

Molecular Similarity and Xenobiotic Metabolism

Samuel Edward Adams



Trinity College

University of Cambridge

This dissertation is submitted for the degree of Doctor of Philosophy

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.

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

Copyright © 2010 Samuel Edward Adams

This work is licensed under a **Creative Commons Attribution-Share Alike 2.0 UK: England & Wales License**.

This means that you are free:

- to copy, distribute, display, and perform the work
- to make derivative works

Under the following condition:

- **Attribution.** You must give the original author credit.
- **Share Alike.** If you alter, transform, or build upon this work, you may distribute the resulting work only under a licence identical to this one.

For any reuse or distribution, you must make clear to others the licence terms of this work. Any of the above conditions can be waived if you get permission from the copyright holder. Nothing in this license impairs or restricts the author's moral rights.

To view the full text of this license, visit <http://www.creativecommons.org>; or, send a letter to Creative Commons, 171 2nd Street, Suite 300, San Francisco, California, 94105, USA.

Summary

Molecular Similarity and Xenobiotic Metabolism

Samuel Edward Adams

MetaPrint2D, a new software tool implementing a data-mining approach for predicting sites of xenobiotic metabolism has been developed. The algorithm is based on a statistical analysis of the occurrences of atom centred circular fingerprints in both substrates and metabolites. This approach has undergone extensive evaluation and been shown to be of comparable accuracy to current best-in-class tools, but is able to make much faster predictions, for the first time enabling chemists to explore the effects of structural modifications on a compound's metabolism in a highly responsive and interactive manner.

MetaPrint2D is able to assign a confidence score to the predictions it generates, based on the availability of relevant data and the degree to which a compound is modelled by the algorithm.

In the course of the evaluation of MetaPrint2D a novel metric for assessing the performance of site of metabolism predictions has been introduced. This overcomes the bias introduced by molecule size and the number of sites of metabolism inherent to the most commonly reported metrics used to evaluate site of metabolism predictions.

This data mining approach to site of metabolism prediction has been augmented by a set of reaction type definitions to produce MetaPrint2D-React, enabling prediction of the types of transformations a compound is likely to undergo and the metabolites that are formed. This approach has been evaluated against both historical data and metabolic schemes reported in a number of recently published studies. Results suggest that the ability of this method to predict metabolic transformations is highly dependent on the relevance of the training set data to the query compounds.

MetaPrint2D has been released as an open source software library, and both MetaPrint2D and MetaPrint2D-React are available for chemists to use through the Unilever Centre for Molecular Science Informatics' website.

Acknowledgements

Firstly I would like to thank my supervisor Professor Robert Glen for giving me the opportunity to undertake these studies, and for all of his help and support throughout the course of my research.

My thanks go to Dr Scott Boyer and the members of his Computational Toxicology group at AstraZeneca, Mölndal, for their welcome and the help they have given me, in particular Lars Carlsson. I would also like to thank Ola Spjuth of Uppsala University for his assistance in working with Bioclipse.

I am grateful to all the members of the Unilever Centre for Molecular Science Informatics for making my time there so interesting and enjoyable. Particular thanks have to go to Charlotte and Phil for keeping the computers working and to Susan and Emma for keeping the centre running!

Finally, I would like to express my gratitude to those who have supported me and borne with me during the writing of this thesis.

This work was funded by Boehringer Ingelheim and Unilever.

Contents

Preface.....	i
Summary.....	ii
Acknowledgements	iii
Contents	iv
1. Introduction	1
1.1 The drug discovery process	2
1.2 The role of computational methods.....	6
1.3 Virtual screening methods	8
1.4 Current challenges and developments.....	28
2. Prediction of xenobiotic metabolism	41
2.1 Introduction.....	41
2.2 Effects of metabolism.....	43
2.3 Mechanisms of metabolism	50
2.4 Predicting xenobiotic metabolism.....	56
2.5 Conclusion	64
3. Development of MetaPrint2D: a tool for predicting sites of xenobiotic metabolism ..	65
3.1 Substrate/Product Occurrence Ratio Calculator	65
3.2 Development of MetaPrint2D	74
3.3 The Symyx [®] Metabolite database	75
3.4 MetaPrint2D's implementation.....	84
3.5 Software availability	114
4. Evaluation and optimization of MetaPrint2D.....	120
4.1 Reaction centre identification	120
4.2 Pre-processing of Symyx [®] Metabolite data	122
4.3 Evaluating metabolic site predictions	126
4.4 Evaluation of MetaPrint2D and the effects of data pre-processing options	132
4.5 Analysis of MetaPrint2D's performance	135
4.6 Speed of predictions.....	143
4.7 Parameterization of MetaPrint2D	144

4.8	Isoform specific models.....	147
4.9	Comparison with other tools.....	152
4.10	Accuracy of the test data.....	153
4.11	Conclusions.....	155
5.	Extension of MetaPrint2D to the prediction of transformation types and the generation of metabolites.....	157
5.1	Introduction.....	157
5.2	Identifying transformations.....	159
5.3	Predicting transformations.....	174
5.4	Generating product structures.....	175
5.5	User interface.....	175
5.6	Evaluation.....	177
5.7	Conclusions.....	182
6.	Retrospective prediction of recently published metabolic schemes	184
6.1	[¹⁴ C]Brasofensine (284).....	185
6.2	¹⁴ C-Brivaracetam (285)	187
6.3	Bicifadine (286).....	189
6.4	N-(2-Hydroxyethyl)-3,5-dinitrobenzamide 2-mustard prodrug (287).....	191
6.5	Dabigatran (288).....	193
6.6	Ligustilide (289)	195
6.7	Lithocholic acid (290).....	197
6.8	Pactimibe (291).....	199
6.9	Seliciclib (292).....	201
6.10	Aryl-propionamide derived selective androgen receptor modulator (293).....	203
6.11	Colchicine (294)	205
6.12	3-Amino-5,6,7,8-tetrahydro-2-{4-[4-(quinolin-2-yl)piperazin-1-yl]butyl}quinazolin-4(3H)-one (295).....	207
6.13	Lasofoxifene (296)	209
6.14	Chenodeoxycholic acid (297).....	211
6.15	Torcetrapib (298).....	213
6.16	Conclusions.....	215
7.	Conclusions and further work	216
8.	Bibliography	220

1. Introduction

This thesis is concerned with the *in silico* prediction of xenobiotic metabolism – the metabolism of compounds such as drugs and environmental chemicals which would not normally be produced by an organism or form part of a normal diet. The first two chapters provide an introduction to the thesis. This chapter introduces current *in silico* molecular similarity and virtual screening techniques, which form the basis of the modelling approaches used later in this work. Chapter two discusses the importance of understanding and predicting xenobiotic metabolism, and reviews current work in this field. Chapters three and four report the development and evaluation of MetaPrint2D – a tool for the prediction of sites of phase I metabolism. Chapter five extends these predictions beyond the identification of sites of metabolism, to prediction of types of transformation and the likely metabolites formed. Finally, the performance of these predictions is assessed in a retrospective analysis of recently published metabolic schemes, reported in chapter six.

The search for substances with the potential to cure sickness and disease has been ongoing since prehistoric times. For thousands of years both organic and inorganic materials such as plants, herbal preparations, animal products, metals and clays have been administered to sick humans and animals (1). With the development and application of scientific methodology, mainly since the late 19th century, medication has become far safer and more effective than in earlier times. Ever increasing demand for better medicinal drugs has led to the formation of a \$600 billion dollar global pharmaceutical industry (2) whose future is dependent on the continual discovery of safe and effective medicines.

Over the past twenty years the pharmaceutical industry has been revolutionized through the introduction of high-throughput screening (HTS) and combinatorial chemistry techniques. Despite these changes, and the far higher speeds of synthesis and screening that they have made possible, the rate of introduction of new drugs to the market place does not seem to be showing any corresponding increase (3). Indeed, the rate of attrition of compounds entering the development process shows no improvement from that of the 1970s and 80s (4), currently estimated by the Pharmaceutical Research and Manufacturers of America (PhRMA) to stand at around 90% (5). The length of time it takes to successfully

develop a new drug is also increasing. Various estimates place the average from 8½ (6,7) to over 14 years (8) – up to 75% longer than during the 1960s (9). Due, in part, to the lengthening of the development process, there has been a spiralling in the costs associated with the development of a novel drug; the average investment required to bring a new drug to market is now thought to stand at between US\$800 million (10) and US\$1.7 billion (11).

In an attempt to combat the growing costs and timescales involved in drug discovery pharmaceutical research is increasingly turning to computational techniques. It is hoped that decision support tools can help to accelerate selection of the most suitable candidate compounds, and elimination of the least suitable. Computational tools are also needed in order to manage and exploit the increasingly large amounts of data that are now being acquired, particularly from the use of high-throughput methods taking advantage of robotics to perform *in vitro* screens of large compound libraries.

An additional factor is the use of virtual screening techniques in the safety assessment of chemical substances required under recently introduced legislation such as the European Union's Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) regulation (12). The potential costs of this testing, combined with the public opposition to, and legal restrictions upon, animal testing, have made computational prediction of a number of key biological properties of molecules, such as toxicity and metabolism, an attractive alternative.

This chapter provides a brief overview of the modern drug discovery process, and describes the roles played by computational methods. The different approaches to computational modelling are described, with particular focus on molecular similarity and Quantitative Structure Activity Relationship (QSAR) techniques. Finally, some of the recent developments and current challenges in virtual screening are reviewed.

1.1 The drug discovery process

The road from the initial decision to discover and develop a new drug to its finally reaching the market place is a long one. The process can be broadly broken down into three phases, shown in Figure 1; the central phase, lead discovery and development, is perhaps the most challenging, involving as it does an exploration of the 'sea of chemical space' in search of an

‘island of activity’, followed by the detailed exploration of this island in order to identify a candidate drug compound (13) – a process known as lead identification and optimisation.

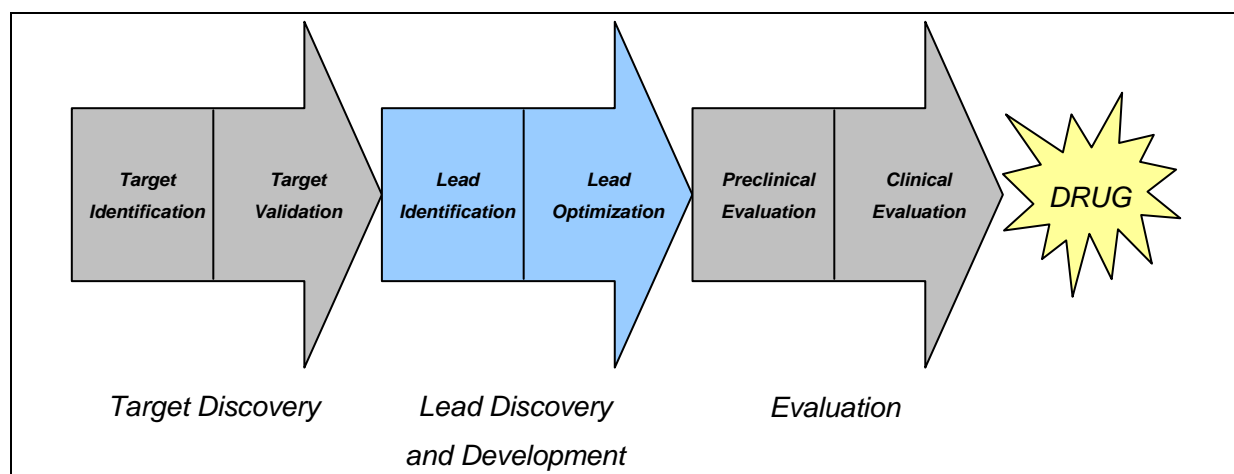


Figure 1: An outline of the drug development process

The first steps in the process of the development of a new drug are target identification and validation (14). Once a decision has been made as to the clinical needs of the new drug, the discovery process commences with a thorough investigation of the mechanism of the disease. Differences between the functioning of the body in both the diseased and healthy states are explored. With the rapid advances in understanding of genetics over the past decade, due in part to programmes such as the Human Genome Project, increasingly the genetic basis of a disease is investigated. The aim in all of this is to pinpoint a mechanism (perhaps involving an enzyme, transport system or receptor) that is the cause of the disease, so that drugs can be designed to target it, whilst minimizing their effects on the rest of the body. Once a target has been identified it is then validated, to confirm its function and effects, and ensure that it is essential to the disease process and is safe to target. At the same time assays to determine whether a molecule binds to the target, and the effect this has on the target, may be developed. Throughout this process bioinformatics and other computational tools are widely used, however these are beyond the scope of this review.

The Human Genome Project has reported that there are in the region of 20000–25000 protein coding genes within in the human genome (15). Through alternative splicing, whereby one gene can code for a number of related proteins, the human proteome is larger still, but in spite of this large number of potential drug targets, to date only 500 or so proteins have been targeted by pharmaceuticals (3).

Once a target has been successfully identified and validated the search for one or more lead compounds begins. Lead compounds are molecules that exhibit some of the desired pharmacological activity against the target, and form the base from which a drug is developed. Traditionally the search for lead compounds started with collections of natural products (14); wide ranges of natural materials such as plants, roots, bark and marine organisms were collected, with as great a biodiversity as possible, and as many chemical compounds extracted as possible. These compounds were then screened against the identified targeted, searching for any that showed some level of activity. More recently pharmaceutical companies have developed in-house screening libraries, typically containing a million or more compounds (16). These may be purchased from external suppliers, produced through combinatorial chemistry programmes or synthesised within the company.

Although lead compounds show some useful pharmaceutical activity against the drug target, they themselves are not usually suitable for therapeutic use. Once their molecular structure has been confirmed, lead compounds undergo cycles of structural modification, or 'optimization', in order to improve their potency and other properties such as solubility and membrane permeability.

Knowledge of the 3-dimensional structure of a drug target enables an alternative approach, whereby molecules are designed to best complement the target's binding site – so called 'rational' drug design. These two approaches are often employed in parallel. Screening can be guided by knowledge of the receptor, and the results of screening can then form the basis for the application of rational techniques to the optimization of the molecule's structure.

Once a candidate compound has been developed, but before it can be tested in humans, it undergoes a series of pre-clinical tests, in order to determine its safety profile (14). This testing is carried out through a mixture of *in vitro* (test-tube) and *in vivo* (animal) studies. The candidate's pharmacodynamic (what the drug does to the body) and pharmacokinetic (what the body does to the drug) profiles are investigated, with the primary aim being to ensure that the candidate compound is safe to test in humans, and determine at what dosage initial testing should take place.

A long series of clinical trials in human volunteers are then carried out to ensure the safety of the candidate drug, and to determine the optimum protocol for its administration. Phase I clinical trials are performed on a small group of normally healthy volunteers, with the primary purpose being to ensure the safety of the drug, since this is the first time that the drug will have been tested on a human body. Volunteers are initially administered very low doses of the drug, and closely monitored for any adverse effects. As the trial proceeds, the doses are increased towards expected therapeutic levels, and further information evaluating the properties of the drug may be obtained.

