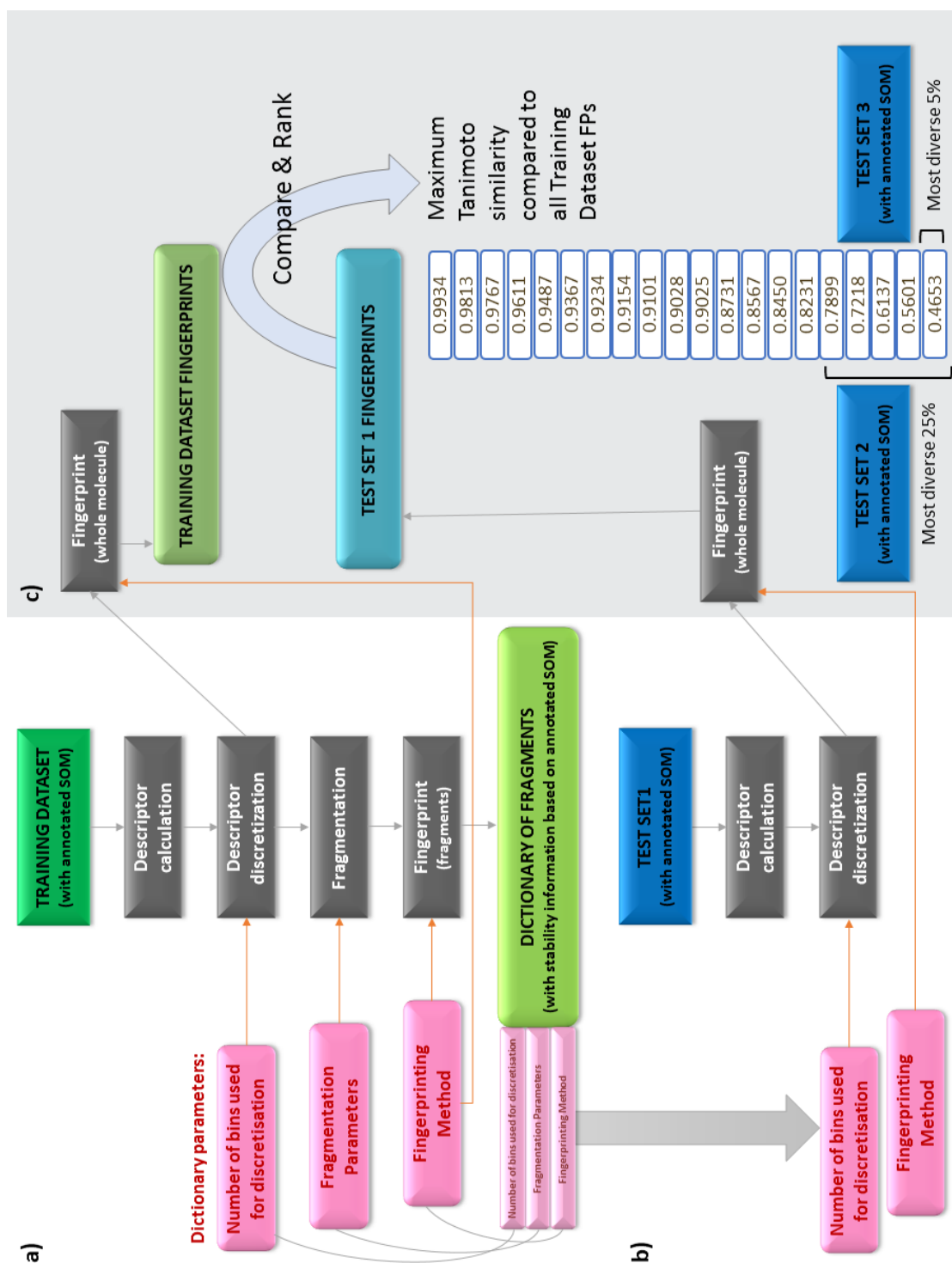


Figure 5.1 a) Workflow for creating a dictionary of fragments from a training dataset to be used for fragment stability prediction and a source of possible replacement fragments. B) The parameters used to create the dictionary are also used to produce fragment fingerprints from structures found in test set 1 – and subsequently c) test set 2 and 3.

After the generation of the dictionary of fragments from the training structures, prediction of metabolic stability can be undertaken (Figure 5.2). Descriptors are first calculated for the query structure, followed by the discretisation of descriptor values and the fragmentation of the query structure. The same discretisation boundaries and fragmentation parameters employed to produce the dictionary of fragments are used here. A set of seven TAP fingerprints are then generated for each query fragments. To generate the stability score for a query fragment, the fragment with the same structure is retrieved from the dictionary of fragments and the associated fingerprints are compared against the fingerprints generated for the query fragment. The similarity score calculation and subsequently the metabolic stability score calculation are detailed in sections 5.3.5 and 5.3.6.



5.3.2 Descriptor Calculation

The seven descriptors used in FamePrint are all atom-based descriptors calculated using the CDK library. FAME¹ version 1.0 uses CDK version 1.4.18 for all descriptor calculations.

5.3.2.1 CDK Version 1.4.18 vs CDK Version 1.5.9

During the early stages of FamePrint development, several issues were identified with the CDK descriptors that are employed by FAME¹ (section 2.1.4) which resulted in incorrect and inconsistent descriptor values being returned by the CDK descriptors.

The first problem was with the CDK EffectiveAtomPolarizabilityDescriptor, which erroneously removed the datum stored in the “number of hydrogen atoms attached” field within the properties of the heavy atom undergoing the calculation. This changed the results of all other descriptors depending on the relative order in which the descriptor calculations occurred. This was due to an issue within the calculation of EffectiveAtomPolarizabilityDescriptor. The bug fix carried out here is included in FAME¹.

The second problem was that the PartialTChargeMMFF94Descriptor method produced incorrect values as the wrong atom types were identified and used. The issue was reported and included the example molecules provided in the Merck Molecular Force Field II paper⁹⁸ as tests to check the results returned by the descriptor, and also attempted to identify the source of the problem. These were then passed onto Dr. Mark Williamson and Dr. John May and steps were taken to correct the atom typing errors within CDK. The majority of atoms in the test cases were allocated the correct atom type when checked against the examples given in the MMFF II paper, however, not all have yet been fixed (in CDK 1.5.10) and it was suspected that the wrong parameters were read from the parameters file.

The FAME software which was used to generate the results published in the literature¹ did not contain any of the fixes mentioned above. When the evaluation of CASSI's performance was first carried out, these issues were not yet identified. Therefore, a comparison of model performance created using the original FAME code (FAME.0), the new FAME code containing the EffectiveAtomPolarizabilityDescriptor fix with CDK version 1.4.18 (FAME.1) and the new FAME code containing the repairs introduced here and employing CDK version 1.5.9 which contained a partial fix for the PartialTChargeMMFF94Descriptor (FAME.2) is carried out.

The same training and test datasets used in CASSI (section 4.2.4) are used in this study. The FAME performance values reported in Table 4.3 were obtained using FAME.0 and are used again here.

Descriptor calculation were carried out using the other two FAME versions (FAME.1 and FAME.2). Each training set produced was then used separately to train a random forest model using Weka version 3.6.9⁴¹ as detailed in FAME¹. Each of the FAME.1 and FAME.2 models were used to carry out predictions on the test dataset separately. The top 1, 2 and 3 scores along with AUC mean and median values are given here:

Version	Top 1	Top 2	Top 3	AUC	
				Mean	Median
FAME.0	66.517	77.817	84.893	0.856	0.941
FAME.1	66.188	77.553	84.915	0.855	0.941
FAME.2	66.473	77.312	84.564	0.853	0.939
p-value	0.880	0.853	0.756	0.634	
F	0.128	0.159	0.280	0.456	
F-critical	2.996	2.996	2.996	2.996	

Table 5.2 FAME.0, FAME.1 and FAME.2 validation results. The percentages of structures which contain at least one SOM atom in the top1, 2 and 3 positions as well as the mean and median of AUC values produced by all test dataset structures were shown. Their respective p-values (5% significance level), F and F-critical values were reported.

Single factor ANOVA tests are carried out in Microsoft Excel. All p-values (pre-set 5% significance level) obtained are much larger than the 5% significance level and all F values are smaller than the F-critical values computed from the data, taking into account the variability within the data. Both of these measures suggests that there are no statistically significant differences between the performances of the three FAME versions. The lack of significant differences in performance despite different descriptor results may be due to the compensation made by the random forest model, as the model is trained on values produced with systematic errors.

CDK version 1.4.18 cannot handle selenium atoms in sp² hybridisation states as the Sybyl atom type Se.2 was not supported by the CDKAtomTypeMatcher. However, this issue is fixed in CDK version 1.5.9, and therefore allowing FAME.2 to process more atoms than FAME.0 and FAME.1. As no statistically significant differences are found in the performance measures of the three FAME versions, CDK version 1.5.9 (FAME.2) is used in FamePrint to improve the types of atoms covered by the method.

5.3.2.2 Handling Invalid Descriptor Values

Descriptors are calculated for all substrate structures in the training and test datasets. Some structures contain atoms that cannot be handled by one or more of the CDK descriptors used. This is

due to the failure to handle certain atom types, mostly by the PartialTChargeMMFF94-Descriptor method and, in some cases, by the PartialSigmaChargeDescriptor method. Fewer than 5% of all structures in the database (training and test dataset combined) contain atoms which produce one or more invalid descriptor values. In these cases, the invalid descriptor values are ignored in the subsequent steps but valid descriptor values from other atoms and descriptors are used to produce valid fingerprints. During a similarity calculation, if one or both fragments contains invalid fingerprints produced by invalid descriptor values, that fingerprint is ignored in the similarity calculation. Instead, the rest of the (valid) fingerprints are used to generate the overall similarity score between the two structures.

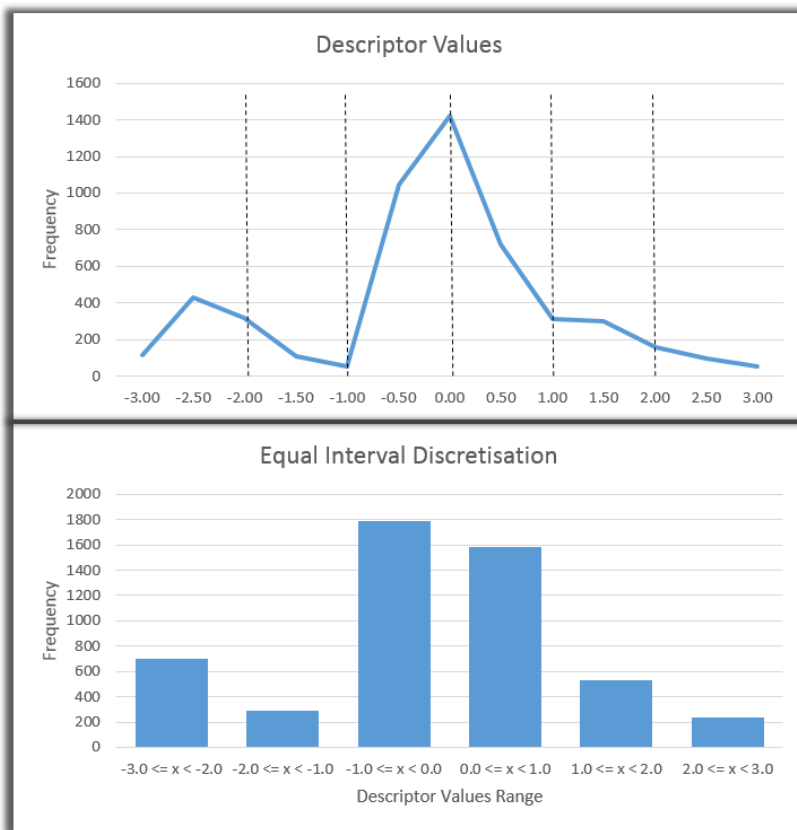
5.3.3 Choice of Discretisation Method

Different types of discretisation methods can be used to place continuous variables into a user specified number of bins. Two commonly used discretisation methods are considered: the equal interval discretisation method, where cut-off values are picked so the data range are partitioned into bins with each bin spanning the same distance, and the equal frequency discretisation, where cut-off values are chosen to produce bins containing equal number of data points after dataset discretisation.

An equal interval discretisation is more straightforward to implement and may result in a quicker discretisation operation. However, as descriptor values are not necessarily evenly distributed, it can easily lead to skewed value distributions (Figure 5.3a). An equal frequency discretisation, on the other hand, will split the descriptor values so that each bin has an equal number of instances (Figure 5.3b). This allows for greater discrimination over descriptor values in ranges containing higher number of instances.

A degree of fuzziness is desired in the method as it allows for the identification of similar but not completely identical fragments (as it is in the case of CASSI) and will be useful when attempting to identify bioisosteric replacements. This fuzziness can be introduced during the discretisation and binning process and both discretisation methods offer a degree of fuzziness. However, using an equal interval discretisation may make the discretisation process extremely sensitive to outliers. If descriptor values are extremely unevenly distributed, such as in the case shown in Figure 5.4, the majority of instances fall into one of two bins when the equal interval discretisation method is used. Equal frequency discretisation offers better discrimination power in this case. Due to its ability to better handle unexpected outliers, equal frequency discretisation is deemed more appropriate and is used for this study.

a) Equal Interval Discretisation



b) Equal Frequency Discretisation

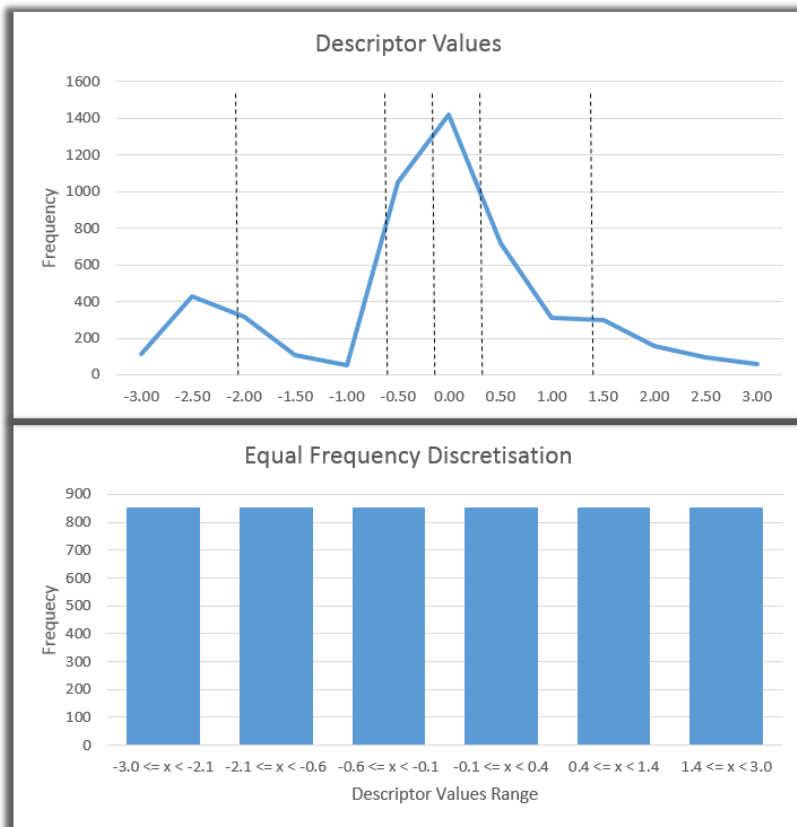
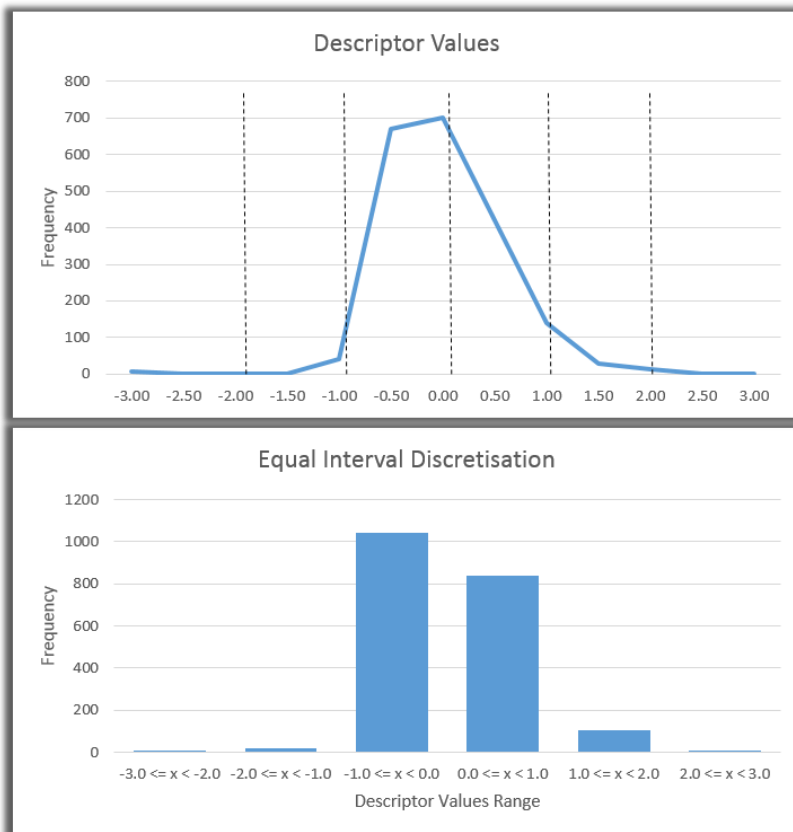


Figure 5.3 Equal interval (a) and equal frequency (b) discretisation examples.

a) Equal Interval Discretisation



b) Equal Frequency Discretisation

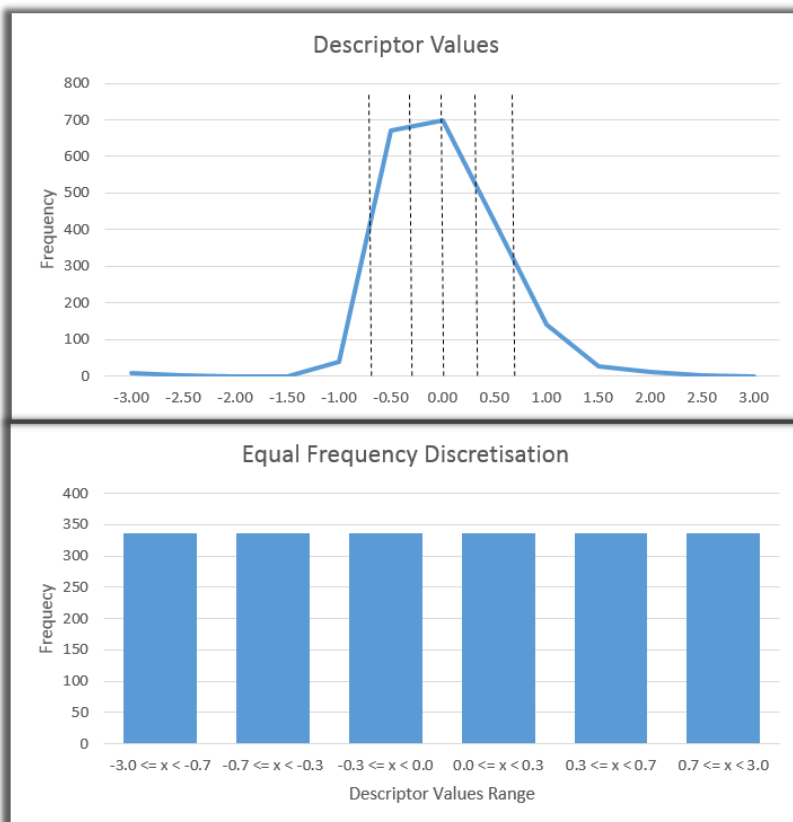


Figure 5.4 Equal interval (a) and equal frequency(b) with skewed distribution of descriptor values.

A range of different bins are tested in the systematic validation process in order to identify the optimal number of bins. The number of bins examined in this investigation ranges from four to eight. Less than or equal to three bins is deemed unlikely to offer enough discrimination. Bin numbers higher than eight are not attempted in the first instance as the resulting fingerprint sizes will very large, resulting in a large dictionary which needed to be kept in memory during the prediction process.

5.3.4 Topological Atom Pair Fingerprint

As all descriptors used are atom-based, a method needs to be devised to link these values together in order to produce a description for a fragment. The information about the distances between atom-based descriptor values need to be expressed in some form in the final description of the fragment. Distances between atoms can be represented by different methods; topological distances (e.g. the number of bonds between two atoms via the shortest path) and Euclidean distances (shortest distance between two atoms, usually applied on 3D structures) are amongst the most used distance measures. Topological distances are the preferred method as this would be fast to compute and only require the 2D structure of a molecule for computation. A Euclidean distances are potentially more useful when bound conformations of ligands are available, not the case here.

A modified topological atom pair fingerprint, based on the concept of atom pairs, has been developed to encapsulate the information required. Atom pairs are originally developed by Carhart *et al.* who defined an atom pair to be a substructure comprised of two heavy atoms in the form of [atom1 description] – [separation] – [atom 2 description], where descriptions used for each atom included information on the element of the atom, the number of heavy atom attachments and the number of electrons involved in π bonding.⁹⁹ This presented a rather simplistic view of the properties of a substructure. A similar concept based on Carhart's atom pairs was explored by Wagener *et al.* who developed a topological pharmacophore fingerprint in an attempt to identify bioisosteric replacements.⁶⁹ Here, the atom pairs are of the following format: [pharmacophore] – [topological distance] – [pharmacophore] where pharmacophores is one of the following: attachment point, hydrogen bond donor, hydrogen bond acceptor, hydrophobe, conjugated atom, aromatic atom, positively charged atom and non-hydrogen atom. All possible atom pairs present were enumerated and transformed into a fingerprint.

The topological atom pair (TAP) fingerprint developed for this study utilises the values from the atom-based descriptors used in FAME¹ (section 2.1.4) as the atom descriptors and topological distances are used by the descriptor as a measure of the separation between atoms. As five out of

the seven atom-based descriptors returned continuous values, an equal frequency discretisation method is required to transform the descriptor results into binned values before the atom pairs are computed.

5.3.4.1 Fingerprints Versions

A simple version of the fingerprint involves producing seven fingerprints for a fragment, one for each of the seven descriptors employed. For each descriptor, each bit of the fingerprint refers to a unique combination of [discretised descriptor value of atom 1] – [topological distance between the two atoms] – [discretised descriptor value of atom 2]; the same bin is identified regardless of the order of the input atoms. For each descriptor fingerprint, all atom pairs of the fragment are examined and used to set the appropriate bits of the fingerprint. This, however, only records the presence or absence of features corresponding to each fingerprint bit and does not take into account the frequency of occurrence of the number of pairs that each bin is responsible for. This version of the fingerprint is termed FamePrint fingerprint version 0 (FP00).

FP00 only considers the descriptor values of atoms in combination with another atom, along with their topological separation, no information is directly stored regarding the descriptor values of each atoms on their own. A second version of the fingerprint (FP01) is created which is identical to FP00, with the addition of bins in each fingerprint responsible for recording the presence or absence of atoms with discretised descriptors values corresponding to the discretised value for which the bin is responsible. This is termed the atom fingerprint layer.

Different options, taking into account the atoms in the fragment which are the connection points for replacement, are also implemented. Connection points are defined as atoms present in a fragment which originally connects to atoms present in the parent structure but not in the fragment (i.e. atoms through which the fragment was connected to the rest of the parent structure before fragmentation occurred). FP02 is based on FP00, but instead of one bit in the fingerprint referencing each unique combination of discretised descriptor pair and topological distance, three bins are assigned for each unique combination. When neither atom within the atom pair is a connection point, the first of the three assigned bits is set. However, if one or both atoms within the atom pairs have been identified as connection points, all three assigned bits in the fingerprints are set instead of one. This results in connection points properties having greater influence on the similarity score during similarity comparison. This is expected to aid the identification of suitable bioisosteric replacement fragments which has compatible connection points as well as suitable physchem properties. This version of the fingerprint is termed FP02.

Based on a similar principle, FP03 also pays extra attention to atom pairs which contain connection points. Instead of only having three assigned bits for each unique combination of descriptor values and topological distance (as is the case in FP02), six bits are assigned to each combination. In the case where neither of the two atoms in the atom pair is a connection point, only one bit of the six assigned bits is set, the same as in FP02. The difference arises when connection point atoms are involved. When only one atom of the atom pairs is a connection point, three bits out of six are set but if both atoms within the pair are connection points, all six assigned bits are set in order to place more importance on the relative distances between the connection points and their properties. This may not improve the performance of the fingerprint when used for metabolic stability prediction, but this is created with the aim of identifying more suitable bioisosteric replacement fragments.

5.3.4.2 Size of Fingerprints

The combination of fingerprint version and the number of discretisation bins determines the number of bins for each fingerprint. The sizes of fingerprints used during evaluation are given in Table 5.3

All combinations of fragmentation parameters (retain/break scaffolds, rings and functional groups) were tested for each fingerprint version with the number of discretisation bins ranging from four to eight (inclusive). Along with these fragmentation parameters, fragmentation depths ranging from zero to six are also tested, for FP00 and FP01. Fragmentation depths are limited to a maximum of 4 for FP02 and FP03 as the size of the resulting dictionaries of fragment are too large to be effectively handled (exceeding 4GB).

Fingerprint Version	Number of discretisation bins	Individual fingerprint sizes	Total fingerprint size
FP00	4	150 (5265)	6165
	5	225 (5265)	6615
	6	315 (5265)	7155
	7	420 (5265)	7785
	8	540 (5265)	8505
FP01	4	154 (5291)	6215
	5	230 (5291)	6671
	6	321 (5291)	7217
	7	427 (5291)	7853
	8	548 (5291)	8579
FP02	4	450 (15795)	18495
	5	675 (15795)	19845
	6	945 (15795)	21465
	7	1260 (15795)	23355
	8	1620 (15795)	25515
FP03	4	900 (31590)	36990
	5	1350 (31590)	39690
	6	1890 (31590)	42930
	7	2520 (31590)	46710
	8	3240 (31590)	51030

Table 5.3 The number of bins for each fingerprint version with the range of discretisation bins tested. Individual fingerprint sizes refer to the number of bits per fingerprint used (per descriptor) for all descriptors (except for SYBYLAtomType, size in brackets). The total fingerprint size refers to the total number of bits created for all seven fingerprints per fragment structure.

5.3.5 Fingerprint Similarity Calculation

When calculating a predicted metabolic stability score for a fragment, a method to compare the similarity of sets TAP fingerprints (from the query structure and structures in the model training dataset) is required to generate the prediction score. In Chemical Similarity Searching¹⁰⁰, Willett *et al.* listed a number of distance metrics and similarity measurements commonly used in Cheminformatics. Aside from Hamming Distance and Euclidean Distance (which would produce equivalent results when measuring similarities between dichotomous variables), all other similarity

measures listed produced very similar if not equivalent results when compared to the Tanimoto similarity, which was chosen for this study.

For a simple, straightforward comparison of two sets of fingerprints (for comparison of two fragments), each descriptor fingerprint from each set of fingerprint is compared to their counterparts in the second set of fingerprints and a Tanimoto similarity score calculated. The average of all seven Tanimoto similarity scores is used as the similarity score between the two fragments (Figure 5.5). Fingerprints from each descriptor are evaluated separately and the average used rather than evaluating the entire set of fingerprints as a whole as there are structures which may produce one or more invalid fingerprints (section 5.3.2.2) and by evaluating each descriptor fingerprints separately, it allows for valid fingerprints from these structures to produce a similarity score.

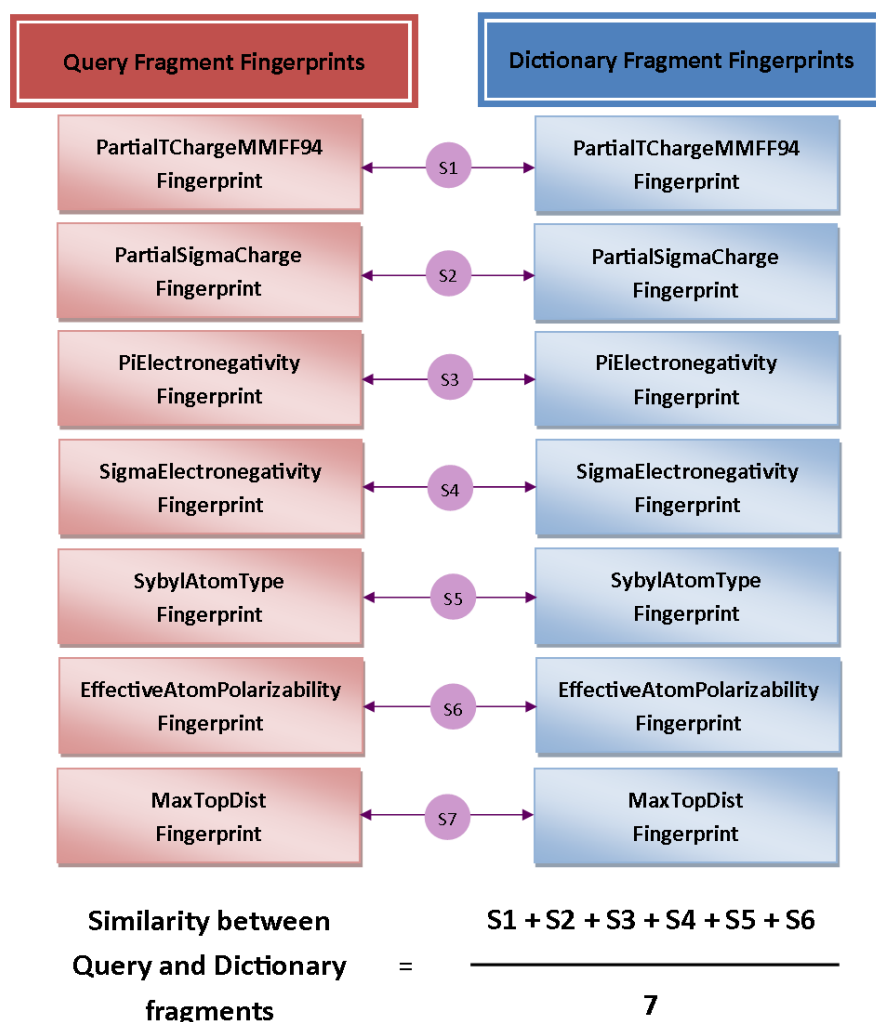


Figure 5.5 Obtaining the non-weighted similarity score between two sets of fragment fingerprints.

5.3.5.1 Fingerprint Weighting

The similarity score calculation shown in Figure 5.5 shows the steps for obtaining the similarity score between fingerprints where all bits and all fingerprints are weighted equally (Weighting: None). Each bit of the fingerprint carries equal weighting of one when two descriptor fingerprints are compared to produce a Tanimoto similarity value between two fingerprints. All seven fingerprints are also considered equally important and an average of the seven similarity values is used to produce the final similarity score between the two fragments.

However, it is shown during the descriptor selection for FAME that the seven descriptors used are not equally important (section 2.1.4.2). The information gain analysis performed for the FAME¹ (Table 2.9) shows that the PartialTChargeMMFF94 descriptor is the most important of the seven descriptors used and MaxTopDist the least.

Descriptor	Information gain	Normalised information gain
PartialTChargeMMFF94 (CDK)	0.0741	1.5595
PartialSigmaCharge (CDK)	0.0661	1.3912
PIElectronegativity (CDK)	0.0608	1.2796
SigmaElectronegativity (CDK)	0.0576	1.2123
SybylAtomType (CDK)	0.0411	0.8650
EffectiveAtomPolarizability (CDK)	0.0180	0.3788
MaxTopDist (Span2End ³⁹)	0.0149	0.3136

Table 5.4 Information gain analysis on the seven descriptors chosen. Figures taken from FAME¹.

These information gain scores from FAME¹ are normalised (Table 5.4) and applied to their respective fingerprint similarity scores for an overall weighted fragment similarity score (Weighting: FPOnly, Figure 5.6).

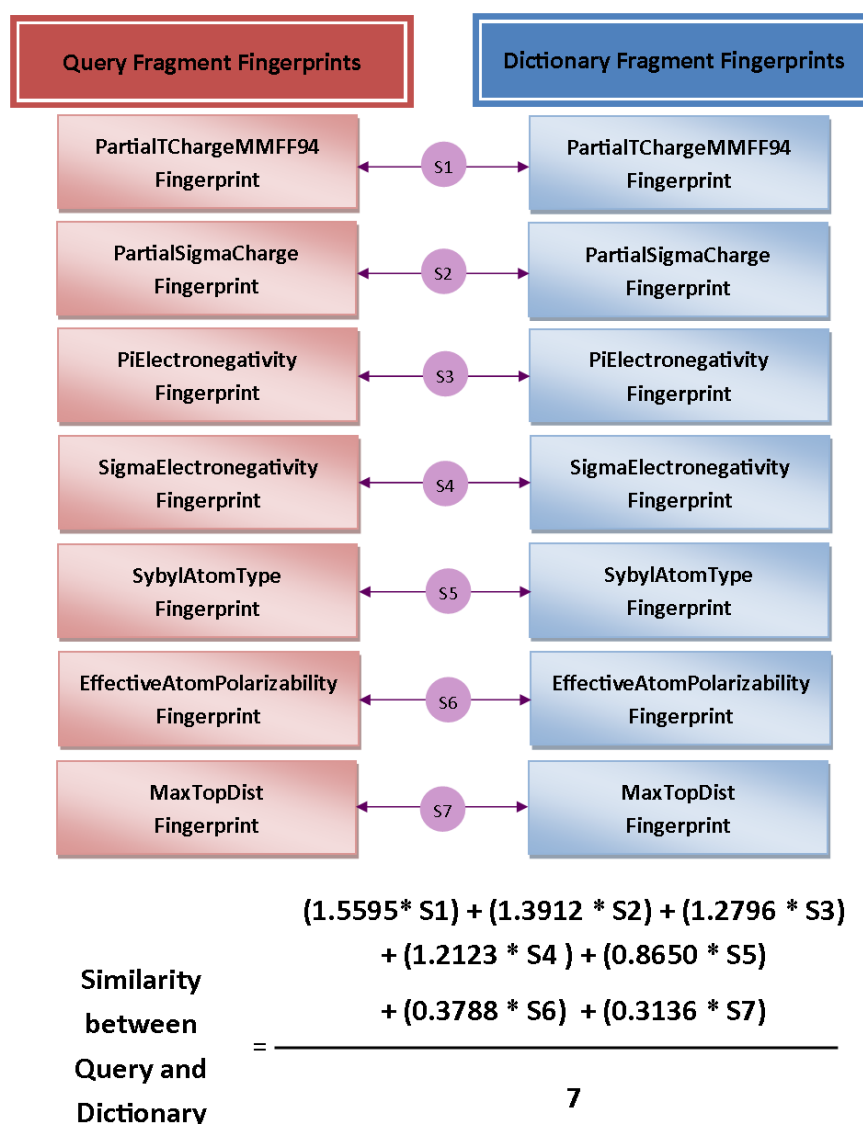


Figure 5.6 Weighted (FPOnly) similarity score calculation

Two other weighting schemes based on information gain for each bit of the fingerprint are implemented and evaluated. For the BitsOnly weighting scheme, the entropy gain of each bit position of each fingerprint is first obtained then these entropy gain values are normalised against all other entropy gain scores within each fingerprint. The entropy gain of each bit position is obtained by subtracting the system entropy, $H(1,0)$ in Equation 5.1 from the entropy of that bit position, $H(\text{bit})$ in Equation 5.1. This is calculated based on all fingerprints present in the model's training dataset and is unique for each combination of training dataset and model parameters. When used for similarity comparison, each bit is weighted by the appropriate information gain score for that given fingerprint position during the Tanimoto similarity score calculation (Figure 5.7). When all seven weighted similarity scores are calculated, the average of these scores is used as the overall similarity score.

The fourth weighting scheme created (Weighting: AllBits) is identical to the BitsOnly weighting scheme except for the normalisation step. After the calculation of entropy gain for each bit position for each fingerprint, these entropy gain values are normalised against all entropy gain values over all seven fingerprints (rather than within individual fingerprints). All four weighting schemes are tested for all dictionaries used in the validation procedure. The calculation for the information gain for each bit is as follows:

$$\text{Information gain} = H(\text{bit}) - H(1,0)$$

where

$$H(\text{bit}) = - \left[\left(\frac{\#bit_{stable}}{\#bit_{total}} \right) * \log \left(\frac{\#bit_{stable}}{\#bit_{total}} \right) + \left(\frac{\#bit_{unstable}}{\#bit_{total}} \right) * \log \left(\frac{\#bit_{unstable}}{\#bit_{total}} \right) \right]$$

$$H(1,0)$$

$$= - \left\{ \begin{aligned} & \frac{\#bit_{stable}}{\#bit_{total}} * \left[\frac{\#bit_{stable,1}}{\#bit_{stable}} * \log \left(\frac{\#bit_{stable,1}}{\#bit_{stable}} \right) + \frac{\#bit_{stable,0}}{\#bit_{stable}} * \log \left(\frac{\#bit_{stable,0}}{\#bit_{stable}} \right) \right] \\ & + \frac{\#bit_{unstable}}{\#bit_{total}} * \left[\frac{\#bit_{unstable,1}}{\#bit_{unstable}} * \log \left(\frac{\#bit_{unstable,1}}{\#bit_{unstable}} \right) + \frac{\#bit_{unstable,0}}{\#bit_{unstable}} * \log \left(\frac{\#bit_{unstable,0}}{\#bit_{unstable}} \right) \right] \end{aligned} \right\}$$

Equation 5.1 Information gain calculation for each bit in a fingerprint.

FP1 normalised entropy gain scores:

FP position	0	1	2	3	4	5	6	7	8	9	10
Entropy gain	0.8	0.1	3.4	0.2	0.4	0.3	0.2	2.2	0.7	0.1	1.6

Query Fragment FP1	0	1	0	0	1	0	0	1	1	0	0
-----------------------	---	---	---	---	---	---	---	---	---	---	---

Dictionary Fragment FP1	1	1	0	1	0	0	1	1	0	1	0
----------------------------	---	---	---	---	---	---	---	---	---	---	---

FP1 intersect	0	1	0	0	0	0	0	1	0	0	0
---------------	---	---	---	---	---	---	---	---	---	---	---

$$Q \cap D = 0.1 + 2.2 = 2.3$$

FP1 union	1	1	0	1	1	0	1	1	1	1	0
-----------	---	---	---	---	---	---	---	---	---	---	---

$$Q \cup D = 0.8 + 0.1 + 0.2 + 0.4 + 0.2 + 2.2 + 0.7 + 0.1 = 4.7$$

$$\begin{aligned} \text{Overall} \\ \text{(Tanimoto)} \\ \text{similarity} \end{aligned} = \frac{Q \cap D}{Q \cup D} = \frac{2.3}{4.7} = 0.4894$$

Figure 5.7 Example of entropy gain weighted (per fingerprint bit) similarity comparison.

5.3.6 Metabolic Stability Score Calculation

The overall predicted metabolic stability score for a fragment may be calculated from the stability scores of structurally identical fragments found in the dictionary, weighted by their similarity to the query fragment:

$$Stability_{Query} = \sum_{i=0}^k Similarity_{Query\ vs.\ i} * Stability_i$$

Equation 5.2 Stability of a query fragment

In Equation 5.2, $Stability_{Query}$ is the predicted stability score of the query fragment, $Similarity_{Query\ vs.\ i}$ is the Tanimoto similarity score (calculated according to 5.3.5) between the set of query fragment fingerprints and one set of dictionary fragment fingerprints and $Stability_i$ refers to the stored stability score of this set of dictionary fragment fingerprints, obtained ultimately from the Accelrys Metabolite Database (section 5.2.1).

During the training of the FamePrint model, a training dataset is fragmented to produce a dictionary of fragments. Each of these fragments will have associated fingerprints and stability scores along for each set of fingerprint (detailed in 5.3.1.1). When computing the predicted metabolic stability score for a query fragment, a search in the dataset of fragments for a fragment of the same structure is carried out. Once found, all fingerprints and stability scores associated with the stored fragment are retrieved. All the retrieved data may be used in the stability prediction of the query fragment using Equation 5.2, where k represents all available fragments. However, it is also possible to only use a subset of the stored fragment fingerprints and stability score pairs where the stored fingerprints are most similar to the query fingerprints (i.e. varying k in Equation 5.2). As part of the performance optimisation of FamePrint, usage of the top 3, 5, 10, 15 most similar as well as all available, relevant stored fingerprints and stability score pairs ($k = 3, 5, 10, 15$ and all) to produce the final query fragment stability score has been tested.

5.4 Graphical User Interface

A Biostere module has been created within the Coralie Cheminformatics Platform allowing the FamePrint methodology to be accessed via a graphical user interface (GUI).

A wizard style tool has been created within the application which loads a dataset (as an SD file). On the first page of the wizard tool (Appendix F – FamePrint Dataset Creation Wizard), the following options are available:

- A) Load in dataset already containing discretised descriptor values
- B) Load in dataset containing continuous descriptor values for:
 - a. Discretisation, bin size specified by user in wizard
 - b. Discretisation, bin boundaries contained in file
- C) Calculate a user specified selection of descriptors with the option for discretisation

The wizard also allows the user to export results as required before passing structures and descriptor values over to be fragmented or fingerprinted.

Once the dictionary of fragments has been created by the wizard, the dictionary file can then be loaded into the FamePrint tab in the Biostere module within the Coralie Cheminformatics Platform (Appendix G – Biostere Tab in Coralie). Once the dictionary has been loaded into memory (held in random-access memory (RAM) for quick access), the fragmentation parameters used to create the dictionary are displayed along with descriptors used in the dictionary.

The “Query” box displays the current query structure and it also allows the structure editor to be selected, where a new query structure of interest may be drawn and submitted for evaluation. Upon submission of the query structure, calculation of the relevant descriptors takes place, followed by the discretisation and fragmentation of the query structure, using the same set of parameters used to create the dictionary of fragments. Once the query fragments have been generated, fragments of the same structure contained within the dictionary of fragments are retrieved and the relevant sets of dictionary fragments fingerprints then used to carry out the metabolic stability prediction on the query fragments.

After the metabolic stability scores have been obtained for all query fragments, these are then ranked according to their predicted metabolic vulnerability. The resulting scores and fragment structures are displayed in the “Query fragments” matrix within the FamePrint tab. Query fragments are displayed in order of their predicted metabolic stability scores – with the most unstable fragment first. The predicted stability scores are displayed above each query fragment in the cell labels, which are also colour coded according to the predicted metabolic stability score.

When a query fragment has been selected, the parent structures which generated fragments of the same structure, and gave rise to the sets of fingerprints used in the metabolic stability prediction of the query structure, are displayed in the “Supporting examples” tab. As mentioned in section 5.3.6, several different metabolic stability prediction methods are investigated, and some methods do not require the use of all fingerprints from all matching dictionary fragments. The selection of a query fragment also initiates the search for suitable replacement fragments. This methodology is described in later sections. The overall metabolic stability of the query structure is taken to be the predicted stability score of the most unstable fragment generated and is displayed in a graphical format within the FamePrint tab. This can be used for visual comparison of the query fragment’s stability score with subsequent structures generated by the replacement of unwanted fragment(s).

5.5 Results and Discussion

A systematic investigation into the metabolic stability prediction performance of FamePrint is carried out on dictionaries of fragments produced using different discretisation and fragmentation parameters for various fingerprint versions, using different versions of the similarity and stability calculations procedures outlined in previous sections in this chapter. This evaluation carried out to identify the optimal combination of parameters and procedures which produces the model with the best metabolic stability prediction performance.

5.5.1 Limitation on Fragment Size

All fragments used for this study are limited to fragments containing two to sixteen heavy atoms (inclusive). The lower boundary is imposed as the method required at least two atoms in order to produce a TAP fingerprint, therefore the fragmentation parameter combination which produced single atoms only (break scaffolds, rings and functional groups, depth = 0) is not used. An upper boundary is set for the fragments in this dictionary for two main reasons. The first is to limit the size of the fingerprints required in order to cover fragments of all sizes in the dictionary. A maximum of 16 heavy atoms in dictionary fragments sets the topological distance upper limit at 15, which keeps fingerprint sizes at a manageable length whilst attempting to reduce the number of discarded fragments due to an imposed upper size limit.

The topological distance and heavy atom limits are also introduced as these dictionaries are intended for both the prediction of metabolic stability as well as for the identification of potential bioisosteric replacements. As with numerous other bioisosteric replacement and MMP studies^{69,79–81}, a substructure/fragment is only considered a suitable replacement or MMP if the substructure is equal in size or smaller than the remaining, unchanged part of the molecule. A maximum of 12 or 15 heavy atoms have previously been employed by other studies as arbitrary cut off points. As the average heavy atom count of structures in the dataset of unique substrate structures is 25, an upper limit of 12 atoms would mean that the fragment is just under 50% of an average sized structure. However, for FamePrint, a maximum of 16 atoms is used as the retention of a larger chemical context may aid the prediction of metabolic stability of compounds as well as locating a suitable replacement. It is also worth bearing in mind that the replacement fragment may be structurally very similar to the query fragment to be replaced, therefore the actual number of atoms altered by the replacement may be significantly lower than the number of atoms contained in the fragment.

5.5.2 Performance Measurements

In order to correctly gauge the performance of FamePrint and the various combinations of parameters and methods, appropriate performance measurements must be employed. For measuring the performance of CASSI, the top- k metrics and AUC values are used. These metrics are widely used in Cheminformatics and allows for direct comparison of the performance of FamePrint with CASSI and other previously reported methodologies from the literature. However, unlike CASSI and FAME¹ where SOM predictions are given for each atom, FamePrint predicts unstable fragments and are therefore not directly comparable to the other two SOM predictors. The true measurement of the success of FamePrint is not as straightforward, as the ranking of fragments rather than atoms complicates matters.

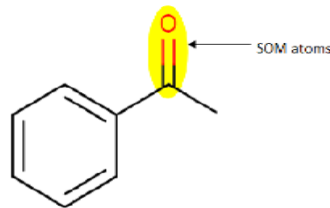
5.5.2.1 Coverage Score

As well as considering whether the top three most unstable fragments identified contain SOM atoms (when calculating the top- k metrics), the size of the fragment selected needs to be taken into account. A measurement of the structure coverage by the selected fragment are included to give an indication as to the fraction of the structure selected by the fragment in the top three positions when a SOM atom is found within the fragment. It was previously mentioned in section 4.4.2, the top- k metric does not take into account the number of SOM atoms vs. the number of atoms in each structure. It therefore does not account for the algorithm randomly selecting the correct atom by chance, therefore the usage of AUC values is included. In the case of FamePrint, a coverage score is absolutely necessary alongside a top- k metric score because, depending on the selected fragmentation parameters, the average size of fragments present in different fragment dictionaries (and selected for in the top three positions) can vary widely. For example, dictionaries generated with the fragmentation parameters break scaffolds, break rings, break functional groups and fragmentation depth of 1 consist entirely of two-atom fragments; on the other hand, dictionaries using the parameters keep scaffolds, keep rings, keep functional groups and a fragmentation depth of 6 contains numerous fragments close to or equal to 16 atoms, the upper limit of atoms allowed in a fragment. This represents an 8-fold difference between the potential sizes of the fragments selected. The selection of half the structure vs. the selection of two atoms out of 30 atoms within a structure makes a significant difference to the interpretation of the top- k scores compared to CASSI (and FAME¹) where the selection of one atom out of 10 is more likely than selecting one out of 30.

When calculating the coverage score, the size of each SOM entry (i.e. number of atoms in each SOM entry) should also be taken into account. If there are four atoms, all of which belong in the same transformation and are annotated as such, if the most unstable fragment contains four atoms and

completely covered all four SOM atoms, the algorithm should not be penalised for selecting a four-atom fragment over a three-atom fragment. Similarly, if there are two fragments of the same size but one which encompassed more SOM atoms than the other, this needs to be taken into account. In the example given in Figure 5.8, the two carbonyl atoms are annotated as SOM and the ideal scenario is for the algorithm to pick the two carbonyl atoms only, as the most unstable fragment, as seen in the first example. In this case, there are no extra (i.e. “redundant”) atoms selected and the coverage score, which represents the amount of extra information required for the selection of a fragment containing SOM atoms, is zero. In the second example, the acetyl group is chosen instead of only the carbonyl atoms, therefore there is an extra carbon atom chosen and the resulting coverage score is 1/7 (one non-SOM atom out of all seven non-SOM atoms in the structure). In the third example, the fragment selected contains only two atoms (the same as the first example), however, in this case there is only one SOM atom in the selected atom. The oxygen atom, as it is present in the same transformation identified by the third fragment, is added to the third fragment to create a pseudo acetyl fragment for the purpose of the coverage calculation. This results in a coverage score of 1/7, the same as the second example where an extra non-SOM carbon atom is also selected in the fragment.

Structure:






Fragment	Coverage
	$1 - \frac{\text{Structure} - \text{Fragment}}{\text{Structure} - \text{Fragment}} = 1 - 1 = 0$
	$1 - \frac{\text{Structure} - \text{Fragment}}{\text{Structure} - \text{Fragment}} = 1 - \frac{6}{7} = \frac{1}{7}$
	$1 - \frac{\text{Structure} - \left(\text{O=C} + \text{CH3} \right)}{\text{Structure} - \text{Fragment}} = 1 - \frac{6}{7} = \frac{1}{7}$

Figure 5.8 Example of coverage calculation and scores.

5.5.2.2 Overlap Score

As well as using the top three (and coverage score) and AUC as performance measurements, another performance measurement is also included. From visual inspection of structures put through the FamePrint tab in the Biostere module, the most unstable fragments predicted often overlap with each other. For instance, in the query structure investigated in Appendix G – Biostere Tab in Coralie, the following fragments have the lowest predicted metabolic stability scores:

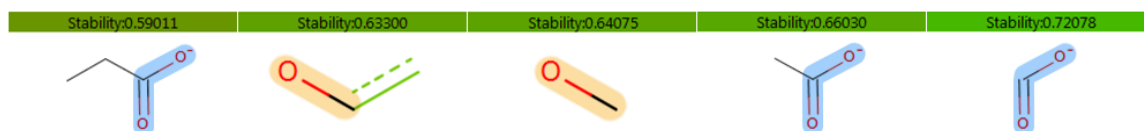


Figure 5.9 The top 5 most unstable fragments from the query structure submitted in Appendix G – Biostere Tab in Coralie for SOM Prediction.

The carboxylic acid substructure (highlighted in blue) is found in three out of the top five query fragments and the methanol substructure (highlighted in orange) is found in the second and third most unstable fragments.

Larger fragments incorporate more information on the original chemical context of the fragment compared to a smaller fragment. Therefore predictions made by larger fragments may provide a more accurate prediction even though the success of a larger fragment is statistically higher. It covers more atoms of the structure, therefore is more likely to have picked a correct SOM due to random chance. However, if a smaller, more specific fragment also has a low predicted metabolic stability score, it is possible both of these fragments are in fact highlighting the same region of the structure and the overlapping areas between the two fragments can be used as additional pointer. As a result, the ratio of (the SOM atoms found):(no SOM atoms found) in the overlapping regions within the top two and the top three most unstable fragments (if present) is also included as an additional performance measure. The coverage scores of the overlapped regions are also included.

5.5.3 Model Evaluation

After all fragments produced from the query structure had undergone metabolic stability prediction, the fragments are then ranked according to their stability scores. For the purpose of this evaluation, if two fragments have the same predicted stability score, the fragment with fewer atoms is prioritised. The top three and overlapping scores (and respective coverage scores) as well as AUC values are calculated after the ranking.

In order to determine the best set of parameters to use for FamePrint, all combinations of the following parameters are tested (except for the combination of break scaffolds, rings and functional group at a depth of 0):

Work flow variables classes	Parameters (variables) tested					
Number of discretisation bins used	4	5	6	7	8	
Fragmentation parameters	P1	P2	P3	P4		
Fragmentation depths	0	1	2	3	4	(5 6)
Fingerprint versions	FP00	FP01	FP02	FP03		
Weighting of fingerprints	None	FPOnly	BitsOnly	AllBits		
Number of most similar FP used	3	5	10	15	all (max possible)	

Table 5.5 List of different parameters tested for each FamePrint workflow variable classes. For definition of fragmentation parameters, see Table 5.6.

#	Retain scaffolds	Retain rings	Retain functional groups
P1	Yes	Yes	Yes
P2	No	Yes	Yes
P3	No	No	Yes
P4	No	No	No

Table 5.6 Key for fragmentation parameter used.

5.5.3.1 Effects of Fragmentation Depths 5 and 6

The increase in fragmentation depth beyond the depth of 4 offers only very slight changes in the overall performance statistics of the method. This is also accompanied by a significant increase in the time and computing resource required to produce the dictionaries, generate prediction results as well as fetching potential replacement fragments. The differences in performance statistics offered by extending the fragmentation depth to beyond 4 do not warrant the increase in computing resources required to produce and use the dictionary of fragments created at these depths. Some of the dictionary files created at fragmentation depth 6 were close to 2GB in size and took up 5GB of

memory when loaded into the FamePrint tab in the Coralie Cheminformatics Platform. It is decided that the dictionaries at depth 5 and 6 will not be produced for FP02 and FP03 (they are produced for FP00 and FP01) as several attempts to produce some of these dictionaries caused the java virtual machine to throw 'OutOfMemory' exceptions (when provided with a maximum of 12GB memory). These depths are excluded from the analysis of performance statistics carried out to identify the optimal set of parameters.

Dictionary Parameters		Dictionary Statistics		
	Depth	#Fragment	#Fingerprints in dictionary	Average fingerprint/ fragment
P1	0	2877	51831	18.02
	1	13238	167900	12.68
	2	39607	344244	8.69
	3	67367	488610	7.26
	4	96534	615042	6.37
P2	0	589	30204	51.28
	1	4197	90965	21.69
	2	17871	190070	10.64
	3	37382	286239	7.66
	4	110734	860458	7.77
P3	0	335	13289	39.67
	1	1225	64235	52.44
	2	4232	156015	36.87
	3	10988	283019	25.76
	4	23936	459833	19.21
P4	1	66	49894	755.97
	2	377	140123	371.68
	3	1277	265876	208.20
	4	3717	441811	118.86

Table 5.7 Number of fragments and fingerprints in each dictionary.

As expected and seen in Table 5.7, the number of fragments in each P1, P2, P3 and P4 dictionary increases with increasing fragmentation depth. It is also not surprising that the P4 dictionaries have the lowest number of fragments and the highest fingerprint/fragment ratio. There is a general decrease in the fingerprint/fragment ratio as the fragmentation depth increases which is a sign of the fragment structures in the dictionaries getting more specific.

It is expected that some workflow variables (Table 5.5) will have more impact on the overall performance of FamePrint than others. In order to identify which variables are more influential, each variable is tested individually. The performance statistics produced by dictionaries which are created with the same combination of parameters (except for the tested variable) are extracted and their performance compared against each other. This is carried out for all dictionaries created with the parameters outlined in Table 5.5. Single factor ANOVA (α level = 0.05) is performed for each performance measurement for each variable tested to identify whether the variables result in statistically significant differences in the performance statistics.

5.5.3.2 Effects of Fragmentation Parameters

Unsurprisingly, the combinations of different fragment parameters, and separately, the changes in fragmentation depths makes the most significant differences to the performance of FamePrint. Fragmentation parameters and fragmentation depth are the only two variables which consistently showed statistically significant differences in all measurements of performance, i.e. where the F value is significantly larger than the F_{crit} in all cases and all p -values < 0.05.

#	Test set	Top 1	Top 2	Top 3	Mean AUC
P1	Test set 1	69.4 (0.34)	76.7 (0.32)	80.1 (0.30)	0.660
	Test set 2	69.8 (0.35)	78.0 (0.33)	81.6 (0.31)	0.649
	Test set 3	69.9 (0.33)	77.8 (0.31)	80.7 (0.30)	0.647
P2	Test set 1	71.7 (0.33)	79.7 (0.32)	83.8 (0.31)	0.712
	Test set 2	71.1 (0.35)	79.9 (0.34)	84.6 (0.32)	0.692
	Test set 3	70.8 (0.31)	79.1 (0.31)	83.5 (0.30)	0.679
P3	Test set 1	63.0 (0.17)	71.0 (0.16)	74.8 (0.15)	0.658
	Test set 2	66.3 (0.21)	73.9 (0.19)	77.7 (0.18)	0.677
	Test set 3	68.4 (0.22)	75.6 (0.21)	78.7 (0.19)	0.688
P4	Test set 1	64.3 (0.14)	72.1 (0.13)	77.6 (0.13)	0.735
	Test set 2	68.5 (0.17)	75.5 (0.16)	80.3 (0.16)	0.748
	Test set 3	72.0 (0.17)	78.7 (0.17)	82.6 (0.17)	0.770

Table 5.8 Averaged performance statistics for dictionaries using each combination of fragmentation parameters. Top 1, 2 and 3 scores shows the % of predictions where a fragment contains a true SOM within the top 1, 2 or 3 positions. In brackets, the coverage score for the fragment in the first, second and third place.

It is interesting to note that for the combinations P1 and P2, both the top- k and AUC metric shows fairly consistent performance across the three test sets. There is sometimes a counter-intuitive

increase in the top-*k* performance statistics when the distance from training to the test dataset structure increases (going from test set 1 to test set 3). This is however often accompanied by an increase in the coverage score.

When the mean AUC values are examined, as expected, the performance dropped as the test structures are increasingly dissimilar to the training dataset. A brief investigation into the average ratio of SOM atoms to structure atoms in each test set also reveals that in all collections of test datasets, structures in test set 2 and test set 3 had an increasingly high percentage of the number of SOM atoms per structure (Table 5.9). This may explain the increase in the top-*k* performance of test set 2 and 3, despite a lowering of the mean AUC value.

Test set	Discretisation bin				
	4	5	6	7	8
1	0.166	0.166	0.166	0.166	0.166
2	0.194	0.191	0.189	0.190	0.188
3	0.236	0.236	0.210	0.199	0.196

Table 5.9 The average ratio of SOM atom(s) : structure atoms in each test set produced.

However, this is not the case for P3 and P4 dictionaries. Both of these combinations produced dictionaries where the performance increases with the distance to the training dataset structures. This is the case for both the top-*k* metrics and the mean AUC measures. In all cases, this increase in performance going from test set 1 to test set 3 is also accompanied by an increase in the coverage of the fragment required to make a correct prediction. This is likely due to the fact that P3 and P4 produce small fragments which cannot sufficiently capture the context required for the accurate recognition of a true SOM; there may be a large number of environments which appear similar but are not considered appropriate when larger contexts are included. As the distance from the training dataset increases, these errors may be less likely to occur as it is less likely for a small, generic fragment to match any fragments generated by the increasingly dissimilar test set structures, therefore leading to an increase in performance statistics.

If only the top-*k* metrics were used as a measure of performance, then the dictionaries produced by breaking scaffolds but retaining rings and functional groups (P2) appear consistently to produce the best performance (except for test set 3, which was slightly outperformed by P4 at the top 1 position). However, if considering the mean AUC as a measurement of performance, then P4 dictionaries produce the best results out of the four combinations tested. The improvement in performance can also be due to the fact that depths of 0 and 1 are not included in the P4 study as

they only produce single atom fragments and two atom fragments which produced single bit fingerprints and the increase in depth brings about a general increase in performance. All performance measurements produced by P4 dictionaries also shows significantly smaller standard deviations compared to P1, P2 and P3 dictionaries.

It is decided that for a fairer comparison, all results from dictionaries produced with depth 0 and 1 are omitted from the analysis and a new comparison is made between P1, P2, P3 and P4:

#	Keep: scaffolds, rings, functional groups		Top 1	Top 2	Top 3	Mean AUC
P1	Yes, yes, yes	Test set 1	75.4 (0.40)	81.3 (0.38)	84.6 (0.37)	0.742
		Test set 2	75.0 (0.40)	81.7 (0.39)	85.3 (0.37)	0.727
		Test set 3	75.5 (0.37)	81.9 (0.36)	84.7 (0.34)	0.731
P2	No, yes, yes	Test set 1	79.4 (0.41)	85.5 (0.40)	89.0 (0.38)	0.796
		Test set 2	77.2 (0.41)	84.2 (0.41)	88.6 (0.39)	0.769
		Test set 3	76.9 (0.36)	83.6 (0.37)	87.4 (0.36)	0.758
P3	No, no, yes	Test set 1	69.3 (0.21)	76.0 (0.21)	80.3 (0.20)	0.767
		Test set 2	71.7 (0.25)	77.8 (0.24)	81.6 (0.23)	0.774
		Test set 3	73.1 (0.25)	78.9 (0.25)	81.9 (0.24)	0.778
P4	No, no, no	Test set 1	64.3 (0.14)	72.1 (0.13)	77.6 (0.13)	0.735
		Test set 2	68.5 (0.17)	75.5 (0.16)	80.3 (0.16)	0.748
		Test set 3	72.0 (0.17)	78.7 (0.17)	82.6 (0.17)	0.771

Table 5.10 Averaged performance statistics for dictionaries using each combination of fragmentation parameters, fragmentation depths 0 and 1 excluded. Top 1, 2 and 3 scores shows the % of predictions where a fragment contains a true SOM within the top 1, 2 or 3 positions. In brackets, the coverage score for the fragment in the first, second and third place.

With the revised selection of dictionaries considered in the analysis, the results show that P2 produces the best performance statistics when structures similar to the training dataset are encountered (test set 1) but P3 appear to have better extrapolation abilities (Table 5.10). However, given that the performance statistics for P2 outperforms all results generated by P3 except for the AUC mean of test set 3, P2 is chosen as the preferred combination of fragmentation parameters. Given the differences in the characteristics of fragments generated, it is not inconceivable that the P2 and P3 dictionaries can be combined to produce a merged dictionary containing different types of fragments to be used for prediction. The range of different fragment types can also prove beneficial in the search for replacement fragments.

5.5.3.3 Effects of Fragmentation Depths

Aside from the combination of fragmentation parameters, the fragmentation depth used during fragmentation also makes a statistically significant impact on the performance of the dictionaries.

Fragmentation depth		Top 1	Top 2	Top 3	Mean AUC
0	Test set 1	51.6 (0.16)	60.8 (0.14)	62.9 (0.13)	0.388
	Test set 2	55.5 (0.17)	65.1 (0.17)	67.1 (0.15)	0.385
	Test set 3	55.8 (0.20)	65.6 (0.17)	67.2 (0.15)	0.391
1	Test set 1	64.5 (0.24)	75.5 (0.20)	81.1 (0.19)	0.690
	Test set 2	66.1 (0.26)	77.5 (0.22)	84.0 (0.21)	0.698
	Test set 3	67.2 (0.25)	77.6 (0.24)	83.6 (0.21)	0.699
2	Test set 1	68.2 (0.25)	75.8 (0.23)	80.7 (0.22)	0.741
	Test set 2	70.3 (0.27)	77.7 (0.25)	82.7 (0.24)	0.741
	Test set 3	72.5 (0.26)	79.3 (0.25)	83.4 (0.24)	0.749
3	Test set 1	72.7 (0.29)	79.2 (0.29)	83.1 (0.27)	0.764
	Test set 2	73.5 (0.31)	80.1 (0.31)	84.0 (0.29)	0.758
	Test set 3	74.8 (0.29)	80.8 (0.30)	84.0 (0.29)	0.762
4	Test set 1	75.4 (0.32)	81.1 (0.32)	84.8 (0.31)	0.775
	Test set 2	75.5 (0.34)	81.6 (0.34)	85.2 (0.33)	0.764
	Test set 3	75.8 (0.31)	82.1 (0.31)	85.1 (0.32)	0.767

Table 5.11 Averaged performance statistics for dictionaries using each tested fragmentation depth. Top 1, 2 and 3 scores shows the % of predictions where a fragment contains a true SOM within the top 1, 2 or 3 positions. In brackets, the coverage score for the fragment in the first, second and third place.

The larger the fragmentation depth (i.e. the bigger the fragment), the better the performance (Table 5.11). All F values are significantly larger than the F_{crit} in all cases and all p -values < 0.05 . Fragmentation depth zero produces undesirable results (mean AUC values below 0.5). A sudden increase in the performance of FamePrint can be seen when fragmentation depth is increased beyond zero. The rise in performance as fragmentation depth increases begins to slow after it reaching a depth of 2 (Figure 5.10). The choice of fragmentation depth can vary depending on the situation. The performance of fragmentation depth of 4 gives the best results, however, the generation and usage of the dictionary of fragments generated at this depth requires more time and resources. Also, the coverage produced by dictionaries using a fragmentation depth of 4 is the highest out of all the fragmentation depths tested. Dictionaries at depth 2 have lower coverage scores and are quicker to generate and to produce predictions. However, they do not perform as

well as their counterparts produced with depth of 3 and 4. Fragmentation depth 3 offers a good balance between performance, speed and the size of fragments selected. This is chosen for use with the optimised set of workflow variables.

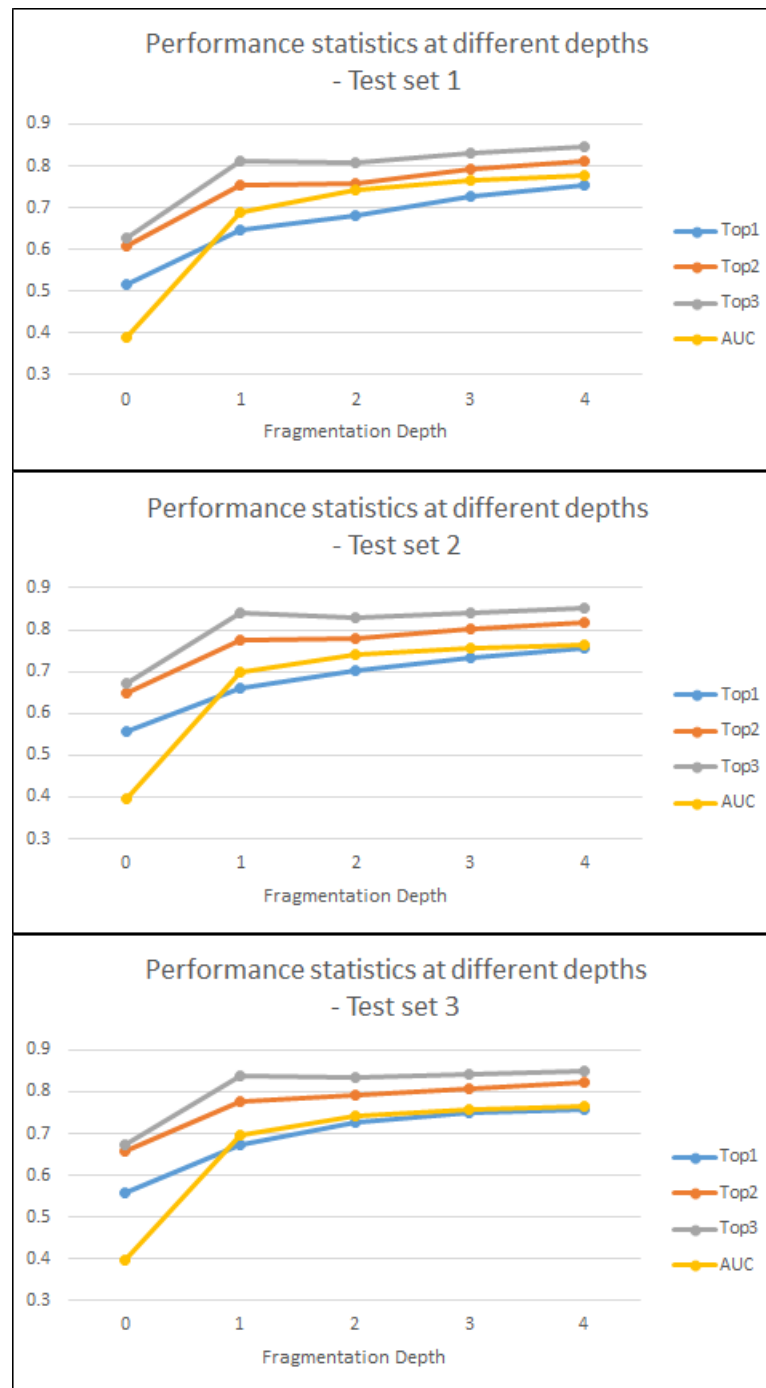


Figure 5.10 Changes in performance statistics as fragmentations depth is varied. Actual performance figures are given in Table 5.11

5.5.3.4 Effects of Fingerprint Versions

After the omission of dictionaries created with fragmentation depths of 0 and 1, statistically significant differences emerges in the comparison of fingerprint versions used during the creation of the dictionary of fragments. Across all performance measurements and all test datasets, FP03 produces superior SOM prediction performance compared to all other fingerprint versions tested.

Fingerprint					Mean
version		Top 1	Top 2	Top 3	AUC
FP00	Test set 1	71.9 (0.29)	78.6 (0.28)	82.7 (0.27)	0.759
	Test set 2	73.0 (0.31)	79.7 (0.30)	83.9 (0.29)	0.754
	Test set 3	74.2 (0.29)	80.6 (0.29)	84.1 (0.28)	0.759
FP01	Test set 1	71.7 (0.29)	78.4 (0.28)	82.6 (0.27)	0.757
	Test set 2	72.8 (0.31)	79.5 (0.30)	83.7 (0.29)	0.752
	Test set 3	74.2 (0.29)	80.5 (0.29)	83.9 (0.28)	0.758
FP02	Test set 1	72.2 (0.29)	78.8 (0.28)	82.9 (0.27)	0.761
	Test set 2	73.2 (0.31)	79.9 (0.30)	84.0 (0.29)	0.755
	Test set 3	74.4 (0.29)	80.8 (0.29)	84.1 (0.28)	0.760
FP03	Test set 1	72.5 (0.29)	79.1 (0.28)	83.2 (0.27)	0.763
	Test set 2	73.4 (0.31)	80.1 (0.30)	84.3 (0.29)	0.757
	Test set 3	74.7 (0.29)	81.1 (0.29)	84.4 (0.28)	0.761

Table 5.12 Averaged performance statistics for dictionaries using each tested version of fingerprint. Top 1, 2 and 3 scores shows the % of predictions where a fragment contains a true SOM within the top 1, 2 or 3 positions. In brackets, the coverage score for the fragment in the first, second and third place.

When using the top- k metric as a measurement or the mean AUC values, fingerprint version FP03 dictionaries clearly gives the best performance. All p -values calculated from the performance statistics are much lower than the α level of 0.05, indicating that the usage of different fingerprint versions alters the performance of the FamePrint workflow. As FP03 is shown to produce the best performance of all fingerprint versions, FP03 is therefore chosen for use and only results produced by this version of fingerprint will be reported in the analyses from here on.

When taking into account the chosen fingerprint version (FP03), fragmentation parameter (P2) and fragmentation depth (3), the results from the relevant dictionaries were compared in order to identify the optimal number of discretisation bins, sets of dictionary fingerprints used for stability calculation (k) and weighting scheme to be applied on fingerprints.

5.5.3.5 Effects of Number of Discretisation Bins

The number of bins used for discretisation shows statistically significant differences, between results from dictionaries with different number of bins, for all performance statistics for all test datasets. However, there is not a universally agreed bin number (Table 5.13): test set 1 appears to favour a higher bin number with 6 being the preferred bin if the top- k metric is used and 8 if the mean AUC value is considered instead. However, test set 2 and 3 mostly prefers the use of 5 bins for discretisation. It may be that as the distance from the training dataset structures increases, the increased level of fuzziness offered by a smaller number of discretisation bin is desirable. If only the mean AUC values are used as a performance measurement, the use of an increasing number of discretisation bins gives an improvement in performance for test set 1. However, as the distance from the training structures increases, all performances drop, with the slowest decrease in performance seen with the use of 5 discretisation bins. As the top- k and mean AUC statistics for test set 1 indicated only a small drop when 5 discretisation bins are used instead of 6, 7 or 8 and 5 bins offers the best performance for test set 2 and 3, the use of 5 discretisation bins is selected.

Number of discretisation bin		Top 1	Top 2	Top 3	Mean AUC
4	Test set 1	80.1 (0.41)	85.7 (0.41)	89.3 (0.40)	0.797
	Test set 2	78.4 (0.42)	84.7 (0.43)	89.2 (0.41)	0.771
	Test set 3	77.5 (0.37)	84.3 (0.38)	88.2 (0.38)	0.762
5	Test set 1	80.2 (0.41)	85.8 (0.41)	89.4 (0.40)	0.799
	Test set 2	77.9 (0.41)	84.9 (0.42)	89.3 (0.40)	0.772
	Test set 3	78.4 (0.37)	84.9 (0.38)	88.7 (0.37)	0.770
6	Test set 1	80.3 (0.41)	85.9 (0.41)	89.4 (0.40)	0.801
	Test set 2	77.9 (0.41)	84.5 (0.42)	89.0 (0.40)	0.771
	Test set 3	77.7 (0.36)	83.8 (0.37)	87.1 (0.37)	0.756
7	Test set 1	80.2 (0.41)	85.8 (0.41)	89.4 (0.40)	0.801
	Test set 2	77.6 (0.42)	84.3 (0.42)	89.0 (0.40)	0.772
	Test set 3	75.4 (0.36)	82.4 (0.38)	86.4 (0.37)	0.755
8	Test set 1	80.3 (0.41)	85.8 (0.41)	89.4 (0.40)	0.802
	Test set 2	77.6 (0.42)	84.3 (0.42)	88.7 (0.40)	0.771
	Test set 3	77.1 (0.36)	83.1 (0.38)	86.9 (0.37)	0.759

Table 5.13 Averaged performance statistics for dictionaries using each tested number of discretisation bin. Top 1, 2 and 3 scores shows the % of predictions where a fragment contains a true SOM within the top 1, 2 or 3 positions. In brackets, the coverage score for the fragment in the first, second and third place.

This will also aid the response time of FamePrint as the use of a smaller number of discretisation bins lowers the size of the fingerprint as well as lowering the amount of memory required to store of the dictionary of fragments.

5.5.3.6 Effects of Number of Most Similar Fingerprints Used

The number of sets of dictionary fingerprints used for stability calculation (k) offers more consistent results (Table 5.14). The usage of the top three or top five most similar sets of fragment fingerprints for stability calculation offers no discernible difference in the performance statistics. These offers the best performance statistics across all performance statistics measurements and therefore stability contribution from the top 5 most similar fingerprint sets will be used in the performance measure ($k = 5$).

k		Top 1	Top 2	Top 3	Mean AUC
3	Test set 1	80.4 (0.41)	86.0 (0.41)	89.5 (0.40)	0.802
	Test set 2	78.0 (0.42)	84.8 (0.42)	89.2 (0.40)	0.773
	Test set 3	77.3 (0.36)	84.0 (0.38)	87.5 (0.37)	0.762
5	Test set 1	80.4 (0.41)	86.0 (0.41)	89.5 (0.40)	0.802
	Test set 2	78.0 (0.42)	84.8 (0.42)	89.2 (0.40)	0.773
	Test set 3	77.3 (0.36)	84.0 (0.38)	87.5 (0.37)	0.762
10	Test set 1	80.1 (0.41)	85.8 (0.41)	89.3 (0.40)	0.800
	Test set 2	77.8 (0.42)	84.4 (0.42)	89.0 (0.40)	0.771
	Test set 3	77.1 (0.37)	83.6 (0.38)	87.3 (0.37)	0.760
15	Test set 1	80.1 (0.41)	85.7 (0.41)	89.3 (0.40)	0.799
	Test set 2	77.8 (0.42)	84.4 (0.42)	88.9 (0.40)	0.770
	Test set 3	77.2 (0.37)	83.5 (0.38)	87.5 (0.37)	0.759
All	Test set 1	80.1 (0.41)	85.7 (0.41)	89.3 (0.40)	0.799
	Test set 2	77.8 (0.42)	84.4 (0.42)	88.9 (0.40)	0.770
	Test set 3	77.2 (0.37)	83.5 (0.38)	87.5 (0.37)	0.759

Table 5.14 Averaged performance statistics for dictionaries using each tested k value. Top 3 Scores: average cumulative performance statistics (average coverage of position, non-cumulative)

5.5.3.7 Effects of Weighting of Fingerprints

The weighting scheme applied to the fingerprints does not make any statistically significant differences to the performance obtained for test set 3. The Weighting: None scheme, where each fingerprint contribute equally appears to give the best performances (Table 5.15) out of all four

weighting schemes tested (exceptions being the top 2 and 3 positions for test set 2 and 3 where FPOnly weighting performed marginally better) and will therefore be chosen for use with the other selected workflow variables.