Once the drug has been demonstrated to be safe in humans, a larger trial is held to determine its safety and effectiveness within its target population. Phase II trials typically involve several hundred patients, one group of whom are administered the new drug, and another group given either the standard treatment or a placebo. The aim of the Phase II trial is to determine the most effective administration regime, varying factors such as dosage, frequency of administration and length of treatment. In order to eliminate any bias, it is common practice to perform a so-called 'double blind' trial, where neither the patients nor clinicians know who is receiving the new treatment, and who is receiving the old treatment or placebo. Phase III trials are then performed in a larger and more diverse group of patients, in order to confirm the drug's effectiveness and detect any less common side-effects, and to compare the effectiveness of the NCE (New Chemical Entity) to currently available therapeutics.

Throughout this process there is a high rate of attrition of candidate compounds. Five years ago the acting commissioner of the US Food and Drug Administration (FDA) reported estimates that just 8% of candidate drugs entering Phase I trials will go on to receive FDA approval, and that only one half of the drug candidates reaching Phase III trials show the necessary safety and effectiveness for approval (11). A more recent study examining cancer trials found that only 25-50% of the new treatments reaching Phase III randomized clinical trials proved successful (17).

Once released, monitoring of a drug's efficacy and side effects continues. Rare adverse reactions may only become apparent once a large population is using the drug, as was the case in the recent widely publicised discovery of an association between an increased risk of

heart attacks amongst patients taking the painkiller Vioxx (18,19) and other cyclooxygenase-2 (COX2) inhibitors; a discovery that led to the withdrawal of a number of high-value products and left Merck defending itself against over 30,000 lawsuits (20) at an estimated cost of US\$4.85 billion (21), due to allegations that it was aware of the risks, but failed to alert users to the possible dangers.

1.2 The role of computational methods

Computational methods, often referred to as *in silico* methods or virtual screening methods, are increasingly being seen as an attractive technique that can be used to complement both traditional and high-throughput screening (HTS) and optimisation strategies (22). As computers have increased in power and decreased in price it has become feasible to carry out computational screening of ever larger databases, using algorithms of increasing sophistication. It is fairly trivial to screen, *in silico*, compound libraries that (while nowhere close to being fully representative of all potential drug-like molecules) are several orders of magnitude larger than even the biggest HTS experiments can handle (23). In order to identify a novel lead molecule with an activity of 1 μ M, a pharmaceutical company will typically have to screen in the order of ten thousand compounds (24), and anything that can be done to reduce this number can significantly increase productivity.

High-throughput screening experiments typically have high rates of false negative and false positive results (25) – active molecules that are missed, or inactive molecules that appear to be active. False positive results cause less of a problem since they are identified during the more reliable secondary or follow-up screens. False negative results, on the other hand, are more serious, since they cause potentially useful hits to be missed.

Virtual screening programmes can be used to help overcome this problem (26). *In silico* screens may be run in parallel to HTS programmes, and the hits from both combined to be used for secondary screening, or alternatively compounds thought to be inactive but similar to active compounds can be added to the HTS hits (27) for further testing.

Not all protein targets are amenable to high-throughput screening. In such cases, possibly resulting from the high cost of an assay, smaller scale iterative screening may be carried out (28). Rather than screening an entire compound collection in one go, a much smaller initial

screen is performed, and its results used to construct a model which is in turn used to select compounds for the next round of screening. This can be repeated until a sufficient number of active compounds have been identified.

Besides virtually screening compounds against a single target, predictions of activity can be made against a large panel of targets in order to identify likely off-target hits, the occurrence of which can lead to toxic side effects, or to suggest novel uses for a compound (29). *In silico* models are also used to predict physicochemical properties such as solubility (30), logP, the octanol/water partition coefficient (31), and pK_a (32), and there is growing use of computational models to make predictions of more complex behaviour of molecules, such as prediction of the metabolic fate of drug molecules (33), of skin penetration (34) and the identification of toxicophores (35,36,37,38) – structural features of molecules indicating likely toxicity.

It has already been mentioned that there is a high rate of attrition of candidate compounds over the course of the drug development process. Figure 2 shows that around two thirds of the failures of drugs reaching clinical development are due to pharmacokinetic problems, animal toxicity and adverse effects in man (39). Historically these issues are often not discovered until late in the development process, by which time significant resources and expense have been incurred. It is hoped that *in silico* tools will enable much earlier identification of potential problems, and reduce the number of lead compound with liabilities that are not discovered until after significant investment has been made in the compound's development, hence reducing overall drug development costs.

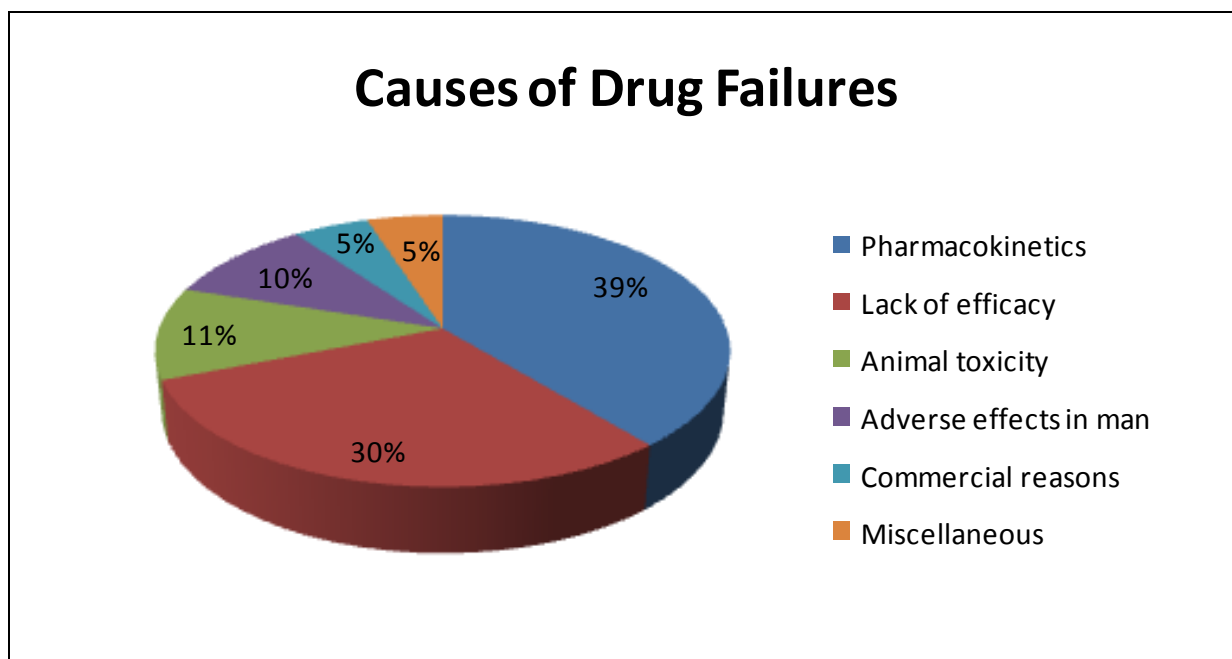


Figure 2: The reasons for the failure of 198 NCEs after reaching clinical development (39). Pharmacokinetic problems, animal toxicity and adverse effects in man together account for almost two thirds of the failures.

1.3 Virtual screening methods

Virtual screening efforts can be broadly divided into two complementary groups of approaches – target- and ligand- based methods. Target-based virtual screening methods generally involve modelling the interactions between the receptor or enzyme active site under investigation and each of the candidate molecules under consideration. This requires a model of the 3-dimensional structure of this target. Ligand-based methods, on the other hand, do not require a model of the target structure; instead they draw on information about molecules with known properties and activities. In ligand-based virtual screening predictions are made on the basis of similarity to known molecules, or on the basis of models constructed around patterns identified in series of compounds. While both approaches have their stories of success and failure, it is ligand similarity based approaches that are the most widely applied, due both to their orders of magnitude greater speed and the far greater availability of suitable data.

Alternatively, *de novo* design techniques, such as SPROUT (40) can be used to design molecules meeting the constraints imposed by the receptor model *in situ*.

1.3.1 Ligand-based virtual screening methods

The concept behind ligand-based virtual screening can be summarised by the ‘similar property principle’ (41): similar molecules are likely to exhibit similar biological activities and properties. Whilst this is a straightforward enough concept, actually deciding how similarity should be measured is a complex matter, and highly dependent on the properties that are under investigation (42).

There are two classes of ligand based virtual screening – molecular similarity searching, which usually has as its aim the identification of potentially active molecules from large databases, and Quantitative Structural Activity/Property Relationships (QSAR/QSPR), which are mainly used in the lead optimisation phase of drug discovery and development, as they are more suited to the detailed analysis of compounds that belong to fairly congeneric chemical series. Both approaches rely on the representation of a molecule through some form of descriptors.

Molecular descriptors

There are many possible representations of a molecule; chemists variously consider a molecule to consist of a collection of atoms and bonds, regions of high and low electron density, or an ensemble of wave-functions, depending on the task at hand. Similarly, there are many ways in which molecules can be represented in a computer. Computers often store molecules as ‘coloured-graphs’ – lists of atoms (nodes) and the bonds (edges) between them, or as a list of atoms together with their coordinates. These representations are not very well suited to mathematical analysis. In order to make comparisons between molecules, descriptors are employed to capture the various properties and features that are thought to be important for modelling molecular interactions, and represent them in a manner that can be understood and manipulated by a computer. Without any clear answer as to how to construct a descriptor that best represents a molecule, an enormous number of different descriptors have been investigated.

Todeschini & Consonni’s Handbook of Molecular Descriptors (43) contains definitions for over 1800 different descriptors, many of which have a number of different implementations. Many software packages, such as Mold2 (44), MOE (45) and SYBYL® (46),

each provide methods for calculating hundreds of descriptors. Despite this huge apparent variety, most descriptors can be assigned to one of three common categories:

- macroscopic physicochemical properties (most often calculated, but sometimes measured) such as the octanol/water partition coefficient (logP) and molecular weight
- substructural fingerprints and feature counts
- shapes and surface properties such as distributions of electrostatic potential

These classifications broadly correspond to the various levels of molecular representation from which the descriptors can be calculated: so-called '1D' descriptors depend only on the formula of the molecule; '2D' descriptors depend on the molecule's connection table – the atoms and the bonds between them; '3D' descriptors depend on the stereochemistry and geometry of the molecule. There are also '4D' descriptors (47) which take into account the wide variety of 3D conformations a molecule can take, and higher dimensionalities accounting for flexibility in protein structures and the induced fit of ligands have also been suggested (48,49).

Examples of the different representations of molecules (1D, 2D, 3D, 4D), and a selection of the descriptors calculable from each are shown in Figure 3.

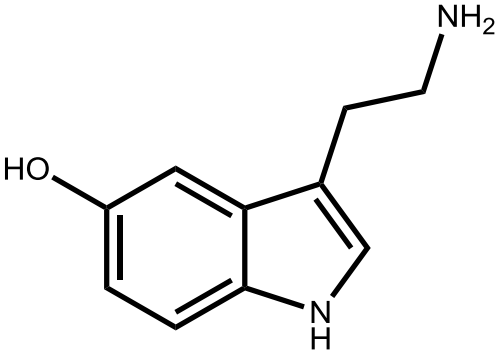
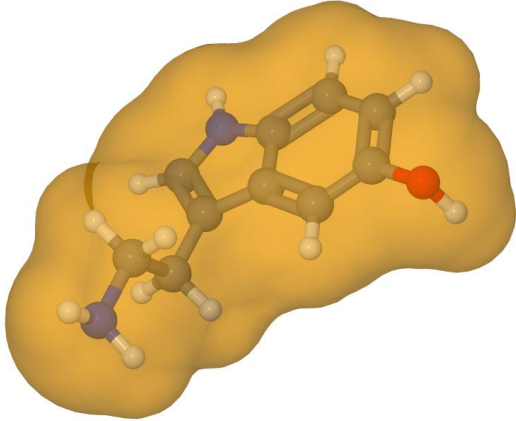
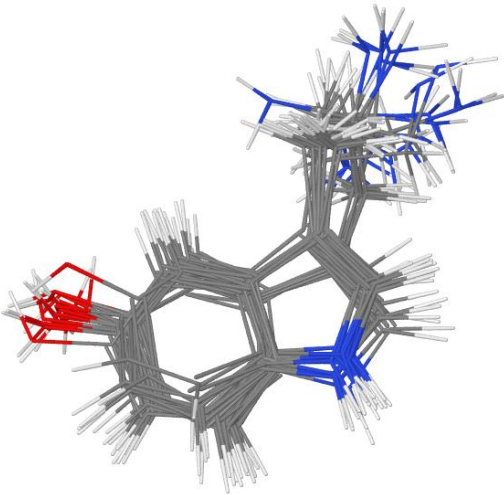
Representation	Descriptors
1D (Formula) $C_{10}H_{12}N_2O$	Molecular weight Heavy atom count Atom counts: carbon, nitrogen, sulphur... Experimental properties (logP, affinities)
2D (Connection Table) 	Hydrogen bond donor/acceptors Number of rotatable bonds Graph invariants Atom additive QSPR (logP, molar refractivity) Substructural fingerprints
3D (Coordinates/Surface) 	Shape Solvent accessible surface area HOMO and LUMO energies Polar volume Dipole moment Pharmacophore fingerprint
4D (Ensemble of Conformations) 	As 3D descriptors, but sampled for different ligand conformations

Figure 3: 1D, 2D, 3D and 4D representations of molecular structure, illustrated with Serotonin, and examples of the descriptors calculable from each representation.

Molecular properties

Molecular properties such as logP and molecular weight are probably the descriptors to have been most widely utilised to express similarity between molecules, and these are still commonly used (50) today. Lipinski's Rule of Five (51), for example, uses four whole molecule properties (molecular weight, logP, and counts of the numbers of hydrogen-bond donors and acceptors) as an indicator of a compound's aqueous solubility, and hence oral bio-availability. Lipinski's Rule states that poor absorption or permeation are likely when a compound violates more than one for the following constraints:

- There are more than 5 hydrogen-bond donors.
- The molecular weight is over 500.
- The logP is over 5.
- There are more than 10 hydrogen-bond acceptors.

This rule of thumb is used throughout the pharmaceutical industry to aid with the selection of compounds for inclusion in screening libraries (although properties required for screening are sometimes at odds to those required for oral bioavailability).

Molecular weight, atom counts and numbers of hydrogen-bond donors/acceptors and rotatable bonds are obvious examples of descriptors for which precise values can be obtained. Values of many more complex descriptors can often be approximated using simple representations of a molecule. Properties such as logP and molar refractivity (MR) can be reliably estimated through atom contribution models, such as XLOGP (52), where each atom is assumed to make an independent contribution to the total logP value, the size of which depends on its local topological environment. An example of such a calculation is given in Figure 4. Surface and volume properties, such as van der Waals areas, which require a 3-dimensional model for a rigorous treatment, can be approximated using analogous methods. Alongside these fairly simple calculations, the results of more complex calculations of molecular properties, such as dipole moments, HOMO and LUMO energies, heats of formation and ionisation potentials derived from quantum mechanical calculations are also often used as descriptors.