Weighting		Top 1	Top 2	Top 3	Mean AUC
None	Test set 1	80.5 (0.41)	86.1 (0.41)	89.6 (0.40)	0.803
	Test set 2	78.1 (0.42)	84.9 (0.42)	89.3 (0.40)	0.774
	Test set 3	77.4 (0.36)	84.2 (0.38)	87.6 (0.37)	0.763
FPOnly	Test set 1	80.3 (0.41)	85.9 (0.41)	89.5 (0.40)	0.802
	Test set 2	77.9 (0.42)	84.8 (0.42)	89.4 (0.40)	0.773
	Test set 3	77.2 (0.37)	84.3 (0.38)	87.9 (0.37)	0.763
BitsOnly	Test set 1	80.3 (0.41)	85.9 (0.41)	89.5 (0.40)	0.801
	Test set 2	78.0 (0.42)	84.7 (0.42)	89.0 (0.40)	0.772
	Test set 3	77.4 (0.36)	83.7 (0.38)	87.3 (0.37)	0.761
AllBits	Test set 1	80.3 (0.41)	85.9 (0.41)	89.5 (0.40)	0.801
	Test set 2	78.0 (0.42)	84.7 (0.42)	89.0 (0.40)	0.772
	Test set 3	77.4 (0.36)	83.7 (0.38)	87.3 (0.37)	0.761

Table 5.15 Averaged performance statistics for dictionaries using each tested fingerprint weighting scheme. Top 1, 2 and 3 scores shows the % of predictions where a fragment contains a true SOM within the top 1, 2 or 3 positions. In brackets, the coverage score for the fragment in the first, second and third place.

5.5.3.8 The Optimised Variables

The final selected set of work-flow variables (Table 5.16) and evaluation of performance with optimised parameters (Table 5.17) are as follows:

Workflow variables	Chosen parameters
Fingerprint version	FP03
Number of discretisation bins	5
Fragmentation parameters	Break scaffolds, retain rings, retain functional groups (P2)
Fragmentation depth	3
Number of fingerprint sets considered for stability prediction (<i>k</i>)	5
Fingerprint weighting for similarity comparison	None

Table 5.16 Optimised selection of FamePrint workflow variables.

Test Set	Top <i>k</i> (coverage)			Overlap (coverage)		AUC	
	1	2	3	1 & 2	1, 2 & 3	mean	median
1	80.5 (0.41)	86.1 (0.41)	89.6 (0.40)	73.6 (0.34)	77.9 (0.40)	0.802	0.873
2	78.1 (0.41)	85.2 (0.42)	89.6 (0.40)	71.6 (0.34)	75.7 (0.41)	0.773	0.850
3	78.2 (0.37)	85.2 (0.38)	88.7 (0.37)	74.1 (0.29)	77.1 (0.38)	0.773	0.889

Table 5.17 Evaluation of FamePrint performance with optimised workflow variables (given in Table 5.16)

The top-*k* performance statistics of FamePrint (Table 5.17) represents the performance of FamePrint when trained on data from all species and metabolism phases available from the Accelrys metabolite database. The same dataset is used to produce the “metabolic phase: Phase I + II, species: all models” of FAME, giving the following results:

Test Set	Top 1	Top 2	Top 3	AUC	
				Mean	Median
1	66.5	77.3	84.6	0.853	0.939
2	64.3	76.1	84.5	0.849	0.925
3	62.9	76.3	84.2	0.851	0.938

Table 5.18 FAME model trained on training dataset used in FamePrint study, both using CDK 1.5.9. Results of predictions carried out on test set 1,2 and 3 (same datasets as Table 5.17) are shown.

As shown in Table 5.18, the FAME model has lower top one to top three scores compared to FamePrint (Table 5.17) but has higher AUC scores. However, it is worth bearing in mind that these

results are not directly comparable as FAME carries out SOM prediction on an atom-by-atom basis whereas FamePrint selects fragments when prediction metabolic stability.

A small selection of structures, all containing counter ions, are kept aside (unused) during the generation of the training and test datasets. After the removal of structures containing metal atoms and metal ion chelators, the 60 structures are used as an external test dataset. These structures have fewer than 100 heavy atoms after their counter ions are removed (counter ions identified as fragment having fewer heavy atoms than the rest of the structure). These are then washed and annotated using the same procedure as the training and test dataset detailed in section 5.2 and used to evaluate the performance of the optimised workflow of FamePrint:

Test Set	Top <i>k</i> (coverage)			Overlap (coverage)		AUC	
	1	2	3	1 & 2	1, 2 & 3	mean	median
external	74.6 (0.32)	81.7 (0.32)	83.1 (0.29)	72.6 (0.22)	73.9 (0.31)	0.797	0.864

Table 5.19 Evaluation of optimised FamePrint workflow performance with an external dataset.

From visual inspection, the structures contained in the external test dataset are quite different from the majority of training dataset structures. The coverage required at the top three positions also suggests that these structures are most similar to those contained in test set 3 (i.e. dissimilar to training dataset structure). Despite a drop in the top-3 statistics, the mean and median AUC values are maintained in the same range as those obtained by test set 1, 2 and 3 (Table 5.17).

5.5.3.9 Merged Model

Dictionaries of different fragmentation methods can be combined to produce a merged dictionary. As previously noted, P2 performed best with structures more similar to the training dataset structures but P3 extrapolated better. The performance statistics of the dictionary of fragments produced by the same combination of parameters as those shown in Table 5.16, except with the breaking of rings during fragmentation, is given here for comparison:

Test Set	Top <i>k</i> (coverage)			Overlap (coverage)		AUC	
	1	2	3	1 & 2	1, 2 & 3	mean	median
1	70.6 (0.21)	77.1 (0.21)	80.9 (0.20)	68.8 (0.16)	67.9 (0.18)	0.776	0.860
2	73.1 (0.25)	78.4 (0.25)	81.8 (0.23)	70.4 (0.19)	70.2 (0.23)	0.783	0.879
3	74.7 (0.25)	79.8 (0.27)	82.3 (0.24)	73.6 (0.21)	73.2 (0.24)	0.788	0.900

Table 5.20 Evaluation of FamePrint performance with optimised workflow variables (given in Table 5.16) (except rings are broken during fragmentation).

These two dictionaries are combined to produce a merged dictionary with fragments produced by both sets of fragmentation parameters. This is then tested on the same test set structures and validation shows very similar top-*k* performance to that produced by the original P2 and P3 dictionaries (Table 5.21). Similar coverage scores are seen in the top two positions and lower coverage scores in the third most unstable fragment position. Overlap scores are also fairly similar to the P2 dictionary, with slightly lower coverage scores. The mean AUC scores from the merged dictionary outperforms the original P2 and P3 dictionary, with noticeable improvement in the mean AUC value of test set 2 and 3. This is also tested on the same external test dataset used on the optimised dictionary.

Test Set	Top <i>k</i> (coverage)			Overlap (coverage)		AUC	
	1	2	3	1 & 2	1, 2 & 3	mean	median
1	79.4 (0.40)	84.9 (0.40)	88.2 (0.38)	73.0 (0.32)	77.1 (0.39)	0.808	0.862
2	77.3 (0.40)	83.9 (0.40)	87.6 (0.38)	71.1 (0.32)	75.6 (0.40)	0.799	0.861
3	78.7 (0.36)	84.5 (0.37)	87.0 (0.35)	73.8 (0.29)	77.0 (0.36)	0.802	0.889
external	76.1 (0.33)	81.7 (0.35)	87.3 (0.30)	71.4 (0.21)	77.5 (0.33)	0.815	0.864

Table 5.21 Evaluation of FamePrint performance with the merged dictionary.

The performance statistics of the merged dictionary are encouraging. Despite the slightly lower top-*k* values (accompanied by lowered coverage scores) the AUC scores have improved. Also the drop in the overlap scores is very slight, accompanied by a more significant drop in the overlap coverage scores (when compared to the top three). The combined dictionary is loaded into the Coralie Cheminformatics Platform for a real time evaluation of query structures submitted via the FamePrint tab. The evaluation of structures as well as the search for replacement structure and subsequent generation of structures using the replacement requires over 9GB in RAM and response time is noticeably slower than the non-merged P2 dictionary (which only requires 6.2GB in RAM).

5.5.3.10 Human-specific Model

All dictionaries created thus far utilises unique structures from the entire Accelrys Metabolite Database. This includes metabolism information from different species, such as human, rat and dog as well as different metabolic phases. Separate models can be generated for each species and/or metabolic phase if enough structures are available. A merged dictionary is created using substrate structures and transformations relevant only to human metabolism, extracted according to the species data field available from the Accelrys Metabolite Database. A list of unique substrate

structures (12891 with 100 or fewer atoms) along with relevant human transformation annotations is extracted from the Accelrys Metabolite Database, carried out using MetaPrint2D (section 2.1.3).

A 70:30 split is carried out on the dataset (same as section 5.2.6) to give the training and test (1) datasets, with 9023 and 3868 structures respectively. Test sets 2 and 3 are generated as detailed in section 5.2.6. A merged (P2 and P3) dictionary with the same workflow variable used by the merged model in 5.5.3.9 is produced and its SOM prediction performance examined using the three test sets produced:

Test Set	Top <i>k</i> (coverage)			Overlap (coverage)		AUC	
	1	2	3	1 & 2	1, 2 & 3	mean	median
1	77.7 (0.36)	83.0 (0.36)	85.9 (0.35)	71.9 (0.29)	75.3 (0.34)	0.802	0.857
2	71.3 (0.36)	78.5 (0.37)	82.4 (0.34)	65.4 (0.29)	69.1 (0.34)	0.770	0.826
3	74.7 (0.38)	81.2 (0.39)	84.9 (0.35)	66.7 (0.32)	71.6 (0.37)	0.777	0.855

Table 5.22 Evaluation of FamePrint performance with merged dictionary of human substrates.

The performance of the human FamePrint model (Table 5.22) has very similar but slightly lower performance statistics compared to the FamePrint model created with the merged dictionary built on the full dataset (Table 5.21). This is also the case for FAME – where the human model performs almost as well as the model created using the training set from all available data. It is also further pointed out in the FAME study that the performance of the models are dependent on the size of training dataset, which is also suggested by the FamePrint results shown here as the human training dataset is only a subset of the full training dataset.

5.6 Conclusion

Keeping in mind that FamePrint provides prediction at the fragment level and FAME at the atom level, FamePrint shows higher top-3 performance compared to the retrained FAME model at all three top three positions for all three test datasets examined (Table 5.17, Table 5.18). However, if the mean AUC value is used, the retrained FAME model appears to perform better than FamePrint. Given its favourable top-*k* and AUC values, plus the frequency overlaps within the top three fragments, FamePrint, with the current set of optimised workflow variables, is appropriate for use to carry out SOM prediction before the replacement of unwanted fragments.

The coverage scores of fragments identified within the top three positions are quite high. However, when taking the frequency of overlaps into account and AUC scores, it suggests that the FamePrint algorithm can discriminate between metabolically stable regions against labile regions. The frequently seen overlaps of the top two or top three fragments suggests that FamePrint can identify regions of metabolic vulnerability, and the inclusion of a larger chemical context when used for stability prediction (larger fragment) gives a more accurate prediction score. Also, as rings are kept intact within this workflow, some fragments will be relatively large (compared to the total structure), especially on smaller structures.

As FamePrint showed favourable performance in SOM prediction, it may be possible to extend this methodology to identifying bioisosteric replacements which can maintain similar metabolic stability profiles but may bring about changes/ improvement to other properties of interest during drug discovery. This will be reported in the next chapter. The merged model shows promising results, however, given the time and resources required to produce prediction and only a limited gain in performance, the non-merged P2 dictionary will be used in chapter 6 for suggesting bioisosteric replacements as the search for replacement fragments is even more computationally intensive than the prediction of metabolic stability.

6. Bioisosteric Replacements with FamePrint

6.1 Introduction

Drug discovery is a multi-objective optimisation problem where many different factors such as metabolic stability, bioavailability, and toxicity have to be taken into account. The use of bioisosteres during the lead optimisation is a particularly powerful technique which attempts to retain the desirable traits the lead structure already possesses whilst improving upon other unwanted properties.

The development of FamePrint, a fragment-based SOM predictor, has been reported in chapter 5. FamePrint shows good performance in predicting metabolically unstable fragments within a query structure. As the methodology operates on a fragment/ substructure level, it can be extended to suggest bioisosteric replacements in cases where the metabolic stability profile of a structure should be maintained but other properties, such as bioavailability, require optimisation. The bioisosteric replacement study reported in this chapter is also developed within the Coralie Cheminformatics Platform (section 4.1.1) and is available in the same Biostere module where FamePrint is implemented (section 5.4).

6.2 Method

This is built upon the FamePrint methodology reported in chapter 5 and utilises the dictionary of fragments that have already been generated for the FamePrint model as a source of fragment structures from which bioisosteric replacement suggestions are sought.

The FamePrint training dataset, a set of unique substrate structures with associated SOM annotations with reaction type information, originates from the Accelrys Metabolite Database (version 2011.2) and has undergone the preparation steps outlined in section 5.2.1. These substrate structures are transformed into a dictionary of fragments according to the steps outlined in section 5.3.1.1, using the optimised set of parameters listed in Table 5.16.

6.3 Workflow

6.3.1 Identification of Fragment for Replacement

This is designed to follow on from FamePrint (chapter 5) where the dictionary of fragments produced for SOM prediction in FamePrint is used here to serve as a source of fragments from which potential bioisosteric replacements for unwanted query fragments can be identified.

Seven atom based descriptors (listed in Table 5.1) are first calculated for each atom of the query structure. These descriptor values are then discretized by an equal frequency discretisation method (section 5.2.3), using the same discretisation boundaries used to create the dictionary of fragments. The query structure is fragmented in the same manner as the dictionary of fragments (section 5.2.4, using the parameters from Table 5.16). Topological atom fingerprints (version FP03, see 5.3.4.1) are then generated for all query fragments. These are the same processes undertaken by a query structure before SOM predictions can be made (Figure 5.2).

6.3.2 Search for Replacement

Once a fragment has been identified as a candidate for bioisosteric replacement, a search for suitable replacement fragments from the dictionary of fragments is initiated. The set of fingerprints produced by the query fragment selected are compared to all sets of fingerprints contained in the dictionary of fragments and similarity scores between fingerprint sets calculated. Similarity score is calculated as detailed in Figure 5.5. Fingerprints produced by fragments with the same structure as the selected query fragment are not considered unless the dictionary fragment has a different set of connection points (atoms connecting the dictionary fragment structure to its originating parent structure) compared to the query fragment (Figure 6.1), as this allows for the rotation of the selected fragment within the query structure and new structures to be generated. Information on each dictionary fragment's connection points is recorded concurrently during the fragmentation step to create the dictionary of fragments and stored in the dictionary along with the fragment's TAP fingerprints and stability score.

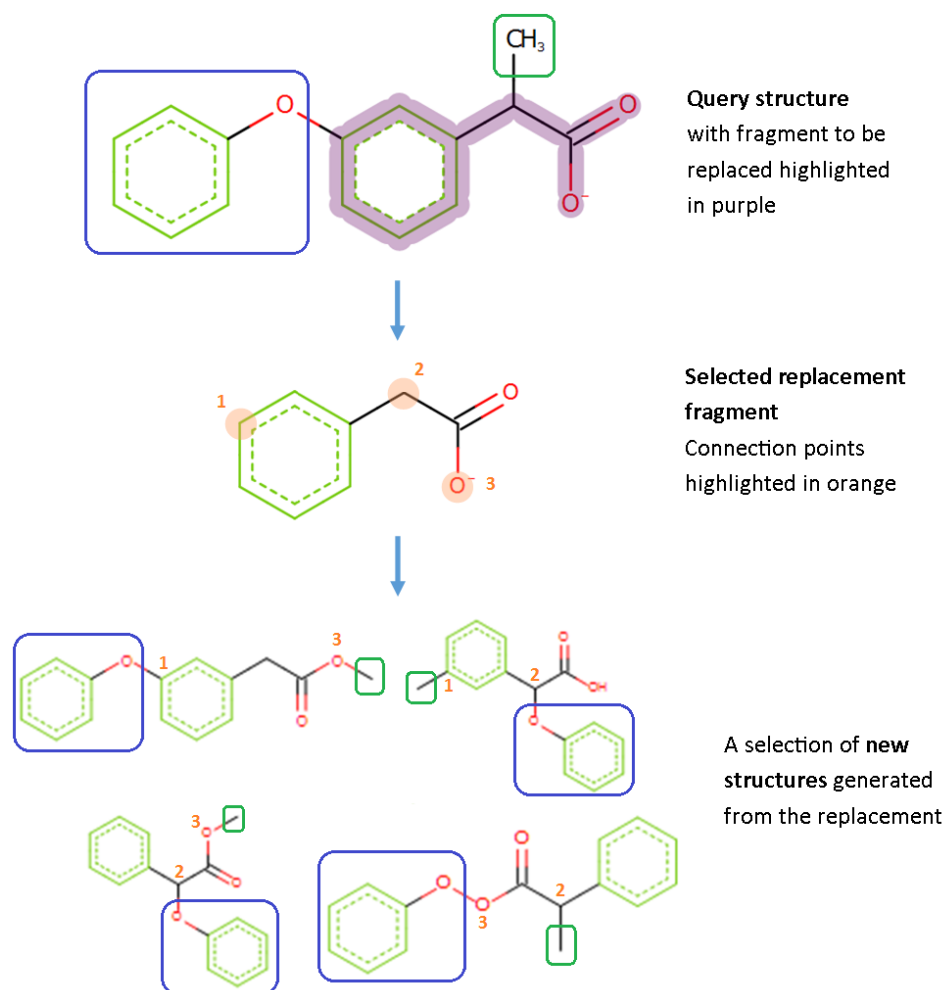


Figure 6.1 Substitution of query fragment with a replacement fragment of identical structure. As the substitution fragment has different connection points, this allows for novel structures to be generated from the substitution.

During the creation of the dictionary of fragments, a maximum limit of 16 heavy atoms for any given fragment has been imposed (section 5.5.1). It is also possible to cap the size of replacement fragments so a fragment no larger than half of the query structure can be selected for replacement. However, even if the unwanted fragment selected for replacement is larger than half of the original structure, the replacement chosen and new structures can be generated in such a way that less than half the structure is ultimately transformed, for example:

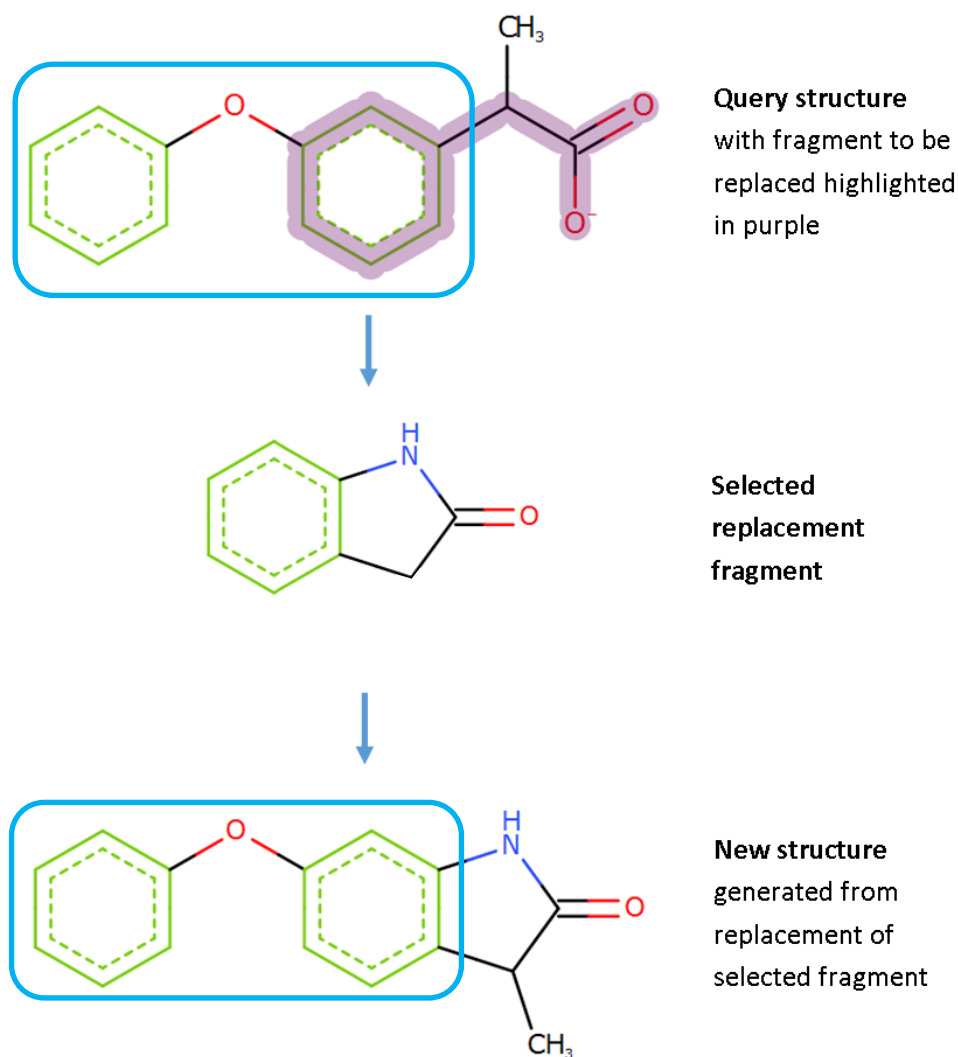


Figure 6.2 Replacement of a large fragment and the generation of a new structure. The substructure highlighted in blue remains unchanged.

6.3.3 Replacement Compatibility

Once similar fragments are identified as potential replacements, they are checked for their compatibility with the query structure. In order to be considered compatible, the dictionary fragment must have at least the same number of connection points as the selected query fragment as well as being able to make at least the same number of the same type of bonds (single, aromatic, double or triple bonds are considered). A list of compatible fragments are then made available in the GUI within the Coralie application, where an optional minimum similarity between the query fragment and the suggested fragments from the dictionary can be used to reduce the number of results returned. Similarity is determined as detailed in Figure 5.5. The selection of suitable fragments to be used for replacement is not an automated step and requires user interaction as there can be a large number of similar, compatible fragments which may lead to an even larger number of new structures generated in the next step.

6.3.4 Generation of New Structure

Once a replacement fragment has been manually selected, new structures are generated by substituting the unwanted query fragment for the selected replacement fragment. If one or more connection points are present on the query fragment and/or the dictionary fragment, all possible substitutions are generated. All compatible combinations of connection points, each governed by their bond order requirement, are used to produce a new structure.

6.4 Graphical User Interface

The bioisosteric replacement functionality has been implemented in the Biostere module (section 5.4, Appendix G – Biostere Tab in Coralie) within the Coralie Cheminformatics Platform along with the FamePrint SOM predictor (chapter 5). The screenshots of the GUI for the bioisosteric replacement functionality is given in Appendix H – Biostere Tab in Coralie for Bioisosteric Replacement.

When a query structure has been submitted and fragmented, all the fragments generated from the query structure will be displayed in the application along with the fragment's predicted stability which is given above the fragment and also indicated by the colour bar on top of the fragment (colour scale given in Appendix H – Biostere Tab in Coralie for Bioisosteric Replacement). Once an unwanted query fragment has been selected, a search for suitable replacement fragments is initiated (section 6.3.2). Compatible replacements (section 6.3.3) are displayed on screen in decreasing similarity to the selected query fragment. The similarity score of the suggested fragments are given above the suggested fragment structure and the suggested fragment's stability score (retrieved directly from the stored stability score of the fingerprints from the dictionary) is indicated by the colour bar on top of the suggested fragment. The same colour scale is used as above. A minimum similarity score between the query fragment and dictionary fragments is available as a slider in the application, allowing user to specify the minimum similarity between the suggested fragments compared to the query fragment selected for replacement (section 6.3.3). The minimum similarity is set at 0.25 by default.

Users can examine the positions of the connection points marked on the replacement fragment by selecting any of the suggested replacement fragments. When a desirable replacement fragment is found, clicking the "Replace!" button will initialise the substitution and generation of new structures with the replacement fragment. For every new structure generated, metabolic stability prediction is carried out on the structure using the FamePrint SOM predictor. The metabolic stability score given for the generated structure is the metabolic stability score of their most unstable fragment. TAP

fingerprints (for the whole structure) will also be created for all newly generated structure(s). These are then compared to the original query and a similarity score between the new structure and the query is calculated. Generated structures are displayed on screen in order of decreasing similarity to the original query structure. The similarity score between the new structure and the original query is displayed above the generated structure and its metabolic stability score is indicated by the colour of the bar above the structure (same colour scale as before).

Selecting one of the newly generated structures will automatically set the chosen structure as the new query within the Biostere module, ready for further examination of its metabolic vulnerabilities and search for potential replacements. The “History Tree” tab within the GUI allows for the exploration of all structures investigated and generated thus far. The new structure can be compared with its parent structure and if the new structure is unsatisfactory, another replacement can be carried out on the newly generated structure or on the original query.

In Appendix H – Biostere Tab in Coralie for Bioisosteric Replacement.7, the structure shown in the Query structure display corresponds to structure entry M_2_1 in the History Tree. M_2_1 is a child of the query structure (M_2), generated by replacement carried out in Appendix H – Biostere Tab in Coralie for Bioisosteric Replacement.4. The original query structure can be accessed by clicking on the M_2 entry in the History Tree (Appendix H – Biostere Tab in Coralie for Bioisosteric Replacement.8). The structure stability score associated with parent structure M_2 is shown below the structure when the structure in the Query display changes.

6.5 Model Evaluation

Due to lack of access to a reasonable sized dataset with experimentally verified bioisosteres, especially ones specifically tested with regards to metabolic stability, it has not been possible to systematically verify the performance of this method.

A number of well-known bioisosteres of carboxylic acid are reported in the literature (Table 2.1). When various searches for replacements are carried out for carboxylic acid fragments, a number of known carboxylic acid bioisosteres are identified and retrieved as potential replacements within the top 20 suggested replacement fragments (Figure 6.3). All of these known bioisosteres have similarity values significantly higher than the default minimum similarity score (set at 0.25). Known bioisosteres are also reported for carbonyl and catechol (Figure 6.3), these are also retrieved and suggested as replacements.

Bioisosteres retrieval for amides and esters are also tested (Figure 6.3). Non-ring bioisosteres are easily retrieved within the top 10 most similar suggestions. It is difficult to identify the ring bioisosteres from non-ring bioisosteres. Once one ring bioisostere is suggested, it is selected for replacement and the remaining ring bioisosteres (within the top 5 suggestions) are returned as well as non-ring amide/ester fragments (within the top 10 suggestions).

The difficulty in identifying ring bioisosteres can potentially be due to the use of fingerprint version FP03, which puts more emphasis on the properties atom pairs where both atoms are connection atoms. The fingerprint also weights atom pairs more if one atom is a connection atom compared to if neither atoms are a connection atom. This can potentially mean that the method is more prone to selecting fragments of similar shapes over fragments with more dissimilar topology.

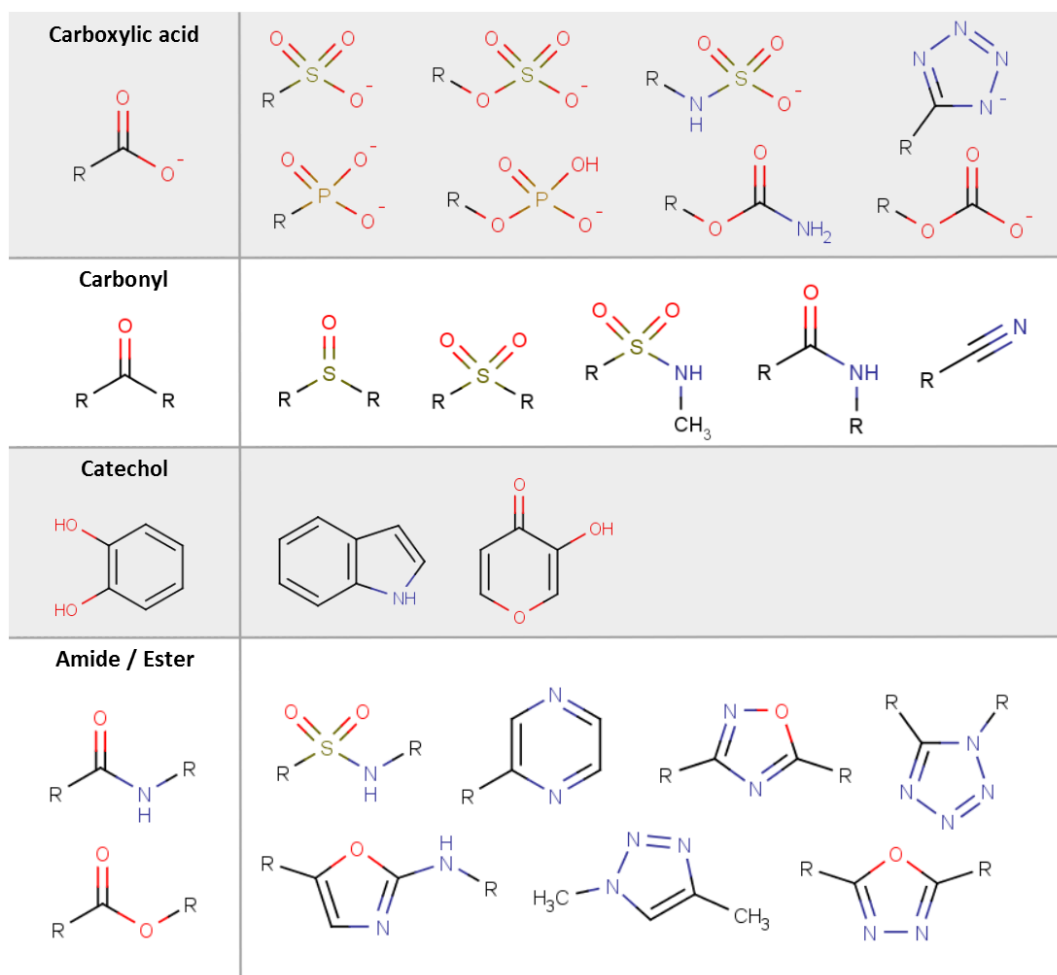


Figure 6.3 Examples of known¹⁰¹ bioisosteres identified by FamePrint.

6.5.1 Retrospective Studies

6.5.1.1 Case 1: Improvement of Prodrug Oral Bioavailability

A thiazole benenesulfonamide derivative has been identified as a β_3 -adrenergic receptor agonist and its metabolites (Figure 6.4) formed in rats are reported by Tang *et al.*¹⁰² The structure (P in Figure 6.4) has been reported to experience issues with low oral bioavailability and hepatic first-pass metabolism in both rats and monkeys.¹⁰³ Structure P has been processed by the two stage P-glycoprotein (PGP) classifiers produced (see chapter 7Improving Bioavailability) and is classified as a PGP substrate.

Tang made substitutions on structure P in an attempt to improve its oral bioavailability by the synthesis of potential prodrugs which, when delivered, will be converted back to structure P. Structure P is found in the test set 1 and 2 and not in the training dataset.

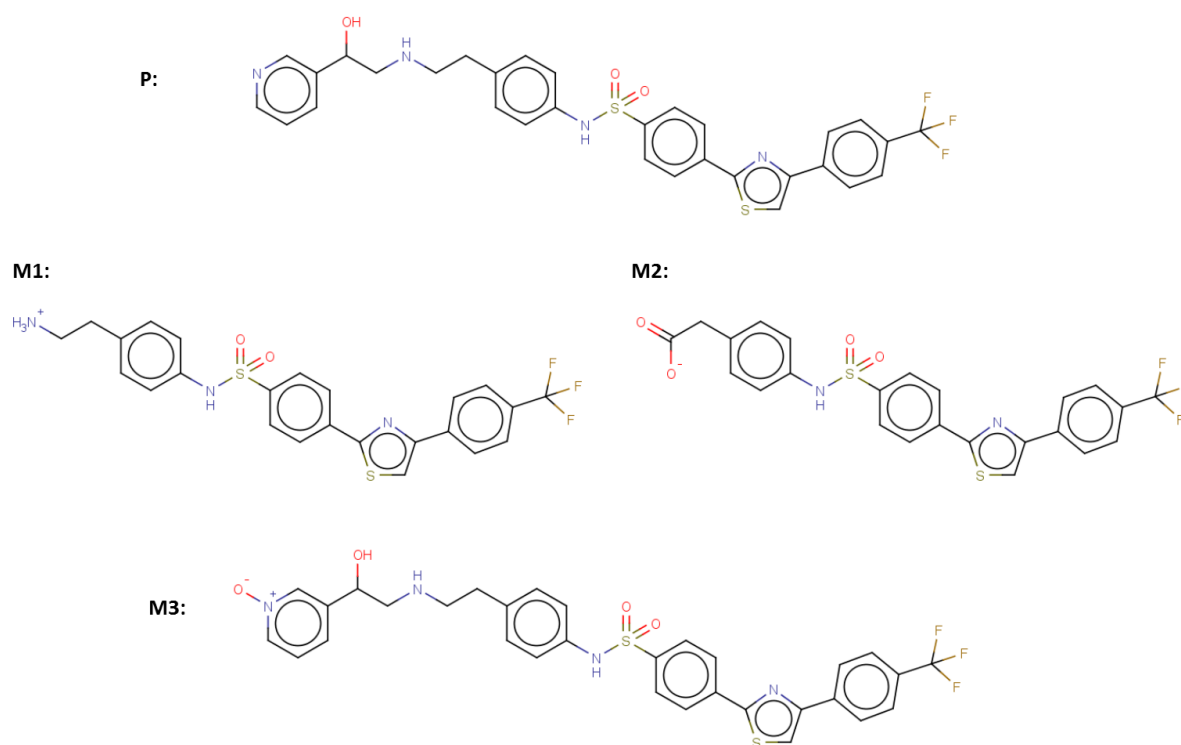
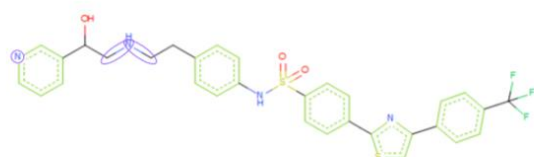


Figure 6.4 (R)-N-[4-[2-[[2-Hydroxy-2-(pyridin-3-yl)ethyl]amino]ethyl]phenyl]-4-[4-(4-trifluoro-methylphenyl)thiazol-2-yl]benzenesulfonamide (P) and its first generation metabolites (M1 – 3).

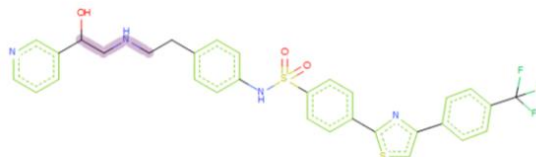
First the identification of metabolically vulnerable regions is carried out on structure P using FamePrint. The top 5 most vulnerable fragments identified (Figure 6.5), out of over 50 fragments,

corresponded to regions which were acted on by CYP enzymes to produce the first generation metabolites given above (M1 – 3 in Figure 6.4).

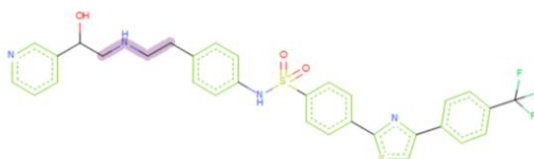
Structure P with experimentally identified SOMs circled in purple:



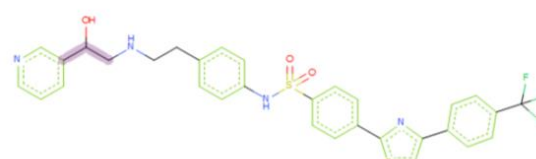
Most unstable fragment predicted:



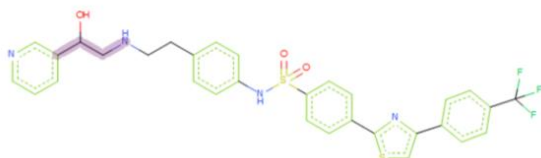
Second most unstable fragment predicted:



Third most unstable fragment predicted:



Fourth most unstable fragment predicted:



Fifth most unstable fragment predicted:

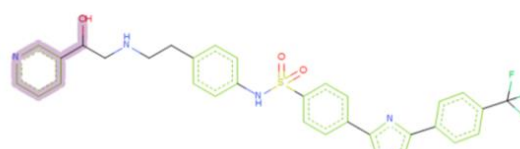


Figure 6.5 The top 5 most metabolically unstable fragments predicted for Structure P.

The fifth most unstable fragment (bottom right, Figure 6.5) contains the pyridine substructure which was modified in ref. ¹⁰³ in order to circumvent the poor oral bioavailability of the prodrug P (as the remaining SOM are untouched in the study). When the pyridine fragment is selected for replacement, the final optimised structure (Figure 6.6) with improved bioavailability in rats and monkeys as given in ref. ¹⁰³ can be generated from the 7th most similar replacement suggestion returned out of over 40 suggestions accessible on screen (Table 6.1).

Similarity	Fragment	Similarity	Fragment
0.795		0.408	
0.593		0.463	
0.552		0.460	
0.503			

Table 6.1 Top 7 most similar replacement suggestions retrieved.

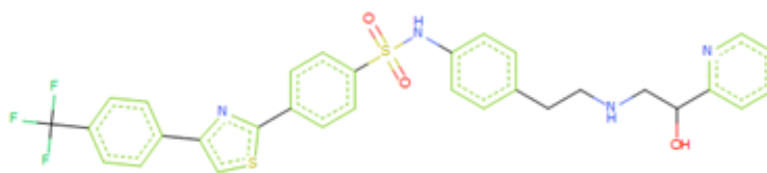


Figure 6.6 Final optimised structure.

A quick SOM prediction analysis of the newly generated structure (Figure 6.6) reveals a reordering of the most vulnerable fragments, compared to Figure 6.5. The predicted metabolic stability score for the 2-hydroxymethylpyridine fragment has increased from 0.37 to 0.54 and it is now ranked behind two additional fragments containing true SOM. The top 7 most unstable fragments of the final structure (Figure 6.5) predicted by FamePrint are given here:

Stability	Fragment	Stability	Fragment
0.086		0.431	
0.110		0.493	
0.171		0.536	
0.416			

Figure 6.7 Top 7 most unstable fragments of the final structure predicted by FamePrint.

Optimisations to structure P was carried out by Stearns¹⁰³, however due to the structure of the changes, and the implementation of the fragmentation algorithm, the reported changes will not be present within the same fragment and therefore a direct comparison between the reported metabolic stability (half-live values) and the order in which FamePrint retrieves the suggestions cannot be made.

6.5.1.2 Case 2: Metabolism-driven Optimisation of a Thrombin Inhibitor

Burgey carried out a metabolism-driven optimisation on thrombin inhibitor, 3-(2-phenethylamino)-6-methylpyrazinone acetamide (Figure 6.8, top) with three primary SOM.¹⁰⁴ The optimisation focused on improving the oral bioavailability and plasma half-life of the thrombin inhibitor (structure P). Structure P, which is not in the training dataset, is submitted as a query in the FamePrint tab for analysis. All SOM are accounted for within the top 13 most unstable fragments out of 45 fragments generated for the structure. There are 10 fragments within the top 13 fragments which contained true SOM.

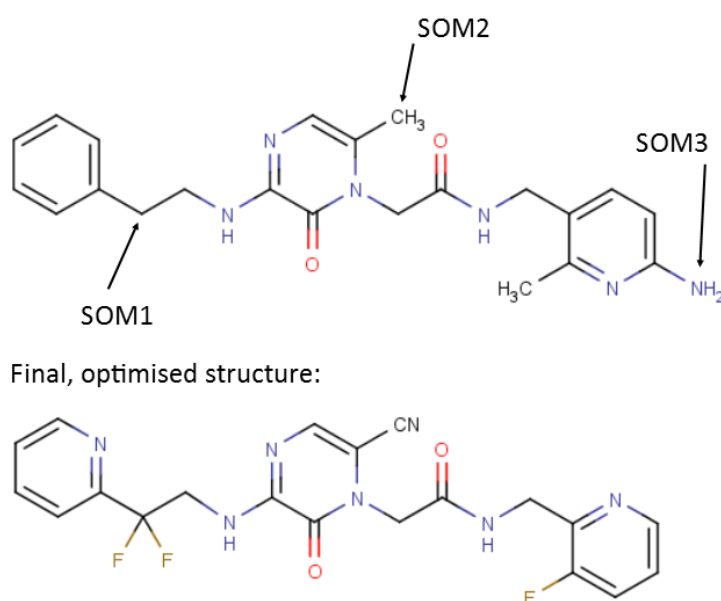


Figure 6.8 3-(2-phenethylamino)-6-methylpyrazinone acetamide, its primary SOM and the final product of metabolism-driven optimisation carried out by Burgey *et al.*¹⁰⁴

In Burgey's study, several optimisation experiments have been reported along with plasma half-life ($t_{1/2}$) values. These are used as an indication of metabolic stability. FamePrint's ability to retrieve fragments containing more similar metabolic stability values to each other before retrieving more dissimilar fragments is examined. Only the top 30 most similar replacement fragments will be retrieved and examined in this study. This limit is put in place to prevent the large number of fragments that will otherwise be retrieved.

6.5.1.2.1. SOM2 Modifications

The following modifications have been made to SOM2 (Figure 6.8), their structure and corresponding $t_{1/2}$ values are as follows:

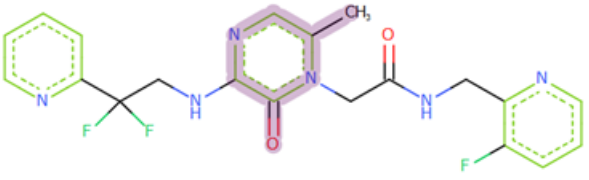
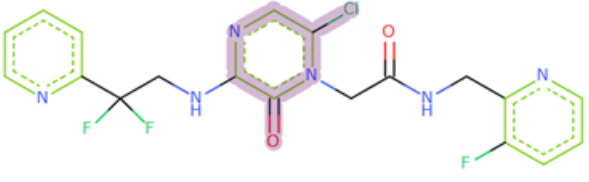
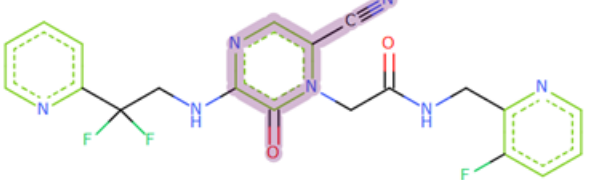
Compound	Structure	$t_{1/2}$
A		3.5
B		6.6
C		9.7

Table 6.2 Modifications to SOM2 along with reported $t_{1/2}$ values (in hours).

When suggestions are requested for the fragment in compound A (highlighted in purple in Table 6.2), the chloro substitution which will transform A into B is suggested as the second most similar suggestion, with a similarity of 0.655. The cyano substitution which will transform A into C is suggested as the 6th most similar suggestion with a similarity of 0.564. In the case of compound A, the transformation which will cause a smaller change to its metabolic stability has been ranked higher than the transformation leading to a larger stability change.

When suggestions are requested for the fragment in compound B (highlighted in purple in Table 6.2), both the methyl substitution (transforming compound B to A) and the cyano substitution (transforming compound B to C) are retrieved. The CN substitution is retrieved as the most similar substitution with a similarity score of 0.662. The methyl substitution is retrieved as the 6th most similar replacement with a similarity score of 0.543. Although both of these transformations will alter the $t_{1/2}$ of compound B by 3.1 hours in either direction, FamePrint suggests the cyano substitution is more similar of the two.

Suggestions are also requested for the fragment in compound C (highlighted in purple in Table 6.2). The chloro substitution (transforming compound C to B) is retrieved as the 4th most similar fragment (similarity score of 0.662) and the methyl substitution (transforming compound C to A) is the 7th

most similar fragment (similarity score of 0.526). In the case of compound C, FamePrint has ranked the substitution leading to a smaller change in metabolic stability as the more similar replacement.

6.5.1.2.2. SOM3 Modifications – Series 1

Several rounds of modifications have also been carried out on the region around SOM3 (Figure 6.8), one series of changes and their corresponding $t_{1/2}$ values are as follows:

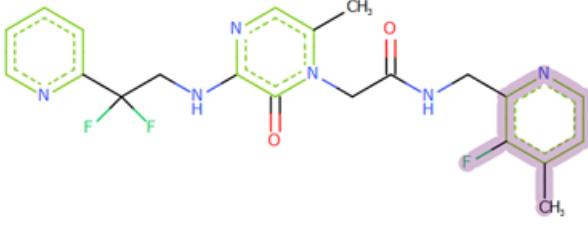
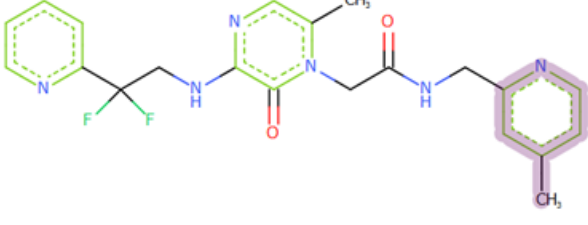
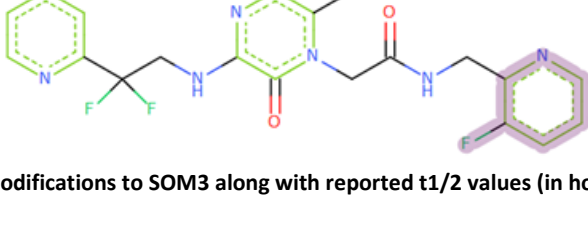
Compound	Structure	$t_{1/2}$
D		1.0
E		2.4
F		3.5

Table 6.3 Series 1 of modifications to SOM3 along with reported $t_{1/2}$ values (in hours).

When suggestions are retrieved for the fragment in compound D (highlighted in purple in Table 6.3), the demethylation substitution (transforming compound D to E) is suggested as the most similar replacement with a similarity score of 0.685 and the removal of the fluoro group (transforming compound D to E) is suggested as the 4th most similar replacement with a similarity score of 0.489. FamePrint has estimated the transformation which will bring about a larger change in metabolic stability (by increasing $t_{1/2}$ by 2.5 hours) as the more similar replacement compared to the change which will increase $t_{1/2}$ by only 1.4 hours.

Fragments in compound E and F (highlighted in purple in Table 6.3) have both been submitted for evaluation. In the case of E, the transformation to F has been retrieved and vice versa. However, as the transformation to D was not found in either case, a comparison cannot be made in this case.

6.5.1.2.3. SOM3 Modifications – Series 2

Another series of modifications have been made to the substructure region around SOM3 (Figure 6.8), these substitutions investigated and their corresponding $t_{1/2}$ values are as follows:

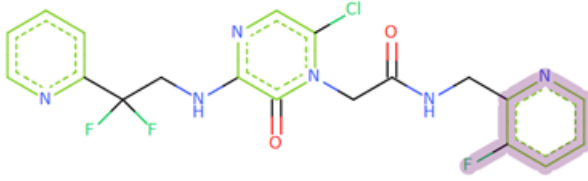
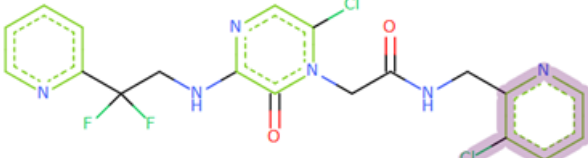
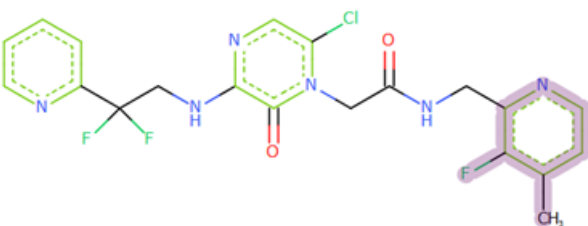
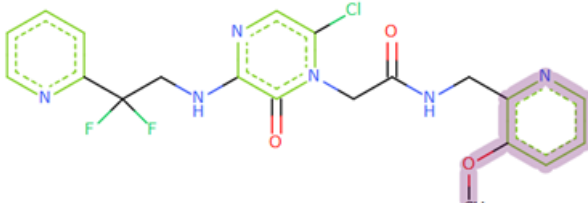
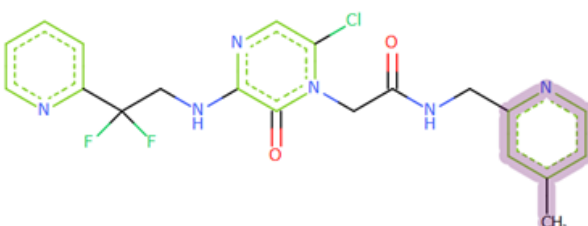
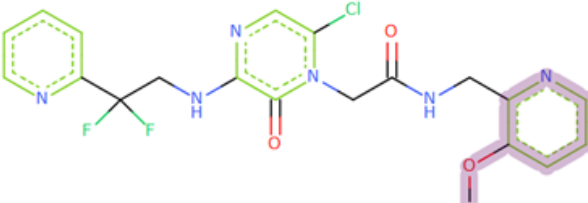
Compound	Structure	$t_{1/2}$
G		6.6
H		5.7
I		3.2
J		2.3
K		2.1
L		2.3

Table 6.4 Series 2 of modifications to SOM3 along with reported $t_{1/2}$ values (in hours).

Replacement suggestions are made for fragment in compound G (highlighted in purple in Table 6.4). The chloro substitution (transforming compound G into H) is retrieved as the 14th most similar fragment with a similarity score of 0.398. The substitution that will transform G into K is retrieved as the 28th most similar fragment with a similarity score of 0.375. Substitutions that will transform G

into I, J or L are not found within the top 30 fragments suggested. In this case, the substitution from G to H which will bring about a $t_{1/2}$ decrease of 0.9 hour is deemed more similar by FamePrint than the substitution from G to K, which brings about a $t_{1/2}$ decrease of 4.5 hours.

When replacement suggestions are retrieved for the fragment in compound H (highlighted in purple in Table 6.4), the fluoro substitution (transforming compound H to G) is the most similar suggestion returned by FamePrint (with a similarity score of 0.712) and the transformation taking H to K is suggested as the 27th most similar replacement with a similarity score of 0.396. Other substitutions transforming H to I, J or L are not found within the top 30 replacement suggestions. In this case, the substitution which will bring about a smaller change in $t_{1/2}$ (+0.9 hour) is also considered more similar by FamePrint than the substitution which will bring about a 3.6 hours change in $t_{1/2}$.

Replacement suggestions are also retrieved for fragment in compound I (highlighted in purple in Table 6.4). The demethylation which will transform I into G is retrieved as the most similar fragment with a similarity score of 0.685. The substitution transforming I into K is retrieved as the 4th most similar replacement (similarity score of 0.468). None of the other substitutions tested are retrieved within the top 30 suggestions. In this case, FamePrint has prioritised the substitution which will bring about a $t_{1/2}$ change of 3.4 hours over the substitution which will only bring about a $t_{1/2}$ change of 1.1 hours.

When suggestions are requested for the fragment in compound J (highlighted in purple in Table 6.4), the substitution taking J to G is retrieved as the 3rd most similar replacement fragment with a similarity score of 0.544. The substitution transforming J to K is retrieved as the 29th most similar replacement with a similarity score of 0.411. No other substitutions tested are found within the top 30 suggestions examined. FamePrint has prioritised the substitution which will bring about a larger metabolic stability change ($t_{1/2}$ increase of 3.4 hours) over the substitution which will only alter $t_{1/2}$ by 0.2 hour.

Replacements suggestions are also retrieved for fragment in compound K and L (highlighted in purple in Table 6.4). Unfortunately, only the substitutions transforming either K or L into G have been retrieved in these cases, therefore a comparison cannot be.

6.5.1.3 Case 3: Metabolic Stability Optimisation of a 5-lipoxygenase Inhibitor

Zileuton is a previously discovered 5-lipoxygenase inhibitor. Bouska *et al.* carried out a metabolic stability and half-life driven optimisation of Zileuton (Figure 6.9, top) by optimising the benzothiophene template, linker and the *N*-hydroxyurea pharmacophore section of the structure separately.¹⁰⁵ Although the hydroxyl of the structure is a known SOM, the *N*-hydroxyurea pharmacophore has previously been identified as being optimal for Zileuton's selectivity and potency, therefore this section was not altered in the optimisation of the structure. Bouska's optimisation effort focused on improving the template and linker.

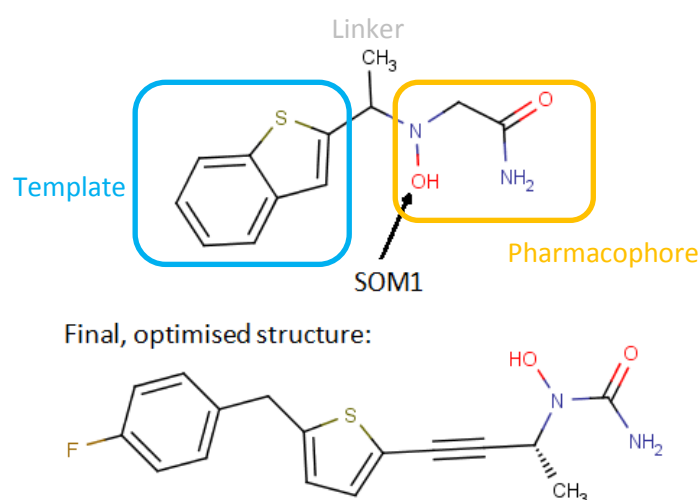


Figure 6.9 Zileuton and the final optimised structure carried out by Bouska *et al.*¹⁰⁵

It is worth noting that the starting structure (Zileuton) is found in the training dataset of the dictionary, however, the optimised structure is not. The SOM is recognised when Zileuton is submitted as the query structure. Two fragments containing the pharmacophore has predicted metabolic stability scores of zero.

The template and linker optimisation experiments have been reported along with plasma half-life ($t_{1/2}$) values. These are used as an indication of metabolic stability. FamePrint's ability to retrieve fragments containing more similar metabolic stability values to each other before retrieving more dissimilar fragments is examined. Only the top 30 most similar replacement fragments will be retrieved and examined in this study. This limit is put in place to prevent the large number of fragment will otherwise be retrieved.

6.5.1.3.1. Template Modification

The following modifications have been made to the template in Figure 6.9, their structure and corresponding $t_{1/2}$ values are as follows:

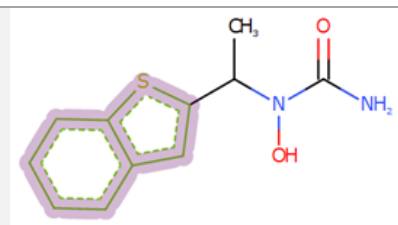
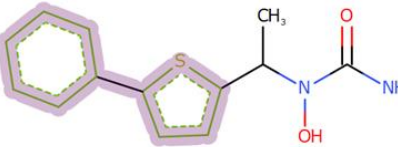
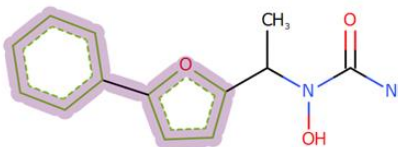
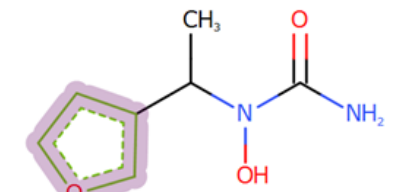
Compound	Structure	$t_{1/2}$
A		0.4
B		0.8
C		1.4
D		3.1

Table 6.5 Template modifications along with reported $t_{1/2}$ values (in hours).

When the fragment (highlighted in purple in Table 6.5) from compound C is selected for replacement, the substitution transforming C to B is retrieved as the most similar replacement fragment with a similarity score of 0.548. The substitution transforming C to A is retrieved as the 15th most similar replacement fragment with a similarity score of 0.432. The replacement transforming C to D is not within the top 30 suggestions retrieved. FamePrint has estimated the C to B transformation (leading to a $t_{1/2}$ reduction of 0.6 hour) as a more similar replacement compared to the C to A transformation which would lead to a larger $t_{1/2}$ reduction of 1 hour.

When the fragments (highlighted in purple in Table 6.5) from compound A, B and D are selected for replacement, none of the tested substitutions shown in Table 6.5 are found within the top 30 suggested replacements. Other template and linker optimisations have been reported by Bouska. However, except for the experiments listed above, FamePrint has only managed to retrieve at most one replacement that was carried out by Bouska with reported $t_{1/2}$ values. This made it impossible to make a direct comparison of the results.

6.6 Conclusion

The ability of FamePrint to identify suitable replacements and generate new structures is tested on examples of optimisations carried out in literature. The replacement procedure is currently only available within the FamePrint tab and requires user input at every stage, from the selection of the fragment to be replaced to the selection of newly generated structures.

Without access to the right dataset, it has not been possible to validate the performance of FamePrint's bioisosteric replacement functionality. It has been shown to prioritise the retrieval of a number of well-known bioisosteres (Figure 6.3). The retrieval of ringed bioisosteres starting using non-ringed fragments was difficult, however the retrieval of non-ringed bioisosteres of ringed fragments is possible and straight-forward.

In the absence of a validation dataset, several retrospective studies are carried out using examples from the literature. Case 1 shows that the SOM prediction by FamePrint correctly highlighted regions. It is also combined with the bioavailability predictor (PGP substrate classifier) to be reported in chapter 7 which gave the correct prediction. It has been possible in this case to reproduce the substitution the authors made to structure P (Figure 6.4) in order to produce the final structure (with a half-life change of 1.8 hours). However, due to the nature of the fragments created by the combination of fragmentation parameters used when creating the dataset for this FamePrint model (Table 5.16), a direct comparison of the reported metabolic stability (half-live values) and the order in which FamePrint retrieves the suggestions cannot be made.

The second and third literature study (Case 2 and 3) compared half-live values ($t_{1/2}$) reported in the literature against the order in which FamePrint retrieved the fragments required to transform one reported literature compound to another. It suffers from the same limitation as case 1 where in some cases, it is simply not possible to make a direct comparison as the changes required for the transformation will not be present in the same fragment. However, in cases where it has been possible to make a direct comparison, it was hypothesised that FamePrint will identify replacement fragments that will create structures which have more similar metabolic stability to the query. Replacement fragments which bring about minimal $t_{1/2}$ changes should be considered more similar, therefore prioritised during retrieval. In the cases where a direct comparison can be made (in case 2 and 3), a mixture of successes and failures is reported.

Due to the lack of an appropriate dataset for validation and the number of cases investigated being too small, it is not possible to determine the feasibility of the bioisosteric replacement methodology reported in this chapter.

7. Improving Bioavailability

Aside from experiencing metabolic stability issues, oral bioavailability often has to be improved upon during drug optimisation. In an effort to understand the cause of poor oral bioavailability, the effects cytochrome P450 3A4 (CYP3A4) and P-glycoprotein (PGP) have on the poor oral bioavailability of drugs were jointly investigated. The former is often implicated in Phase I metabolism of xenobiotic compounds and the latter can be responsible for the efflux of compounds. Together, these two detoxifying systems are a significant cause of low oral bioavailability of some drugs. It has also been identified that CYP3A4 and PGP exhibit overlap in their substrate spectra and transcriptional regulation as well as tissue expression patterns and gene expression.¹⁰⁶ The potential interplay and synergistic actions of CYP3A4 and PGP will be examined and an attempt will be made to produce a classification model for CYP3A4 substrates and PGP substrates.

The aim of this chapter is to develop *in silico* models that will predict the possibility that a compound is a substrate of PGP and/or CYP3A4. These results could then be used in conjunction with FamePrint to produce a tool that is capable of predicting a structure's metabolic vulnerabilities and its oral bioavailability, as well as methods, through structural modification, to improve upon these properties.

Elucidating how CYP3A4 and PGP work together to limit the oral bioavailability of a large number of compounds is not only of scientific interest, but will also help to reduce the amount of time and cost spent on developing a new drug. Given the overlap of their broad substrate spectra and the similarities observed between the systems in terms of their tissue-specific expression patterns and (up)regulation control, the initial hypothesis is that CYP3A4 and PGP interplay does exist and that CYP3A4 metabolites are better substrates for PGP than CYP3A4 substrates.

7.1 Origin of Low Bioavailability

Oral bioavailability is a time-dependent measurement of the fraction of the orally administered dose that reaches the systemic circulation.¹⁰⁷ For orally administered drugs, this largely reflects the extent of gastrointestinal tract absorption. The poor oral bioavailability of many drugs is often due to insufficient solubility in the gastrointestinal fluids, poor gut membrane permeability and/or extensive hepatic first-pass elimination.

The liver is the most important organ in drug metabolism, where most cytochrome-dependant Phase I oxidative reactions and Phase II conjugation reactions take place. However, it is worth noting that drug metabolising enzymes are also present in other tissues such as the gastrointestinal mucosa. The

concomitant metabolism of drugs in the intestine by CYP enzymes and the efflux action of drug transporting proteins is also increasingly recognised to be a major contributor towards significantly lowering the bioavailability of orally administered compounds. This has been supported by clinical studies of a wide variety of orally administered drugs where intestinal metabolism has significantly reduced oral bioavailability, such as erythromycin, tamoxifen, fluoxetine, midazolam, ritonavir, verapamil and raloxifene.¹⁰⁸

In humans, CYP3A4 is the most abundant isoform of the CYP enzymes present in the small intestine^{109,110} and it has been shown to function as a barrier against drugs and xenobiotic compounds in the small intestine.¹¹¹ Inhibition, induction and saturation of CYP3A4 significantly changes the bioavailability of compounds that are CYP3A4 substrates, showing its importance as a first-pass metabolising enzyme.¹¹² Aside from CYP3A4-mediated metabolism, the bioavailability of many compounds is also limited by the efflux action of the transporter, PGP. PGP is present in large quantities in the apical membrane of the intestinal epithelium¹¹³ and it transports drugs and other xenobiotic substances from the intestinal epithelial cells back into the intestinal lumen. Both detoxifying systems contribute towards lowering the bioavailability of xenobiotic compounds and the possible interplay between the two systems will be investigated in more detail in this study.

7.1.1 Cytochrome P450 3A4

CYP3A4 is a member of the CYP superfamily of haem-thiolate monooxygenases and is one of the four members of the CYP3A subfamily found in humans. The CYP3A4 protein contains 503 amino acid residues and can be found tethered to endoplasmic reticulum and microsomal membranes where it performs NADPH-dependent oxidation reactions. The haem prosthetic group is covalently bound to the protein via an axial sulphur atom from a cysteine residue. As is true of all CYP enzymes, CYP3A4 is involved in Phase I and not Phase II biotransformations. The wide range of shapes and sizes of the substrates metabolised by CYP3A4 raises the question of how one active site recognises so many different structures.

CYP3A4 exhibits the same fold, tertiary structure and catalytic cycle as other CYP enzymes. As with a number of other CYPs, the haem prosthetic group in the active site of CYP3A4 is buried inside the protein with no obvious access channels for larger substrates from the cytoplasm.¹¹⁴ Three channels have been found from the various crystal structures; channel 1 passes through the B-C loop as labelled in Figure 7.1, channel 2 passes between β sheet 1, the B-B' loop and the F'-G' loops, and channel 3 passes through the phenylalanine cluster located just above the haem group, made up of seven Phe residues (4 from the F-F' loop, one on B', one from the G-G' loop and one on the I helix).

The Phe-cluster residues interact with each other via π - π stacking to create a hydrophobic barrier through which compounds have to pass if the channel is to be used.

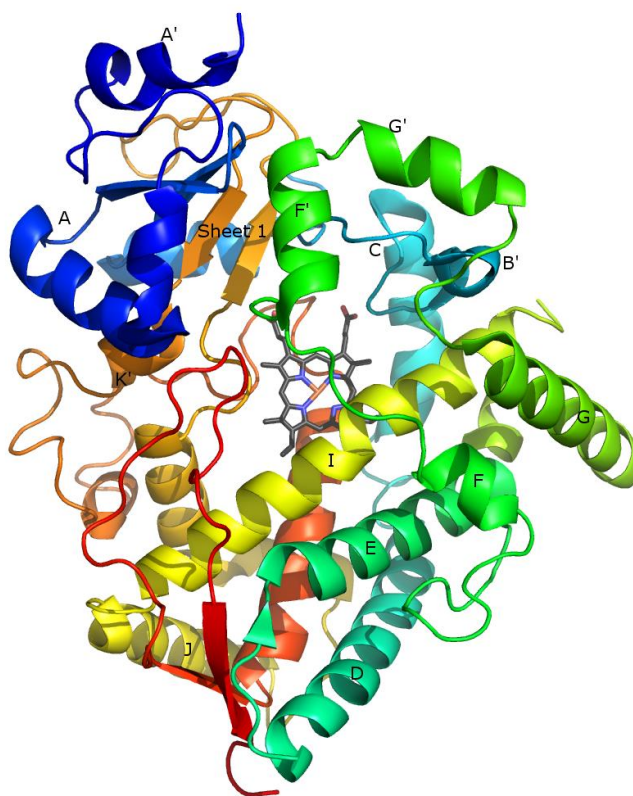


Figure 7.1 Cytochrome P450 3A4 structure. The structure is shown in ribbons (coloured blue at the N terminus to red at the C terminus) and the haem group shown in sticks. The image was drawn using PyMOL, and secondary structures were labelled according to the scheme used in the Williams' study.¹¹⁵

Denisov *et al*¹¹⁶ carried out an all-atom molecular dynamics (MD) simulation on CYP3A4 in water without the N-terminal helix. A separate simulation was also carried out on CYP3A4 with its N-terminal transmembrane helix retained and inserted into a 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC) lipid bilayer with the enzyme in contact with only the polar glycerol head groups and not the fatty acid (except for the N-terminal transmembrane helix). The study showed that both channel 1 and 2 were open whilst the protein was in water, but that only channel 2 remained open in the membrane simulation as the F and G helix shifted and blocked off channel 1. This substrate channel opened directly into the bilayer and contained “additional space in the active site for direct binding and release of bulky hydrophobic substrates and products” into the bilayer.¹¹⁶

CYP3A4 exhibits the broadest substrate specificity out of all human CYP isoforms and is responsible for the metabolism of about 50% of all marketed drugs. The active site of the protein has been shown to undergo dramatic changes, including an increase in volume by >80% upon binding to ketoconazole and erythromycin (Figure 7.2), as well as showing two distinct conformations;¹¹⁷ these

are both significantly different when compared to the shape of the active site when smaller ligands are bound (e.g. metyrapone and progesterone¹¹⁵) where very different active site volumes were observed. Docking studies with dirloapide¹¹⁸ and kinetics and equilibrium studies using ritonavir¹¹⁹ as substrates confirmed significant conformational changes upon binding to CYP3A4. These findings strongly suggest that the enzyme is capable of multiple binding modes, and this can potentially help explain the broad substrate specificity of CYP3A4.

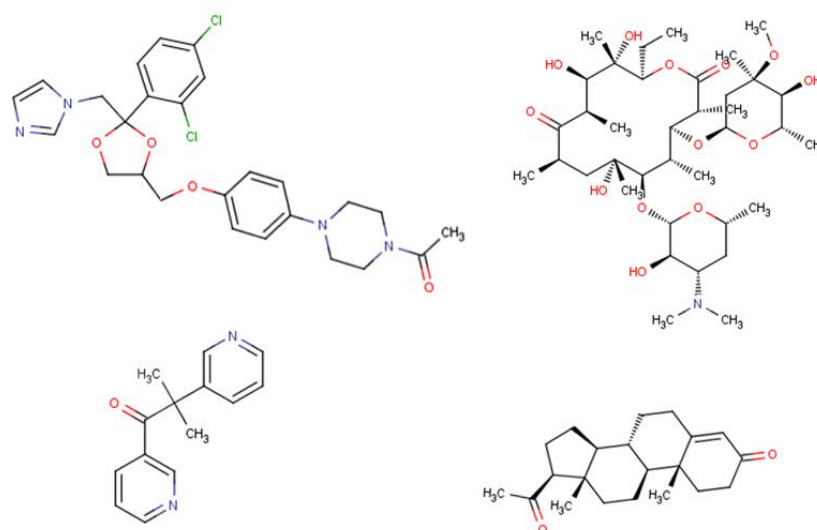


Figure 7.2 Ligands used during CYP3A4 crystallisation. Ketoconazole (top left) and erythromycin (top right), M_r 531 Da and 734 Da respectively, were used in the Ekroos study.¹¹⁷ Metyrapone (bottom left) and Progesterone (bottom right), M_r 226 Da and 314 Da respectively, were used in the Williams study.^{115''}

7.1.1.1 Atypical Substrate Binding Kinetics

Michaelis-Menten kinetics can be used to describe the rate of metabolism of a substrate by an enzyme in many instances for a first-order reaction. This is the one of the best known and simplest enzyme kinetics models. However, atypical, non-Michaelis-Menten kinetic profiles have been observed for an increasing number of CYP3A4 substrates. Multiple binding site models and CYP3A4 allosterism have been shown to account for non-hyperbolic kinetic profiles.¹²⁰ Five binding modes were proposed that could account for the different kinetic profiles observed: activation, auto-activation, partial inhibition, substrate inhibition and biphasic saturation.¹²¹

Experimental studies with testosterone have shown that testosterone exhibits homotropic cooperativity (auto-activation) upon binding to CYP3A4; the enzyme requires the binding of two testosterone molecules for product formation.¹²² Heterotropic cooperativity was shown in the case of α -naphthoflavone binding to CYP3A4's peripheral binding site, facilitating testosterone binding in the proximal binding site and allowing oxidation of testosterone to occur.¹²³ Similarly, the well-known CYP3A4 substrate midazolam also exhibits homo- and hetero-tropic cooperativity. Simulation

studies showed a stacked configuration of two midazolam molecules or one carbamazepine and one midazolam molecule in the CYP3A4 active site.¹²⁴ The authors also speculated that the stacked motif may be a common theme for cooperativity between CYP3A4 binders.¹²⁴ The different binding modes purposed for CYP3A4 is shown here:

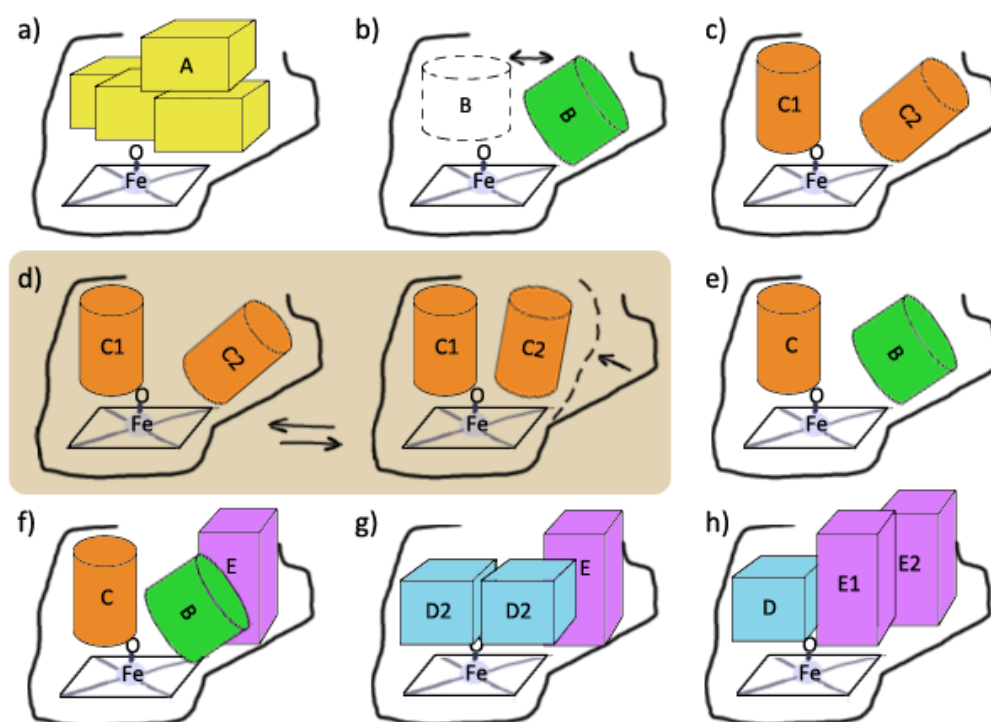


Figure 7.3 Multiple binding modes of CYP3A proposed. The diagram shows the different binding modes suggested^{121,125–138} to account for kinetic data observed. a) Single occupancy of large substrate A. b) B occupies the pocket with different possible orientations and water molecules occupy the rest of the pocket. c) Two identical molecules simultaneously occupy two distinct sites. d) Two identical molecules simultaneously occupy the pocket with more than one possible orientation of the binding pocket. e) Two different molecules simultaneously occupy two distinct sites. f) Three different molecules, one of which may be an effector (inducer/inhibitor), bind simultaneously g) Two identical molecules and an effector bind simultaneously, each occupying a unique site. h) Two effectors and a substrate bind simultaneously, with one effector acting as substrate. (This diagram was reproduced based on image from Ekins *et al.*¹³⁹ and the summary was from the original diagram is used here.)

7.1.1.2 Existing CYP3A4 Models

As CYP3A4 is one of the most important enzymes responsible for the metabolism of clinically relevant compounds and an enzyme of great scientific interest, it is no surprise that despite the uncertainty about substrate binding modes, there have been multiple attempts to generate *in silico* models for the classification of CYP3A4 substrates (as well as inhibitors and activators).

Early molecular modelling and docking studies showed that a H-bonding interaction occurs frequently between the substrate and Asn74 of CYP3A4, along with π - π stacking of the substrate with Phe72.^{140,141} One substrate pharmacophore contains a hydrogen bond acceptor 3Å from the ferryl oxygen and 5.5-7.8Å from the SOM.¹⁴⁰ Another study by Ekins pointed out the importance of the presence of one hydrophobic region, one H-bond donor and two H-bond acceptors in the substrate. The pharmacophore model produced is shown in Figure 7.4.

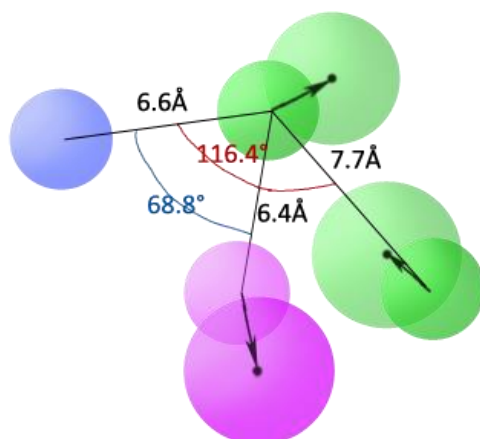


Figure 7.4 CYP3A4 substrate pharmacophore model by Ekins¹³⁴ Hydrophobic area (blue), H-bond donor (red) and H-bond acceptors (green). Diagram reproduced based on image from Ekins.¹³⁴

Although not a QSAR approach, docking studies can also provide valuable insight into the interaction between substrates and enzyme residues. A recent study employed a multi-conformational docking system and used conformations of the protein with small to large significant changes in the active site pocket.¹⁴² This supports the evidence that the residues Arg105, Arg202, Glu374, Ser119, Thr309, Phe213, Phe215 and Phe304 are important in ligand binding and the Phe residues noted here are also part of the Phe-cluster mentioned in section 7.1.1.

The development of a CYP3A4 inhibition model (compounds with $IC_{50} < 3\mu M$) has been reported which is based on high-throughput screening results from 4470 proprietary compounds.¹⁴³ Different molecular fingerprints and topological indices were used as molecular descriptors and a number of machine-learning methods including Naïve Bayes classifier, logistic regression, k-nearest neighbour classification and SVM were used. Every combination of descriptor and machine-learning method was tried and the three best models gave 82%, 82% and 81% accuracy (Barnard Chemical Information fingerprints/SVM, MDL/SVM, topological indices/recursive partitioning).¹⁴³ Although the models made were for predicting CYP3A4 inhibition, it is feasible that the same techniques could be applied to produce a classifier for CYP3A4 substrates.

As previously stated, a number of SOM predictors exist which focus solely on predicting CYP metabolism. However, these models operate on any submitted query structure with the assumption that the structure would be a CYP substrate. A substrate/non-substrate classification model could be applied prior to the prediction of likely SOM.