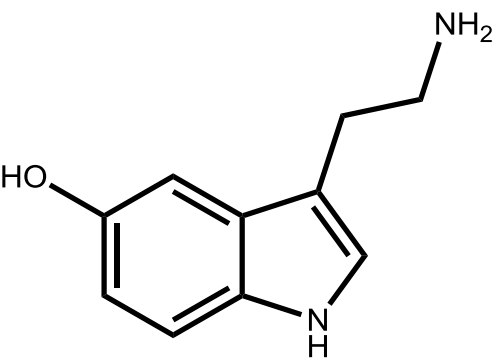
	Atomic Grouping	Contribution
	1 x C3 (2° heteroatom)	-0.2035
	1 x C10 (2° aromatic)	-0.0516
	1 x C13 (aromatic heteroatom)	-0.5443
	4 x C18 (aromatic)	0.1581
	2 x C19 (aromatic bridgehead)	0.2955
	1 x C21 (4° aromatic)	0.1360
	1 x N1 (1° amine)	-0.3239
	1 x N10 (aromatic)	-0.2893
	1 x O2 (alcohol)	-1.0190
	8 x H1 (hydrocarbon)	0.1230
	1 x H2 (alcohol)	-0.2677
	3 x H3 (amine)	0.2142
	LogP	0.287

Figure 4: Illustration of logP calculation using an atomic contribution model (31).

Substructural descriptors

The descriptors most widely used in molecular similarity searching are substructural descriptors. These consist of sets of atoms and/or bonds describing regions of a molecule. Substructural descriptors originally used dictionaries of predefined structural fragments, such as Symyx® MACCS keys (53), to identify the features contained in a molecule. This has the drawback that fragments not considered important by the designers of the dictionary are ignored, when in fact they could prove vital to a particular interaction. This makes the techniques highly dependent on the quality and appropriateness of the particular dictionary that they employ.

Various techniques for automatically identifying fragments have been used to overcome this limitation. Initially these tended to produce fairly small, simple fragments (e.g. augmented atoms, formed from a central atom and its immediate neighbours) (54), but as computer power has increased so have the fragment sizes and complexities that can reasonably be handled.

Descriptors of this type include Daylight's fingerprints (55) which consist of an exhaustive list of all the paths of atoms and bonds that can be traced through the molecular graph,

generally up to a maximum length of seven atoms, and Tripos' UNITY fingerprints (56) which combines paths calculated in a similar fashion to those of Daylight's fingerprints with added counts of certain chemical elements and of additional features such as ring systems.

Molecular fingerprints encode the presence or absence of each substructural feature with a 1 or 0 at a position in a binary bit-string. In the case of key-based systems each feature can be assigned to a specific position in the fingerprint, but this is not possible when the feature set is dynamically generated. Instead, a hashing function is applied to each feature to determine which position in the bit-string to set. This has the disadvantage of reducing the interpretability of the generated fingerprints; many different features may hash to the same fingerprint position, so it is no longer possible to identify the specific features present from the fingerprint's bit-string.

Atom-centred hierarchical fragments of molecules, illustrated in Figure 5 below, form a further class of substructural descriptor. Hierarchically ordered spherical environment (HOSE) and hierarchically ordered ring description (HORD) codes (57) were originally proposed in 1978 for use in the prediction of ^{13}C NMR chemical shifts and indexing files of molecular structures. A more modern description of hierarchical fragment-type descriptors are the Signature Molecular Descriptors of Faulon (58,59).

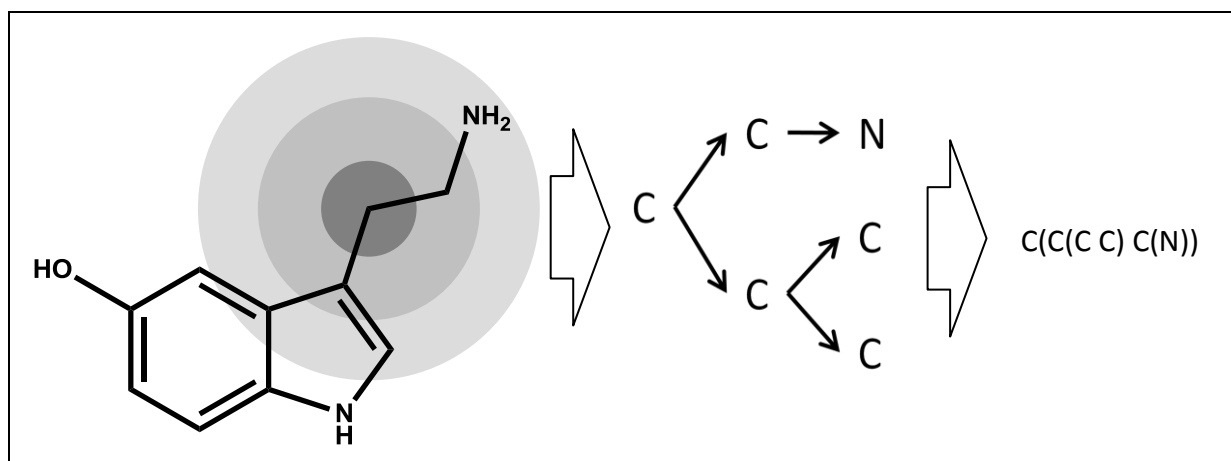


Figure 5: Illustration of the hierarchical structure of fragments and circular atom environments. This specific example is of the generation of a Signature (58,59) -like atom environment fragment description. The region of the molecule on which the fragment is centred is shown, along with the tree structure of the fragment and the descriptor generated.

Most recently, circular atom environments have been introduced, with the aim of encoding the electronic environment surrounding each atom in a molecule, rather than the exact connectivity of a substructural fragment. Their use was proposed by Xing *et al.* (32,60) for the prediction of physicochemical properties such as pK_a and $\log P$ through partial-least-squares regression models based on contributions from the occurrences of SYBYL® (46) atom types at successive topological distances from a central atom, as illustrated in Figure 6 below. This descriptor, combined with a Naive Bayesian classifier and Information Gain based feature selection, forms the basis of Bender *et al.*'s MOLPRINT-2D molecular similarity searching technique (61,62).

A similar concept has been used by SciTegic in the development of their Extended Connectivity and Functional Class Fingerprints (ECFP/FCFP) (63,64), where each atom is assigned a numeric description representing its class, and this classification is augmented with the classes of atoms' neighbours using a series of iterations in a procedure similar to that of the Morgan algorithm (65). R-Group descriptors (66) are generated through the same Morgan-like approach, but are based upon the values of a number of atomic properties including atomic weight, hydrophobicity, molar refractivity and polar surface area, rather than the elemental or pharmacophoric descriptions (pharmacophores are described on page 18) used by SciTegic.

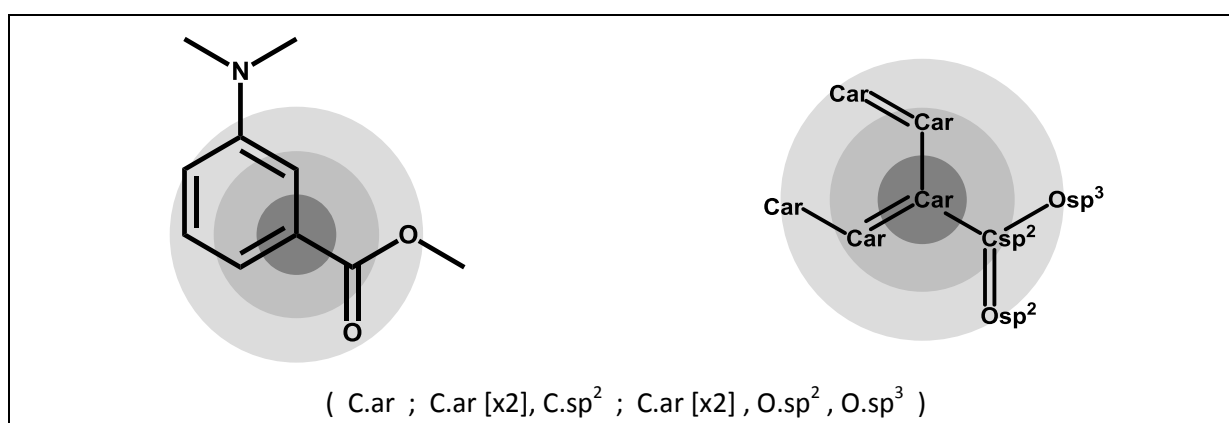


Figure 6: Illustration of the calculation of a circular atom environment fingerprint, using SYBYL® atom types. To the left is a molecule with the fingerprinted region identified, and to the right is the fingerprinted region, showing the SYBYL® atom types. The fingerprint consists of the count of each atom type at each hierarchical depth from the central atom. In this representation layers in the fingerprint are separated by semicolons. 'C.ar' represents an aromatic carbon atom, and 'X.sp²'/'X.sp³' represent atoms of the specified element and hybridization.

A number of descriptors representing longer range features than are usually captured by substructure fragments have also been proposed. Carhart *et al.* introduced atom pairs (67) in 1985. These consist of a pharmacophore-based representation of each pair of atoms in the molecule, along with the length of the shortest path between them. A similar concept, named REX, was developed by Judson (68), where the length of every path was included, not just the shortest. Melville and Hirst (69) have reported the use of partial charges, molar refractivity, logP and logS in the calculation of topological autocorrelation descriptors, which capture the distribution of these physicochemical properties through pair-wise combinations of atoms, and the shortest path between them. Young *et al.* (70) proposed the use of augmented atom pairs – a combination of atom pair and environment approaches, where each end of an atom pair is described in terms of its chemical environment, and a range-based, rather than exact, measure of the distance between the atoms is used in model generation. Nigsch and Mitchell (71) introduced Molecular Orthogonal Sparse Bigrams, which have the potential to described correlation between different regions of a molecule through the pairing of selected atoms' circular fingerprint descriptors, but do not capture the distance between the features.

Fingerprints may be constructed from the combination of several types of feature. As mentioned above, Tripos' UNITY (56) fingerprints are the result of the concatenation of a hashed path fingerprint and a fingerprint representing the counts of certain chemical elements and features such as ring systems. Other fingerprints have combined substructural descriptors with non-structural properties, for instance having bits indicating whether the molecule has a logP value within a particular range.

Certain types of descriptors are better suited to particular applications. Originally the major use of descriptors was in database searching and chemical registration (72), where fingerprints based on a descriptor are used to quickly identify and rank similar molecules, or refine queries by reducing the number of molecules for which it is necessary to carry out computationally much more expensive graph matching against the query compound. Only certain types of descriptor, such as path fingerprints and certain structural keys can be used to refine database searches. Hierarchical fragments and circular atom environments are not suitable for this task, but have shown better performance than path-based descriptors when used to predict biological activities (73,74).

Example calculations of a number of common classes of substructural descriptor are illustrated in Figure 7.

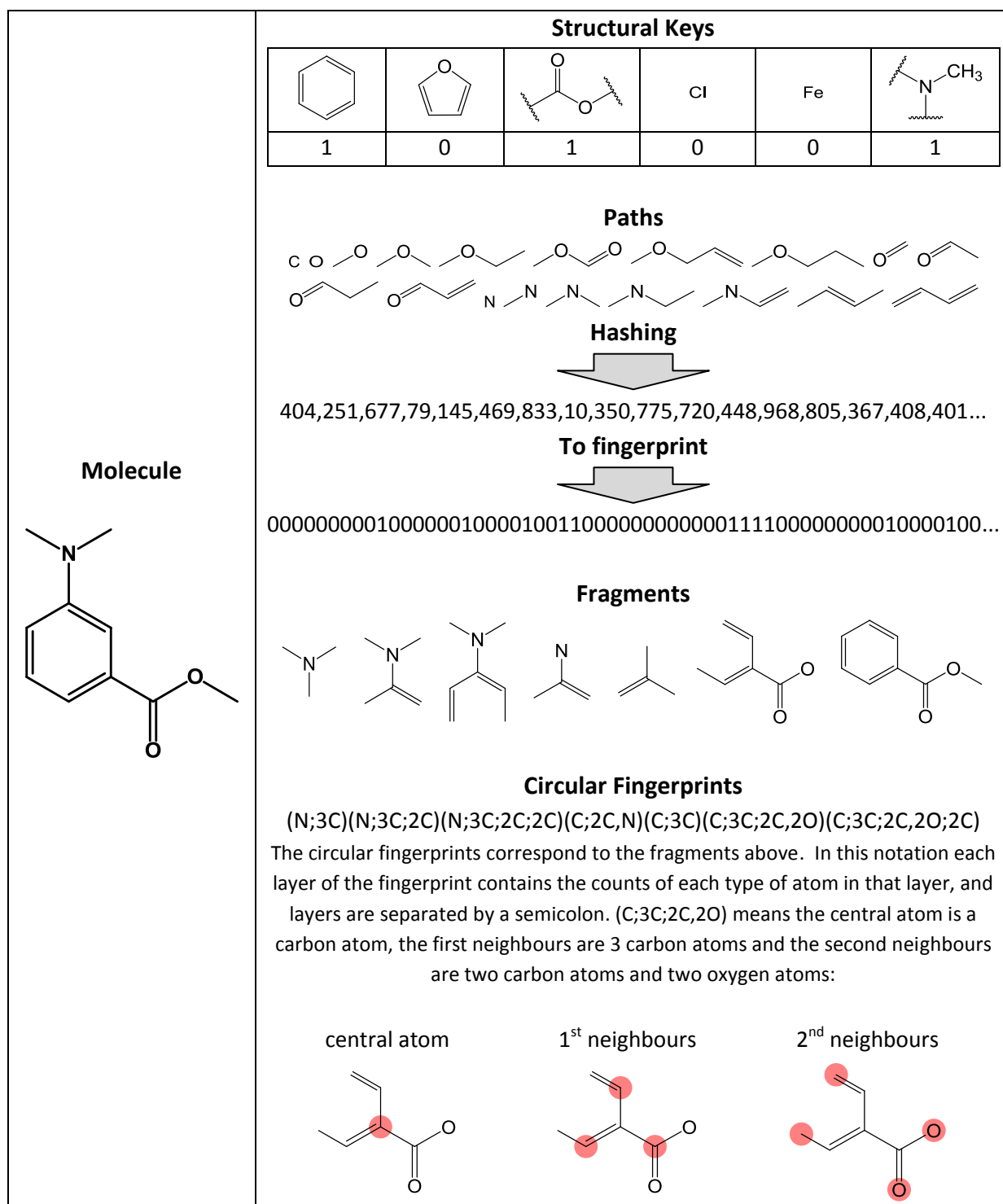


Figure 7: Example calculation of structural keys and path, fragment and circular atom environment substructural fingerprints.

Pharmacophores

Pharmacophores are descriptions of the spatial arrangement of molecular features that are believed to be necessary for biological activity (75) (similar to the 'lock and key' hypothesis). Generally there are six feature types from which pharmacophore models are built – hydrogen-bond donors, hydrogen-bond acceptors, basic groups, acidic groups, aromatic groups and hydrophobic groups. Pharmacophoric features can be identified through the use of simple rules, such as 'primary and secondary amines, and hydroxyl groups are hydrogen-bond donors' (75). Pharmacophores can be constructed from the features common to a number of molecules that are known to bind to a target, or through manual inspection of likely modes of binding.

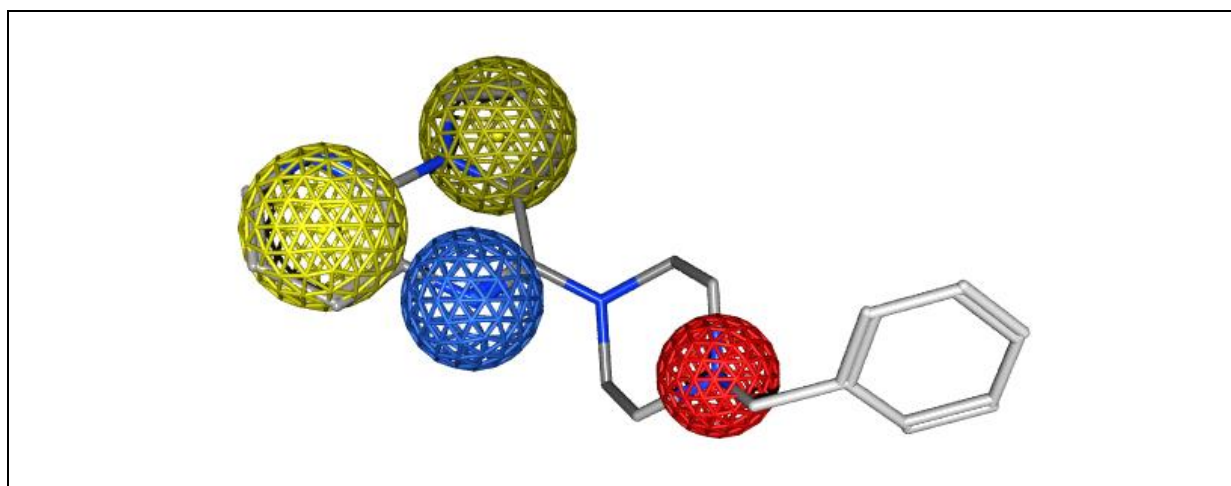


Figure 8: An example pharmacophore model superimposed on the template structure from which it was generated. The yellow spheres represent aromatic regions, the blue sphere a hydrogen–bond acceptor and the red sphere a hydrogen–bond donor and cation.

Pharmacophore fingerprints are an extension of the pharmacophore approach to the generation of 3D fingerprints. The fingerprints are based on combinations of all the potential pharmacophore points identified in the molecule, together with the distances between them (76). Two to four point pharmacophores are constructed from the potential pharmacophore points, and the pharmacophore types together with the distances between the points are hashed into a position in the fingerprint bit string.

As well as pharmacophore fingerprints based on 3-dimensional geometric fingerprints, topological pharmacophores, similar to the atom-pair descriptors mentioned on page 16,

can be constructed. Schneider *et al.* (77) suggested that pharmacophore models generated using topological distances rather than geometric ones, a technique they named Chemically Advanced Template Search (CATS), could overcome one of the main limitations of screening with substructure-based descriptors – the tendency models have to ‘learn’ the core scaffolds that they have been trained on. Unfortunately models constructed with this approach have been found to perform poorly (74).

Shape and surface descriptors

Whilst a large number of descriptors rely on the comparison of the structural framework of molecules it is well known that molecules with different core scaffolds can interact with the same biological site in similar ways. This is because molecules interact *via* their electronic properties, with the point atoms and rigid bonds picture that chemists typically employ being merely a convenient representation of the structure. Substructure based descriptors have been found to exhibit a tendency to ‘learn’ these core scaffolds, restricting their use for scaffold-hopping between chemotypes. 3D descriptors aim to surmount this shortcoming by describing the shape and surface properties of molecules independently of their connection tables.

As has already been mentioned, approximations to various surfaces can be calculated quite simply; however more precise surfaces can be calculated from a 3D representation of the molecule. Commonly observed surface descriptors include the solvent accessible surface area of the molecule, the van der Waals surface area, and the proportions of these areas that are acidic, basic, hydrophobic, polar, or hydrogen-bond accepting (45).

Molecular shapes have been described in terms of assemblies of standard geometrical objects, or using mathematical functions. Morris *et al.* (78) have proposed the use of spherical harmonics to describe the shapes of molecules, which presents a straightforward method of comparison through use of the spherical harmonic expansion coefficients; and Ballester and Richards (79) have proposed representing molecular shapes through the moments of the distribution of atom’s distances from key points in the molecule.

Other common molecular shape and surface related approaches include molecular interaction field (80) based descriptors, where molecules are first aligned and then various probes, measuring steric and electrostatic interactions, are moved over the surfaces of the

molecules, or through regularly spaced grids containing the molecules. Most commonly force field based methods, such as GRID (81) and CoMFA (82) (Comparative Molecular Field Analysis) are used to score these interactions, though Quantum Similarity (83) approaches making use of electron probability density functions are also found, however the latter can be very time consuming to calculate.

Typically, once interaction fields have been determined for a number of known active molecules (training set) they are overlaid and common regions detected. In CoMFA searches this is carried out through partial least squares (PLS) regression between the interaction energy at each grid point and the molecules' activity. These conserved features are assumed to be responsible for the interactions involved in binding to the target. Test-set molecules are examined to see whether they possess the same features. These types of techniques are particularly suited to data sets consisting of relatively rigid structures, since these make it much easier to generate good alignments.

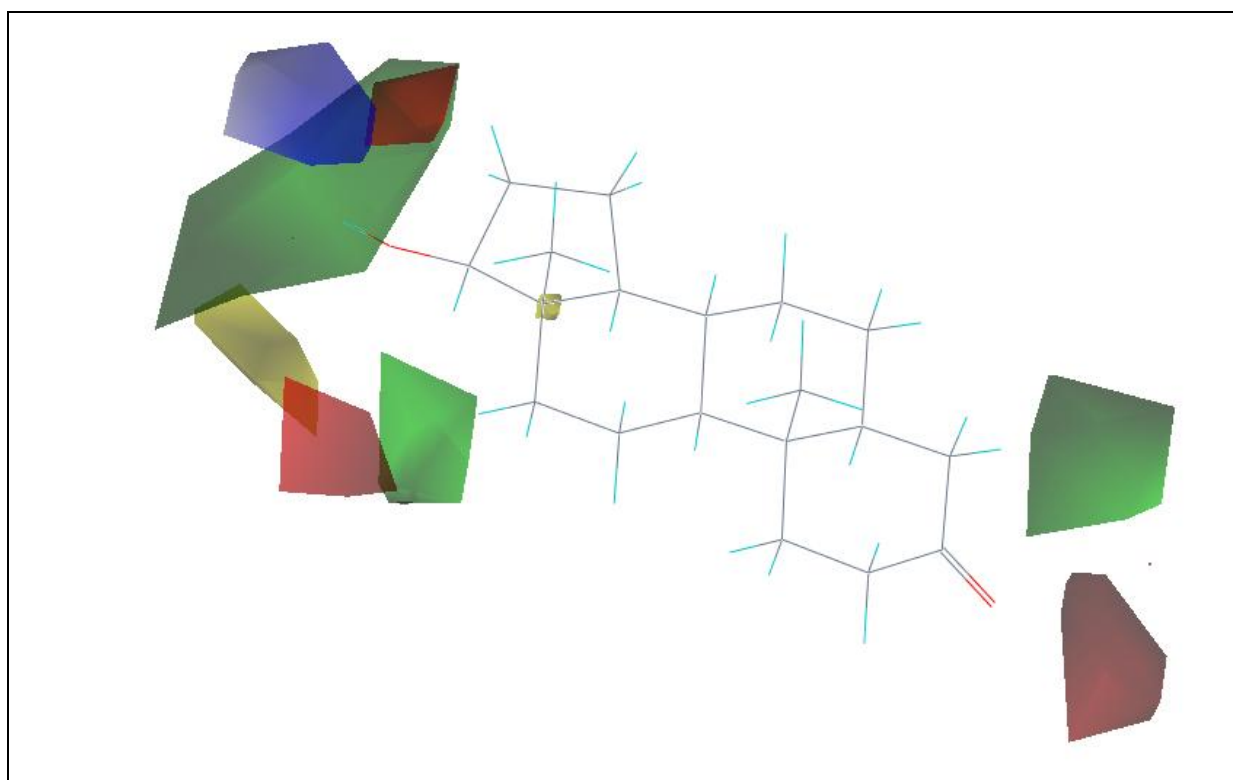


Figure 9: Example of a CoMFA model constructed from a set of steroid molecules, overlaid on the most active molecule from the training set. The green and yellow regions indicate areas that are favourable or unfavourable with respect to steric effects, and the red and blue regions indicate where positive and negative charges are favourable.

Cheeseright *et al.* (84) have developed an alternative analysis of field approaches, based around the idea of electrostatic, steric and hydrophobic field extrema. They propose that previous efforts in modelling what is 'seen' during molecular interactions have failed due to inadequate definition of charge distribution, and the large quantity of data required to describe surface properties. To resolve this they have developed an improved molecular mechanics charge model – the eXtended Electron Distribution (XED) force field, which moves away from the conventional atom centred charge monopole towards a more distributed model. To surmount the second hurdle they have proposed using a pharmacophore point like model, working only with points at the local extrema of the fields, as opposed to trying to process the whole surface or grid, though these models are not restricted to the three or four points generally used in pharmacophore models.

Cheeseright *et al.* (84) have shown that 'molecular field overlays' generated using this approach can identify experimentally observed conformations of the ligands without requiring knowledge of the active site. This was achieved by generating a number of conformations of each of two or three active site substrates, and detecting consistent molecular field overlays between the different molecules. A consistent molecular field overlay is a set of field extrema which are conserved between two or more ligands; all combinations of molecular field extrema generated from the different conformations of each pair of ligands (each conformation of a ligand produces a different molecular field) is examined to determine the best matches. When a consistent molecular field overlay can be found for the set of known ligands, it is assumed that this reflects the required features for binding to the active site.

The main problem with many 3D techniques is their high dependence on the conformation and alignment of the molecules. In order to make analysis as fast as possible, conformations are usually generated using a rule based system, such as CORINA (85) or CONCORD (86). Coordinates may subsequently be optimised using a force field calculation, but there is no guarantee that the biologically active conformation is close to that found through optimisation in a vacuum. When molecules have a high degree of flexibility or are substantially different it can be very difficult to generate an alignment (87).

In order to increase the speed and improve the accuracy of these methods, attempts have been made to remove the need to align molecules before performing grid and field based analysis. GRid-INdependent Descriptors (GRIND) (88) select a small, representative subset of the tens or hundreds of thousands of grid points generated in a typical analysis of a drug-sized molecule, based on the strength of interaction and distance from other representative points. These representative points are encoded on the basis of their pair-wise distances and energies, giving a representation independent of the molecule's alignment in space.

Bender *et al.* proposed the MOLPRINT-3D (89) technique, where the results from GRID probes are assembled into surface patches a few Ångströms in diameter, and the distribution of scores at increasing distances from the centre of the patch binned and recorded, in a manner analogous to the generation of 2D atom environments. These descriptors were found to perform at a mid range level when compared to a variety of standard substructural techniques, though they did detect actives with a wider variety of chemotypes.

Unfortunately such methods take far longer to calculate, so are currently less feasible for use in large scale screening as 2D fingerprinting techniques.

Other descriptors

Many other descriptors have been tried, and only a few of them will briefly be mentioned here. Many topological and other graph based indices have been proposed, examples of which are described in Todeschini & Consonni's Handbook of Molecular Descriptors (43). A range of molecular spectra (90) – X-ray, electron diffraction, infra-red and NMR (91) have also been used, with mixed reports of success (92,93).

Rather than generating fingerprints on the basis of the occurrence of structural features in a molecule, they can be constructed based on the binding affinities when screened against a panel of uncorrelated reference targets. Use of these fingerprints is based on the hypothesis that compounds binding to the reference proteins in a similar manner are likely to bind in a target protein in a similar manner too. Both *in vitro* (using experimentally determining binding affinities) (94) and *in silico* (based on docking experiments) (27,95) affinity fingerprints have been investigated. Compounds are screened against the reference panel, and their activity profiles, representing the response of each target to the compound,

generated. For novel targets, a subset of the profiled compounds is screened, and then models constructed using the compounds' activity profiles.

Molecular similarity searching

Molecular similarity searching has developed out of tools originally created to enable the searching of databases of chemical structures. Initially these relied on the matching of a single query structure against the database contents, with presence of the query structure in a database molecule resulted in the entry being flagged for retrieval (96). This was subsequently developed into searches for molecules containing a number of substructural fragments. As searches became more complex it became desirable for them to return results matching some but not all of the substructural features specified, and these results needed ranking depending on how many of the substructural features they contained. This led on to search systems where an entire query molecule could be entered, and all the nearest-neighbour matches identified through a similarity measure based on substructures common to the query compound and the database molecules.

In order to perform a basic similarity search, a query molecule – such as a known binder to the target under investigation – is specified, and substructural fingerprints generated. This fingerprint is compared with the fingerprint of each of the compounds in the database, and the similarity determined using some metric, the best known of which is the Tanimoto (97) coefficient (also known as Jaccard's "coefficient of community" (98,99)), which scores similarity as the ratio of the number of features the two molecules have in common to the total number of distinct features found between them. There are a wide variety of such similarity metrics available (96) and since their scores are generally highly correlated (100) the results are quite insensitive to the choice of metric.

$$t = \frac{\text{number of features common to both molecules}}{\text{total number features found in either molecule}}$$

Figure 10: Calculation of the Tanimoto coefficient.

Various modifications to the basic bit-string approach can also be made, such as setting multiple bit positions to represent different numbers of occurrences of a feature, rather than simply recording its presence or absence (101,102). Binary fingerprints have also been extended to produce feature count vectors, known as Molecular Holograms (103). A study

by Fetchner *et al.* (104) found that models constructed using holographic fingerprints rarely yielded significantly higher enrichment factors than models constructed using their binary equivalents. It is however interesting to note that the lists of individual molecules returned by the holographic and binary fingerprints did differ considerably, so there could be advantages found in amalgamating lists of actives produced through both methods, in order to increase the molecular diversity of the compounds retrieved.

The use of bit-string based similarity can cause a number of problems; in particular there can be a bias towards larger molecules, which will tend to have more bits set, and there is quite a large 'twilight zone' (105) in which it can be difficult to know whether the similarity score calculated indicates that the molecules are really similar, or not. This can be overcome by deciding up front how many molecules the search should return, and picking enough of the top scoring results to fit this.

Tanimoto coefficient based similarity searches are still used, but there is also a wide range of machine learning and regression techniques that are often applied. Recent similarity techniques such as MOLPRINT-2D and SciTegic's fingerprints make use of machine learning techniques such as Naïve Bayesian classifiers to compare molecules. These offer a number of advantages over the older similarity metrics, most noticeably the ability to train models on a large number of both active and inactive molecules, selecting the relevant descriptors from each, and also producing a more meaningful output than similarity coefficients generally do – the relative likelihood of the molecule under investigation being active or inactive, given that it contains the features that it does, rather than a more abstract 'similarity score'.