7.1.2 P-glycoprotein

PGP, also called multidrug resistance protein 1, is encoded by the *ABCB1* gene (structure shown in Figure 7.5). The protein contains 1280 amino acid residues,¹⁴⁴ has a mass of 170kDa and is located on the apical cell membrane of intestinal epithelial cells. PGP is a member of the superfamily of ATP-binding cassette (ABC) transporters. It is an ATP-dependent drug efflux pump for a wide variety of xenobiotic compounds and can cause multidrug resistance in cancer cells.

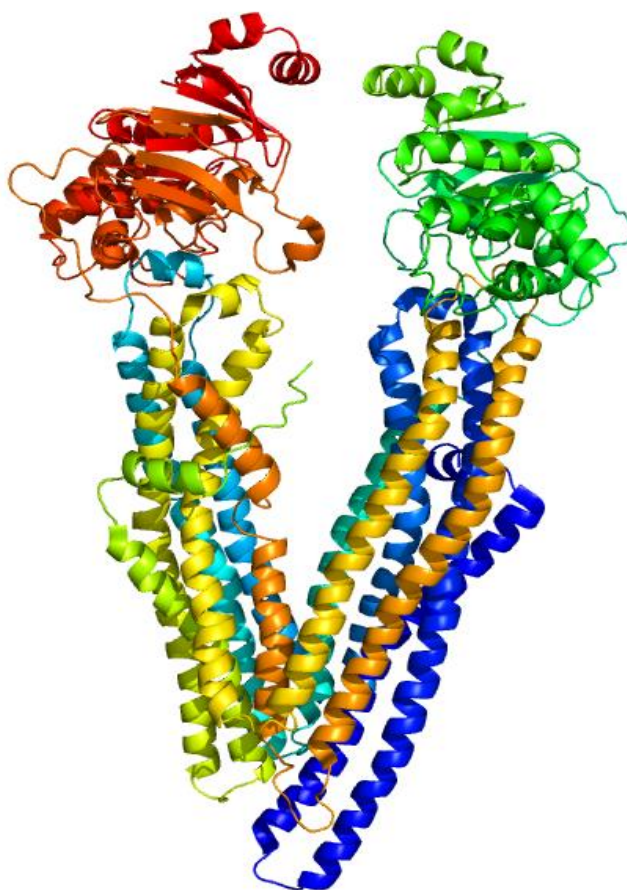


Figure 7.5 PGP structure. The structure is shown in ribbons (coloured blue at the N terminus to red at the C terminus). The image was drawn using PyMol from the PDB structure 3G5U.¹⁴⁵

PGP has the typical architecture of an ABC transporter. The transporter is made up of four subunits all present on the same polypeptide chain: two transmembrane domains (TMDs), which sit in the cell membrane, and two cytoplasmic nucleotide-binding domains (NBDs), which are attached to the base

of their respective TMD.¹⁴⁶ The mechanism of how PGP transports substrate across the membrane and how this transportation is potentially coupled to ATP-hydrolysis is still under investigation. Based on the similarities of the structures of all ABC transporters, importers and exporters, an 'alternating access' mechanism has been suggested¹⁴⁷ and is supported by recent studies performed on analogous transporter proteins with the same coupling helix between the TMDs and NBDs.^{148,149} Due to the difficulties of crystallising membrane proteins, the NBDs are better understood than the TMDs and hence also the substrate binding sites.

7.1.2.1 Transport Mechanism

PGP is similar to CYP3A4 in the sense that both proteins exhibit very promiscuous substrate specificity. It is an accepted hypothesis that this promiscuity is the result of an 'induced-fit' mechanism.^{150,151} The initial binding of a substrate by PGP has been shown to take place close to the cytoplasmic boundary within the membrane.^{152,153} This would mean that the drug molecule can effectively be incorporated into the transporter whilst diffusing across the cell membrane and before reaching the cytoplasm. This makes PGP an excellent safeguard for the cell against unwanted xenobiotics.

However, the mechanistic details of how PGP couples ATP hydrolysis to drug efflux are controversial. Some studies suggested a step-wise mechanism in which one ATP molecule is required for substrate translocation and a second is required to restore the transporter to its resting state¹⁵⁴ whilst others hypothesise that a concerted process is more appropriate.¹⁵⁵ Another hypothesis is that the ATP binding event powers the translocation instead of the hydrolysis.¹⁵⁶ More recent studies seem to be in favour of an efflux mechanism where substrate and ATP binding events are essential prerequisites to the substrate being transported across the membrane, and the energy of ATP hydrolysis is required to reset the resting state of the transporter prior to transportation.

A recent MD simulation study using a mouse PGP crystal structure (PDB 3G5U, Figure 7.5) embedded in POPC lipids found that there is only one channel with access to the substrate binding site that remains open during the simulation. This channel allows substrates that are in the membrane to access the substrate binding pocket of PGP found in the TMDs. The simulation suggests that the binding of a substrate molecule causes long-range conformational changes in the NBDs which then lead to the movement of the protein. The movement observed is consistent with that expected for the translocation of substrates and subsequent hydrolysis of the bound ATP molecules.¹⁵⁷

These results have yet to be verified by experiment. The mechanism as to how translocation couples to ATP hydrolysis is still open to debate and the answer will probably require evidence from thermodynamic studies as well as crystal structures of PGP as it progresses through the cycle.¹⁵⁸

7.1.2.2 Existing PGP Models

A number of PGP recognition patterns have been identified so far. The importance of the presence of a nitrogen atom, van der Waals forces and hydrophobic interactions between the substrates and transporter has been noted.¹⁵⁹ Seelig suggested that “well-defined structural elements are required for an interaction with P-glycoprotein”.¹⁶⁰ It has been predicted that molecules are likely to be substrates if they contain one or more instances of three groups: “type I units” with two electron-donating groups separated by 2.5 Å, and “type II units” which contain two electron-donating groups 4.6 Å apart with an optional electron-donating group between the two. This was confirmed by Ecker, who also pointed out the interaction between a nitrogen atom (frequently seen in PGP substrates) and PGP is not an ionic interaction but based more on its ability to act as an electron donor,¹⁶¹ making it unlikely that the nitrogen will interact with PGP in its positively charged form.

Xue *et al.* produced a PGP substrate classifier using an SVM, which gave a prediction accuracy of 81% for PGP substrates and 79% for non-substrates.¹⁶² The authors found that the SVM gave better prediction accuracies than other machine-learning methods used. However, the dataset Xue used is quite small and only contains 116 PGP substrates. Another study carried out by Wang *et al.* also produced a PGP substrate classifier using an SVM and produced a prediction accuracy of 88%.¹⁶³ The study was carried out on a larger dataset containing 332 structures (206 PGP substrates, 126 non-substrates) and used a combination of ADRIANA.Code¹⁶⁴ and MOE³⁸ descriptors.

A 3D QSAR approach based on Grid-Independent Descriptors was used in another study.¹⁶⁵ Both pharmacophore-based and physicochemical descriptors were tested and a pharmacophore was identified. The pharmacophore was more appropriate for predicting PGP inhibitors than PGP substrates and for predicting the initial interaction with the transporter rather than with the binding pockets for translocation.

As with CYP3A4, PGP studies also have to deal with problems caused by the flexibility and promiscuity of the binding site. A number of recent QSAR studies on PGP concentrated on just one series of homologous compounds with similar scaffolds and chemistry.^{166–168} Several studies^{169,170} pointed out the importance of the presence and distances between aromatic groups and H-bond acceptors, as well as the lipophilicity and molar refractivity of the molecule. A number of PGP pharmacophore models for specific series of compounds have been reviewed by Ecker.¹⁷¹

7.1.3 CYP3A4 and PGP interplay

Both CYP3A4 and PGP exhibit promiscuous substrate binding specificity and have been shown to be responsible for reducing the oral bioavailability of many xenobiotic compounds, including drugs. Very extensive overlaps in the substrate spectrum and the tissue-specific expression of the two proteins have been noted.¹⁷² The additional finding that both CYP3A4 and PGP are subject to regulation by the same prototypical nuclear xenobiotic receptors (e.g. pregnane X receptor)¹⁷³ strongly suggests a functional relationship between the two detoxifying systems. The systems may well work together to limit the bioavailability of a large number of xenobiotic compounds giving, on the whole, a more extensive xenobiotic metabolism.¹⁷⁴ A number of mechanistic proposals have been presented based on evidence from studies performed on CYP3A4 and PGP; three of these are listed here and summarised in Figure 7.6.

In hypothesis 1, a substrate drug, which is a common substrate to both CYP3A4 and PGP, attempts to enter the cell from the intestinal lumen. If, upon entry, the substrate drug encounters PGP, the substrate would get transported back into the lumen. The efflux action of PGP keeps the intracellular concentration of the substrate below the concentration at which CYP3A4 is saturated. In other words, PGP works to keep the intracellular substrate concentration within the range where a change in substrate concentration will produce a linear response in the metabolising capacity of CYP3A4.¹⁷⁴ This could slow down the rate of metabolism but should give rise to a more extensive drug metabolism.¹⁷⁴

In hypothesis 2, it was proposed that the efflux action of PGP on a common substrate drug combined with the reuptake of the substrate into the same or adjacent cells simply afforded CYP3A4 more time to metabolise the substrate drug. In this case, the saturation of CYP3A4 is not taken into account.¹⁷²

In hypothesis 3, it was suggested that the metabolite of CYP3A4 metabolism of a substrate drug may be a better substrate for PGP than the substrate drug itself.^{175,176} This would help prevent the product inhibition of CYP3A4 and may offer a partial explanation to the proximity and potential positional synergy of the two detoxifying systems.¹⁷⁴

An understanding of the interplay between the two systems in controlling drug absorption, metabolism and efflux from the intestinal mucosa into the intestinal lumen will help facilitate the determination of the extent to which intestine mucosa contribute to first-pass metabolism.

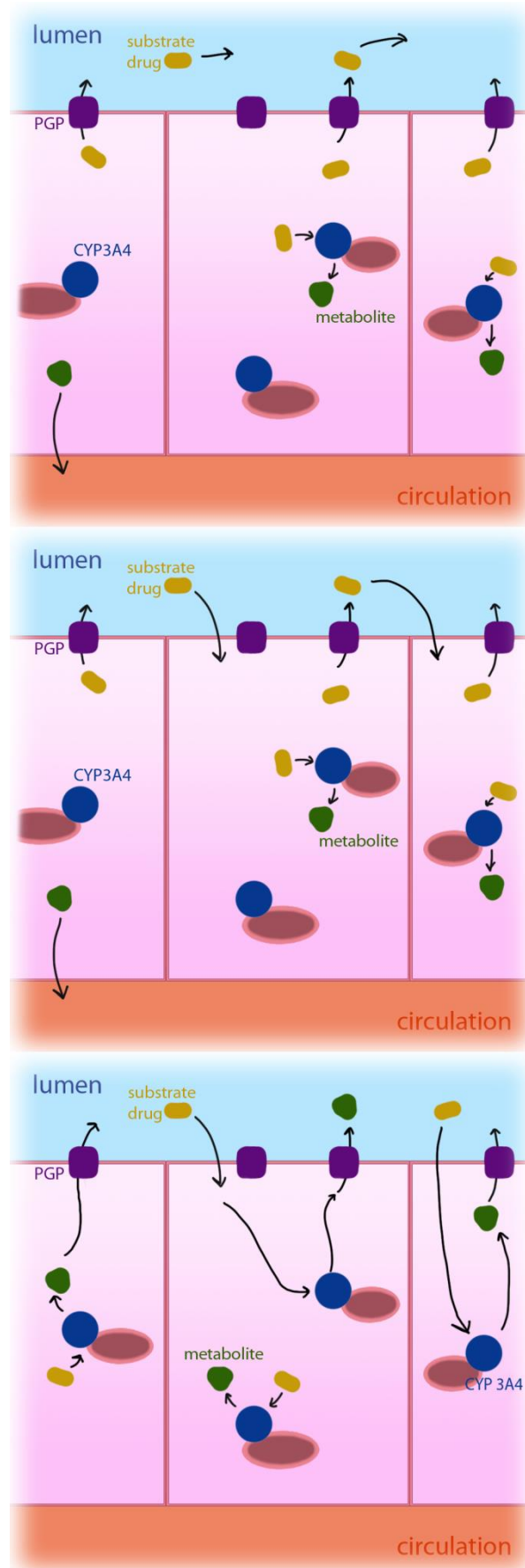


Figure 7.6 CYP3A4 and PGP interplay hypotheses. Hypothesis 1 (top), hypothesis 2 (middle) and hypothesis 3 (bottom).

7.2 Structure-activity Relationship Models

As is the case with SOM predictors and bioisosteric replacement methodologies, various computational approaches have been used to produce bioavailability prediction tools, using ligand-based and structure-based methods. As the ultimate aim is to incorporate the classifiers produced into FamePrint to provide additional guidance on the bioavailability of newly generated structures, it is desirable to have efficient computational methods to allow for real-time predictions to be made in an interactive tool. Therefore, for this investigation, a ligand-based approach was taken for compatibility with FamePrint. A data-mining approach in conjunction with machine-learning methods will be used to produce substrate/non-substrate structure-activity relationship (SAR) classifiers for CYP3A4 and PGP.

7.2.1 Structure-activity Relationship and Machine-learning Approaches

In a study of SAR, it is assumed that similar molecules have similar biological activities.¹⁷⁷ Much like SOM predictors and bioisosteric replacement methods, an appropriate way to describe a molecule is a prerequisite. Different descriptors can be used to capture different aspects of molecular properties, such as topology, spatial arrangement and conformation.¹⁷⁸ In SAR modelling, molecules are traditionally represented by different classes of descriptions such as fingerprints, molecular descriptors (2D and 3D), electron density and various molecular fields.¹⁷⁹ It was found in a study performed on fingerprint descriptors that the aspects of a molecule captured depends much more on the class of the descriptor used rather than the parameterisation of the descriptor.¹⁸⁰ This is a good opportunity to compare the performance of FamePrint with other descriptors; both molecular descriptors and FamePrint fingerprints were used in this study.

Researchers working to create PGP and CYP3A4 substrate predictions often encounter similar problems – both proteins exhibit extremely broad substrate specificities. Given the uncertainty about the substrate-binding pocket, ligand-based approaches are particularly suitable for substrate prediction as they focus on the properties of the substrates. The downside of ligand-based methods is the lack of information on the accessibility of the binding pocket and the steric constraints imposed on the ligands by the pocket itself. It is difficult to tell whether any subsequent modifications suggested will lead to steric clashes with residues in the binding site. However, active compounds are assumed to include all physical properties that allow them to access the binding site.⁹ Also, given the flexibility of the PGP and CYP3A4 binding pockets, this may not be a major problem.

In conjunction with QSAR modelling, machine-learning methods are frequently used to give medicinal chemists immediate feedback on their query compounds. These methods are computationally less demanding than e.g. MD simulations and can therefore give a quicker response. For our problem of trying to classify CYP3A4 and PGP substrates, supervised machine-learning methods would be appropriate. However, unlike MD simulations, QSAR models and machine-learning methods can be limiting in their extrapolation ability and applicability domain, which has to be defined for each model.

7.2.2 Classification Workflow Design

One of the objectives is to study the interplay between CYP3A4 and PGP, as well as producing classification models to categorise substrates and non-substrates of the two targets. It would be useful to compile a ‘negative’ dataset which could represent, as far as possible, the accessible chemical space (including purchasable compounds, metabolites, bioactive structures and drugs). This dataset can be used to represent the general ‘background’ chemical space occupied by small molecules and also for visualisation to observe areas occupied by PGP substrates compared to CYP3A4 substrates. It can also be used to create a two-stage classifier for PGP and CYP3A4 substrates, with the first stage classifier attempting to identify whether the query structure is within the correct region of chemical space to interact with PGP or CYP3A4; and if that is the case, the second stage classifier can then classify the structure as a substrate/non-substrate of PGP or CYP3A4 (separately). However, this background dataset has to be used with caution. Although care was taken whilst building this background dataset, it is not possible to say with absolute certainty that none of the structures in the negative (i.e. neither CYP nor PGP binding) dataset was not a substrate of either CYP3A4 or PGP.

7.3 Data Sources and Preparation

7.3.1 Data Sources

7.3.1.1 Background Dataset

A diverse background dataset was provided by Dr. Andreas Bender. This dataset was used in Peironcelly’s study of metabolite space and metabolite-likeness.¹⁸¹ Compounds were collected from the Human Metabolome Database (HMDB, version 2.5), ZINC (release 8), ChEMBL (release 8) and DrugBank (release 2.5), to represent the chemical space occupied by human metabolites, purchasable compounds (“all”), bioactive compounds and drugs. The assumption is that compounds which interact with CYP3A4 and PGP will fall under part of the chemical space represented by the

data points corresponding to human metabolites, bioactive compounds and drugs, which is a subset of the chemical space represented by the entire background dataset. All structures collected by Peironcelly were represented by Extended-Connectivity Fingerprints ECFP¹⁸², with each atom neighbourhood described by the atom's connections up to 4 bonds away. The similarities between fingerprints were calculated using Tanimoto coefficients.¹⁸³ The structures were then clustered using the maximal dissimilarity partitioning algorithm through the "Cluster Molecules" component in Pipeline Pilot¹⁸⁴. The Tanimoto similarity score between each structure within a cluster and the structure at the cluster centre must be higher than 0.4 for that structure to be a valid member of the cluster. Each data source was clustered individually and all structures that act as cluster centres were combined to form the diverse dataset representing each data source.¹⁸¹ All five diverse datasets were combined and the resulting dataset, containing 83,018 compounds, was used in this study. These compounds were checked against the CYP3A4 and PGP substrate dataset and any substrate structures also in the background dataset were removed. A random subset was chosen to represent background chemical space in the first stage classifier (Table 7.1).

7.3.1.2 CYP3A4 and PGP Substrates Dataset (Stage 1 Classifier)

ChEMBL is a publically available database maintained by the European Bioinformatics Institute (EBI) containing small drug-like bioactive molecules along with their binding, functional and ADMET information.⁵² ChEMBL release 12 contains 1,222,969 compound records, 1,077,189 unique molecules, 596,122 assays, 5,654,847 bioactivity listings and 8,703 targets. This was used as a source of CYP3A4 substrate and PGP substrate information. Compounds with the ChEMBL target IDs CHEMBL340 (CYP3A4) and CHEMBL4302 (PGP) were extracted by Dr. Lora Mak from the ChEMBL database. There were a total of 119 structures annotated as "substrate" under the structure's activity comment and these were taken to form the CYP3A4 substrate dataset from ChEMBL. There were a total of 8 structures annotated as "transporter substrate" under the activity comment retrieved using the ID CHEMBL4302. Due to the small number of PGP substrates found via these annotations, literature sources for 66 compounds listed as "active" against PGP were each examined and 10 more substrates were identified. Thus, a total of 18 PGP substrates were identified these formed the PGP substrate dataset from ChEMBL.

DrugBank 3.0¹⁸⁵ contains 6771 drug structures and is a freely available database containing drug compounds along with their target information. 149 CYP3A4 and 121 PGP substrates were extracted from the DrugBank XML file and their respective substrate datasets from ChEMBL augmented.

SuperCYP is a freely available database containing information on different isoforms of CYP enzymes.¹⁸⁶ Information found on the database was mostly extracted from the original literature

and a list of CYP3A4 substrates was already available on the SuperCYP webpage. However, the structures were not in any downloadable format. Andrew Howlett (PhD student), provided a list of SMILES extracted using their CAS IDs using command line scripting. The 342 substrate structures extracted were added to the CYP3A4 substrate dataset mentioned above.

The Accelrys Metabolite Database (section 3.1) was also used as a source of CYP3A4 substrate structures. All compounds which were listed as undergoing CYP3A4-mediated Phase I biotransformations were extracted and 1418 CYP3A4 substrates were found and appended to the CYP3A4 substrate dataset.

TP-search (2007 release) is a publicly available database of compounds that are substrates, inhibitors or inducers of transporters.¹⁸⁷ It contains information on transporters found in humans, mice, rats, rabbits, pigs and winter flounders. TP-search has data on 33 human transporters, one of which is PGP. 485 PGP substrates entries were listed on the website. Dr. Lora Mak provided the extracted PGP substrate structures (obtained by parsing the compound's name into OPSIN¹⁸⁸ and NCI¹⁸⁹). There were 63 entries where the structure name was misspelt or could not be found. These were manually extracted from the original literature and SMILES were produced using Daylight Depict¹⁹⁰. These compounds were then added to the PGP substrate dataset.

When all structures from all data sources were combined to form the substrate datasets for CYP3A4 and PGP, duplicate entries were removed. All structures were then charged using the MOE washing protocol as detailed in section 3.5.

7.3.1.3 CYP3A4 and PGP Substrates/Non-substrate Datasets (Stage 2 Classifier)

The CYP3A4 stage 2 classifier dataset was obtained from the Yap and Chen study.¹⁹¹ The study collected inhibitor and substrate structures for CYP3A4, CYP2D6 and CYP2C9 from various data sources, resulting in a dataset of 368 CYP3A4 substrates. The authors noted that negative results (non-substrates) are rarely reported in the literature; only 6 non-substrates were identified for CYP3A4 after a comprehensive literature search. The study undertaken by Molnar and Keseru¹⁹² used CYP3A4 substrates as CYP3A4 non-inhibitors as they were evaluated against the target with no reported inhibitory effect. The same rationale and approach was taken by Yap and Chen in identifying non-substrate and non-inhibitor structures for their study and a total of 390 CYP3A4 non-substrates were identified. As only the names were available in the published dataset, the same script used to extract SMILES strings for compounds contained in the SuperCYP database was used to obtain structures from this study.

Metrabase is a database created in the Centre for Molecular Informatics, comprise of substrates and modulators of protein targets which are involved in transport and xenobiotic metabolism.¹⁹³ The majority of Metrabase database entries came from literature sources (Figure 7.7) with less than a third of entries originating from ChEMBL and TP-search. PGP substrate and non-substrate structures were obtained from Metrabase to form the dataset for the stage 2 classifier. Metrabase defines transporter substrates as structures that are transported by a transporter and transporter non-substrates as structures that were experimentally tested against the target transporter but showed no activity. The extracted dataset contained consisted of 446 PGP substrates and 486 non-substrates (accessed July 2013).

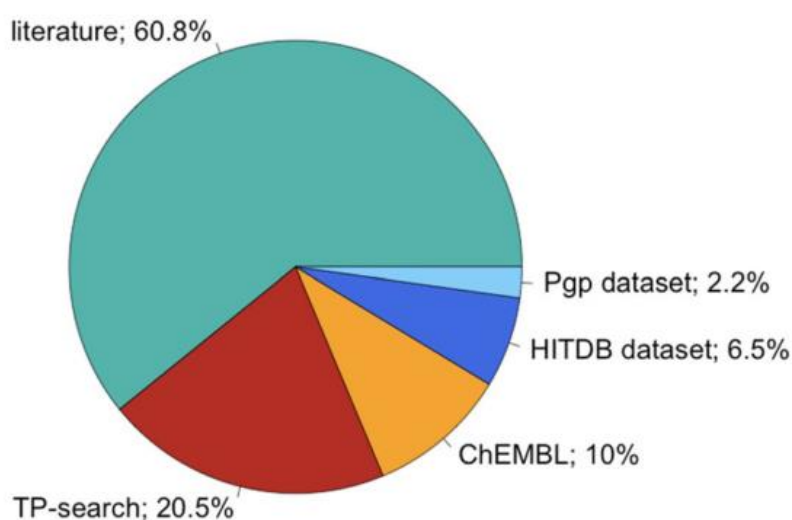


Figure 7.7 Metrabase structure records breakdown by data source

7.3.2 Data Quality and Limitations

A large amount of biological data on chemical compounds has been generated by academic as well as commercial laboratories, especially from high-throughput screening by the latter. Some of these results are available in public and commercial databases. However, these data have to be used with caution, as different databases have different annotation styles, and, regardless of data source, they all contain structural errors (varying between 0.1% - 3.4%).¹⁹⁴ Also, as databases contain information from different assays and laboratories, it is difficult to directly compare activity data, such as IC_{50} , EC_{50} , K_i , K_m and V_{max} values, as they are dependent on assay conditions. It is also difficult to compare values obtained from different assays examining the same compounds and targets.

A recent study on the experimental uncertainty in K_i measurements in public databases found a mean absolute error of around 0.45 pK_i units,¹⁹⁵ which highlighted the need to take experimental uncertainties into account. In an ideal situation, all activity data would be obtained from the same

assay with the same experimental conditions. However, due to the shortage of data, different experimental measures tend to be included in the same model.

The amount of structural overlap of compounds between different databases has also been investigated¹⁹⁶ and it was found that the proportion of compounds unique to each database varies widely across sources investigated; therefore the utilisation of information from multiple sources is advantageous as more data can be obtained. Different databases often contain data points from the same scientific studies. However, it is not uncommon to find discrepancies between entries citing the same scientific publication in different databases, so the use of different data sources is encouraged, not only to increase the amount of information available but also for validation of the accuracy of common data points.

7.3.3 Dataset Curation

The quality of a model built on chemical data can only be as good as the quality of the chemical data in the database; therefore data curation is also an important step before any model building can happen. Also, as structures in this study were obtained from different sources, it is crucial that all duplicate entries were identified and only registered once for each classifier dataset. There are a number of free and commercial dataset curation tools available; a good summary was provided by Tropsha *et al.*¹⁹⁷

Counter-ions (such as Cl⁻ and Br⁻) were first removed from all structures as otherwise two identical structures with different counter-ions could be entered twice into the dataset without being noted as duplicates. This was done using the wash function in MOE (section 3.5) which also protonates strong acids and deprotonates strong bases in structures.

All compounds must also be converted to a standard tautomer before duplicate removal can be carried out, as they could be represented differently depending on which database the structure originated from. The ChemAxon Standardizer was used for dataset preparation.¹⁹⁸ Standardizer is part of ChemAxon's Java-based JChem package and is designed to canonicalise computer generated chemical structures. Standardizer comprises a set of conversion operations (such as aromatisation, dearomatisation, tautomer conversion etc.) which can be customised and specified by the user prior to each (batch) conversion of structures. The following conversion rules were applied to the datasets:

1. Clear stereo: removes absolute stereo configuration from tetrahedral and double bond stereocentres

2. Clear isotopes: converts isotopes to non-isotopic form
3. Remove absolute stereo: removes the absolute stereo flag (chiral flag)
4. Remove water: preforms water removal if possible
5. Tautomerize: generates a canonical tautomeric form of the molecule; the canonical tautomer can be effectively used for duplicate identification. This also dearomatises the molecule
6. Unmap: removes all map numbers from the atoms
7. Remove explicit hydrogens: converts explicit hydrogens to implicit; lonely, isotopic, charged, radical, mapped or wedged explicit hydrogens are excluded

After canonicalization, the SMILES strings of all structures in each dataset were compared and duplicates removed. Stage 1 datasets contain background chemical structures as well as an amalgamation of CYP3A4 or PGP substrates collected from various data sources (as described in section 7.3.1.2) and will be used to produce a classifier to separate out structures which are within the region of chemical space to interact with CYP3A4 or PGP from those that are not. Stage 2 datasets are gathered from specific literature studies (as described in section 7.3.1.3) and will be used to create classifier which will separate CYP3A4 or PGP substrates from their respective non-substrates. A random 80:20 split was carried out using the Coralie Cheminformatic Platform on the both stage 1 and stage 2 datasets to produce the following training and test datasets:

Dataset		Training		Test	
Stage		Substrates	Background/ Non-substrates	Substrates	Background/ Non-substrates
1	CYP3A4	1126	1126	294	294
	PGP	373	373	93	93
2	CYP3A4	292	266	73	66
	PGP	373	389	93	97

Table 7.1 Number of substrate and background (for stage 1)/ non-substrate (for stage 2) structures in the CYP3A4 and PGP training and test datasets.

7.4 Materials and Methods

7.4.1 Descriptor Calculations

7.4.1.1 MOE Descriptors

There are 334 molecular descriptor implementations available in MOE and 186 of them are 2D descriptors. These cover a range of different properties and all 186 2D descriptors were calculated for all molecules in this study.

11 out of 186 2D descriptors are physical properties based on the atom connectivity of the molecule, which contains information on atomic numbers, hydrogen count and neighbour properties. 18 descriptors based on Wildman and Crippen's work¹⁹⁹ were calculated based on the amount of accessible van der Waals surface area approximated from the connectivity data. 28 descriptors are based on atom and bond counts of the molecule. 16 descriptors based on Kier & Hall's work on molecular connectivity chi indexes and Kappa Shape Indexes²⁰⁰ aim to capture different aspects of the shape of a molecule. 8 descriptors are based on descriptors calculated using information on distances of heavy atoms based on a selection of work on topological indices and graphs.^{201–204} 12 descriptors aim to capture the pharmacophore features of the molecule and assign different feature labels such as donors and acceptors to each heavy atom in the molecule. 30 of the 186 descriptors are partial charge descriptors.²⁰⁵

7.4.1.2 FamePrint Fingerprints

Aside from building classification models using molecular descriptor values mentioned above, another set of models were built using FamePrint fingerprint as a description of the structures. Other commonly used fingerprinting methods, such as ECFP, can also be used. However, as these, along with MOE descriptors have already been used in other similar studies¹⁶³, it was decided that FamePrint fingerprint here offers more novelty and information for future research. As these are structures instead of fragments, the length of the fingerprint will be longer than those used in the FamePrint study. For this reason, only version FP00 was used. Fingerprint depths 4, 5, 6, 7, 8 and 9 were used and the results were compared.

First the calculation of all 7 FAME descriptors (using CDK version 1.5.9) for each atom of each structure in the datasets was carried out. The resulting descriptor values underwent discretisation using the adapted equal frequency discretisation method mentioned in section 5.2.3. All seven FP00 fingerprints were then generated for each structure using the procedure outlined in section 5.3.4.1.

All steps from the calculation of descriptors through to producing the fingerprints were carried out on the Coralie Cheminformatics Platform (see section 5.4).

7.4.2 Information Gain Analysis

7.4.2.1 On MOE Descriptors

Information gain is a measure of how much the uncertainty in the class of a structure is reduced once the value of a particular descriptor is known. For the two sets of descriptors calculated using MOE, each of the 186 2D descriptors provide different amounts of useful information. Information gain was used to select the descriptors which would provide the most significant discriminating power, given the end goal.

To obtain the amount of information gain of each descriptor, the training datasets were loaded into Weka (version 3.6.12)⁴¹. The `weka.attributeSelection.InfoGainAttributeEval` attribute evaluation was used to calculate the amount of information gained by each descriptor and rank them accordingly. This was used alongside the `weka.attributeSelection.Ranker` which allows the specification of the minimum information gain threshold value. Descriptors which provided information gain below the specified threshold at discriminating substrates from background or non-substrate structures were removed.

7.4.2.2 On FamePrint Fingerprints

Information gain was also applied to the fingerprints. Fingerprint sizes depend on the longest topological distance between any given pair of atoms within a structure, and the length of the fingerprint can get rather large (Table 7.2). However, there are some positions/bits along the fingerprint in each dataset that would not be set by any structure, and potentially some that are set by all structures. These bits will provide no information gain for the purpose of discriminating positive structures from negative ones and hence were removed to reduce fingerprint sizes.

Each training dataset was loaded into Weka (version 3.6.12)⁴¹ and all bits in the fingerprint containing a constant value through all training structures were removed. This was done using the `weka.filters.unsupervised.attribute.RemoveUseless` filter. The `maximumVariancePercentageAllowed` ($\text{number_of_distinct_values}/\text{total_number_of_values} * 100$) was set to 100, such that no bits other than those containing constant values would be removed. All bits which contained (calculated information gain) information were kept.

Dataset		Fingerprint depth				
		4	5	6	7	8
Classifier 1	CYP3A4	38223 (31831)	41013 (33679)	44361 (35931)	48267 (38621)	52731 (41732)
	PGP	38223 (31828)	41013 (33432)	44361 (35379)	48267 (37819)	52731 (40527)
Classifier2	CYP3A4	11508 (7646)	12348 (7806)	13356 (8013)	14532 (8247)	15876 (8594)
	PGP	38223 (30673)	41013 (32018)	44361 (33823)	48267 (36084)	52731 (38646)

Table 7.2 Length of fingerprint for each dataset. Number of bits per fingerprint with zero information included in parenthesis.

7.4.3 Machine-learning Methods

All machine-learning algorithms used in this study were run within Weka (version 3.6.12)⁴¹. Random forest, SVM and Naïve Bayes were chosen as a starting point as they are three of the most used machine-learning methods and tend to give good results from examples given in the literature. The J48 decision tree within Weka is also chosen, as it is straightforward to produce and results from initial investigations showed some promise.

SVM is a popular classification technique. SVM generates a hyperplane that separates different classes from each other and it aims to maximise the distance between each data point and the plane. The LibSVM a popular implementation of SVM²⁰⁶ which is also embedded in Weka. There are two classification SVM formulations available in LibSVM, C-SVC and nu-SVC. The two differ only in that nu-SCV provides an upper bound on the fraction of training errors and a lower bound on the fraction of support vectors.²⁰⁶ Both were attempted but nu-SVC produced better results on initial attempts. Different combinations of the relevant adjustable parameters (such as cost, gamma and nu) were altered to produce different SVM models.

Random forest is a classification technique first developed by Breiman;²⁰⁷ it is essentially an ensemble of classification trees. Each classification tree in the forest is developed from a randomly selected subset of descriptors and the descriptor that best reduces the uncertainty in the class label is chosen. The classification by RF is based on the majority vote from all tree classifiers in the forest. Different combinations of maximum depth of tress, number of attributes used at each node and the number of trees to generate when building a model were attempted.

A Naïve Bayes classifier is a probabilistic classifier that assumes that the presence or absence of any feature in a class is completely independent of the presence or absence of any other feature in the same class. Relative frequency was used as a probability and for conditional probability estimation.

J48 classifier is also used in the study as initial tests revealed promising results. The J48 classifier implemented by WEKA generates a pruned or un-pruned C4.5 decision tree. The C4.5 decision tree considers one attribute at each split, and for each split, the attribute which provides the greatest amount of information gain with regards to the discrimination of classes under test is used. Both the pruned and un-pruned trees were tested, along with varying the required minimum number of examples per node.

All machine-learning methods mentioned above along with different parameterisations of each were tested on the training set for each dataset. The models which produced the best result from 10-fold cross validation carried out on the training data were then applied to the test dataset and the results are reported below. The generation of the 10-fold cross validation datasets using the training data as well as the evaluation of the cross validation results were all carried out in Weka.

7.4.4 Multidimensional Scaling

Multidimensional scaling (MDS) aims to find a lower-dimensional projection of a higher dimensional dataset. The use of MDS in this study was to help visualise the distance between each structure in all dimensions on a 2D plot and to see if either of the CYP3A4 or PGP substrate datasets occupied a different and distinctive region compared to each other and the background chemical space (using the datasets for stage 1 classifiers). In an MDS plot, the axes are dimensionless.

An MDS calculation requires Euclidean distances as input, which can be calculated from discretised data. MDS implemented in Orange canvas (version 2.7.8) was used in this study. Discretisation was performed using the “Entropy-MDL discretization” and “Use default discretization for all attributes” options available under the Discretize widget. The discretisation aims to maximize the information gain from each splitting until the gain is below a minimum threshold. Manhattan distances were then calculated and as a measure of distance between structures when carrying out MDS calculations.

Each optimisation of the MDS graph was carried out with the stopping condition triggered when either i) the minimum average stress changed was less than 0.00005 compared to the previous optimisation step or ii) when 5000 steps had been carried out. It is likely that the first optimisations will lead to local minima rather than the global minimum; therefore after each optimisation step, the “Jitter” functionality was employed. This moves each point in a random direction for a short distance and is useful in trying to escape from local minima. Optimisation was then repeated until the average stress of the plot was of the order of 1.

7.5 Results and Discussion

7.5.1 Classifiers

Various machine-learning methods and parameterisations were evaluated and the model which produced the highest AUC value during the 10-fold cross validation carried out on the training dataset was then used to carry out testing on the test dataset. The results are given in the tables below (Table 7.3, Table 7.5, Table 7.7 and Table 7.9) along with the information gain cut off applied to the datasets being tested and the machine-learning method used in each case.

FamePrint fingerprint version FP00 was also used to evaluate the same training and test dataset to produce two stage 1 and stage 2 classifiers. The results are given in tables below (Table 7.4, Table 7.6, Table 7.8, and Table 7.10) along with the fingerprint depth used in each case.

7.5.1.1 Stage 1 Classifier Results

Stage 1 classifiers attempt to determine whether the query structure is within the correct chemical space occupied by a CYP3A4 or a PGP substrate amongst general background chemical space.

7.5.1.1.1. CYP3A4

Information gain cut-off	Number of descriptors	10-fold cross validation			External test set			Method
		Accuracy	AUC	MCC	Accuracy	AUC	MCC	
0.00001	55	98.0	0.996	0.961	74.2	0.816	0.483	RF
0.02	44	98.0	0.996	0.950	74.1	0.815	0.479	RF
0.03	21	98.0	0.996	0.937	72.1	0.802	0.415	RF
0.039	8	98.0	0.996	0.929	69.8	0.763	0.392	RF

Table 7.3 CYP3A4 stage 1 classifier results using molecular descriptors.

Fingerprint depth	10-fold cross validation			External test set			Model
	Accuracy	AUC	MCC	Accuracy	AUC	MCC	
4	82.5	0.825	0.603	72.5	0.737	0.527	SVM
5	82.2	0.823	0.660	75.7	0.766	0.562	SVM
6	83.3	0.833	0.675	74.8	0.758	0.550	SVM
7	83.2	0.832	0.666	74.1	0.751	0.536	SVM
8	83.2	0.832	0.664	77.5	0.780	0.570	SVM

Table 7.4 CYP3A4 stage 1 classifier results using FamePrint fingerprints.

7.5.1.1.2. PGP

Information gain cut-off	Number of descriptors	10-fold cross validation			External test set			Model
		Accuracy	AUC	MCC	Accuracy	AUC	MCC	
0.00001	159	97.0	0.979	0.940	94.6	0.947	0.893	J48
0.055	128	95.5	0.957	0.909	97.3	0.956	0.948	RF
0.09	98	92.6	0.972	0.852	84.6	0.953	0.692	RF
0.12	72	92.9	0.974	0.859	85.9	0.955	0.719	RF
0.16	43	93.9	0.974	0.880	87.9	0.950	0.759	RF
0.18	30	93.4	0.976	0.869	88.6	0.953	0.747	RF
0.25	11	92.9	0.965	0.859	89.3	0.956	0.787	RF
0.27	5	92.3	0.968	0.846	90.6	0.930	0.815	RF

Table 7.5 PGP stage 1 classifier results using molecular descriptors.