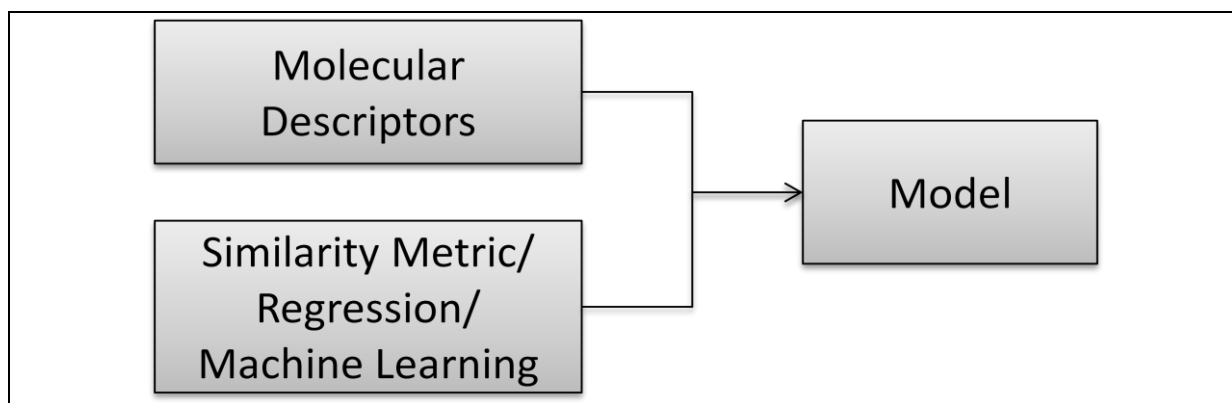


Figure 11: Outline of molecular similarity techniques. Molecules are represented to computers in a numerical or binary form. Models are constructed from these descriptors using similarity metrics or machine learning methods.

Toolkits such as the Weka Machine Learning Workbench (106), RapidMiner (formerly YALE) (107) and the R Project for Statistical Computing (108) provide straightforward access to large numbers of machine learning and statistical methods.

By no means all similarity searching is carried out using substructural features – any descriptors can be used. Many machine learning methods can take real-valued descriptors as inputs, but if not, these can either be binned, to form binary bit-strings, or vectors of values can be used along with distance measures such as the Euclidian distance (length of the line whose ends are at the points represented by the two vectors of descriptors).

QSAR/QSPR modelling

Once a lead series has been identified, interest often moves away from basic molecular similarity measures to the construction of more quantitative models of the activity and properties of molecules in that chemical series. The original QSAR and QSPR models were pioneered by Hansch (109,110) and by Free and Wilson (111) in the 1960s, using linear-regression models based on a small number of molecular property-type descriptors with clear physicochemical meaning (112) to predict activities and properties for specific chemical series.

Similar techniques are still employed today, particularly for the prediction of physicochemical and pharmacokinetic properties such as solubility, with a continuing stream of new publications in the area; one study reports the identification of over 18,800 QSAR and QSPR models (113). Modern models tend to use a much greater range of descriptors,

often including far less clearly interpretable descriptors than the earlier models did. In tandem, newer statistical techniques such as partial least squares and principal component analysis regression, or machine learning methods such as neural networks and support vector machines, are being employed. These methods have the advantage that they can better model the non-linear relationships that are often found to occur between descriptors and activities or properties, but this can come at the cost of a loss of interpretability.

When generating models from a large number of descriptors, particularly when the quantity of training data is relatively small, the selection of relevant descriptors and features is essential (114) if over-fitting of the model to the training data is to be avoided. This selection process can also enable the model to run faster, and produce models that do not contain more complexity than is necessary, making them more clearly understandable (115).

Descriptors that are relevant in one circumstance may be useless in another – one study reported finding that logP (one of the most widely used descriptors in QSAR modelling) was no more useful than random numbers when predicting the biological activity of certain chemical series (13). The cost of calculating descriptors may also be borne in mind – often a complex descriptor may be strongly correlated to a much simpler one (112,116), leaving the time spent calculating advanced descriptors wasted.

Descriptor and feature selection can be an integral component to a machine learning method, as is the case with Random Forests, or can be applied as a filter prior to the model's generation, as with the Naïve Bayesian classifier/Information Gain (117) filter employed by MOLPRINT-2D.

1.3.2 Target-based virtual screening methods

Docking

The most commonly used form of target-based virtual screening is docking, which is the subject of many reviews, examples being those by Kitchen *et al.* (118) and Mohan *et al.* (119). Docking experiments investigate how a candidate ligand could potentially bind to the active site of an enzyme or receptor. The possible conformations and orientations of a small molecule are sampled, and each is placed into the binding pocket of a biological target. The

fit for each docking pose is measured using a *scoring function* typically based on the interactions between the ligand and its target. The highest scoring poses are taken to have the strongest interactions between the ligand and receptor, and postulated to be the biologically relevant fit.

Attempts have been made to correlate docking scores to experimentally determined binding affinities, though this has proven to be challenging (120), likely due to the complex nature underlying the thermodynamics associated with the weak non-covalent interactions involved in protein-ligand binding, particularly entropic and solvent effects. Errors are also likely to be, at least in part, due to deficiencies in the receptor model. As will be discussed later, there is no guarantee that the 3-dimensional structure used for docking is in the biologically relevant conformation. This is compounded by the little or no flexibility afforded to the protein structure, unlike the ligand, due to the high computational cost that would be involved.

A range of docking programs such as GOLD (121), FlexX (122,123) and Glide (124,125) are regularly used in drug discovery processes. These are most successful when consensus methods, combining the results of a number of different scoring functions for each pose, are used. This has been found to reduce the occurrence of false-positives (126), though obviously at a cost of increasing the computational effort required, which reduce the technique's through-put.

Receptor-based pharmacophores

Pharmacophores, which were discussed in more detail on page 18, are arrangements of generalised molecular features representing the main interactions required for the binding of a ligand to a protein. Pharmacophore models are usually constructed from known ligands, however Meagher and Carlson have reported (127) the development of pharmacophore models through analysis of the flexibility found in a collection of unliganded protein structures. This technique was found to be able to discriminate between known inhibitors of HIV protease and drug-like non-inhibitors, and offers the promise of providing more flexible models than traditional pharmacophore techniques where only a single structure is considered.

Structures

As has already been mentioned, in order to carry out target-based virtual screening a 3-dimensional model of the target's structure is required. The most common sources of these are X-ray crystallographic models, followed by NMR and homology modelling and simulation. Some of the greatest challenges associated with target-based virtual screening arise from the quality and the limitations of the structural data available. For many potential targets no structural data is available – this is particularly true of membrane proteins which are notoriously difficult to crystallise, though structures are starting to appear as new crystallization techniques are developed (128), such as the recently published human G protein-coupled receptor (GPCR) structures (129,130,131). Even in cases where 'good quality' crystallographic data is available, mistakes are often made; PDBREPORT (132) estimates that as many as 15% of deposited structures contain errors. Some of the problems associated with the use of crystallographic data are due to users not appreciating that X-ray crystallographic structures are "one crystallographer's subjective interpretation of an electron density map" (133), rather than the direct output of X-ray crystallographic experiments. The inherently static nature of a crystal makes it difficult to appreciate the dynamic nature of structures under native conditions, which can be particularly important when a ligand is bound through an 'induced-fit' mechanism, and the crystallisation process can also lead to instances where the crystal structure does not accurately reflect the conformation of a protein when in solution.

1.4 Current challenges and developments

In spite of the huge growth of virtual screening over the past decade, the high rate of attrition in drug development has continued. Recently there has been a growing feeling among practitioners that *in silico* screening is not performing as well as was expected (134,135). It is increasingly apparent that the reported performance of models, generally from cross-validation or hold-out data at the time of construction, are not being achieved when the models are applied to novel data (136).

1.4.1 Activity cliffs

In part these problems are caused by deficiencies in the modelling techniques. There is a growing appreciation of the appearance of 'activity cliffs' (136) in QSAR data. QSAR models

are based on the ‘similar property principle’ and exhibit ‘neighbourhood behaviour’ (13) – the assumption that activity landscapes can be compared to gently rolling hills, where small changes in molecular structure lead to a small change in activity (137). For many series of compounds this holds true, however instances of the ‘similarity paradox’ (138) where a seemingly small structural modification leads to a large change in biological activity are also common. Examples of this are shown below. Figure 12 shows how successive lengthening of the alkyl chain in a series on morphine analogues moves the compound’s activity from potent agonist to potent antagonist and then back to a potent agonist (139).

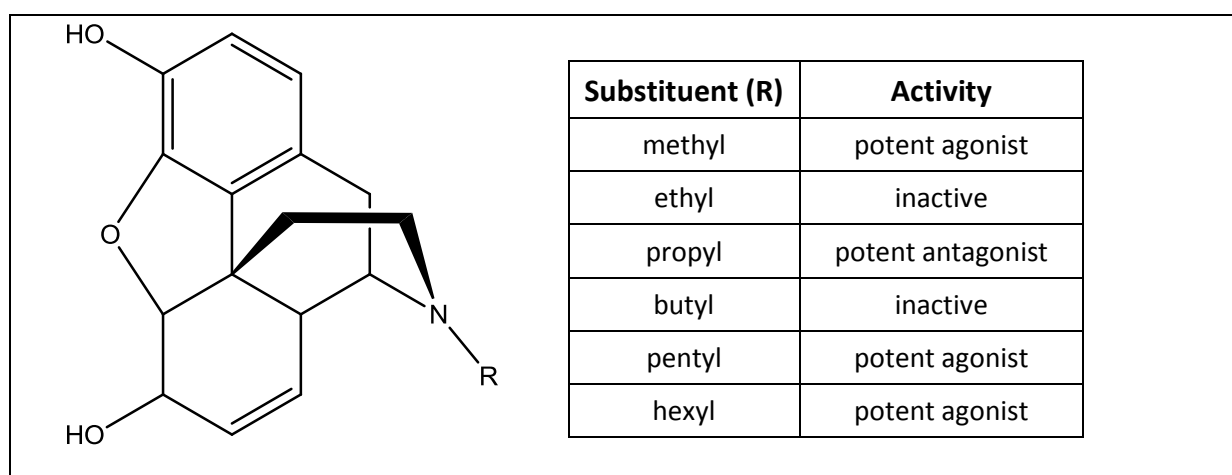


Figure 12: The activities of Morphine (R=methyl) and a series of analogues. Small changes to the structure of the compound can dramatically alter its activity.

The presence of activity cliffs can often be rationalised if the binding mode of the ligand is known; it could be, for example, that the addition of a methyl group leads to an unfavourable steric interaction. If an appropriate descriptor is used to represent the molecule then this behaviour may be captured in the model, otherwise an activity cliff appears. The occurrence of activity cliffs can also be receptor dependent – ligands that appear very similar in some circumstances can behave quite differently in others. Figure 13 shows how a small structural change can lead to small changes in activity against some targets, but an order of magnitude increase in binding to another.

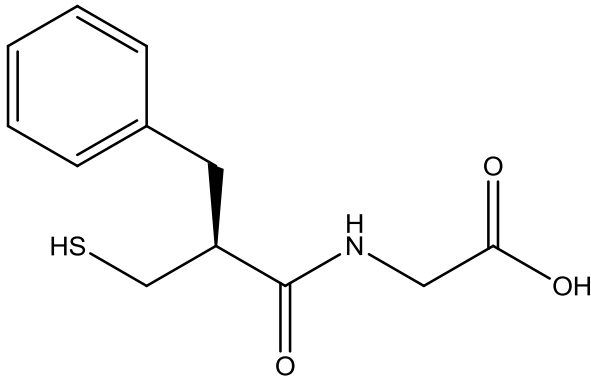
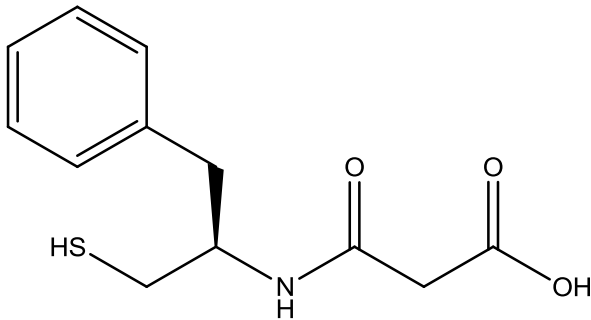
Ligand	Activity (K_i / μM)		
	Thermolysin	NEP 24.11	ACE
 <chem>O=C(O)CN[C@@H](Cc1ccccc1)CCS</chem>	1.8	0.0019	0.14
 <chem>O=C(O)CC(=O)N[C@@H](Cc1ccccc1)CCS</chem>	2.3	0.0023	10

Figure 13: Example of molecules simultaneously illustrating both the ‘similar property principle’ (when binding to thermolysin or NEP 24.11) and the ‘similarity paradox’ (when binding to ACE) (140,141).

1.4.2 Simple models

There have recently been a number of reports of relatively simple models being found to perform as well as much more sophisticated ‘state of the art’ methods. In one comparison of virtual screening tools (116) predictions based on ‘dumb’ atom count descriptors, consisting of the total number of atoms in a structure, the number of heavy atoms and the numbers of each of ten commonly occurring elements (boron, bromine, carbon, chlorine, fluorine, iodine, nitrogen, oxygen, phosphorus and sulphur), were compared with a number of much more sophisticated methods. On average the best performing method in the study (MOLPRINT-2D) achieved enrichment factors that, although 70% higher than the ‘dumb’ atom counts, were much lower than the often reported 10× or higher enrichment relative to random compound selection. For two of the eleven datasets described in the study the atom counts achieved higher enrichment factors than UNITY fingerprints, and as was noted

by the authors, generally found hits with a wide variety of scaffolds, avoiding the biases commonly associated with many substructural descriptors.

Gillett *et al.* (142) have reported the use of simple descriptors to detect compounds from the World Drug Index in a much larger collection of molecules. They found that even single descriptor models using features such as the number of hydrogen-bond donors in a structure gave enrichment factors of up to 4.6× higher than random selection.

Another recent study (143) evaluated a number of logP prediction methods using two small public datasets (223 and 43 molecules) and two larger in-house datasets (882 molecules from Nycomed and 95809 molecules from Pfizer). Thirty methods of predicting logP were evaluated using the public datasets, and 18 against the in-house datasets. Included in the comparison were two ‘dumb’ models: an *Arithmetic Average Model* (AAM), which assigned all molecules the same logP (the mean of the dataset), and a very simple QSPR, based only on the number of carbon atoms (NC) and the number of hetero-atoms (neither carbon nor hydrogen atoms – NHET) in the molecule:

$$\text{LogP} = 1.46 + 0.11 \times \text{NC} - 0.11 \times \text{NHET}$$

When evaluated against both of the in-house datasets no more than half of the tested methods performed better (assessed by RMSE in predictions) than the Arithmetic Average Model, and the simple NC+NHET QSAR was among the best performing of all the methods. Against the public datasets the majority of prediction methods performed better than the AAM and NC+NHET models. The authors suggest that this could be due to the paucity of publicly available data making it likely that many of the tools would have included data from the public datasets in their development.

It has also been shown that simplistic 3D models can perform on a par with more sophisticated ones. Manchester and Czermiński (144) compared CoMFA with a significantly less sophisticated alternative they termed *Simple Atom Mapping Following Alignment* (SAMFA). Like CoMFA, the SAMFA method is dependent on an alignment of structures, but rather than comparing steric and electrostatic field values at points on a regular grid containing each structure, SAMFA compares the occurrence of particular elements and pharmacophoric atom features at each point occupied by an atom in any of the aligned

structures. In a comparison of the two techniques' performance over nine data sets the authors found that the SAMFA method performed "as good, or slightly better" than COMFA. They surmised that the number of simplifications and approximations inherent in QSAR modelling techniques overshadowed any problems due to the simplifications made in the SAMFA method, and felt that the model's atom-centred basis made it straightforward to interpret.