Fingerprint depth	10-fold cross validation			External test set			Model
	Accuracy	AUC	MCC	Accuracy	AUC	MCC	
4	98.0	0.979	0.960	96.0	0.968	0.920	J48
5	97.3	0.967	0.946	95.8	0.966	0.920	J48
6	97.3	0.959	0.946	95.3	0.953	0.906	J48
7	98.7	0.986	0.973	95.0	0.954	0.899	J48
8	98.0	0.992	0.960	95.5	0.963	0.909	J48

Table 7.6 PGP stage 1 classifier results using FamePrint fingerprints.

7.5.1.2 Stage 2 Classifier Results

The stage 2 classifiers attempt to determine whether the query structure is a substrate or non-substrate of CYP3A4 or PGP.

7.5.1.2.1. CYP3A4

Information gain cut-off	Number of descriptors	10-fold cross validation			External test set			Model
		Accuracy	AUC	MCC	Accuracy	AUC	MCC	
0.00001	166	64.0	0.702	0.277	65.5	0.677	0.306	RF
0.03	56	65.2	0.673	0.301	64.0	0.647	0.276	RF
0.04	36	64.0	0.662	0.276	61.9	0.612	0.233	RF
0.048	19	64.2	0.642	0.280	61.9	0.605	0.233	RF
0.055	11	64.9	0.645	0.294	59.7	0.593	0.189	SVM
0.064	7	65.1	0.646	0.298	62.6	0.622	0.247	SVM

Table 7.7 CYP3A4 stage 2 classifier results using molecular descriptors.

Fingerprint depth	10-fold cross validation			External test set			Model
	Accuracy	AUC	MCC	Accuracy	AUC	MCC	
4	65.2	0.698	0.297	61.6	0.644	0.228	RF
5	65.6	0.703	0.305	62.3	0.674	0.243	RF
6	63.7	0.701	0.267	65.9	0.658	0.317	RF
7	65.0	0.725	0.293	66.7	0.687	0.332	RF
8	66.9	0.705	0.334	64.5	0.668	0.289	RF
9	65.4	0.709	0.304	62.3	0.681	0.245	RF

Table 7.8 CYP3A4 stage 2 classifier results using FamePrint fingerprints.

7.5.1.2.2. PGP

Information gain cut-off	Number of descriptors	10-fold cross validation			External test set			Model
		Accuracy	AUC	MCC	Accuracy	AUC	MCC	
0.00001	167	90.3	0.960	0.805	87.1	0.943	0.743	RF
0.05	122	90.1	0.960	0.803	86.6	0.949	0.731	RF
0.105	81	89.9	0.959	0.795	88.2	0.953	0.763	RF
0.14	56	90.1	0.955	0.794	88.2	0.953	0.763	RF
0.185	29	87.7	0.939	0.754	85.5	0.936	0.710	RF
0.2	18	87.8	0.942	0.756	84.9	0.930	0.700	RF
0.21	11	88.1	0.941	0.761	84.9	0.936	0.700	RF

Table 7.9 PGP stage 2 classifier results using molecular descriptors.

Fingerprint depth	10-fold cross validation			External test set			Model
	Accuracy	AUC	MCC	Accuracy	AUC	MCC	
4	84.4	0.930	0.685	81.1	0.878	0.639	RF
5	85.0	0.928	0.697	76.9	0.890	0.559	RF
6	84.0	0.921	0.676	78.7	0.872	0.587	RF
7	85.0	0.929	0.697	80.5	0.900	0.628	RF
8	84.7	0.931	0.691	80.5	0.900	0.628	RF
9	84.9	0.931	0.695	80.5	0.894	0.624	RF

Table 7.10 PGP stage 2 classifier results using FamePrint fingerprints.

7.5.1.3 Discussion

The results show that all the CYP3A4 stage 1 classifiers (using 2D molecular descriptors and FamePrint fingerprints) perform much better than their stage 2 counterparts. This is as expected as background structures represent a large chemical space, including a diverse range of chemical properties, as opposed to the chemical space occupied by non-substrates.

The performance of stage 1 CYP3A4 classifiers based on molecular descriptors shows better cross validation results than the corresponding classifiers produced by FamePrint fingerprints. However, when the models selected (based on the highest AUC value) are used to perform prediction on test dataset structures, the performances produced by molecular descriptor classifiers are not significantly higher than the ones produced by FamePrint fingerprints. This suggests that perhaps the molecular descriptor models are over-fitted to the training dataset and the FamePrint fingerprint is less prone to overfitting, at least in this case.

Stage 1 PGP classifiers also produced better performance compare to their stage 2 counter parts. As is the case with the CYP3A4 models, this is also to be expected as the division between substrates and a diverse range of chemical structures should be clearer than the division between substrates and non-substrates. The differences in performance between molecular descriptor classifiers and FamePrint classifiers are smaller compared to the differences seen in the stage 1 CYP3A4 classifiers. Performance in the stage 1 PGP classifiers shows a gradual decrease with a decreasing number of molecular descriptors (due to an increase in the information gain threshold). In contrast, the performance stays relatively constant with different fingerprint depths used.

All stage 1 PGP classifiers performed better than all stage 1 CYP3A4 classifiers. This is no surprise either. This further endorses the inference that CYP3A4 has broader substrate specificity, leading to a fuzzier chemical region which both the 2D molecular descriptors used here and FamePrint fingerprints are less well able to define. As mentioned in 7.1.1., the substrate pocket of CYP3A4 is known to be extremely flexible, attributing to the broad substrate specificity of the enzyme. This is expected to make distinguishing CYP3A4 substrates from general background chemical space (and indeed non-substrates) more challenging, as reflected by the performances of stage 1 classifiers.

Stage 2 CYP3A4 classifiers show the poorest performance. There are only small differences between the classifiers produced by molecular descriptors and FamePrint fingerprint. Neither the 2D molecular descriptors nor the fingerprints proved adequate in producing classifiers with good discriminatory powers to distinguish CYP3A4 substrates from its non-substrates.

When comparing the fraction of FamePrint fingerprint bits that are used by both stage 1 and stage 2 CYP3A4 and PGP classifiers (Table 7.11), the spread of information across the whole fingerprint is significantly higher in the case of the stage 2 CYP3A4 model when compared to the fingerprints used by the rest of the classifiers. This suggests that there is a large amount of variation in the description of the structures which perhaps made it more difficult in locating the discriminating features between CYP3A4 substrates and non-substrates.

Dataset		Fingerprint depth				
		4	5	6	7	8
Classifier 1	CYP3A4	16.7%	17.9%	18.8%	20.0%	20.9%
	PGP	16.7%	18.5%	20.2%	21.6%	23.1%
Classifier2	CYP3A4	33.6%	36.8%	40.0%	43.2%	45.9%
	PGP	19.8%	21.9%	23.8%	25.2%	26.7%

Table 7.11 Percentages of fingerprint containing information, calculated from Table 7.2)

The stage 2 PGP classifiers on the other hand show good performance in discriminating PGP substrates from non-substrates. The molecular descriptor classifiers in particular show promising results, although the same gradual decrease in performance with increasing information gain threshold is also seen here. The stage 2 PGP classifiers produced by FamePrint fingerprints did not perform as well as their molecular descriptor equivalents in this case (as was the case with stage 1 PGP classifiers).

In terms of the performance differences between 2D molecular descriptors chosen for this study and FamePrint fingerprints (FP00), both perform equally badly when attempting to distinguish a CYP3A4 substrate from a non-substrate. Both produced similarly successful predictions when discriminating PGP substrates from general background chemical space. FamePrint fingerprints did not perform as well as the 2D molecular descriptors during the cross validation of the training dataset carried out for stage 1 CYP3A4 classifiers and stage 2 PGP classifiers. The stage 2 PGP FamePrint fingerprints classifiers also did not perform as well as the 2D molecular descriptor when carrying out predictions on the test datasets. However, this is not the case for the stage 1 CYP3A4 classifiers. Despite poorer performance during cross validation, stage 1 CYP3A4 FamePrint fingerprint classifiers produced a similar prediction performance to its molecular descriptor counterpart.

It is also interesting to note that the best performing model in each “block” of classifiers is always produced by the same machine-learning method, with slight differences in parameterisation. This suggests that perhaps one type of machine-learning method (as opposed to purely parameterisation) is better able to describe a particular set of differences in the dataset, and

performance of classification models depends more on the type of machine-learning model used than how they are parameterised. This could certainly be an interesting investigation.

7.5.2 Multidimensional Scaling

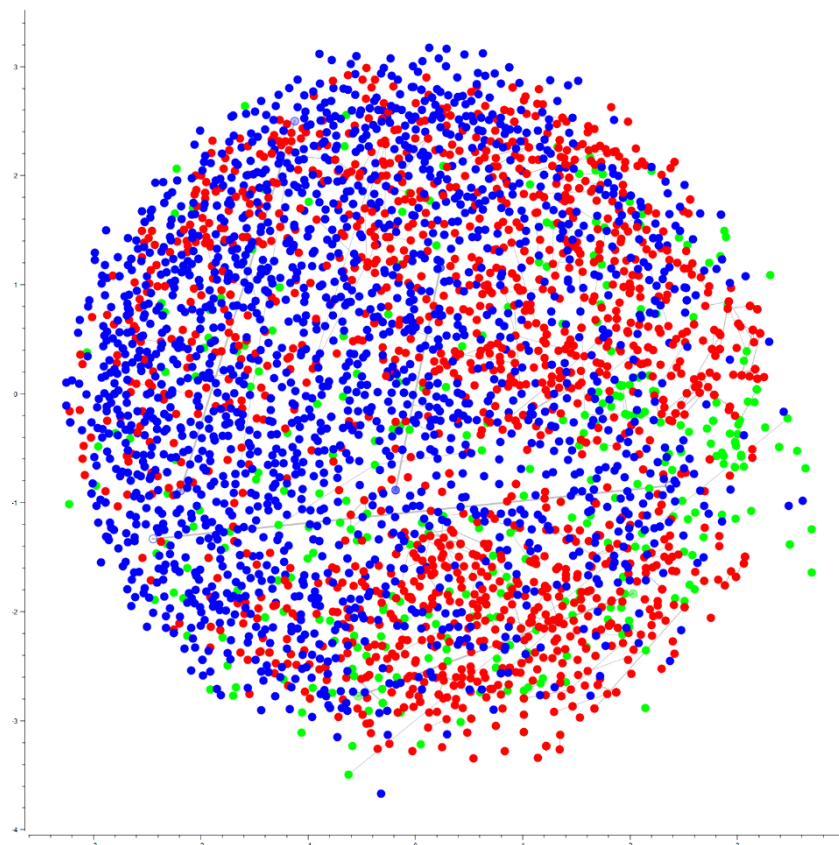


Figure 7.8 MDS plot of background (blue), CYP3A4 substrate (red) and PGP substrate (green) structures show overlap of CYP3A4 and PGP substrate chemical space.

The flexibility of these proteins is such that they can accommodate a very wide range of compounds. This can be seen, especially in the case of CYP3A4 substrates, where the area occupied by substrate structures very much blends into the area occupied by the background structures. In terms of clustering or classification, where the objective is to identify subsets of compounds that are common or unique to each protein, this presents severe challenges for current approaches using ligand-based methods.

As can be seen in the graph above, there is significant overlap between the red and blue regions, representing CYP3A4 substrate and background structures respectively, although a slightly higher concentration of red instances on the right hand side and blue instances on the left hand side can be seen. Comparatively, despite the smaller number of structures, PGP substrates occupy much more distinctive regions and have less overlap with background structure regions when compared to the CYP3A4 substrate structures. This would explain the difficulty in producing good classification

models for CYP3A4 substrates compared to classifiers for PGP substrates. The diversity of compounds in the CYP3A4 dataset makes this much more challenging to describe with a classification model compared to PGP substrates.

Although a conclusion regarding the form that a synergistic relationship may take between CYP3A4 and PGP cannot be reached from analysis of the visualisation above, it is clear that there are major overlaps in the chemical space occupied by the two sets of substrate structures. The results from the MDS study support experimental data which suggest that there is significant overlap between substrates that bind to CYP3A4 and PGP.

7.6 Conclusion

This chapter reported the development of machine-learning based two-stage classifiers for CYP3A4 and PGP substrates. The first stage classifiers attempts to differentiate structures which are within the correct chemical space to interact with either CYP3A4 or PGP. Structures will only be passed through to the second stage classifiers, which differentiate between substrates and non-substrates of CYP3A4 and PGP, if they have been evaluated as being in the appropriate chemical space by the first stage classifier.

The performance of the first stage classifiers which differentiates CYP3A4 or PGP substrates against the much wider chemical space occupied by the background dataset (consisting of human metabolites, purchasable compounds, bioactive compounds and drugs) is acceptable. This is not too surprising as even with the promiscuity of CYP3A4 substrate binding pocket, as the background dataset is expected to cover a much broader chemical space.

The second stage CYP3A4 classifiers did not produce comparable performance to the study conducted by Yap¹⁹¹, from which the dataset for the second stage CYP3A4 classifiers was obtained. The best performing stage 2 model produced by this study has an accuracy of 66.7% (MCC = 0.33), compared to the accuracy of 98.2% (MCC = 0.899) reported by Yap.

The second stage PGP classifiers produced slightly better performance compared to what has been reported in literature. The best performing classifier from this study has an accuracy of 88.2% (MCC = 0.76), compared to the study conducted by Wang¹⁶³ (from which the dataset for the second stage PGP classifiers was obtained) where the best performing model gave an accuracy of 88% and MCC score of 0.73.

8. Conclusions and Future Work

This thesis has reported the development of Coralie Atom-based Statistical SOM Identifier (CASSI) and FamePrint, along with a series of CYP3A4 and PGP substrate/ non-substrate classifiers. Both CASSI and FamePrint have been implemented within the Coralie Cheminformatics platform created by Lhasa Limited. and are available for use via the Coralie application GUI.

CASSI and FamePrint were trained and tested on the same data source used by MetaPrint2D, MetaPrint2D-React and FAME. The SOM prediction performance using different parameterisations of CASSI and FamePrint were evaluated. A few parameterisations of CASSI produced similar performance statistics to MetaPrint2D and MetaPrint2D-React but did not perform as well as FAME. The SOM prediction performance of different parameterisations of FamePrint were evaluated and the final selection of parameters produced SOM prediction performance that outperformed FAME in all performance measures tested (except for median AUC value).

The performance of CASSI could potentially be improved by setting a minimum number of examples required for a fragment. This would be the minimum number of times a fragment has been generated during the training dataset fragmentation before the fragment structure could be used for the prediction of metabolic stability.

As well as being a SOM predictor, FamePrint has been extended to suggest bioisosteric replacements which potentially maintain the substructure's metabolic stability profile. When a query fragment has been selected for replacement, FamePrint will retrieve a number of compatible replacement suggestions from a dataset of fragments based on the properties of the query fragment. If a replacement suggestion is chosen, new compound(s) will be created using the replacement. The GUI created for FamePrint also provides users with the ability to generate a SOM predictor based on their own structures using the FamePrint method if a structure file with SOM information is provided.

A current drawback of the FamePrint method is the amount of memory required to store the data required for SOM prediction and bioisosteric replacement identification. To significantly reduce the memory required to execute the application, this data could be stored in a database. The application would then be enhanced to retrieve information needed for its prediction and replacement tasks from this database, making the application usable on less powerful machines or open up the possibility of using larger fingerprints. However, a quicker fingerprint comparison method may be required in order to produce a rapid response for real-time results, especially if a larger dataset was used. This would require indexing of the fingerprints stored in the database.

Further improvements to the FamePrint fingerprints could also be carried out. One limitation of the fingerprints is their length. Each structure/fragment is represented by 7 fingerprints, each represented as a bit string. One limitation of the current versions of FamePrint fingerprints is that all 7 fingerprints are used independently of each other during similarity scoring. This is due to the fact that combining all 7 would require an extremely long fixed length fingerprint, making storage and operation of fingerprints in memory impractical, especially on less powerful machines. A better data storage/ access solution and indexing function could also allow the combination of all 7 fingerprints into one. This approach may even allow fuzzy fingerprints to be produced (which is currently hindered by the length of the fingerprint that is required). As 'fuzzy' implementations have previously been shown to reduce categorisation error introduced by distance bins, this would be expected to improve performance.^{177,208}

The bioisosteric replacement process within the Biostere module currently involves user input at every step. It is possible this could be automated – replacements could be generated, for example, for the top 5% of the most unstable fragments using the top 10 most similar fragments identified. However, this poses significant challenges as it can quickly lead to a combinatorial explosion of new structures generated. Also, as observed in several examples given and in the high overlap scores obtained during systematic evaluation of dictionaries, the most unstable fragments often overlap. It may therefore be redundant to use all of these overlapping fragments for starting points for replacement. It would be possible to use only the maximum common substructure of the overlapping fragments to begin the replacement process. A backtracking algorithm can also be put in place to remove any structure that have previously been processed to avoid redundant loops.

It is important to point out that despite compatibility checks carried out prior to the generation of new structures by FamePrint, the structures created by replacement of a query fragment with the suggested replacement fragment may not always be synthetically accessible. This is a common problem with similarity-based methods. Knowledge-based approaches are less likely to encounter this particular pitfall. However, as ligand-based methods are capable of identifying previously unused potential replacements, this disadvantage is often overlooked. A workaround is to include a synthetic accessibility filter after the generation of new structures have been carried out. A number of methodologies are already available and can be adopted.^{82,209,210}

Another possible improvement for FamePrint (and CASSI) is to alter the SOM annotations carried out by MetaPrint2D-React (section 2.1.3.4) by injecting awareness of catalytic mechanism of metabolic enzymes. Currently, atoms are labelled as SOM if they were altered in the transformation listed in the Accelrys Metabolite Database. This does not account for the catalytic mechanism carried out by

metabolising enzymes and so is only a quick approximation of true SOM and is not always representative of which site the transformation is initialised at. For example, for the hydrolysis of an amide, the carbon atom of the carbonyl and the nitrogen atom would be annotated together as one SOM entry containing two atoms by MetaPrint2D-React, when it is more likely that the carbonyl carbon atom is where the reaction is initialised. Inclusion of catalytic mechanism awareness into the SOM annotation process should improve the performance of FamePrint in cases like this. The MACiE database available from the EBI containing enzyme reaction mechanisms could be used for this purpose.²¹¹

In the present work, the study of synergy between CYP3A4 and PGP was severely hampered by the lack of appropriate data. It was originally planned to use the Accelrys metabolite database as a source of CYP3A4 metabolite structures. It was thought that there was a sufficient number of CYP3A4 metabolite structures which would also be covered by the PGP substrate/non-substrate dataset. However, even after the inclusion of PGP data from both stage 1 and stage 2 classifiers, there were only 14 structures which are both a CYP3A4 metabolite and a PGP substrate. A more extensive literature search could be carried out to identify further structures which can be used to study the synergy between the two targets.

From the results of the present work, it is clear that neither the 2D molecular descriptors nor FamePrint fingerprints could produce classifiers that distinguish CYP3A4 substrates from their non-substrates with any significant degree of success. It is also possible that the CYP3A4 dataset is too diverse and that good, predictive models cannot be produced over the entire dataset without employing 3D descriptors or structure-based methods. The CYP3A4 datasets can be clustered by their shapes to account for the different binding modes used by different ligands and separate models (using different descriptors if necessary) can be produced to better differentiate properties unique to each set of substrates. This is currently hampered by the lack of data but is possible in theory.

It would also be interesting to investigate the distribution of logP values and the volumes and shapes of CYP3A4 substrates. It has been suggested that there are multiple access channels leading to the CYP3A4 active site and one is associated with substrate entry from the membrane and the other from solvent (cytoplasm).^{116,212} The membrane access channel shows a larger active site volume. An appropriate description of all CYP3A4 substrates should be found (perhaps using ROCS, 3D molecular descriptors or shaped based descriptors) and the substrate structures clustered. The logP values of all clusters should then be calculated and the spread per cluster examined to see if there is a correlation between the volume/shape of a CYP3A4 substrate structures and their logP values. Also,

if there is enough data, the same analysis should be carried out for CYP3A4 metabolites and PGP substrates and compared with each other (as well as CYP3A4 substrates) to determine if there are any patterns and whether they fit the hypotheses outlined in 7.1.3.

The stage 1 and stage 2 PGP classifiers, if suitably improved, can be integrated into the FamePrint SOM predictor in Coralie. Supervised deep-learning neural network classification could be used to attempt to improve the performance as these methods have been proven to be successful in discovering patterns in high-dimensional data, although more data may be required than that collected for this study.²¹³ Once classifiers have been integrated into Coralie, users will then have the option of gauging whether the compound(s) they are interested in will be transported by PGP and whether their original query compound or compounds created using a bioisosteric replacement fragment are better. Classification models for other transporters could also be produced and integrated in the same fashion. It may be possible to automatically provide the PGP substrate prediction results when generating a new structure using a bioisosteric replacement. This would be displayed on screen alongside the structure stability score when a generated structure is selected.

The performance produced by stage 1 FamePrint fingerprint classifiers for PGP proved to be extremely promising. This can be used to filter out structures to be passed over to the stage 2 classifier. The FamePrint fingerprint classifiers are the preferred option here (rather than the 2D molecular descriptors) as the fingerprint for the newly generated structures would already have been computed for similarity comparison to the original query structure. This will save computational time and effort, reducing the time required to generate a response.

9. Appendices

Appendix A – Fragmentation Tab in Coralie's SOM Module

1) Dataset loaded and structures displayed

The screenshot displays the 'Fragmentation' tab in the Coralie SOM Module. The interface is divided into several sections:

- Top Navigation:** Includes tabs for 'Fragmentation', 'Prediction', 'Analysis', and 'Validation'. Below these is a label 'Fragments (index: fragment stability)'.
- Left Panel:** A large table with 5 columns and 10 rows, currently empty, intended for displaying the loaded dataset.
- Right Panel (Examples):** A section titled 'Examples' showing four chemical structures in a 2x2 grid, each labeled with an index and '<Unknown>':
 - Top-left: A molecule with a central carbon atom bonded to a chlorine atom (Cl), two fluorine atoms (F), and a hydroxyl group (OH). Label: '<0 : Unknown>'.
 - Top-right: A naphthalene derivative with two hydroxyl groups (OH) at the 1 and 2 positions. Label: '<1 : Unknown>'.
 - Bottom-left: A naphthalene derivative with two hydroxyl groups (OH) at the 1 and 2 positions, with a different substitution pattern than the top-right molecule. Label: '<2 : Unknown>'.
 - Bottom-right: A naphthalene derivative with two hydroxyl groups (OH) at the 1 and 2 positions, with a different substitution pattern. Label: '<3 : Unknown>'.
- Fragmentation Parameters Panel (Bottom Right):** Contains several controls:
 - Checkboxes: 'Keep scaffolds' (unchecked), 'Ignore unknown reactions' (unchecked), 'Keep ring systems' (checked), and 'Keep function groups' (checked).
 - Slider: A 'Depth' slider set to 1.
 - Buttons: 'Fragment' and 'Export'.
- Footer:** A navigation bar with links: 'Som', 'Matrix', 'Search', 'org.lhasalimited.coralie.core.editor.example', 'Sohn', 'Fragment'.

2) Fragmentation options panel: fragmentation parameters could be specified & fragmentation process initialised.

3) Fragments created were displayed.

Fragmentation

Prediction

Analysis

Validation

Fragments (index : fragment stability)

200.30	210.59	220.80	230.58
240.61	250.82	260.63	270.72
280.67	290.78	300.74	310.56
320.71	330.71	340.70	350.55
360.82	370.82	380.74	390.79

Fragmentation Parameters

☐ Keep scaffolds
 ☐ Ignore unknown reactions

☒ Keep ring systems
 ☒ Keep function groups

Depth:

Fragment

Export

Selected Fragment

Examples

<0 : Unknown>

<1 : Unknown>

<2 : Unknown>

<3 : Unknown>

Cell format:

fragment index: transformation likelihood (max = 1)

Stable

Unstable

4) Selected fragment of interest

Fragmentation Prediction Analysis Validation
Fragments (index : fragment stability)

200.30	210.39	220.80	230.58
240.61	250.82	260.63	270.72
280.67	290.78	300.24	310.56
320.71	330.71	340.70	350.55
360.82	370.82	380.74	390.37

Selected Fragment

Transformation Likelihood Atoms count

Transformation	Likelihood	Atoms	count
Total	1	0	161
[45] Reduction(=/-)	0.32	(2, 3)	52
[0] Unknown	0.25	(3)	41
None	0.24	0	38
[15] Reduction(-/+)	0.22	(0, 1)	36

Examples

<0 : Unknown>

<1 : Unknown>

<2 : Unknown>

<3 : Unknown>

Fragmentation Parameters

☐ Keep scaffolds ☒ Keep function groups

☐ Ignore unknown reactions

Depth: 1

Fragment

Som | Matrix | Search | org.jhaslimited.coralie.core.editor.example | Sohn | Fragment

5) Metabolic stability information relating to the fragment displayed along with (parent) structures which produced the fragment

Appendix B – Prediction Tab in Coralie's SOM Module

- 1) Double clicked on the Query structure panel to bring up the Edit structure box where a new query structure could be drawn.
(SMILES or MOL file are also accepted)

The screenshot displays the 'Prediction' tab in Coralie's SOM Module. The main interface includes a 'Query' panel on the left, a 'Som' panel on the right, and a 'Ranking Methods' section at the bottom. The 'Query' panel contains a 'Transformation' table with columns for 'Likelihood', 'count', and 'total'. The 'Som' panel shows a chemical structure of a pyridine derivative with a methyl group and a hydroxyl group. The 'Ranking Methods' section lists 'Reaction specific: all atoms', 'Reaction specific: top 3', 'Atom specific: all atoms', and 'Atom specific: top 3'. The 'Edit structure' dialog box is open, showing a chemical structure of a pyridine derivative with a methyl group and a hydroxyl group. The dialog box has 'OK' and 'Cancel' buttons. The 'Query' panel also shows a chemical structure of a pyridine derivative with a methyl group and a hydroxyl group. The 'Som' panel shows a chemical structure of a pyridine derivative with a methyl group and a hydroxyl group. The 'Ranking Methods' section lists 'Reaction specific: all atoms', 'Reaction specific: top 3', 'Atom specific: all atoms', and 'Atom specific: top 3'.

Fragmentation Prediction Analysis Validation

Query

Som

Transformation

Likelihood count total

Ranking Methods

Selecting Ranking Method:

Reaction specific: all atoms

Reaction specific: top 3

Atom specific: all atoms

Atom specific: top 3

Rank

OK Cancel

SMILES or MOL file are also accepted

- 2) Upon submission, the query structure would be fragmented and matching fragments from the dictionary used to calculate the stability of each atom, using the Reaction specific ranking algorithm (selected by default).

The screenshot displays the Som Matrix web application interface. At the top, a navigation bar includes tabs for Fragmentation, Prediction, Analysis, and Validation. The main content area is divided into several sections:

- Query:** A chemical structure of a substituted benzene ring is shown. The atoms are color-coded: red for high instability, orange for medium instability, yellow for low instability, and green for stable. The structure includes a benzene ring with a methyl group (H₃C) and two hydroxyl groups (OH).
- Ranking Methods:** A dropdown menu is open, showing options: "Reaction specific: all atoms", "Reaction specific: top 3", "Atom specific: all atoms", and "Atom specific: top 3".
- Atom highlights:** A legend on the right side of the interface defines the color coding:
 - Red circle: = High instability
 - Orange circle: = Medium instability
 - Yellow circle: = Low instability
 - Green circle: = Stable
 - Grey circle: = Unknown
- Table:** A table with columns for Transformation, Likelihood, count, and total. The table is currently empty.

Arrows indicate the flow of information: one arrow points from the "Atom highlights" legend to the chemical structure, and another points from the "Ranking Methods" dropdown to the chemical structure.

3) To view the top 3 most unstable atoms, select the option from the list of ranking methods and press Rank to show the updated results.

The screenshot shows the SOM software interface. The main window displays a chemical structure of a substituted benzene ring. The structure has a central benzene ring with a dashed green outline. Substituents include a hydroxyl group (OH) at the top, a hydroxyl group (HO) at the bottom left, a methyl group (H₃C) at the bottom right, and a hydroxyl group (OH) at the top right. Three atoms are highlighted: a red circle on the top-left carbon, an orange circle on the top-right carbon, and a yellow circle on the bottom-right carbon. The interface includes a 'Query' tab, a 'Ranking Methods' dropdown menu, and a 'Rank' button. The 'Ranking Methods' dropdown is currently set to 'Reaction specific: all atoms'. The 'Rank' button is highlighted with a blue arrow. The 'Atom highlights' legend is located at the bottom right of the interface.

Atom highlights:

- = Most unstable atom
- = Second most unstable atom
- = Third most unstable atom

- 4) When an atom was selected, its associated transformation records would be displayed. Under the Reaction specific ranking algorithm, the glucuronidation entry was solely responsible for the stability score of the selected oxygen atom as the transformation with the highest likelihood.

Fragmentation
Prediction
Analysis
Validation

Query

Som

Transformation	Likelihood	count	total
[85] Glucuronidation	0.27	30	111
[30] Sulfation	0.15	17	111
[0] Unknown	0.05	6	111
[4] Phosphorylation	0.01	1	111
[65] Methylation	0.01	1	111
[83] Glucosidation(+X)	0	61	29099
[57] Acylation	0	51	29099
[55] Acetylation	0	39	29098
[66] Alkylation	0	2	29099
[91] Glycosidation(+XP)	0	1	29099
[28] Nitrosation	0	1	29099
[18] Hydroxylation	0	1	29099

Ranking Methods
Selecting Ranking Method:

Reaction specific: all atoms

Reaction specific: top 3

Atom specific: all atoms

Atom specific: top 3

Rank

Examples

Som
Matrix
Search
org.hasalimited.coralie.core.editor.example
Sohn
Fragment

5) Transformation of interest selected. The responsible parent structures were displayed.

Fragmentation

Prediction

Analysis

Validation

Query

Som

Transformation	Likelihood	count	total
[85] Glucuronidation	0.27	30	111
[30] Sulfation	0.15	17	111
[0] Unknown	0.05	6	111
[4] Phosphorylation	0.01	1	111
[65] Methylation	0.01	1	111
[83] Glucosidation(+X)	0	61	29099
[57] Acylation	0	51	29099
[55] Acetylation	0	39	29098
[66] Alkylation	0	2	29099
[91] Glycosidation(+XP)	0	1	29099
[28] Nitrosation	0	1	29099
[18] Hydroxylation	0	1	29099

Ranking Methods

Selecting Ranking Method:

Reaction specific: all atoms

Reaction specific: top 3

Atom specific: all atoms

Atom specific: top 3

Rank

Examples

<0 : Unknown>

<1 : Unknown>

<2 : Unknown>

<3 : Unknown>

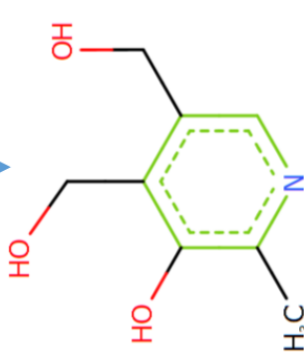
6) Switching over to the Analysis tab allowed for each fragment generated by the query structure to be analysed.

Appendix C – Analysis Tab in Coralie's SOM Module

7) Query structure and its generated fragments displayed in the Analysis tab.

Fragmentation Prediction Analysis Validation





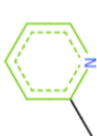
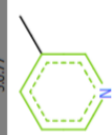

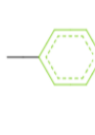

Query



Som

Examples

Fragments

00.82	10.81	20.81
		
30.81	40.78	50.77
		
60.74	70.71	80.74
		

Som Matrix Search org.hassalimited.coralie.core.editor.example Sohn Fragment

Fragmentation

Prediction

Analysis

Validation

Query

Som

Transformation	Likelihood	Atoms	count
Total	1	0	111
None	0.54	0	60
[85] Glucuronidation	0.27	(1)	30
[30] Sulfation	0.15	(1)	17
[18] Hydroxylation	0.05	(2)	6
[0] Unknown	0.05	(1)	6
[0] Unknown	0.05	(3)	6
[0] Unknown	0.05	(2)	5
[0] Unknown	0.05	(6)	5
[0] Unknown	0.05	(0)	5
[19] Hydroxylation	0.03	(6)	3
[86] Glutathionation(+Sx)	0.01	(4)	3
[18] Hydroxylation	0.01	(3)	1
[4] Phosphorylation	0.01	(1)	1
[0] Unknown	0.01	(4)	1
[65] Methylation	0.01	(1)	1
[74] Ring opening	0.01	(5, 6)	1

Fragment

Matrix

Search

org.hasalimited.coralie.core.editor.example

Sohn

Fragment

Examples

<0 : Unknown>	<1 : Unknown>	<2 : Unknown>	<3 : Unknown>

8) Fragment of interest selected and all its associated records and parent structures were displayed.
Substructure corresponding to the fragment of interest highlighted in parent structures.

205

- 9) One of the fragment's transformation record of interest selected and the corresponding atom highlighted in the query structure.
 Examples matrix updated to display parent structures responsible for the selected transformation record.

Fragmentation | Prediction | Analysis | Validation

Query

Som

Transformation	Likelihood	Atoms	count
Total	1	0	111
None	0.54	0	60
[85] Glucuronidation	0.27	(1)	30
[30] Sulfation	0.15	(1)	17
[18] Hydroxylation	0.05	(2)	6
[0] Unknown	0.05	(1)	6
[0] Unknown	0.05	(3)	6
[0] Unknown	0.05	(2)	5
[0] Unknown	0.05	(6)	5
[0] Unknown	0.05	(0)	5
[19] Hydroxylation	0.03	(6)	3
[86] Glutathionation(+Sx)	0.03	(4)	3
[18] Hydroxylation	0.01	(3)	1
[4] Phosphorylation	0.01	(1)	1
[0] Unknown	0.01	(4)	1
[65] Methylation	0.01	(1)	1
[74] Ring opening	0.01	(5, 6)	1

Examples

<0 : Unknown>	<1 : Unknown>	<2 : Unknown>	<3 : Unknown>

Fragments

10.81	20.81	30.81
40.78	50.77	60.74
70.71	80.74	90.54

Som

Matrix

Search

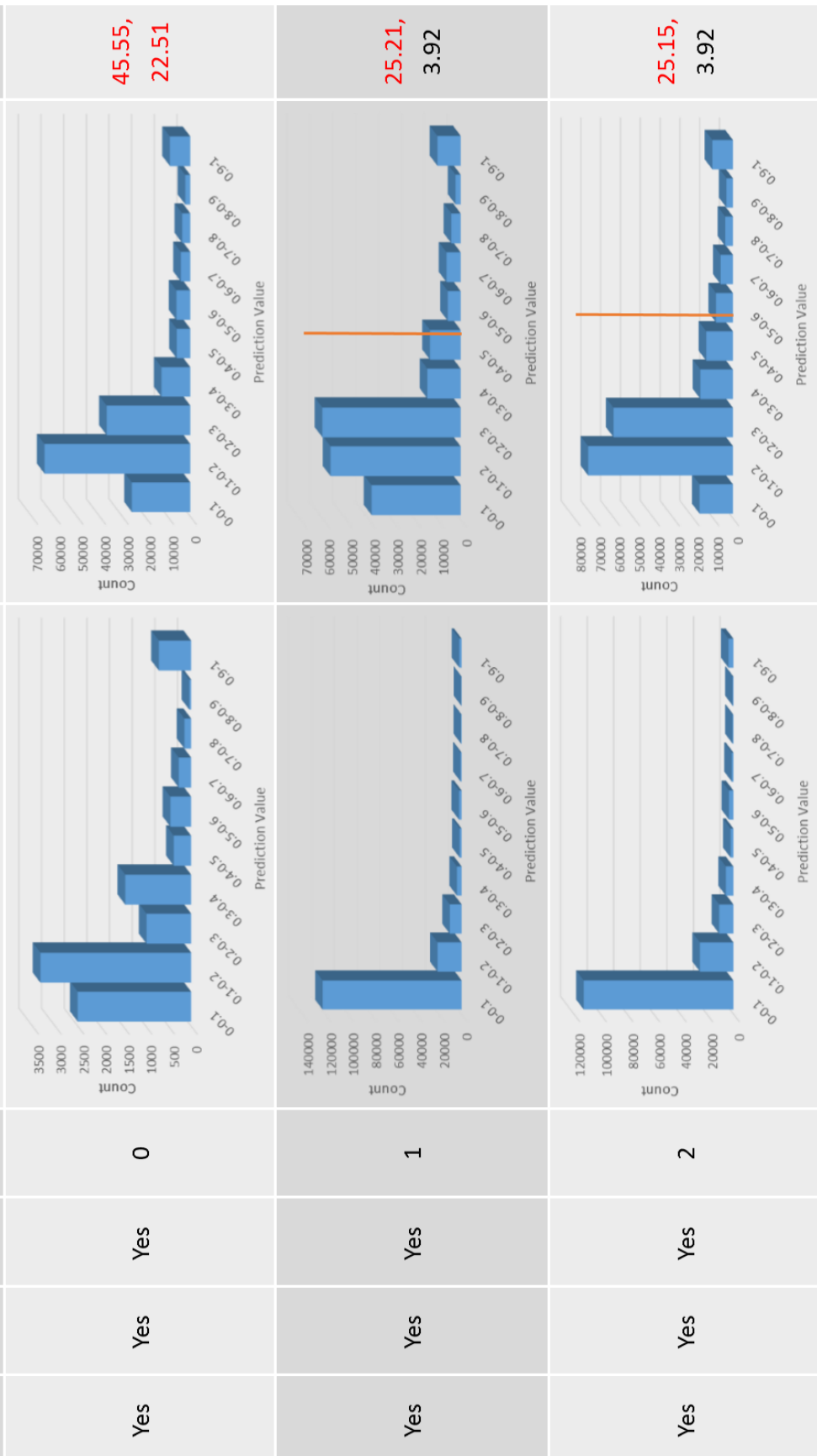
org.hasalimited.coralie.core.editor.example

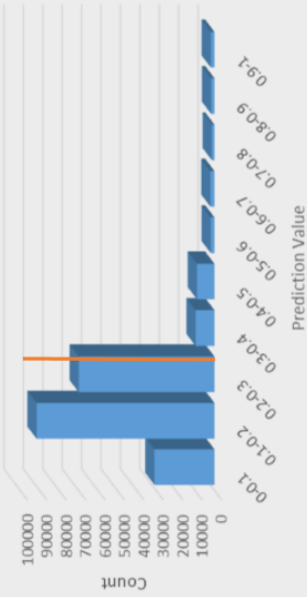
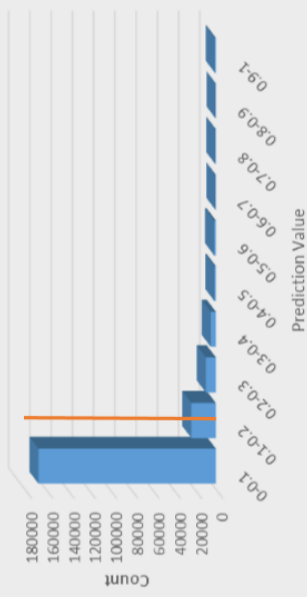
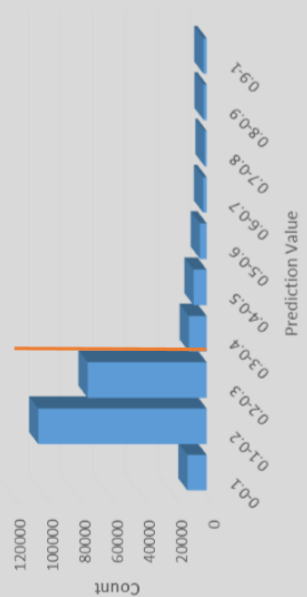
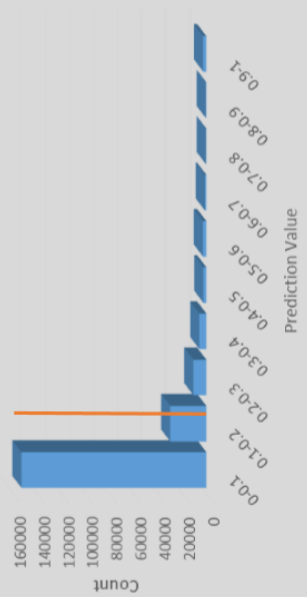
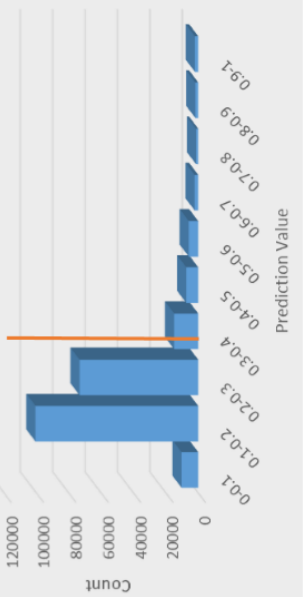
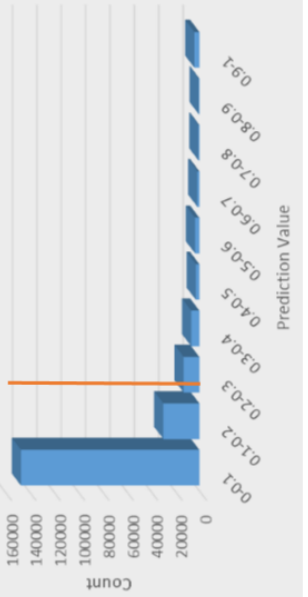
Sohn

Fragment

Appendix D – Frequency Distribution of CASSI Prediction on Test Dataset

RANKING	% Unknown	
	Atom Specific	
	Reaction Specific	
Fragmentation Depth	0	2
Retain Functional Groups	Yes	Yes
Retain Ring Systems	Yes	Yes
Retain Scaffolds	Yes	Yes



RANKING	% Unknown	
	Atom Specific	Reaction Specific
Fragmentation Depth	1	1
Retain Functional Groups	Yes	Yes
Retain Ring Systems	Yes	Yes
Retain Scaffolds	No	No
		
		
		
	5.49, 0.38	5.43, 0.38
	5.43, 0.37	

RANKING	% Unknown	
	Atom Specific	Reaction Specific
Fragmentation Depth	3.04, 0.39	2.98, 0.39
Retain Functional Groups	Yes	Yes
Retain Ring Systems	No	No
Retain Scaffolds	No	No

RANKING	% Unknown	
	Atom Specific	Reaction Specific
Fragmentation Depth	0.07, 0.01	0.03, 0.01
Retain Functional Groups	No	No
Retain Ring Systems	No	No
Retain Scaffolds	No	No

The values in the “% Unknown” column refer to the percentage of atom which produced no prediction result (reaction specific, atom specific ranking).

Appendix E – Validation Tab in Coralie’s SOM Module

1) Locate and select test dataset SD file

Fragmentation | Prediction | Analysis | **Validation**

FAME Validation:

Coralie Validation:

Test Set:

Specify test set file:

Coralie Ranking Methods:

Selecting Ranking Method:

Reaction specific: all atoms

Reaction specific: top 3

Atom specific: all atoms

Atom specific: top 3

Test Set Structures:

Som | Matrix | Search | org.lhasalimited.coralie.core.editor.example | Sohn | Fragment

2) Structures contained in the specified SD file loaded into the matrix. Upon selection...

Fragmentation

Prediction

Analysis

Validation

Coralie Validation:

FAME Validation:

TestSet:

Specify test set file:

C:\Users\admin\Dataset\TestSet.sdf

Browse

Coralie Ranking Methods:

Selecting Ranking Method:

Reaction specific: all atoms

Reaction specific: top 3


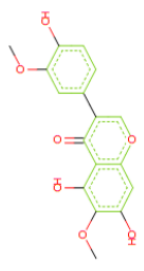
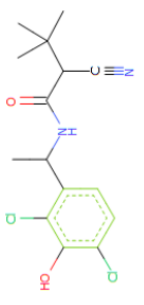
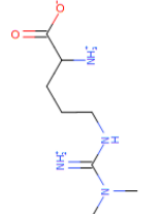
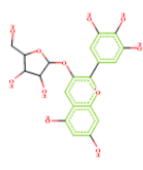
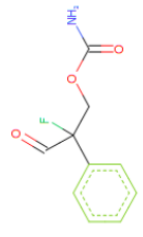
Atom specific: all atoms

Atom specific: top 3

Rank

Batch it:

Test Set Structures:

<div><0 : Unknown></div> 	<div><4 : Unknown></div> 	<div><2 : Unknown></div> 
<div><3 : Unknown></div> 	<div><4 : Unknown></div> 	<div><5 : Unknown></div> 

Som

Matrix

Search

org.lhasalimited.coralie.core.editor.example

Solm

Fragment

- 3) The selected structure would be displayed. Atoms with associated SOM annotations obtained from the Accelrys Metabolite Database were highlighted in purple.

The screenshot shows the FAME Validation software interface. The main window displays a chemical structure of a pyridine derivative with atoms highlighted in purple (9, 11, 4) and yellow (8, 12). The structure is labeled "FAME Validation:". Below the structure, there are tabs for Fragmentation, Prediction, Analysis, and Validation. The Validation tab is active. To the right, there is a "Test Set Structures" section with a list of structures and their corresponding SOM annotations. The structures are labeled <0 : Unknown>, <1 : Unknown>, <2 : Unknown>, <3 : Unknown>, <4 : Unknown>, and <5 : Unknown>. The interface also includes a "Specify test set file" field, a "Browse" button, and a "Coralie Ranking Methods" section with a "Selecting Ranking Method" dropdown menu. The dropdown menu is open, showing options: "Reaction specific: all atoms", "Reaction specific: top 3", "Atom specific: all atoms", and "Atom specific: top 3". The "Rank" and "Batch it!" buttons are visible at the bottom right.

- 4) Atoms in the Coralie Validation structure display were highlighted according to the same rules in the Prediction tab. Atoms in the FAME Validation structure display had the top 3 most unstable atoms according to Fame prediction highlighted in red, amber and yellow in decreasing instability.

Fragmentation | Prediction | Analysis | Validation

Coralie Validation:

FAME Validation:

TestSet:

Specify test set file:
C:\Users\admin\Dataset\TestSet.sdf
Browse

Coralie Ranking Methods:
Selecting Ranking Method:
Reaction specific: all atoms
Atom specific: top 3
Atom specific: top 3

Rank
Batch it!

Test Set Structures:

<0 : Unknown>	<1 : Unknown>	<2 : Unknown>	<3 : Unknown>	<4 : Unknown>	<5 : Unknown>

Som | Matrix | Search | org.lhasalimited.coralie.core.editor.example | Sohn | Fragment

- 4) (cont.) Atoms in the Coralie Validation structure display were highlighted according to the same rules in the Prediction tab. Highlights change according to the ranking method selected.

Fragmentation

Prediction

Analysis

Validation

Coralie Validation:

FAME Validation:

TestSet:

Specify test set file:

C:\Users\admin\Dataset\TestSet.sdf

Browse

Coralie Ranking Methods:

Selecting Ranking Method:

Reaction specific: all atoms

Atom specific: top 3

Atom specific: all atoms

Atom specific: top 3

Rank

Batch It!

Test Set Structures:

<0 : Unknown>	<1 : Unknown>	<2 : Unknown>	<5 : Unknown>
<3 : Unknown>	<4 : Unknown>		

Som

Matrix

Search

org.lhasalimited.coralie.core.editor.example

Sohn

Fragment

5) All test dataset structure could be inspected by selecting the structure of interest in the matrix.

Appendix F – FamePrint Dataset Creation Wizard

1)

Option A: load in pre-calculated, discretised descriptors stored in SD tag in structures contained in the SD file

Descriptors Calculation
Use descriptors present in current file OR select descriptors required. Optional output to .sdf file available.

Calculate/ Load descriptors:

- ☒ Use discretized descriptors from file
- ☐ Load continuous descriptors from file for discretization. (Export/ Overwrite)
 - ☐ Use existing boundaries from file (as SD tag in first structure)
- ☐ Calculate descriptors
 - ☐ Discretize

Select descriptors required:

Descriptor	Specify bin sizes:
<input type="checkbox"/> FAME descriptors	n/a
<input type="checkbox"/> SYBYL atom types	4
<input type="checkbox"/> Effective Atom Polarizability (CDK)	4
<input type="checkbox"/> Partial Sigma Charge (CDK)	4
<input type="checkbox"/> Partial Total Charge - MMFF94 (CDK)	4
<input type="checkbox"/> Pi Electronegativity (CDK)	4
<input type="checkbox"/> Sigma Electronegativity (CDK)	4
<input type="checkbox"/> Maximum Topological Distance (Span2End)	4

Select output destination (optional):
Output file:

Buttons: Export, Calculate Descriptors/ Load from File, < Back, Next >, Finish, Cancel

2)

Descriptors Calculation
Use descriptors present in current file OR select descriptors required. Optional output to .sdf file available.

Calculate/ Load descriptors:

- ☐ Use discretized descriptors from file
- ☒ Load continuous descriptors from file for discretization. (Export/ Overwrite)
 - ☐ Use existing boundaries from file (as SD tag in first structure)
- ☐ Calculate descriptors
 - ☐ Discretize

Select descriptors required:

Descriptor	Specify bin sizes:
<input type="checkbox"/> FAME descriptors	n/a
<input type="checkbox"/> SYBYL atom types	4
<input type="checkbox"/> Effective Atom Polarizability (CDK)	4
<input type="checkbox"/> Partial Sigma Charge (CDK)	4
<input type="checkbox"/> Partial Total Charge - MMFF94 (CDK)	4
<input type="checkbox"/> Pi Electronegativity (CDK)	4
<input type="checkbox"/> Sigma Electronegativity (CDK)	4
<input type="checkbox"/> Maximum Topological Distance (Span2End)	4

Select output destination (optional):
Output file:

Buttons: Export, Calculate Descriptors/ Load from File, < Back, Next >, Finish, Cancel

Option B: load in pre-calculated, continuous descriptors stored in SD tag in structures contained in the SD file.

The continuous descriptors can be discretised and the number of discretisation bins to be used can be specified in the wizard.

3)

Descriptors Calculation
Use descriptors present in current file OR select descriptors required. Optional output to .sdf file available.

Calculate/ Load descriptors:

- ☐ Use discretized descriptors from file
- ☒ Load continuous descriptors from file for discretization. (Export/ Overwrite)
 - ☒ Use existing boundaries from file (as SD tag in first structrue)
- ☐ Calculate descriptors
- ☐ Discretize

Select descriptors required:

Descriptor	Specify bin sizes:
<input type="checkbox"/> FAME descriptors	n/a
<input type="checkbox"/> SYBYL atom types	n/a
<input type="checkbox"/> Effective Atom Polarizability (CDK)	4
<input type="checkbox"/> Partial Sigma Charge (CDK)	4
<input type="checkbox"/> Partial Total Charge - MMFF94 (CDK)	4
<input type="checkbox"/> Pi Electronegativity (CDK)	4
<input type="checkbox"/> Sigma Electronegativity (CDK)	4
<input type="checkbox"/> Maximum Topological Distance (Span2End)	4

Select output destination (optional):
Output file:

Export

Calculate Descriptors/ Load from File

< Back Next > Finish Cancel

Option B: load in pre-calculated, continuous descriptors stored in SD tag in structures contained in the SD file.

A SD file with continuous descriptors a pre-calculated list of discretisation boundaries (stored in SD tag of the first structure in file) can also be loaded in and the boundaries in file used for discretisation. This is used for discretisation of the test data sets (boundaries obtained from training set).

4)

Option C: calculate continuous descriptors of structures in the SD file. The descriptors to be calculated can be selected — all seven FamePrint descriptors or a subset of it.

Descriptors Calculation
Use descriptors present in current file OR select descriptors required. Optional output to .sdf file available.

Calculate/ Load descriptors:

- ☐ Use discretized descriptors from file
- ☐ Load continuous descriptors from file for discretization. (Export/ Overwrite)
 - ☒ Use existing boundaries from file (as SD tag in first structrue)
- ☒ Calculate descriptors
 - ☐ Discretize

Select descriptors required:

Descriptor	Specify bin sizes:
<input checked="" type="checkbox"/> FAME descriptors	n/a
<input checked="" type="checkbox"/> SYBYL atom types	n/a
<input checked="" type="checkbox"/> Effective Atom Polarizability (CDK)	4
<input checked="" type="checkbox"/> Partial Sigma Charge (CDK)	4
<input checked="" type="checkbox"/> Partial Total Charge - MMFF94 (CDK)	4
<input checked="" type="checkbox"/> Pi Electronegativity (CDK)	4
<input checked="" type="checkbox"/> Sigma Electronegativity (CDK)	4
<input checked="" type="checkbox"/> Maximum Topological Distance (Span2End)	4

Select output destination (optional):
Output file:

Export

Calculate Descriptors/ Load from File

< Back Next > Finish Cancel

5)

Option C: calculate continuous descriptors of structures in the SD file. The option to discretise the descriptors being calculated is available and the number of bins to be used for the discretisation can be specified.

Descriptors Calculation
Use descriptors present in current file OR select descriptors required. Optional output to .sdf file available.

Calculate/ Load descriptors:

- ☐ Use discretized descriptors from file
- ☐ Load continuous descriptors from file for discretization. (Export/ Overwrite)
- ☒ Use existing boundaries from file (as SD tag in first structrue)
- ☒ Calculate descriptors
- ☒ Discretize

Select descriptors required:

- ☒ FAME descriptors
- ☒ SYBYL atom types
- ☒ Effective Atom Polarizability (CDK)
- ☒ Partial Sigma Charge (CDK)
- ☒ Partial Total Charge - MMFF94 (CDK)
- ☒ Pi Electronegativity (CDK)
- ☒ Sigma Electronegativity (CDK)
- ☒ Maximum Topological Distance (Span2End)

Specify bin sizes:

Descriptor	Bin Size
FAME descriptors	n/a
SYBYL atom types	4
Effective Atom Polarizability (CDK)	4
Partial Sigma Charge (CDK)	4
Partial Total Charge - MMFF94 (CDK)	4
Pi Electronegativity (CDK)	4
Sigma Electronegativity (CDK)	4
Maximum Topological Distance (Span2End)	4

Select output destination (optional):
Output file:

Export

Calculate Descriptors/ Load from File

< Back Next > Finish Cancel

6)

The option to export the calculated and/ or discretised descriptors to file rather than keeping it in memory for fingerprinting in the next wizard page.

Descriptors Calculation
Use descriptors present in current file OR select descriptors required. Optional output to .sdf file available.

Calculate/ Load descriptors:

- ☐ Use discretized descriptors from file
- ☐ Load continuous descriptors from file for discretization. (Export/ Overwrite)
- ☒ Use existing boundaries from file (as SD tag in first structrue)
- ☒ Calculate descriptors
- ☒ Discretize

Select descriptors required:

- ☒ FAME descriptors
- ☒ SYBYL atom types
- ☒ Effective Atom Polarizability (CDK)
- ☒ Partial Sigma Charge (CDK)
- ☒ Partial Total Charge - MMFF94 (CDK)
- ☒ Pi Electronegativity (CDK)
- ☒ Sigma Electronegativity (CDK)
- ☒ Maximum Topological Distance (Span2End)

Specify bin sizes:

Descriptor	Bin Size
FAME descriptors	n/a
SYBYL atom types	4
Effective Atom Polarizability (CDK)	4
Partial Sigma Charge (CDK)	4
Partial Total Charge - MMFF94 (CDK)	4
Pi Electronegativity (CDK)	4
Sigma Electronegativity (CDK)	4
Maximum Topological Distance (Span2End)	4

Select output destination (optional):
Output file:

C:\Users\admin\coralie-folder\Dataset.sdf

Export

Calculate Descriptors/ Load from File

< Back Next > Finish Cancel

7)

Descriptors Calculation

Use descriptors present in current file OR select descriptors required. Optional output to .sdf file available.

Calculate/ Load descriptors:

☒ Use discretized descriptors from file

☐ Load continuous descriptors from file for discretization. (Export/ Overwrite)

☐ Use existing boundaries from file (as SD tag in first structrue)

☐ Calculate descriptors

☐ Discretize

Select descriptors required:

Select descriptors required:	Specify bin sizes:
<input type="checkbox"/> FAME descriptors	n/a
<input type="checkbox"/> SYBYL atom types	4
<input type="checkbox"/> Effective Atom Polarizability (CDK)	4
<input type="checkbox"/> Partial Sigma Charge (CDK)	4
<input type="checkbox"/> Partial Total Charge - MMFF94 (CDK)	4
<input type="checkbox"/> Pi Electronegativity (CDK)	4
<input type="checkbox"/> Sigma Electronegativity (CDK)	4
<input type="checkbox"/> Maximum Topological Distance (Span2End)	4

Select output destination (optional):

Output file:

C:\Users\admin\coralie-folder\Dataset.sdf

Export

Calculate Descriptors/ Load from File

< Back Next > Finish Cancel

When calculation and/ or discretisation has finished and file exported (if the export option is enabled), the next page will be available.

8)

Coralie Fragmentation can be initiated with a set of user specified fragmentation parameters.

Fragmentation

Specify fragmentation algorithm and parameters. Generate Fingerprints

Select fragmentation algorithm:

☒ Coralie Fragmentation

☐ Keep scaffolds

☒ Keep ring systems

☒ Keep function groups

☒ Include unknown reactions

Fragment Depth:

3

Fingerprint

☐ Fingerprint whole molecule

< Back Next > Finish Cancel

The option to produce whole structure fingerprints is also available. This can be used to produce the fingerprints used to compare the training and test dataset structures to extract structures for test dataset 2 and 3.

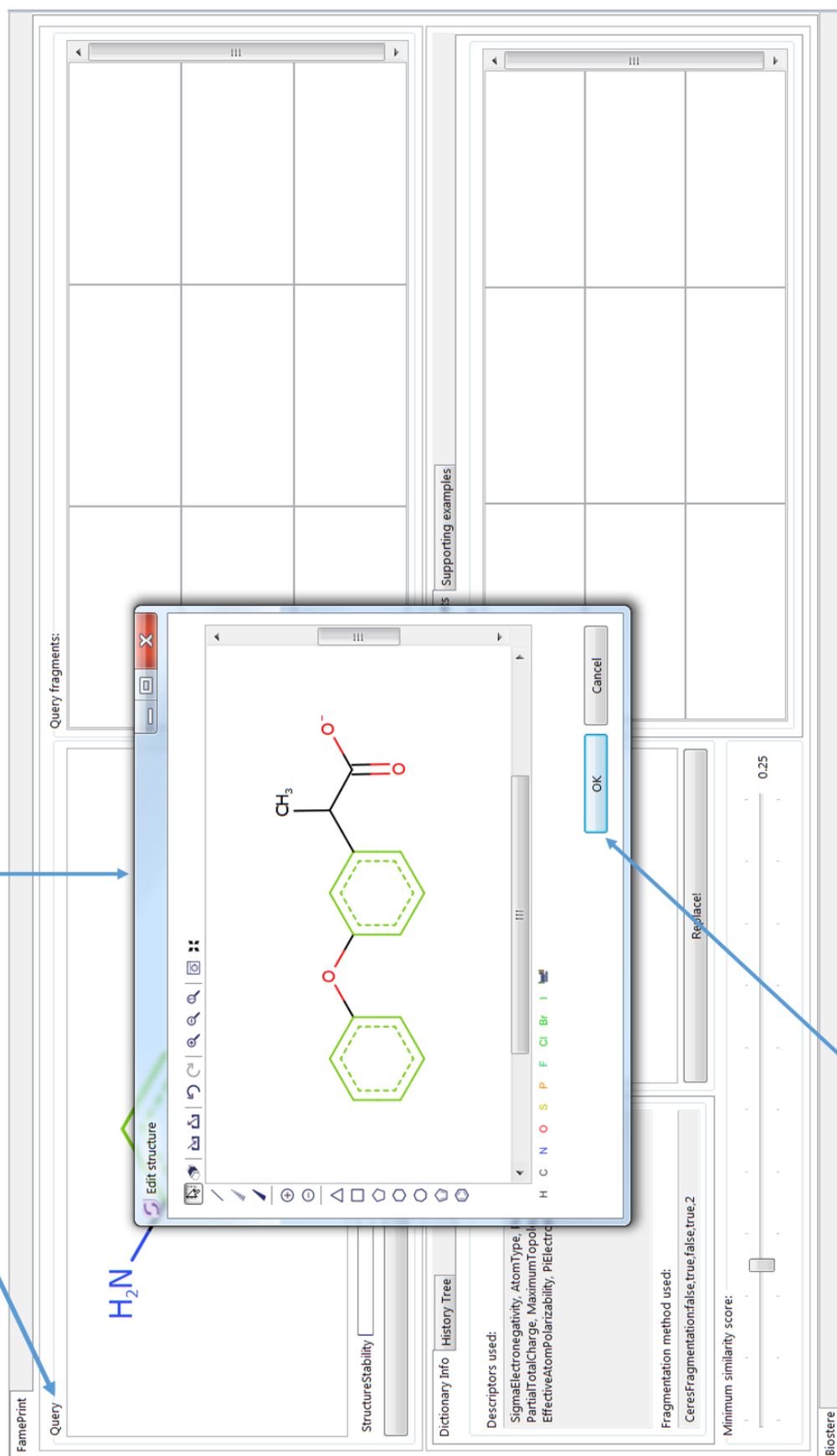
Appendix G – Biostere Tab in Coralie for SOM Prediction

- 1) The Dictionary Info tab contains information regarding the descriptors and fragmentation parameters used to create the dictionary of fragment. These information are displayed once the whole dictionary has been loaded into memory and is ready for use.

The screenshot displays the Biostere software interface. At the top, a chemical structure of a query molecule is shown, consisting of two benzene rings connected by an ether oxygen, with a carboxylic acid group attached to one of the rings. Below the structure is a 'Query' section containing two empty tables: 'Query fragments' and 'Replacement fragments'. To the right of these tables is a 'Dictionary Info' section. This section includes a 'History Tree' tab, a 'Descriptors used' list (SigmaElectronegativity, AtomType, PartialSigmaCharge, PartialTotalCharge, MaximumTopologicalDistance, EffectiveAtomPolarizability, PEElectronegativity), a 'Fragmentation method used' dropdown menu (set to 'CeresFragmentationfalse,true,false,true,2'), and a 'Minimum similarity score' slider (set to 0.25). A blue box highlights the 'Descriptors used' and 'Fragmentation method used' sections, with an arrow pointing to the 'Fragmentation method key' text.

Fragmentation method key:
method name, retain rings, retain functional group, retain scaffold, include unknown transformation, fragmentation depth.

- 2) Double click on the Query box to bring up the structure editor where a new query structure could be drawn. SMILES or MOL structure input are also accepted.

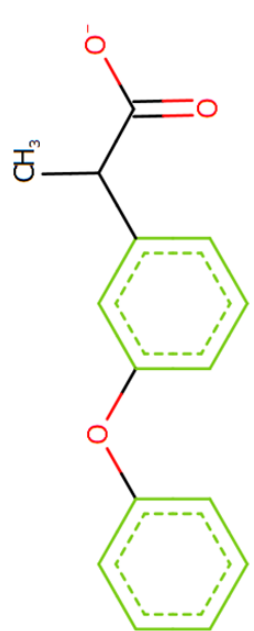


- 3) Upon submission of a new query structure, the metabolic stability prediction process will begin.

- 4) Once metabolic stability prediction on all query fragments has been completed, the fragments will be displayed along with their predicted scores. These are sorted in order of increasing metabolic stability and the cell labels are coloured according to the stability score.

FramePrint

Query



StructureStability

Analysed

Query fragments:

Stability:0.59011	Stability:0.63300	Stability:0.64075
Stability:0.66030	Stability:0.72078	Stability:0.73797
Stability:0.74827	Stability:0.76135	Stability:0.80422

Replacement Fragments

Supporting examples

Matched fragments:

Dictionary Info

History Tree

Descriptors used:

SigmaElectronegativity, AtomType, PartialSigmaCharge, PartialTotalCharge, MaximumTopologicalDistance, EffectiveAtomPolarizability, PElectronegativity,

Fragmentation method used:

CeresFragmentation:false,true,false,true,2

Minimum similarity score:

Replace!

Biostere

Fragment stability score key:

Stable (1.0)

Labile (0.0)

- 5) When a fragment of interest has been selected, corresponding atoms are highlighted. Supporting examples tab displays all parent structures which generated the selected fragment structure. Not all may be used for stability prediction, depending on the similarity calculation used.

The screenshot displays the Biostere software interface. At the top, a query molecule is shown with a highlighted fragment (a benzene ring connected to a carbonyl group). Below the query molecule, a 'StructureStability' bar indicates the overall stability score. To the right, a 'Query fragments' table lists several fragments with their corresponding stability scores:

Fragment Structure	Stability
	Stability: 0.59011
	Stability: 0.63300
	Stability: 0.72078
	Stability: 0.73797
	Stability: 0.74827
	Stability: 0.76135
	Stability: 0.80422

Below the query fragments table, a 'Supporting examples' tab displays four molecules (1, 2, 3, 4) which are used for comparison with the query structure. The 'Dictionary' tab shows the descriptor used for the stability calculation: 'SigmaElectronegativity, AtomType, PartialSigmaCharge, PartialTotalCharge, MaximumTopologicalDistance, EffectiveAtomPolarizability, PLElectronegativity'. The 'Fragmentation method used:' is 'CeresFragmentation: false, true, false, true, 2'. The 'Minimum similarity score:' is set to 0.25. The 'Matched Fragment Connection Points:' section shows a 'Replace!' button. The 'Biostere' logo is visible in the bottom right corner.

- 6) The overall stability of the query structure is also displayed for comparison with new structures once a fragment substitution has been made.

Appendix H – Biostere Tab in Coralie for Bioisosteric Replacement

- 1) When a query fragment has been selected, the search for a replacement will begin. Once found, dictionary fragments with similar fingerprints will be displayed in the Replacement Fragments tab (similarity score displayed in writing, stability information given by cell label colour).

Query fragments:

Stability: 0.59011	Stability: 0.63300	Stability: 0.64075
Stability: 0.66030	Stability: 0.72078	Stability: 0.73797
Stability: 0.74827	Stability: 0.76135	Stability: 0.80422

Replacement Fragments:

Matched fragments:

Stability: 0.73469	Stability: 0.61565	Stability: 0.60131
Stability: 0.57367	Stability: 0.54077	Stability: 0.53840
Stability: 0.53642	Stability: 0.51200	Stability: 0.46955

Fragment stability score key:

Stable (1.0) Labile (0.0)

Dictionary Info | **History Tree** | **Matched Fragment Connection Points:**

Descriptors used:
 SigmaElectronegativity, AtomType, PartialSigmaCharge, PartialTotalCharge, MaximumTopologicalDistance, EffectiveAtomPolarizability, PElectronegativity.

Fragmentation method used:
 CeresFragmentation: false, true, false, true, 2

Minimum similarity score:

Replace:

Biostere

Fragment stability score key:

Stable (1.0)

Labile (0.0)

- 2) The required minimum similarity score (set by user, default 0.25) between the selected query fragment and any dictionary fragments found during search for replacement.

- 3) When a replacement fragment of interest is selected, connection points on the fragment will be highlighted (in orange) on the structure shown in the Matched Fragment Connection Points structure display.

The screenshot shows the Biostere software interface. The main window displays a chemical structure of a substituted benzene ring. A dashed green box highlights a phenyl group. A blue box highlights the 'Matched Fragment Connection Points' section, which shows a list of descriptors and a 'Replace!' button. Below this, a 'Minimum similarity score' slider is set to 0.25. The 'Query fragments' section on the right shows a list of fragments with their stability values. The 'Replacement fragments' section on the left shows a list of fragments with their similarity values. A blue arrow points from the 'Replace!' button to the 'Matched Fragment Connection Points' section.

Query fragments:

Fragment	Stability
	Stability: 0.59011
	Stability: 0.63500
	Stability: 0.73797
	Stability: 0.72078
	Stability: 0.66030
	Stability: 0.74827
	Stability: 0.76135
	Stability: 0.80422

Replacement fragments:

Fragment	Similarity
	similarity: 0.73469
	similarity: 0.61565
	similarity: 0.53840
	similarity: 0.54077
	similarity: 0.51200
	similarity: 0.53642
	similarity: 0.46955

Matched Fragment Connection Points:

Descriptors used:
 SigmaElectronegativity, AtomType, PartialSigmaCharge,
 PartialTotalCharge, MaximumTopologicalDistance,
 EffectiveAtomPolarizability, piElectronegativity,

Fragmentation method used:
 CeresFragmentation: false, true, false, true, 2

Minimum similarity score: 0.25

Replace!

- 4) If the selected replacement fragment has satisfactory connection point, clicking the "Replace!" button will initialise the substitution process, generating new structures from all compatible combinations of connection points between the query structure and replacement fragment.

6) Once a structure generated by replacement has been selected, it will automatically be submitted as the new query structure (structural metabolic stability displayed below). The scoring of its fragments and search for potential replacements can be initialised by clicking the "Analyse!" button.

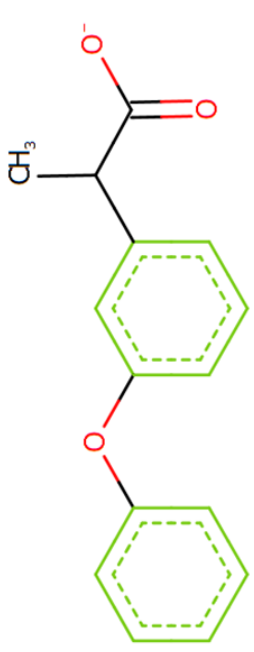
The screenshot displays a chemical structure analysis software interface. The main window is divided into several panels. On the left, the 'Query' tab shows a chemical structure of a substituted benzene ring with a phosphate group. The 'StructureStability' bar is highlighted. Below the structure, the 'Analyse!' button is visible. To the right, the 'Dictionary info' tab shows a list of fragments (M.1, M.2, M.2.1). The 'History Tree' tab shows a tree structure. The 'Matched Fragment Connection Points' section has a 'Replace!' button. The 'Minimum similarity score' slider is set to 0.25. The 'Supporting examples' section shows a table with columns for 'Replacement Fragments' and 'Matched fragments'.

7) The "History Tree" tab shows the generations of structures created from the original query by replacement.

8) Selection of the parent structure in the History Tree brings the original query structure back.

FramePrint

Query



StructureStability

Analyse!

Query fragments:

Dictionary Info History Tree

M.1 M.2 M.2.1

Matched Fragment Connection Points:

Replace!

Minimum similarity score:

0.25

Biostere

Replacement Fragments Supporting examples

Matched fragments:

10. References

- (1) Kirchmair, J., Williamson, M. J., Afzal, A. M., Tyzack, J. D., Choy, A. P. K., Howlett, A., Rydberg, P., and Glen, R. C. (2013) FAsT MEtabolizer (FAME): A Rapid and Accurate Predictor of Sites of Metabolism in Multiple Species by Endogenous Enzymes. *J. Chem. Inf. Model.* 53, 2896–2907.
- (2) Bunnage, M. E. (2011) Getting pharmaceutical R&D back on target. *Nat. Chem. Biol.* 7, 335–339.
- (3) Ng, R. (1994) From Discovery to Approval, in *Drugs: From Discovery to Approval*, pp 1–8. John Wiley & Sons, Inc.
- (4) Ng, R. (2003) Drug Development and Preclinical Studies, in *Drugs: From Discovery to Approval*, pp 25–31. John Wiley & Sons, Inc.
- (5) Ng, R. (1996) Clinical Trials, in *Drugs: From Discovery to Approval*, pp 130–148. John Wiley & Sons, Inc.
- (6) Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., and Schacht, A. L. (2010) How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov* 9, 203–214.
- (7) Coleman, M. (2010) Drug Biotransformational Systems – Origins and Aims, in *Human Drug Metabolism: An Introduction*, pp 13–22. John Wiley & Sons, Ltd.
- (8) Ng, R. (2010) Induction of Cytochrome P450 Systems, in *Human Drug Metabolism: An Introduction*, pp 65–92. John Wiley & Sons, Ltd.
- (9) Kirchmair, J., Williamson, M. J., Tyzack, J. D., Tan, L., Bond, P. J., Bender, A., and Glen, R. C. (2012) Computational Prediction of Metabolism: Sites, Products, SAR, P450 Enzyme Dynamics, and Mechanisms. *J Chem Inf Model* 52, 617–648.
- (10) Afzelius, L., Hasselgren Arnby, C., Broo, A., Carlsson, L., Isaksson, C., Jurva, U., Kjellander, B., Kolmodin, K., Nilsson, K., Raubacher, F., and Weidolf, L. (2007) State-of-the-art Tools for Computational Site of Metabolism Predictions: Comparative Analysis, Mechanistical Insights, and Future Applications. *Drug Metab. Rev.* 39, 61–86.
- (11) Hennemann, M., Friedl, A., Lobell, M., Keldenich, J., Hillisch, A., Clark, T., and Göller, A. H. (2009) CypScore: Quantitative prediction of reactivity toward cytochromes P450 based on semiempirical molecular orbital theory. *ChemMedChem* 4, 657–669.
- (12) Dewar, M. J. S., Zebisch, E. G., Healy, E. F., and Stewart, J. J. P. (1985) AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* 107, 3902–3909.
- (13) ParaSurf. Cepos InSilico.
- (14) MetaSite. Molecular Discovery.
- (15) Cruciani, G., Carosati, E., De Boeck, B., Ethirajulu, K., Mackie, C., Howe, T., and Vianello, R. (2005) MetaSite: Understanding metabolism in human cytochromes from the perspective of the chemist. *J. Med. Chem.* 48, 6970–6979.

- (16) Zamora, I., Afzelius, L., and Cruciani, G. (2003) Predicting drug metabolism: A site of metabolism prediction tool applied to the cytochrome P450 2C9. *J. Med. Chem.* 46, 2313–2324.
- (17) Rydberg, P., Gloriam, D. E., Zaretski, J., Breneman, C., and Olsen, L. (2010) SMARTCyp: A 2D method for prediction of cytochrome P450-mediated drug metabolism. *ACS Med. Chem. Lett.* 1, 96–100.
- (18) StarDrop P450 metabolism. Optibrium.
- (19) Gasteiger, J., Rudolph, C., and Sadowski, J. (1990) Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Comput. Methodol.* 3, 537–547.
- (20) ADMET Predictor - Metabolism Module. Simulations Plus, Inc.
- (21) <http://www.drugbank.ca/about>.
- (22) Zaretski, J., Bergeron, C., Rydberg, P., Huang, T. W., Bennett, K. P., and Breneman, C. M. (2011) RS-Predictor: A new tool for predicting sites of cytochrome P450-mediated metabolism applied to CYP 3A4. *J. Chem. Inf. Model.* 51, 1667–1689.
- (23) Zaretski, J., Bergeron, C., Huang, T. W., Rydberg, P., Joshua Swamidass, S., and Breneman, C. M. (2013) RS-WebPredictor: A server for predicting CYP-mediated sites of metabolism on drug-like molecules. *Bioinformatics* 29, 497–498.
- (24) Adams, S. E. (2010) Molecular Similarity and Xenobiotic Metabolism. University of Cambridge.
- (25) Carlsson, L., Spjuth, O., Adams, S., Glen, R. C., and Boyer, S. (2010) Use of historic metabolic biotransformation data as a means of anticipating metabolic sites using MetaPrint2D and Bioclipse. *BMC Bioinformatics* 11, 362.
- (26) Accelrys Inc. (2008) Accelrys Metabolite Database version 2008.1. San Diego, CA.
- (27) MetabolExpert. CompuDrug.
- (28) Klopman, G., Dimayuga, M., and Talafoos, J. (1994) META. 1. A program for the evaluation of metabolic transformation of chemicals. *J. Chem. Inf. Comput. Sci.* 34, 1320–1325.
- (29) Meteor Nexus. Lhasa Limited.
- (30) Judson, P. N., Marchant, C. A., and Vessey, J. D. (2003) Using Argumentation for Absolute Reasoning about the Potential Toxicity of Chemicals. *J. Chem. Inf. Comput. Sci.* 43, 1364–1370.
- (31) Ridder, L., and Wagener, M. (2008) SyGMa: Combining expert knowledge and empirical scoring in the prediction of metabolites. *ChemMedChem* 3, 821–832.
- (32) Metabolizer. ChemAxon.
- (33) Piechota, P., Cronin, M. T. D., Hewitt, M., and Madden, J. C. (2013) Pragmatic approaches to using computational methods to predict xenobiotic metabolism. *J. Chem. Inf. Model.* 53, 1282–1293.
- (34) Bemis, G. W., and Murcko, M. a. (1996) The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* 39, 2887–93.
- (35) Krissinel, E. B., and Henrick, K. (2004) Common subgraph isomorphism detection by

backtracking search. *Softw. - Pract. Exp.* 34, 591–607.