1.4.3 2D versus 3D

Despite 2D based virtual screening methods generally containing no information on the shape or stereochemistry of a molecule, they have widely been found to perform as well as, or better than, either descriptor- or docking- based 3D methods (145,146,147,148). Combined with their generally much higher throughput, this has led to 2D models being used much more regularly than 3D approaches. Although there is considerably more information lost in the generation of 2D models, the construction of 3D models introduces more noise into the data. While it is well known that stereochemistry is important to molecular recognition (149), a method of including stereochemistry in 2D descriptors is yet to find widespread acceptance.

As previously discussed, ligand and protein flexibility are often only partially considered, or completely ignored, by docking programs. If they are fully included then this can result in too many degrees of freedom in the system for the problem to be computationally tractable with current resources and techniques. Ligand flexibility is also often ignored in 3D modelling. The ligand conformations used in 3D model generation and docking are often taken to be the coordinates of an idealised energy minimum structure, while it is known that binding often occurs in higher energy conformations. Feher and Williams (150) have reported that despite docking tools usually allowing for ligand flexibility, their generated poses and scores are highly sensitive to the ligands' input conformations. In an investigation into the effects of input geometry of ligands on docking calculations, Feher and Williams found that none of the sources of coordinates evaluated (X-ray crystal structures, force field minimized CORINA (85) generated structures and conformational searches) consistently produced better results than the alternative sources.

A further complication arising with docking approaches is that a docking program relies on two distinct components: a method for generating ligand (and possibly protein) conformations and a scoring function that evaluates the binding of each ligand conformation to the target. Scoring functions are generally intended to reflect binding pose, providing a relative ordering of ligand/protein complex conformations. Unfortunately current scoring functions are not considered to be very reliable in predicting binding affinity (151); one recent investigation of docking (152) reported that “comparative studies indicate that none of the docking programmes truly outperforms the others”. This is likely to be at least in part due to scoring functions being based mainly on the strength of interactions between the ligand and protein, while the free energy of binding depends on many other factors, such as the hydrophobic effect, destabilisation of the unbound protein or ligand or changes to proteins’ normal modes (heat capacity), and hence entropy, on binding.

Specific biomolecular systems can be investigated through atomistic simulations, but this is not tractable in a high-throughput manner due to the computational resources required. There are also a number of challenges associated with such simulations; bespoke force field parameterization is often necessary for uncommon ligands, and there are difficulties associated with modelling protein quaternary structures, cooperative binding and lipid bilayers that have not yet been fully solved.

1.4.4 Local versus global models

Early QSAR/QSPRs were local models, describing changes in activity, or the variation of a property, within a single chemical series – a collection of structurally related compounds. Modern *in silico* models tend to be global, able to make predictions for any compound. The predictions of these models are, however, only reliable within the regions of chemical space in which the model was trained (the applicable domain). Some of the dissatisfaction with current *in silico* models arises from attempts to extrapolate beyond these limits, which generally results in poor predictions (135). A number of studies have shown that compromises between local and global approaches, such as the construction of sub-models or the application of local corrections, can lead to improvements over a single global model.

While predicting molecules’ pK_a , Xing *et al.* (60) constructed separate models for subsets of chemical space; acids were subdivided into four broad categories: aromatic acids, aromatic

alcohols, aliphatic alcohols and aliphatic acids, and bases were similarly separated into different classes of molecule. The combination of these simple sub-models gave much more accurate predictions than their previously published global models for the pK_a of all acids and all bases (32). Similarly, when modelling solubility, Bergström *et al.* (145) found that by generating sub-models for acids, bases and ampholytes (molecules containing both acidic and basic groups) more accurate predictions were possible than when using a single global model.

Rather than pre-selecting which sub-models to generate, local models can be generated on-the-fly, based only on the subset of the training data most relevant to a query compound (153,154). Based on the assumption that similar compounds will be subject to similar errors in prediction, local model corrections can also be applied, adjusting the value of predictions made from global models according to the mean error in prediction of the k-nearest neighbouring molecules from the training data (155) or from data acquired after the model is constructed (156,157). It has been reported that this type of approach can be much more accurate than use of a single global model. However, there are conflicting reports as to the circumstances under which it is appropriate to apply this technique:

“[Local regression] can also lead to larger prediction errors when compared to ordinary global regression. This is especially true when the training data is sparse.” (153)

“[Locally weighted linear regression] appears to be especially well-suited for the development of highly predictive models for the sparse or unevenly distributed data sets.” (154)

Similar results can be achieved through the use of certain machine learning techniques, such as decision trees, which inherently divide the model space in a (hopefully) optimal manner.

QSARs, and other *in silico* models, have tended to be static, generated and evaluated by an expert, and then left unchanged for long periods. Rodgers *et al.* have shown that the performance of these models can exhibit time dependent behaviour (158). They reported the results of constructing and evaluating models for Human Plasma Protein Binding on a monthly basis over a two year period, with a portion of the new data collected each month

held back from model generation and used for testing. The accuracy of predictions of earlier test data were found to be stable as the models were updated. However, the updated models were better at predicting both new and future test data. This suggests that over time the focus of current research can shift away from the compounds with which models were constructed, and into new regions of chemical space. By ensuring that models are updated regularly to reflect this, the accuracy of predictions can be maintained.

1.4.5 Consensus methods and data fusion

Consensus methods, basing predictions on the combined output of many different approaches, have been found to perform well in a wide variety of fields. There are many reports on the ‘wisdom of crowds’ – the average of many independent estimates made by humans being more accurate than those of individuals, even those of specialists (159), and the same is often true of computational models. The recently announced winning entry of the Netflix Prize for the machine learning algorithm that best predicts subscribers’ ratings of movies makes its predictions through the combination of a number of diverse approaches (160).

Increasing computer power has made combining the results of multiple models ever more feasible. A recent review of data fusion methods by Willet (161) did not find them to be any more effective than the best individual predictor in most studies, but their results were comparable to the best individual functions, and were robust to changes, while the best predictor varied from experiment to experiment.

The logP study (143) discussed above included a consensus model based on the mean value of the predictions made by the other models. This was more accurate (predicted logP values had a lower RMSE) than any of the tools individually for three of the four datasets used in the study, and was close to the best performing model in the fourth.

In virtual screening experiments it is often not only the accuracy of the results that is important, but also the diversity of the structures identified. The output of screening approaches with little correlation, such as affinity and structural fingerprints (162), are often complementary to each other – many of the hits returned by each method are missed by the other (27). Combining the results of different similarity searching methods leads to the inclusion of more hits than are identified by any one method alone (163). Similarly, it has

been found that in docking experiments use of consensus scoring functions substantially improves tools' performance in most cases (164,165).

1.4.6 Model interpretability and inverse QSAR

Ideally, models and the factors contributing to each prediction are easily interpreted by a chemist, enabling them to appreciate how alterations to the molecule's structure will affect the model's predictions. Linear models tend to provide this interpretability, however non-linear models and machine learning methods are often found to offer the highest predictivity, but at the expense of losing interpretability. Non-linear models are often considered to be 'black boxes', giving little or no indication of the basis for a prediction.

This is not true of all machine learning and non-linear techniques, for example Bayesian classifiers can be interrogated to determine the contribution of each feature to a prediction. When using other machine learning techniques, such as Random Forests, it may be possible to extract the importance of each variable within a model, but not assess its contribution to an individual prediction.

Carlsson *et al.* (166) have recently proposed a novel method of assessing the importance of each input variable of a model to an individual prediction. They proposed the generation of a locally linear approximation to non-linear or black-box models by either analytical or numerical calculation of the partial derivative with respect to each variable about the point at which a prediction is made. Assuming that the function is sufficiently smooth, the gradient of each variable reflects its importance to that particular prediction, and enables rational exploration of chemical space in the local neighbourhood about a molecule, by indicating how the predicted property will vary with minor changes to that variable.

1.4.7 Data quality

One of the major issues that virtual screening research has faced, especially in academia, is access to and quality of data. Much chemical data is commercially sensitive, so never gets published. Many of the available datasets are fairly small, and published as supporting information to papers (e.g. Briem and Lessel (27), Jacobsson *et al.* (167) and Fontaine *et al.* (168)). As a result many virtual screening studies report building and evaluating their models on far smaller data sets than the libraries to which they intend the models to be applied.

A number of databases abstracting details of active ligands from the literature are available. The Symyx® Molecular Drug Data Report (MDDR) (169) and the World of Molecular Bioactivity (WOMBAT) (170) database are commercial offerings that have been available for some years, and recently the EBI has acquired a similar resource, the StARlite database, and made it freely available rebranded as ChEMBL (171). While useful resources, these databases do have limitations; they often contain a relatively small number of molecules tested against the majority of reported targets, and being aggregated from literature generally report only active compounds.

A further problem with data aggregated from a number of sources is that results are not necessarily comparable between experiments. Where quantitative results are available, some measures, such as IC50 values, depend on the experimental conditions, and even measurements of a property as apparently straightforward as solubility can vary wildly (172).

Due to the relative rarity of activity, it is assumed in many analyses that any compound not reported to be active is inactive, but it could be that this compound has just not been tested, and also problems can arise if promiscuous binders are not identified.

The NIH Molecular Libraries Initiative (173) is now making publically available the results of high-throughput screening (HTS) programmes through the PubChem Bioassay service (174), providing data on both activity and inactivity. So far there have been few reports of models constructed using this data, possibly because, as with all HTS data, this brings with it issues regarding quality and noise.

Due to the large imbalance in much of this data, with many more inactive (or presumed inactive) compounds than active ones, very high prediction accuracies can be achieved simply by ignoring the presence of active compounds altogether (175) – if only one in a thousand molecules is active, a model can correctly classify 99.9% of compounds simply by predicting that all of them are inactive! This can be a particular problem when models are being generated and evaluated in an automated manner.

Many screening libraries and other datasets contain series of close analogues designed to bind to a particular target, which can also skew the results of evaluations of virtual screening techniques.

Some datasets contain inherent biases. In the α_{1A} agonist dataset published by Jorisson and Gilson (176) active molecules are considerably larger than the inactive compounds*, meaning that any model including a measure of size as a descriptor can easily discriminate between actives and inactives. Inherent anomalies such as this are not discoverable through cross-validation, but only become clear when the model generated is applied to alternative data and not found to perform as expected.

The relative scarcity of data means that many supposedly independent models have in fact been trained using much of the same data. As with the case of the logP models discussed earlier, not knowing exactly which data was used to develop a model can make reliable evaluation of its performance difficult.

1.4.8 Applications of virtual screening

Molecular similarity, QSARs and other virtual screening methods are widely used in the prediction of activities and properties of compounds (such as logP and solubility) but additionally these approaches are increasingly being applied to other more complex problems.

Originally virtual screening experiments were conducted with the aim of testing as many compounds as possible against a single target. Multi-target models are now being generated which enable compounds to be evaluated against a large panel of potential targets at once (e.g. BioPrint from Cerep (177)). It is hoped that this will lead to early identification of off-target effects and potential drug-interactions, and may help to better understand the mode of action of multi-target drugs.

A further application of multi-target predictions is the identification of novel therapeutic uses of existing drugs. While some additional trials are necessary, extending the use of an existing drug to a novel therapeutic area is much less expensive than developing a new drug

* The mean length of the shortest path between the most distant atoms in each inactive compound is 9.3 bonds, but for the actives it is 16 bonds – an increase of 71%.

from scratch – the compound will already have been developed into a therapeutically useful form, and have undergone extensive safety testing.

A recent study has reported the *in silico* screening of 3,665 approved small-molecule drugs and other pharmaceutical compounds (29) against a large panel of targets, using molecular similarity techniques. This predicted a number of previously unknown targets for many of the compounds. The authors tested 30 of the predicted drug-target interactions, and 23 of these were confirmed (29).

In a similar manner, several groups have used virtual screening approaches to investigate traditional Chinese medicines (TCMs) (178,179,180). The constituent compounds from TCM ingredients such as ginger and ginseng have been screened against models constructed for panels of drug targets, generating ‘bio-prints’ of the compounds’ likely activity. These bio-prints have been related to the TCM’s therapeutic use in order to elucidate possible modes of action, and potentially identify new lead compounds for pharmaceutical development. Evidence has been found supporting a number of the modes of action of TCM ingredients predicted using this approach (179).

QSAR models often perform well within a series of closely related compounds, however it is often desirable to identify molecules with a novel scaffold (core structure) but offering similar properties to a query compound. This may be due to a desire to avoid a liability identified with a particular scaffold such as toxicity or promiscuous binding, or to avoid regions of chemical space infringing on a competitor’s patents. A number of studies e.g. (181) have investigated the scaffold hopping potential of different tools, particularly 3D methods.

There have been a number of advances in automated model generation (182,112). This can offer the potential to explore many combinations of descriptors and modelling techniques in order to identify the optimal combination. While the risks of models over-fitting their training data are believed (at least for the users of such models) to be fairly well understood, the huge numbers of descriptors and machine learning and regression

techniques available to an automated system increases the likelihood of discovering chance correlations*.

Virtual screening tools are also being applied to the improvement of screening library collections, with the aim of generating more drug-like hit and lead compounds, reducing the effort required to move from lead to candidate (183). *In silico* models are increasingly being used to identify potential ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) liabilities as early in the discovery process as possible, in order to reduce the number of hugely expensive late-stage failures.

The prediction of xenobiotic metabolism, which is of great importance to the pharmacokinetics, efficacy and toxicity of New Chemical Entities, forms the basis for the remainder of this thesis. The next chapter provides an introduction to xenobiotic metabolism, and current approaches to its prediction. The development and evaluation of MetaPrint2D, a new tool for predicting sites of xenobiotic metabolism, is then discussed, and the extension of MetaPrint2D to prediction of types of metabolic transformation and the likely metabolites formed described in subsequent chapters.

* Assuming that there is a 0.1% probability of a model exhibiting chance correlation, only 693 independent models must be generated for there to be a better than 50% likelihood that one will show chance correlation ($0.999^{693} = 0.4999$).

2. Prediction of xenobiotic metabolism

The remainder of this thesis describes work investigating prediction of the metabolism of xenobiotic compounds. The *Substrate/Product Occurrence Ratio Calculator* (SPORCalc) method for predicting sites of xenobiotic metabolism has been re-designed with a number of enhancements, thoroughly evaluated, and its performance increased to the point where it can be used in an interactive or high-throughput manner. The new software (MetaPrint2D) is available as a freely distributable library and includes a number of example applications, enabling more wide-spread use of the method.

This chapter provides an introduction to xenobiotic metabolism and its effects, and to the current computational approaches used for its prediction. The following two chapters describe the development and evaluation of MetaPrint2D, and Chapter 5 describes MetaPrint2D-React: an extension of MetaPrint2D extending it from site of metabolism prediction, to prediction of the metabolic transformations and metabolites formed. Both MetaPrint2D and MetaPrint2D-React have been extensively evaluated, and this is described in their respective chapters. Finally a retrospective analysis of recently published metabolic pathways is reported in Chapter 6.

2.1 Introduction

2.1.1 Xenobiotic metabolism

Xenobiotics are compounds that are introduced into an organism, but which would not normally be produced by the organism or form part of a normal diet. These can include, for example, drugs and food additives together with environmental chemicals, such as agrichemicals and personal and household products, to which the organism has been exposed. These compounds must often be removed from the organism to prevent their producing any adverse effects, and this is achieved through their metabolism.

Xenobiotic metabolism is generally considered to occur in two phases (184,185). Phase I transformations act to ‘functionalise’ the xenobiotic in order to prepare it for phase II reactions, where the compound is conjugated to groups that will aid in its clearance from the organism. Phase I transformations may add new functional groups to the compound,

increase the polarity of existing groups, or unmask existing but protected ones; reactions such as hydroxylation and hydrolysis are common. Phase II transformations conjugate functional groups of the parent compound or its phase I metabolites to highly polar endogenous molecules such as glucuronic acid, sulphate and glutathione, which increase the hydrophilicity of the compound, facilitating its excretion. An example of a metabolic pathway showing both phase I and phase II transformations is shown in Figure 14, below.

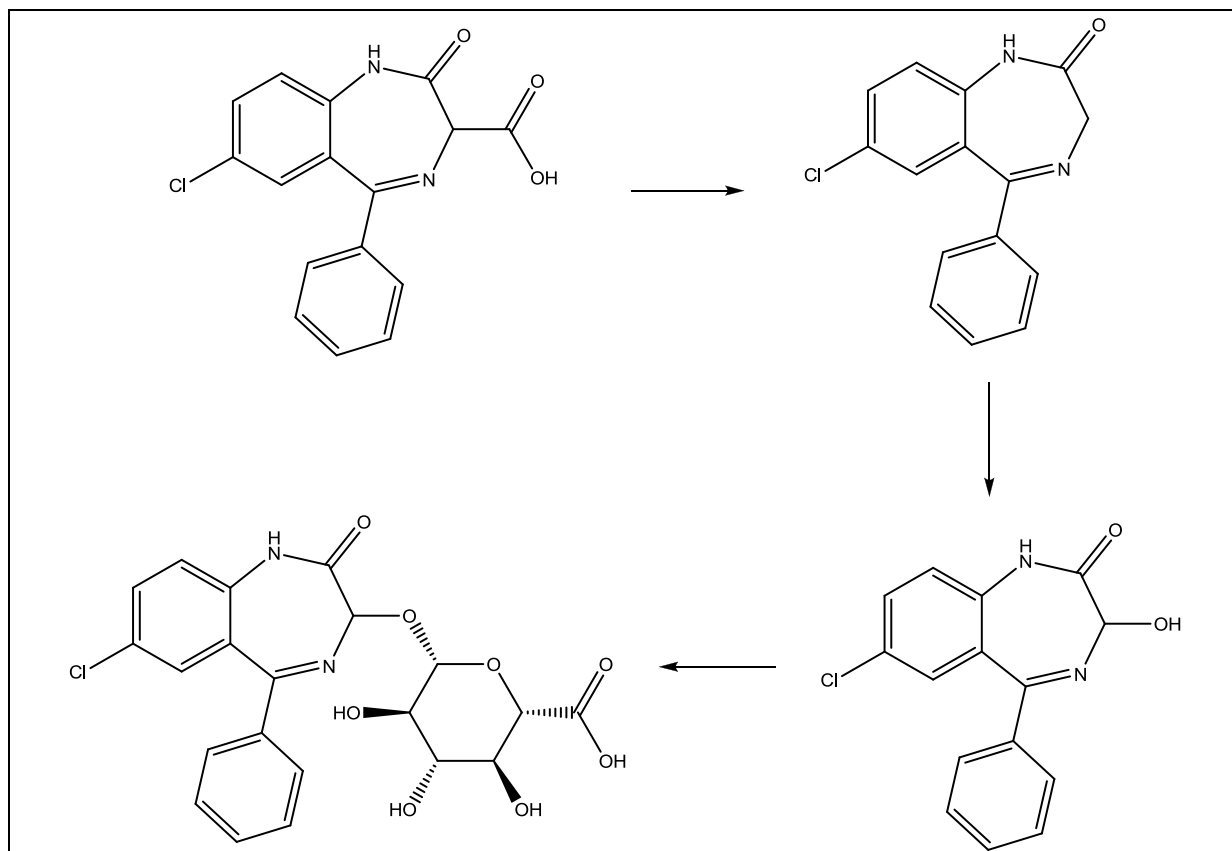


Figure 14: One metabolic pathway of the sedative Clorazepate, from the 2008.1 release of the Symyx® Metabolite database. In the first two steps the compound is undergoing phase I transformations, resulting in the introduction of a hydroxyl group. The metabolite formed then undergoes a phase II transformation, conjugating to a glucuronic acid molecule. The highly polar product formed will be rapidly excreted.

Understanding how xenobiotics are metabolised is of great interest both within the pharmaceutical industry and the wider chemical community. Metabolic transformations may reduce the bioactivity of a compound – deactivating the therapeutic properties of a drug, or detoxifying an environmental compound. Alternatively they may increase a compound's bioactivity, which can be exploited through the development of prodrugs, but this can also lead to the formation and build up of toxic metabolites. These effects can only

be predicted if the biotransformations that a molecule will undergo can be understood or anticipated. Additionally, prediction of likely biotransformations can help to guide experimental design when trying to identify a compound's metabolites.

Over the past decade pharmacokinetic problems, including metabolic liabilities, have been recognised as a major cause of failures in the development of new pharmaceuticals, particularly in the later stages of the drug development process where failures are most expensive. It is now recognised that potential ADME and Toxicology problems should be addressed as early in the development cycle as possible (186,187) – ideally when selecting and optimising lead compounds. At these early stages in the drug discovery process it is often not practical or economical to exhaustively experimentally determine the ADME profile of candidate compounds, so computational models are used instead, enabling the results of high-throughput screening programmes to be prioritized, and even the selection and elimination of compounds pre-synthesis.

2.2 Effects of metabolism

2.2.1 Toxicity, bioavailability and clearance

In order for a drug to exhibit its desired pharmaceutical effect it must be present at a concentration within the drug's therapeutic window. At too low a concentration the drug will not have its desired effect, but conversely at too high a concentration the drug will likely exhibit adverse side-effects. It is important that pharmaceutical compounds are removed from the body after their administration, in order to prevent their accumulation to toxic levels. At the same time they require a certain degree of metabolic stability, in order to persist long enough to be able to achieve their therapeutic effect. This is particularly true of orally administered compounds which have to survive the harsh conditions of the digestive system and first-pass metabolism in the liver, before they are able to enter the bloodstream.

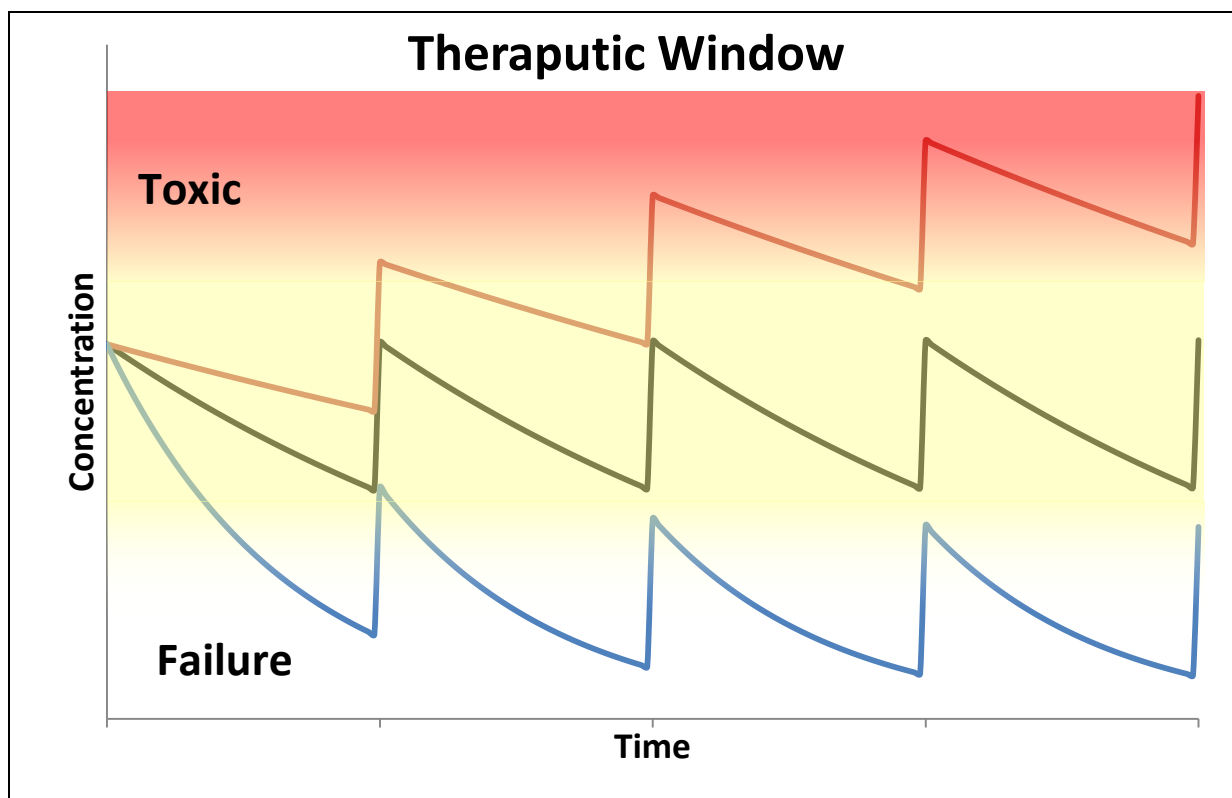


Figure 15: In order to be effective (black line) a drug's concentration must remain within its therapeutic window (yellow region). The drug's concentration rises on dosing, and then drops as the compound is metabolised and excreted. If the compound is metabolised too slowly (red line) then repeated dosing causes its concentration to rise to toxic levels (red region). On the other hand, if the compound is metabolised too quickly (blue line) then its concentration will fall below that required for the drug to be effective (below the yellow region). Figure adapted from (188).

Understanding a drug's metabolism can enable adjustment of the compound's pharmacokinetic profile through the blocking of major sites of metabolism, or addition of functional groups facile to metabolism. Where a molecule has very poor bioavailability, caused by its rapid metabolism and clearance, this can be resolved through the identification of the major route of metabolic degradation, and subsequent modification of the compound in order to block this pathway. A successful example of this is illustrated in Figure 16.

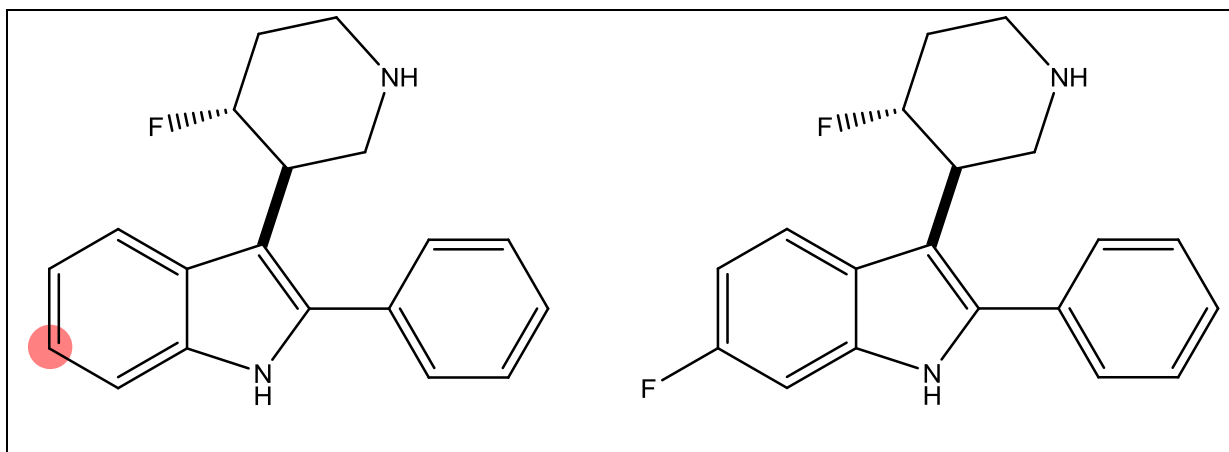


Figure 16: Identification of the major site of metabolism of the 5-HT_{2A} antagonist 3-(4-fluoropiperidin-3-yl)-2-phenyl-1*H*-indole (compound on left, major site of metabolism is highlighted), and its subsequent blocking with a fluorine atom (compound on right) reduced the rates of first pass metabolism and clearance, leading to the bioavailability in rats increasing from 18% to 80%, and the half-life from 1.4 to 12 hours (189).

Knowledge of a drug's metabolism is also necessary in order to determine safe dosage levels and warn of drug-drug interactions. Consideration must be given to the possibility that a drug, which on its own is perfectly safe to take, may inhibit or induce the metabolism of other drugs if taken in combination. Induction of a drug's metabolism will lead to increased rates of clearance, lowering its concentration, possibly below effective levels. Inhibition, on the other hand, can result in the accumulation of the drug to toxic concentrations. The antihistamine Terfenadine was withdrawn for this reason (190).

Terfenadine is metabolised in the gut wall, so usually has a very low systemic concentration. It was found that when Terfenadine's rate of metabolism is decreased, through competition with or inhibition by other drugs, the increased concentration of Terfenadine in the blood stream led to a risk of cardiac arrhythmia. Investigation of Terfenadine's metabolites found that one of them – Fexofenadine (structures shown below in Figure 17) – was in fact the major active compound, while not exhibiting the adverse effects. Fexofenadine is now prescribed in place of Terfenadine, as a safe alternative (191).

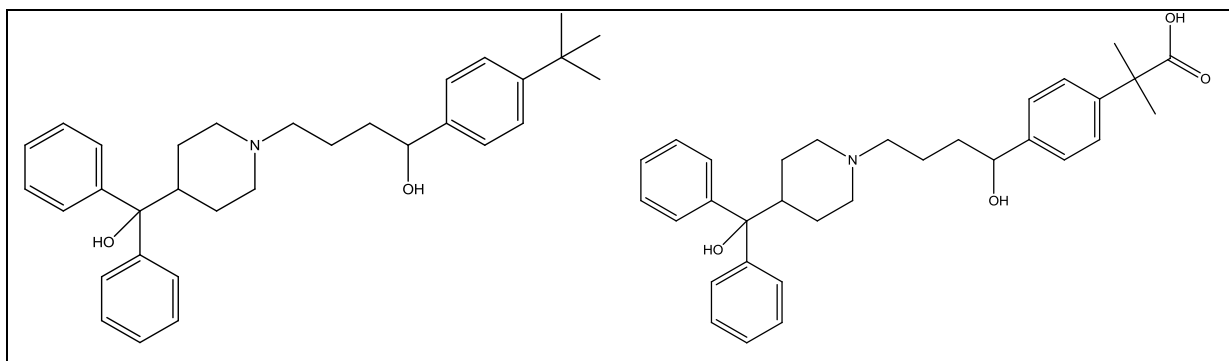


Figure 17: The antihistamine Terfenadine (left) has been withdrawn from sale due to safety considerations. Fexofenadine (right), one of Terfenadine's metabolites (the product of oxidation of a methyl group to carboxylic acid) was found to be the active compound and is now prescribed in place of Terfenadine.