(36) Tripos. SYBYL Atom Types.

(37) Accelrys Inc. (2011) Accelrys Metabolite Database version 2011.2. San Diego, CA.

(38) (2011) Molecular Operating Environment (MOE). Chemical Computing Group Inc.

(39) Rydberg, P., and Olsen, L. (2012) Predicting Drug Metabolism by Cytochrome P450 2C9: Comparison with the 2D6 and 3A4 Isoforms. *ChemMedChem* 7, 1202–1209.

(40) Long, A., and Rydberg, P. (2013) Enrichment of True Positives from Structural Alerts Through the Use of Novel Atomic Fragment Based Descriptors. *Mol. Inform.* 32, 81–86.

(41) Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, I. H. W. (2009) The WEKA Data Mining Software: An Update. *SIGKDD Explor.* 11, 10–18.

(42) Schneider, G., Neidhart, W., Giller, T., and Schmid, G. (1999) “Scaffold-Hopping” by topological pharmacophore search: A contribution to virtual screening. *Angew. Chemie - Int. Ed.* 38, 2894–2896.

(43) Böhm, H. J., Flohr, A., and Stahl, M. (2004) Scaffold hopping. *Drug Discov. Today Technol.* 1, 217–224.

(44) Sun, H., Tawa, G., and Wallqvist, A. (2011) Classification of scaffold hopping approaches. *Drug Discov. Today* 17, 310–324.

(45) Schuffenhauer, A. (2012) Computational methods for scaffold hopping. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 2, 842–867.

(46) Lima, L. M., and Barreiro, E. J. (2005) Bioisosterism: A Useful Strategy for Molecular Modification and Drug Design. *Curr. Med. Chem.* 12, 23–49.

(47) Langdon, S. R., Ertl, P., and Brown, N. (2010) Bioisosteric Replacement and Scaffold Hopping in Lead Generation and Optimization. *Mol. Inform.* 29, 366–385.

(48) Vainio, M. J., Kogej, T., Raubacher, F., and Sadowski, J. (2013) Scaffold Hopping by Fragment Replacement. *J. Chem. Inf. Model.* 53, 1825–1835.

(49) Devereux, M., and L.A. Popelier, P. (2010) In Silico Techniques for the Identification of Bioisosteric Replacements for Drug Design. *Curr. Top. Med. Chem.* 10, 657–668.

(50) Papadatos, G., and Brown, N. (2013) In silico applications of bioisosterism in contemporary medicinal chemistry practice. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 3, 339–354.

(51) Sheridan, R. P. (2002) The Most Common Chemical Replacements in Drug-Like Compounds. *J. Chem. Inf. Comput. Sci.* 42, 103–108.

(52) Gaulton, A., Bellis, L. J., Bento, a P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., and Overington, J. P. (2012) ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40, D1100–D1107.

(53) Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y., Maclejewski, A., Arndt, D., Wilson, M., Neveu, V., Tang, A., Gabriel, G., Ly, C., Adamjee, S., Dame, Z. T., Han, B., Zhou, Y., and

Wishart, D. S. (2014) DrugBank 4.0: Shedding new light on drug metabolism. *Nucleic Acids Res.* 42, 1091–1097.

(54) Digital Chemistry Ltd. (2015) Bioster 15.1. Sheffield, UK.

(55) Ertl, P. (1998) World Wide Web-based system for the calculation of substituent parameters and substituent similarity searches. *J. Mol. Graph. Model.* 16, 11–13.

(56) Ertl, P. (2003) Cheminformatics analysis of organic substituents: Identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups. *J. Chem. Inf. Comput. Sci.* 43, 374–380.

(57) Holliday, J. D., Jelfs, S. P., Willett, P., and Geddeck, P. (2003) Calculation of Intersubstituent Similarity Using R-Group Descriptors. *J. Chem. Inf. Comput. Sci.* 43, 406–411.

(58) Martin, E. J., Blaney, J. M., Siani, M. A., Spellmeyer, D. C., Wong, A. K., and Moos, W. H. (1995) Measuring diversity: experimental design of combinatorial libraries for drug discovery. *J. Med. Chem.* 38, 1431–1436.

(59) Kier, L. B., and Hall, L. H. (2004) Bioisosterism: quantitation of structure and property effects. *Chem. Biodivers.* 1, 138–51.

(60) Devereux, M., Popelier, P. L. a, and McLay, I. M. (2009) Quantum isostere database: A web-based tool using quantum chemical topology to predict bioisosteric replacements for drug design. *J. Chem. Inf. Model.* 49, 1497–1513.

(61) Birchall, K., Gillet, V. J., Willett, P., Ducrot, P., and Luttmann, C. (2009) Use of reduced graphs to encode bioisosterism for similarity-based virtual screening. *J. Chem. Inf. Model.* 49, 1330–46.

(62) Ujváry, I. (1997) BIOSTER-a database of structurally analogous compounds. *Pestic. Sci.* 51, 92–95.

(63) Wermuth, C. G., Ganellin, C. R., Lindberg, P., and Mitscher, L. A. (2009) Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998). *Pure Appl. Chem.* 70, 1129–1143.

(64) Leach, a R., Green, D. V, Hann, M. M., Judd, D. B., and Good, a C. (2000) Where are the GaPs? A rational approach to monomer acquisition and selection. *J. Chem. Inf. Comput. Sci.* 40, 1262–1269.

(65) Lewell, X. Q., Jones, A. C., Bruce, C. L., Harper, G., Jones, M. M., McLay, I. M., and Bradshaw, J. (2003) Drug rings database with Web interface. A tool for identifying alternative chemical rings in lead discovery programs. *J. Med. Chem.* 46, 3257–3274.

(66) Pearlman, R. S. (1987) Rapid generation of high quality approximate 3D molecular structures. *Chem Des Aut News* 2, 1–6.

(67) Pearlman, R. S. (1993) 3D Molecular Structures: Generation and Use in 3D Searching, in *3D QSAR in Drug Design - Theory Methods and Applications*, pp 41–79.

(68) Stiefl, N., Watson, I. a., Baumann, K., and Zaliani, A. (2006) ErG: 2D Pharmacophore descriptions for scaffold hopping. *J. Chem. Inf. Model.* 46, 208–220.

(69) Wagener, M., and Lommerse, J. P. M. (2006) The quest for bioisosteric replacements. *J. Chem.*

Inf. Model. **46**, 677–85.

(70) Schuffenhauer, A., Gillet, V. J., and Willett, P. (2000) Similarity searching in files of three-dimensional chemical structures: analysis of the BIOSTER database using two-dimensional fingerprints and molecular field descriptors. *J. Chem. Inf. Comput. Sci.* **40**, 295–307.

(71) Fujita, T. (1995) Quantitative Structure—Activity Analysis and Database-Aided Bioisosteric Structural Transformation Procedure as Methodologies of Agrochemical Design, in *Classical and Three-Dimensional QSAR in Agrochemistry* (Hansch, C., and Fujita, T., Eds.), pp 13–34. American Chemical Society, Washington, DC.

(72) Ujváry, I., and Hayward, J. (2012) BIOSTER: A Database of Bioisosteres and Bioanalogus, in *Bioisosteres in Medicinal Chemistry*, pp 55–74. Wiley-VCH.

(73) Floersheim, P., Pombo-Villar, E., and Shapiro, G. (1992) Isosterism and Bioisosterism Case Studies with Muscarinic Agonists. *Chim. Int. J. Chem.* **46**, 323–334.

(74) Watson, P., Willett, P., Gillet, V. J., and Verdonk, M. L. (2001) Calculating the knowledge-based similarity of functional groups using crystallographic data. *J. Comput. Aided. Mol. Des.* **15**, 835–57.

(75) Thormann, M., Klamt, A., Hornig, M., and Almstetter, M. (2006) COSMOsim: Bioisosteric similarity based on COSMO-RS ?? profiles. *J. Chem. Inf. Model.* **46**, 1040–1053.

(76) Stewart, K. D., Shiroda, M., and James, C. a. (2006) Drug Guru: a computer software program for drug design using medicinal chemistry rules. *Bioorg. Med. Chem.* **14**, 7011–22.

(77) Papadatos, G., Bodkin, M. J., Gillet, V. J., and Willett, P. (2012) Mining for Context-Sensitive Bioisosteric replacements in Large Chemical Databases, in *Bioisosteres in Medicinal Chemistry*, pp 103–127. Wiley-VCH.

(78) Hussain, J., and Rea, C. (2010) Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *J. Chem. Inf. Model.* **50**, 339–48.

(79) Papadatos, G., Alkarouri, M., Gillet, V. J., Willett, P., Kadirkamanathan, V., Luscombe, C. N., Bravi, G., Richmond, N. J., Pickett, S. D., Hussain, J., Pritchard, J. M., Cooper, A. W. J., and Macdonald, S. J. F. (2010) Lead optimization using matched molecular pairs: inclusion of contextual information for enhanced prediction of HERG inhibition, solubility, and lipophilicity. *J. Chem. Inf. Model.* **50**, 1872–86.

(80) Wassermann, A. M., and Bajorath, J. (2011) Identification of target family directed bioisosteric replacements. *Medchemcomm* **2**, 601–606.

(81) Wassermann, A. M., and Bajorath, J. (2011) Large-scale exploration of bioisosteric replacements on the basis of matched molecular pairs. *Future Med. Chem.* **3**, 425–436.

(82) Ertl, P., and Schuffenhauer, A. (2009) Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **1**.

(83) Gedeck, P., Rohde, B., and Bartels, C. (2006) QSAR—how good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets. *J. Chem. Inf. Model.* **46**, 1924–36.

- (84) Oprea, T. I. (2002) On the information content of 2D and 3D descriptors for QSAR. *J. Braz. Chem. Soc.* 13, 811–815.
- (85) Accelrys. (2005) CTFile Formats.
- (86) Khalili, F., Henni, A., and East, A. L. L. (2009) pKa values of some Piperazines at (298, 303, 313, and 323) K. *J. Chem. Eng. Data* 54, 2914–2917.
- (87) Kumler, W. D., and Eiler, J. J. (1943) The Acid Strength of Mono and Diesters of Phosphoric Acid. The n-Alkyl Esters from Methyl to Butyl, the Esters of Biological Importance, and the Natural Guanidine Phosphoric Acids. *J. Am. Chem. Soc.* 65, 2355–2361.
- (88) Storer, A. C., and Cornish-Bowden, A. (1976) Concentration of MgATP²⁻ and other ions in solution. Calculation of the true concentrations of species present in mixtures of associating ions. *Biochem. J.* 159, 1–5.
- (89) Lemke, T. L., and Williams, D. A. (2007) Foye's Principles of Medicinal Chemistry. Lippincott Williams & Wilkins, Philadelphia.
- (90) Yamane, N., Tozuka, Z., Kusama, M., Maeda, K., Ikeda, T., and Sugiyama, Y. (2011) Clinical relevance of liquid chromatography tandem mass spectrometry as an analytical method in microdose clinical studies. *Pharm. Res.* 28, 1963–1972.
- (91) Germani, M., Crivori, P., Rocchetti, M., Burton, P. S., Wilson, A. G. E., Smith, M. E., and Poggesi, I. (2007) Evaluation of a basic physiologically based pharmacokinetic model for simulating the first-time-in-animal study. *Eur. J. Pharm. Sci.* 31, 190–201.
- (92) ChemAxon Fragmenter API. ChemAxon.
- (93) OpenEye makefraglib. OpenEye Scientific Software.
- (94) Hanser, T., Barber, C., Rosser, E., Vessey, J. D., Webb, S. J., and Werner, S. (2014) Self organising hypothesis networks: a new approach for representing and structuring SAR knowledge. *J. Cheminform.* 6, 21.
- (95) Mason, S. J., and Graham, N. E. (2002) Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves : Statistical significance and interpretation. *Q. J. R. Meteorol. Soc.* 128, 2145–2166.
- (96) Swets, J. A. (1988) Measuring the accuracy of diagnostic systems. *Science* 240, 1285–1293.
- (97) cms.waikato.ac.nz. weka.filters.unsupervised.attribute.Discretize.
- (98) Halgren, T. a. (1996) Merck molecular force field. {II.} {MMFF94} van der Waals and electrostatic parameters for intermolecular interactions. *J. Comput. Chem.* 17, 520–552.
- (99) Carhart, R. E., Smith, D. H., and Venkataraghavan, R. (1985) Atom Pairs as Molecular Features in Structure-Activity Studies : Definition and Applications. *J Chem Inf Comput Sci* 25, 64–73.
- (100) Willett, P., Barnard, J. M., and Downs, G. M. (1998) Chemical Similarity Searching. *J. Chem. Inf. Model.* 38, 983–996.
- (101) Brown, N. (2012) Classical Bioisosteres, in *Bioisosteres in Medicinal Chemistry*, pp 15–29.

Wiley-VCH.

- (102) Tang, W., Stearns, R. a., Miller, R. R., Ngui, J. S., Mathvink, R. J., Weber, A. E., Kwei, G. Y., Strauss, J. R., Keohane, C. a., Doss, G. a., Chiu, S. H. L., and Baillie, T. a. (2002) Metabolism of a thiazole benzenesulfonamide derivative, a potent and selective agonist of the human β 3-adrenergic receptor, in rats: Identification of a novel isethionic acid conjugate. *Drug Metab. Dispos.* 30, 778–787.
- (103) Stearns, R. A., Miller, R. R., Tang, W., Kwei, G. Y., Tang, F. S., Mathvink, R. J., Naylor, E. M., Chitty, D., Colandrea, V. J., Weber, A. E., Colletti, A. E., Strauss, J. R., Keohane, C. A., Feeney, W. P., Illiff, S. A., and Chiu, S. H. L. (2002) The pharmacokinetics of a thiazole benzenesulfonamide β 3-adrenergic receptor agonist and its analogs in rats, dogs, and monkeys: Improving oral bioavailability. *Drug Metab. Dispos.* 30, 771–777.
- (104) Burgey, C. S., Robinson, K. A., Lyle, T. A., Sanderson, P. E. J., Lewis, S. D., Lucas, B. J., Krueger, J. A., Singh, R., Miller-Stein, C., White, R. B., Wong, B., Lyle, E. A., Williams, P. D., Coburn, C. A., Dorsey, B. D., Barrow, J. C., Stranieri, M. T., Holahan, M. A., Sitko, G. R., Cook, J. J., McMasters, D. R., McDonough, C. M., Sanders, W. M., Wallace, A. A., Clayton, F. C., Bohn, D., Leonard, Y. M., Detwiler, T. J., Lynch, J. J., Yan, Y., Chen, Z., Kuo, L., Gardell, S. J., Shafer, J. A., and Vacca, J. P. (2003) Metabolism-Directed Optimization of 3-Aminopyrazinone Acetamide Thrombin Inhibitors. Development of an Orally Bioavailable Series Containing P1 and P3 Pyridines. *J. Med. Chem.* 46, 461–473.
- (105) Bouska, J. J., Bell, R. L., Goodfellow, C. L., Stewart, A. O., Brooks, C. D. W., and Carter, G. W. (1997) Improving the in vivo duration of 5-lipoxygenase inhibitors: Application of an in vitro glucuronosyltransferase assay. *Drug Metab. Dispos.* 25, 1032–1038.
- (106) Wacher, V., Wu, C., and Benet, L. (1995) Overlapping Substrate Specificities and Tissue Distribution of Cytochrome P450 3A and P-Glycoprotein : Implications for Drug Delivery and Activity in Cancer Chemotherapy MULTIDRUG RESISTANCE AND DRUG METABOLISM 134, 129–134.
- (107) Sugano, K. (2012) Bioavailability, in *Biopharmaceutics Modeling and Simulations: Theory, Practice, Methods, and Applications*, pp 464–465. John Wiley & Sons, Inc.
- (108) Zhou, S., Chan, E., Lim, L. Y., Boelsterili, U. A., Li, S. C., Wang, J., Zhang, Q., Huang, M., and Xu, A. (2004) Therapeutic drugs that behave as mechanism-based inhibitors of cytochrome P450 3A4. *Curr Drug Metab* 5, 415–442.
- (109) Zhang, Q., Dunbar, D., Ostrowska, A., Zeisloft, S., Yang, J., Kaminsky, L. S., and York, N. (1999) Characterization of Human Small Intestinal Cytochromes P-450. *Drug Metab Dispos* 27, 804–809.
- (110) Paine, M. F., Hart, H. L., Ludington, S. S., Haining, R. L., Rettie, A. E., Zeldin, D. C., Carolina, N., Hill, C., Pharmaceutical, B., Virginia, W., Park, T., and Z, N. C. D. C. (2006) The Human Intestinal Cytochrome P450 “ Pie .” *Drug Metab Dispos* 34, 880–886.
- (111) Tachibana, T., Kato, M., and Sugiyama, Y. (2011) Prediction of Nonlinear Intestinal Absorption of CYP3A4 and P-Glycoprotein Substrates from their In Vitro Km Values. *Pharm. Res.* 29, 651–668.
- (112) Kato, M. (2008) Intestinal first-pass metabolism of CYP3A4 substrates. *Drug Metab.*

Pharmacokinet. 23, 87–94.

(113) Thiebaut, F., Tsuruo, T., and Hamada, H. (1987) Cellular localization of the multidrug-resistance gene product P-glycoprotein in normal human tissues. *Proc Natl Acad Sci USA* 84, 7735–7738.

(114) Ortiz de Montellano, P. R. (2005) Cytochrome P450 Structure, Mechanism, and Biochemistry 3rd ed. Springer.

(115) Williams, P. a, Cosme, J., Vinkovic, D. M., Ward, A., Angove, H. C., Day, P. J., Vonnrhein, C., Tickle, I. J., and Jhoti, H. (2004) Crystal structures of human cytochrome P450 3A4 bound to metyrapone and progesterone. *Science* (80-.). 305, 683–686.

(116) Denisov, I. G., Shih, A. Y., and Sligar, S. G. (2011) Structural differences between soluble and membrane bound cytochrome P450s. *J Inorg Biochem* 108, 150–158.

(117) Ekroos, M., and Sjogren, T. (2006) Structural basis for ligand promiscuity in cytochrome P450 3A4. *Proc Natl Acad Sci USA* 103, 13682–13687.

(118) Sun, H., Bessire, A. J., and Vaz, A. (2012) Dirlotapide as a model substrate to refine structure-based drug design strategies on CYP3A4-catalyzed metabolism. *Bioorg Med Chem Lett* 22, 371–376.

(119) Sevrioukova, I., and Poulos, T. (2010) Structure and mechanism of the complex between cytochrome P4503A4 and ritonavir. *Proc Natl Acad Sci USA* 107, 18422–18427.

(120) Atkins, W. M. (2004) Implications of the allosteric kinetics of cytochrome P450s. *Drug Discov Today* 9, 478–484.

(121) Hutzler, J. M., and Tracy, T. S. (2002) Atypical kinetic profiles in drug metabolism reactions. *Drug Metab Dispos* 30, 355–362.

(122) Denisov, I. G., Baas, B. J., Grinkova, Y. V., and Sligar, S. G. (2007) Cooperativity in cytochrome P450 3A4: linkages in substrate binding, spin state, uncoupling, and product formation. *J Biol Chem* 282, 7066–7076.

(123) Roberts, A., and Atkins, W. M. (2007) Energetics of heterotropic cooperativity between α -naphthoflavone and testosterone binding to CYP3A4. *Arch Biochem Biophys* 463, 89–101.

(124) Roberts, A. G., Yang, J., Halpert, J. R., Nelson, S. D., Thummel, K. T., and Atkins, W. M. (2012) The Structural Basis for Homotropic and Heterotropic Cooperativity of Midazolam Metabolism by Human Cytochrome P450 3A4. *Biochemistry* 50, 10804–10818.

(125) Korzekwa, K. R., Krishnamachary, N., Shou, M., Ogai, A., Parise, R. A., Rettie, A. E., Gonzalez, F. J., and Tracy, T. S. (1998) Evaluation of atypical cytochrome P450 kinetics with two-substrate models: evidence that multiple substrates can simultaneously bind to cytochrome P450 active sites. *Biochemistry* 37, 4137–4147.

(126) Ekins, S., Ring, B. J., Binkley, S. N., Hall, S. D., and Wrighton, S. A. (1998) Autoactivation and activation of cytochrome P450s. *Int. J. Clin. Pharmacol. Ther.* 36, 642–651.

(127) Wang, R., Newton, D., Liu, N., Atkins, W., and Lu, A. (2000) Human cytochrome P-450 3A4: in vitro drug-drug interaction patterns are substrate-dependent. *Drug Metab Dispos* 28, 360–366.

- (128) Domanski, T. L., He, Y. A., Khan, K. K., Roussel, F., Wang, Q., and Halpert, J. R. (2001) Phenylalanine and tryptophan scanning mutagenesis of CYP3A4 substrate recognition site residues and effect on substrate oxidation and cooperativity. *Biochemistry* 40, 10150–10160.
- (129) Hosea, N. A., Miller, G. P., and Guengerich, F. P. (2000) Elucidation of Distinct Ligand Binding Sites for Cytochrome P450 3A4. *Biochemistry* 39, 5929–5939.
- (130) Shou, M., and et.al. (1994) Activation of CYP3A4: evidence for the simultaneous binding of two substrates in a cytochrome P450 active site. *Biochemistry* 33, 6450–6455.
- (131) Shou, M., Dai, R., Cui, D., Korzekwa, K. R., Baillie, T. a, and Rushmore, T. H. (2001) A kinetic model for the metabolic interaction of two substrates at the active site of cytochrome P450 3A4. *J Biol Chem* 276, 2256–2262.
- (132) Schrag, M. L., and Wienkers, L. C. (2001) Covalent Alteration of the CYP3A4 Active Site: Evidence for Multiple Substrate Binding Domains. *Arch Biochem Biophys* 391, 49–55.
- (133) Domanski, T. L., Liu, J., Harlow, G. R., and Halpert, J. R. (1998) Analysis of Four Residues within Substrate Recognition Site 4 of Human Cytochrome P450 3A4 : Role in Steroid Hydroxylase Activity and a -Naphthoflavone Stimulation. *Arch Biochem Biophys* 350, 223–232.
- (134) Ekins, S., Bravi, G., Wikel, J. H., and Wrighton, S. A. (1999) Three-Dimensional-Quantitative Structure Activity Relationship Analysis of Cytochrome P-450 3A4 Substrates. *J Pharmacol Exp Ther* 291, 424–433.
- (135) Koley, A. P., Robinson, R. C., Markowitzt, A., and Friedman, F. K. (1997) Drug-drug interactions: effect of quinidine on nifedipine binding to human cytochrome P450 3A4. *Biochem Pharmacol* 53, 455–460.
- (136) Atkins, W., Wang, R., and Lu, A. (2001) Allosteric behavior in cytochrome P450-dependent in vitro drug–drug interactions: A prospective based on conformational dyamics. *Chem Res Toxicol* 14, 338–347.
- (137) Xue, L., Wang, H. F., Wang, Q., Szklarz, G. D., Domanski, T. L., Halpert, J. R., and Correia, M. A. (2001) Influence of P450 3A4 SRS-2 Residues on Cooperativity and / or Regioselectivity of Aflatoxin B 1 Oxidation. *Chem Res Toxicol* 14, 483–491.
- (138) Domanski, T., He, Y., Harlow, G., and Halpert, J. (2000) Dual Role of Human Cytochrome P450 3A4 Residue Phe-304 in Substrate Specificity and Cooperativity. *J Pharmacol Exp Ther* 293, 585–591.
- (139) Ekins, S., Stresser, D. M., and Andrew Williams, J. (2003) In vitro and pharmacophore insights into CYP3A enzymes. *Trends Pharmacol Sci* 24, 161–166.
- (140) Lewis, D. F., Eddershaw, P. J., Goldfarb, P. S., and Tarbit, M. H. (1996) Molecular modelling of CYP3A4 from an alignment with CYP102: Identification of key interactions between putative active site residues and CYP3A-specific chemicals. *Xenobiotic* 26, 1067–1086.
- (141) Tomlinson, E., Lewis, D., Maggs, J., Kroemer, H., Park, B., and Back, D. (1997) In vitro metabolism of dexamethasone (DEX) in human liver and kidney: The involvement of CYP3A4 and CYP17 (17, 20 LYASE) and molecular modelling studies. *Biochem Pharmacol* 54, 605–611.

- (142) Teixeira, V. H., Ribeiro, V., and Martel, P. J. (2010) Analysis of binding modes of ligands to multiple conformations of CYP3A4. *Biochim Biophys Acta* 1804, 2036–2045.
- (143) Arimoto, R., Prasad, M.-A., and Gifford, E. M. (2005) Development of CYP3A4 inhibition models: comparisons of machine-learning techniques and molecular descriptors. *J Biomol Screen* 10, 197–205.
- (144) Vigié F. (1998) <http://AtlasGeneticsOncology.org/Genes/PGY1ID105.h>. *Atlas Genet Cytogenet Oncol Haematol*.
- (145) Aller, S. G. (2009) <http://www.pdb.org/pdb/explore/explore.do?structureId=3G5U>.
- (146) Rees, D. C., Johnson, E., and Lewinson, O. (2009) ABC transporters: the power to change. *Nat Rev Mol Cell Biol* 10, 218–227.
- (147) Jardetzky, O. (1966) Simple allosteric model for membrane pumps. *Nature* 211, 969–970.
- (148) Locher, K. P., Lee, A. T., and Rees, D. C. (2002) The E. coli BtuCD structure: a framework for ABC transporter architecture and mechanism. *Science* (80-.). 296, 1091–1098.
- (149) Dawson, R. J. P., Hollenstein, K., and Locher, K. P. (2007) Uptake or extrusion: crystal structures of full ABC transporters suggest a common mechanism. *Mol Microbiol* 65, 250–257.
- (150) Pawagi, A. B., Wang, J., Silverman, M., Reithmeier, R., and Deber, C. (1994) Transmembrane aromatic amino acid distribution in P-glycoprotein. A functional role in broad substrate specificity. *J Mol Biol* 235, 554–564.
- (151) Loo, T. W., Bartlett, M. C., and Clarke, D. M. (2003) Substrate-induced conformational changes in the transmembrane segments of human P-glycoprotein. Direct evidence for the substrate-induced fit mechanism for drug binding. *J Biol Chem* 278, 13603–13606.
- (152) Lugo, M., and Sharom, F. (2005) Interaction of LDS-751 with P-glycoprotein and mapping of the location of the R drug binding site. *Biochemistry* 44, 643–655.
- (153) Stein, W. (1997) Kinetics of the multidrug transporter (P-glycoprotein) and its reversal. *Physiol Rev* 77, 545–590.
- (154) Ambudkar, S. V., Kimchi-Sarfaty, C., Sauna, Z. E., and Gottesman, M. M. (2003) P-glycoprotein: from genomics to mechanism. *Oncogene* 22, 7468–7485.
- (155) Qu, Q., Chu, J. W. K., and Sharom, F. J. (2003) Transition state P-glycoprotein binds drugs and modulators with unchanged affinity, suggesting a concerted transport mechanism. *Biochemistry* 42, 1345–1353.
- (156) Higgins, C. F., and Linton, K. J. (2004) The ATP switch model for ABC transporters. *Nat Struct Mol Biol* 11, 918–926.
- (157) Ferreira, R. J., Ferreira, M. U., and Santos, D. J. V. A. (2012) Insights on P-Glycoprotein's Efflux Mechanism Obtained by Molecular Dynamics Simulations. *J Chem Theory Comput* 8, 1853–1864.
- (158) Sarkadi, B., Homolya, L., Szakacs, G., and Varadi, A. (2006) Human Multidrug Resistance ABCB and ABCG Transporters : Participation in a Chemoimmunity Defense System. *Physiol Rev* 86, 1179–

1236.

(159) Zheleznova, E., Markham, P., Edgar, R., Bibi, E., Neyfkh, A., and Brennan, R. (2000) A structure-based mechanism for drug binding by multidrug transporters. *Trends Biochem Sci* 25, 39–43.

(160) Seelig, A. (1998) A general pattern for substrate recognition by P-glycoprotein. *Eur J Biochem* 251, 252–261.

(161) Ecker, G., Huber, M., Schmid, D., and Chiba, P. (1999) The importance of a nitrogen atom in modulators of multidrug resistance. *Mol Pharmacol* 56, 791–796.

(162) Xue, Y., Yap, C. W., Sun, L. Z., Cao, Z. W., Wang, J. F., and Chen, Y. Z. (2004) Prediction of P-glycoprotein substrates by a support vector machine approach. *J Chem Inf Comput Sci* 44, 1497–1505.

(163) Wang, Z., Chen, Y., Liang, H., Bender, A., Glen, R. C., and Yan, A. (2011) P-glycoprotein substrate models using support vector machines based on a comprehensive data set. *J Chem Inf Model* 51, 1447–1456.

(164) Molecular Networks GmbH: Erlangen, G. (2011) ADRIANA.Code.

(165) Cianchetta, G., Singleton, R. W., Zhang, M., Wildgoose, M., Giesing, D., Fravolini, A., Cruciani, G., and Vaz, R. J. (2005) A pharmacophore hypothesis for P-glycoprotein substrate recognition using GRIND-based 3D-QSAR. *J Med Chem* 48, 2927–2935.

(166) Pajeva, I. K., and Wiese, M. (2009) Structure-activity relationships of tariquidar analogs as multidrug resistance modulators. *AAPS* 11, 435–444.

(167) Gadhe, C. G., Madhavan, T., Kothandan, G., Lee, T.-B., Lee, K., and Cho, S.-J. (2011) Various Partial Charge Schemes on 3D-QSAR Models for P-gp Inhibiting Adamantyl Derivatives. *Bull Korean Chem Soc* 32, 1604–1612.

(168) Gadhe, C. G., Madhavan, T., Kothandan, G., and Cho, S. J. (2011) In silico quantitative structure-activity relationship studies on P-gp modulators of tetrahydroisoquinoline-ethyl-phenylamine series. *BMC Struct Biol* 11:5.

(169) Raub, T. J. (2006) P-glycoprotein recognition of substrates and circumvention through rational drug design. *Mol Pharm* 3, 3–25.

(170) Pleban, K., and Ecker, G. (2005) Inhibitors of P-Glycoprotein - lead identification and optimisation. *Mini Rev Med Chem* 5, 153–163.

(171) Ecker, G. (2010) QSAR Studies on ABC Transporter - How to Deal with Polyspecificity, in *Transporters as Drug Carriers: Structure, Function, Substrates*, pp 195–214. Wiley-VCH.

(172) Wachter, V. J., Wu, C. Y., and Benet, L. Z. (1995) Overlapping substrate specificities and tissue distribution of cytochrome P450 3A and P-glycoprotein: implications for drug delivery and activity in cancer chemotherapy. *Mol Carcinog* 13, 129–134.

(173) Urquhart, B., Tirona, R., and Kim, R. (2007) Nuclear Receptors and the Regulation of Drug-Metabolizing Enzymes and Drug Transporters: Implications for Interindividual Variability in Response

to Drugs. *J Clin Pharmacol* 47, 566–578.

(174) Waterschoot, R. A. B. Van, and Schinkel, A. H. (2011) A Critical Analysis of the Interplay between Cytochrome P450 3A and P-Glycoprotein : Recent Insights from Knockout and Transgenic Mice. *Pharmacol Rev* 63, 390–410.

(175) Christians, U., Schmitz, V., and Haschke, M. (2005) Functional interactions between P-glycoprotein and CYP3A4 in drug metabolism. *Expert Opin Drug Metab Toxicol* 1, 641–654.

(176) Watkins, P. (1997) The barrier function of CYP3A4 and P-glycoprotein in the small bowel. *Adv Drug Deliv Rev* 27, 161–170.

(177) Sheridan, R. P., Miller, M. D., Underwood, D. J., and Kearsley, S. K. (1996) Chemical Similarity Using Geometric Atom Pair Descriptors. *J. Chem. Inf. Comput. Sci.* 36, 128–136.

(178) Nikolova, N., and Jaworska, J. (2003) Approaches to Measure Chemical Similarity– a Review. *QSAR Comb Sci* 22, 1006–1026.

(179) Jaworska, J., and Nikolova, N. (2004) Review of methods for assessing the applicability domains of SARS and QSARS. Paper 4: SAR applicability Domain.

(180) Bender, A., Jenkins, J. L., Scheiber, J., Sukuru, S. C. K., Glick, M., and Davies, J. W. (2009) How similar are similarity searching methods? A principal component analysis of molecular descriptor space. *J Chem Inf Model* 49, 108–119.

(181) Peironcelly, J. E., Reijmers, T., Coulier, L., Bender, A., and Hankemeier, T. (2011) Understanding and classifying metabolite space and metabolite-likeness. *PLoS One* 6, e28966.

(182) Rogers, D., and Hahn, M. (2010) Extended-connectivity fingerprints. *J Chem Inf Model* 50, 742–754.

(183) Rogers, D. J., and Tanimoto, T. T. (1960) A Computer Program for Classifying Plants. *Science* (80-.). 132, 1115–1118.

(184) Accelrys Inc. (2010) Accelrys Pipeline Pilot. San Diego, CA.

(185) Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., Djoumbou, Y., Eisner, R., Guo, A. C., and Wishart, D. S. (2011) DrugBank 3.0: a comprehensive resource for “omics” research on drugs. *Nucleic Acids Res* 39, D1035–D1041.

(186) Preissner, S., Kroll, K., Dunkel, M., Senger, C., Goldsobel, G., Kuzman, D., Guenther, S., Winnenburg, R., Schroeder, M., and Preissner, R. (2010) SuperCYP: a comprehensive database on Cytochrome P450 enzymes including a tool for analysis of CYP-drug interactions. *Nucleic Acids Res* 38, D237–D243.

(187) Sugiyama, Y., Kusahara, H., and Maeda, K. <http://125.206.112.67/tp-search/>.

(188) Lowe, D. M., Corbett, P. T., Murray-Rust, P., and Glen, R. C. (2011) Chemical name to structure: OPSIN, an open source solution. *J Chem Inf Model* 51, 739–753.

(189) <http://cactus.nci.nih.gov/chemical/structure>.

(190) http://www.daylight.com/daycgi_tutorials/depict.cgi.

- (191) Yap, C. W., and Chen, Y. Z. (2005) Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. *J. Chem. Inf. Model.* 45, 982–92.
- (192) Molnar, L., and Keseru, G. M. (2002) A neural network based virtual screening of cytochrome P450 3A4 inhibitors. *Bioorg. Med. Chem. Lett.* 12, 419–21.
- (193) Mak, L., Marcus, D., Howlett, A., Yarova, G., Duchateau, G., Klaffke, W., Bender, A., and Glen, R. C. (2015) Metrabase: a cheminformatics and bioinformatics database for small molecule transporter data analysis and (Q)SAR modeling. *J. Cheminform.* 7, 31.
- (194) Young, D., Martin, T., Venkatapathy, R., and Harten, P. (2008) Are the Chemical Structures in Your QSAR Correct? *QSAR Comb Sci* 27, 1337–1345.
- (195) Kramer, C., Kalliokoski, T., Gedeck, P., and Vulpetti, A. (2012) The Experimental Uncertainty of Heterogeneous Public K(i) Data. *J Med Chem* 55, 5465–5173.
- (196) Tiikkainen, P., and Franke, L. (2011) Analysis of Commercial and Public Bioactivity Databases. *J Chem Inf Model* 52, 319–326.
- (197) Tropsha, A. (2010) Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol Inf* 29, 476–488.
- (198) ChemAxon. <http://www.chemaxon.com>.
- (199) Wildman, S. a., and Crippen, G. M. (1999) Prediction of Physicochemical Parameters by Atomic Contributions. *J Chem Inf Comput Sci* 39, 868–873.
- (200) Hall, L., and Kier, L. (2007) The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling, in *Reviews in Computational Chemistry* (Lipkowitz, K., and Boyd, D., Eds.), pp 367–422. John Wiley & Sons, Inc, Hoboken, NJ, USA.
- (201) Balaban, A. T. (1979) Chemical Graphs: Five new topological indices for the branching of tree-like graphs. *Theor Chem Acc* 53, 355–375.
- (202) Balaban, A. T. (1982) Highly discriminating distance-based topological index. *Chem Phys Lett* 89, 399–404.
- (203) Petitjean, M. (1992) Applications of the radius-diameter diagram to the classification of topological and geometrical shapes of chemical compounds. *J Chem Inf Comput Sci* 32, 331–337.
- (204) Wiener, H. (1947) Structural determination of paraffin boiling points. *J Am Chem Soc* 69, 17–20.
- (205) ChemComp. <http://www.chemcomp.com/journal/descr.htm>.
- (206) Chang, C., and Lin, C. (2011) LIBSVM: a library for support vector machines. *ACM TIST* 2, 1–39.
- (207) Breiman, L. (2001) Random forests. *Mach. Learn.* 45, 5–32.
- (208) Stiefl, N., and Baumann, K. (2003) Mapping property distributions of molecular surfaces: Algorithm and evaluation of a novel 3D quantitative structure - Activity relationship technique. *J. Med. Chem.* 46, 1390–1407.

- (209) Bonnet, P. (2012) Is chemical synthetic accessibility computationally predictable for drug and lead-like molecules? A comparative assessment between medicinal and computational chemists. *Eur. J. Med. Chem.* 54, 679–689.
- (210) SYLVIA - Estimation of the Synthetic Accessibility of Organic Compounds. *Molecular Networks*.
- (211) Holliday, G. L., Andreini, C., Fischer, J. D., Rahman, S. A., Almonacid, D. E., Williams, S. T., and Pearson, W. R. (2012) MACiE: Exploring the diversity of biochemical reactions. *Nucleic Acids Res.* 40, 783–789.
- (212) Krishnamoorthy, N., Gajendrarao, P., Thangapandian, S., Lee, Y., and Lee, K. W. (2010) Probing possible egress channels for multiple ligands in human CYP3A4: a molecular modeling study. *J Mol Model* 16, 607–614.
- (213) LeCun, Y., Bengio, Y., and Hinton, G. (2015) Deep learning. *Nature* 521, 436–444.