2.2.2 Toxic metabolites

The most serious effect of xenobiotic metabolism is the formation of toxic metabolites. The most frequent reason for the withdrawal from market of an approved drug is drug-induced liver injury (192) and this is often found to be due to a metabolite, rather than the xenobiotic compound itself (193). If potential metabolites can be predicted, then these predictions can be linked with computer systems for the prediction of toxicity, of which a number are commercially available such as Derek (36) and TOPKAT (37,194), enabling *in silico* screens for such liabilities to be carried out (195,196).

There are various biological pathways through which metabolites can generate adverse drug reactions. Metabolites may exhibit pharmacological activity, which can be towards the same target as the parent drug, increasing the effects to those that would occur if the drug was administered at much higher concentrations, or may be off-target, affecting other systems in unintended ways. A further possibility is the formation of reactive metabolites which bind to other proteins and enzymes, or damage DNA. One mechanism through which this occurs is *via* the formation of reactive oxygen species, such as peroxides, oxides and oxygen radicals. Despite the body's mechanisms to deal with such toxins, these can lead to serious cell damage.

Not all toxicity is due to compounds' reactivity and activity. 'Non-specific' toxicology is the result of a general disruption of cell membranes and biochemical processes by a xenobiotic (197). As drug development extends into new areas of medication, particularly the use of

biological agents as pharmaceuticals, there is the possibility of other types of toxicity. Monoclonal antibody therapies carry a risk of over-stimulating the immune system, with serious consequences, as was recently the case in the widely reported adverse reactions during the 'first-in-man' trials of TeGenero's rheumatoid arthritis and leukaemia drug TBN1412 (198).

2.2.3 Prodrugs

The traditional approach to overcoming barriers to a drug's bioavailability has been to search for analogues of the drug – i.e. an alternative compound that delivers similar activity, but providing different pharmacokinetic properties. An alternative approach is the use of a prodrug (199,200), where chemical modification of a drug molecule, or the attachment of an extra moiety, renders the molecule inactive but allows it to overcome the barrier to bioavailability – conceptually similar to the use of a protecting group during an organic synthesis. Once the prodrug is absorbed, the moiety is removed by the organism's metabolic pathways, restoring the drug molecule to its active form. Recently there has been a growing trend in the development of prodrugs; approximately 15% of the new drugs approved in 2001 and 2002 were prodrugs, and they are now thought to comprise from 5-7% of the total drugs approved worldwide (201).

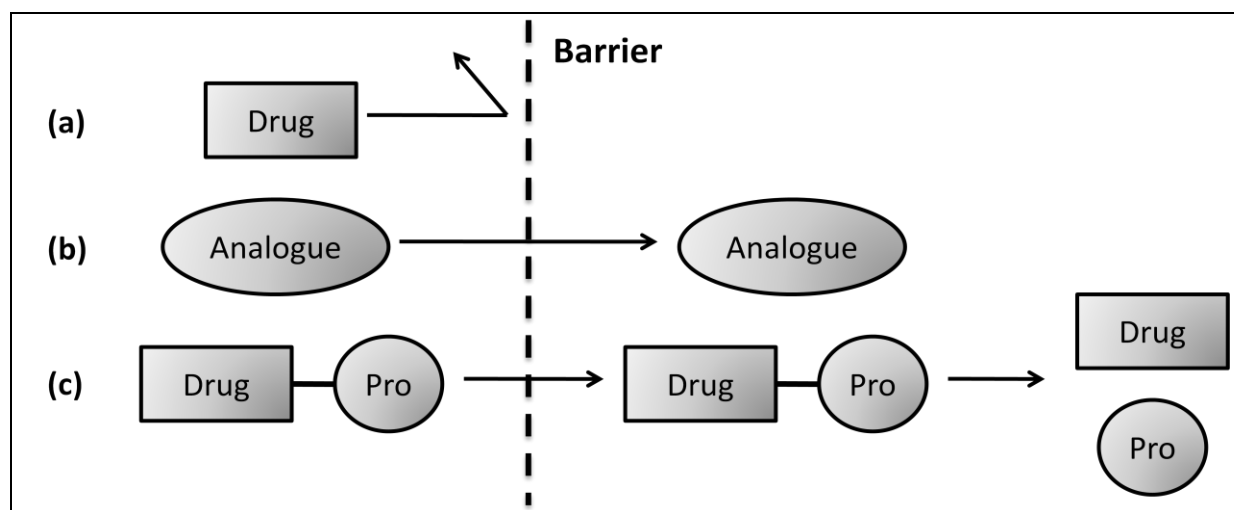


Figure 18: Illustration of the concept of prodrugs; in situations where a drug molecule cannot pass some barrier to bioavailability (a) there are two possible solutions; an analogue (b) may be found – that is an alternative compound exhibiting the required activity, but with different properties, or a prodrug (c) may be developed.

The prodrug approach can be used to overcome a number of barriers to bioavailability. The well known influenza drug Tamiflu (Oseltamivir ethylester) is in fact an ester prodrug of the active compound Oseltamivir carboxylate (see Figure 19). The modified compound is orally bioavailable and overcomes the poor intestinal mucosal permeability of the active drug (202).

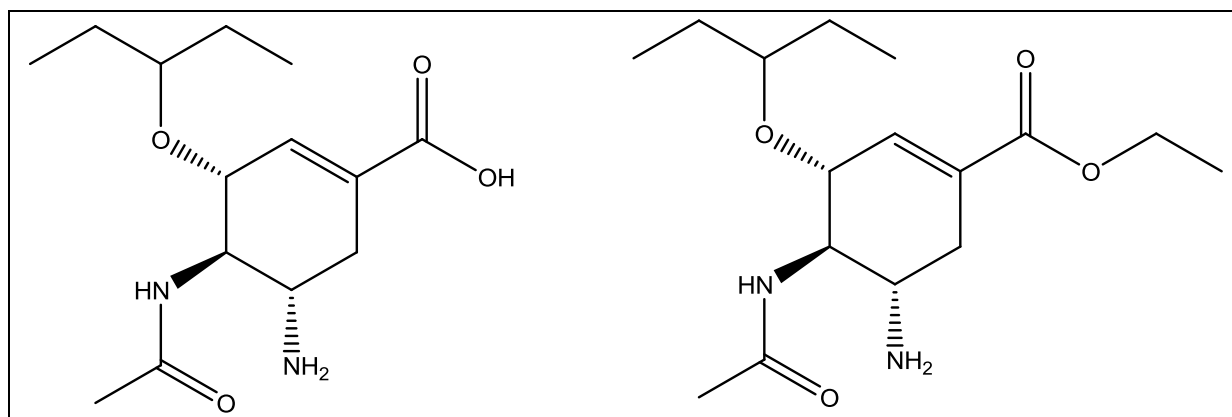


Figure 19: Oseltamivir carboxylate (left) an antiviral, and (right) its prodrug Tamiflu (Oseltamivir ethylester), which was designed to improve intestinal mucosal permeability.

The same prodrug approach can be undertaken for many other purposes (203), including improving solubility, aiding in the targeting of an active compound to a particular organ in the body and controlling drugs' rates of release.

2.2.4 CYP450 mediated drug-drug interactions

A common cause of adverse drug reactions has been found to be the modulation of one drug's metabolism by another. Inhibition of a drug's metabolic pathway by a co-administered pharmaceutical may lead to the drug accumulating to toxic levels, as in the case of Terfenadine, mentioned previously. Similar problems can arise when grapefruit juice or, to a lesser extent, red wine are consumed in combination with certain pharmaceuticals, since these contain flavonoids and other compounds which inhibit cytochrome P450 3A4 (204). Competition between drugs metabolised by the same enzyme can also reduce their rates of clearance. Alternatively, one xenobiotic can increase the rate of clearance of another, by induction of the enzymes in its metabolic pathway. This can lead to the drug's concentration falling below therapeutic levels, as can occur during co-administration of oral contraceptives with St John's Wort (188).

2.2.5 Metabolite elucidation

Studies into the metabolic fate of drug compounds are performed at several stages in the drug development process. During lead optimization, knowledge of the propensity of a drug candidate to undergo metabolic transformation can help identify candidate molecules with undesirable ADME characteristics and thus guide the selection of which compounds to commit substantial resources for further development (205). As development of the drug candidate proceeds any metabolites are investigated for signs of toxicity, alongside the parent compound.

A range of experimental techniques are employed for the investigation of a compound's metabolites, but structural elucidation without additional data on the metabolites likely to be formed is challenging. *In silico* systems can suggest metabolites unexpected or overlooked by human experts (206). Liver microsomal preparations, plasma and excreta from *in vivo* studies and extracts obtained from necropsy can all be examined for the presence of metabolites. Covalent protein binding assays are carried out to test for potential liver toxicity. High resolution liquid chromatography-mass spectrometry (LC-MS) is used to determine accurate masses of ions, which can be used to calculate the elemental composition of compounds. LC-MS reveals which fragment of a compound has undergone metabolism, but there can be several atoms within that fragment at which the metabolic transformation could be centred, and tools for predicting sites of metabolism can help resolve this ambiguity.

A further challenge is the identification of which components of complex biological mixtures are in fact metabolites of the compound under investigation. Labelling of the parent compound with radioisotopes such as tritium can facilitate this. However, radiolabelling experiments require the time-consuming and expensive synthesis of a labelled compound, and when this is not possible, prediction of potential metabolites is necessary.

Of particular difficulty is the experimental detection of reactive metabolites. Stable drug metabolites can be isolated, purified and identified using standard experimental techniques; however reactive metabolites are generally too short-lived for the same to apply. This is an area where *in silico* predictions of the structure of metabolites can be particularly useful.

2.3 Mechanisms of metabolism

Orally delivered drugs are first subjected to metabolism in the gastrointestinal tract. From here the parent compound and any metabolites formed are absorbed through the mucous membrane of the small intestine or through the wall of the stomach, and carried to the liver where they encounter further metabolic enzymes. Finally the remaining drug and its metabolites enter systemic circulation. Metabolism occurring before the drug has first entered systemic circulation is termed 'first-pass metabolism', and in some instances can reduce the bioavailability of a drug to such a degree that alternative routes of administration are required. As a protein, insulin, administered to Type I diabetics, is catabolised in the gastrointestinal tract so must be administered through subcutaneous injections in order to reach circulation without degradation. The corticosteroid beclometasone dipropionate, until recently administered to asthmatics, does not enter the blood stream in detectable levels when taken orally due to its high rate of clearance through first-pass metabolism (207), so is instead administered as a nasal spray (Beconase) or through an inhaler (Becotide).

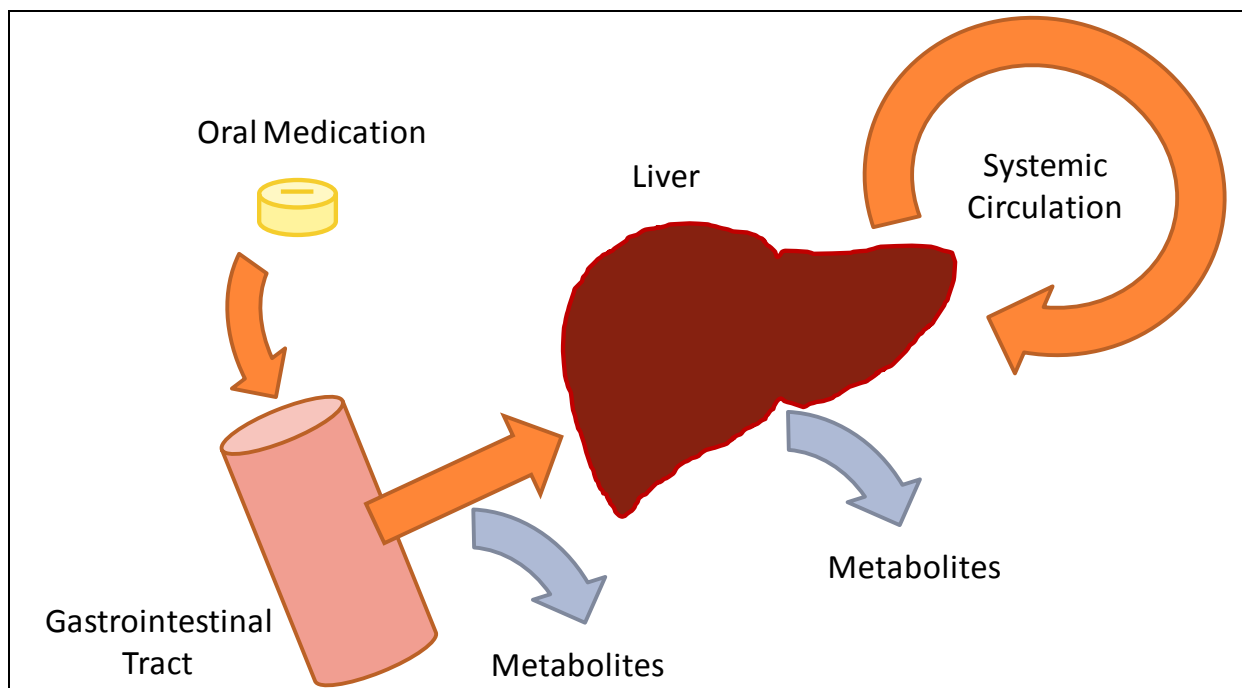


Figure 20: First pass metabolism: drugs encounter metabolising enzymes in the gastrointestinal tract and the liver, before entering systemic circulation. This can considerably decrease the bioavailability of the drug.

There are a wide variety of pathways and mechanisms through which xenobiotic compounds are metabolised (184). Many compounds can be metabolised through a number of competing pathways, leading to the generation of a variety of metabolites, in differing concentrations. The degree of formation of each metabolite depends on a number of factors, such as the availability of enzymes and cofactors, and competition with other xenobiotics. This means that the metabolic profile of a drug varies with both environment and genetics. A drug's metabolism can differ between species, between individuals of the same species but different gender and age, and even within one individual at different instances in time.

The pathways through which Phase I metabolism occurs are generally divided into those involving cytochromes P450 (CYP450), and those which do not. Cytochromes P450 are a large family of enzymes, involved in the majority of drug metabolisms. A study carried out by Pfizer examining the top 200 drugs prescribed in the United States in 2002 found that cytochromes P450 were involved in two-thirds of the metabolic clearance pathways (208). The human genome project has identified 57 CYP450 genes (209), which give rise to a variety of different CYP450 enzymes, known as isoforms. Each isoform can bind to a number of substrates. Some are very promiscuous, metabolising a wide variety of molecules.

Cytochromes P450 can catalyse a range of reactions, some examples of which are illustrated in Figure 21, below. The most common cytochrome P450 catalysed transformation is monooxygenase hydroxylation, inserting a single oxygen atom into an R-H bond, producing R-OH (210). This can be the final product of the transformation, or may lead to a dealkylation, as in the case of the oxidative deamination reaction shown below. Cytochromes P450 can also oxidise heteroatoms, such as nitrogen, and form epoxides of alkenes (209).

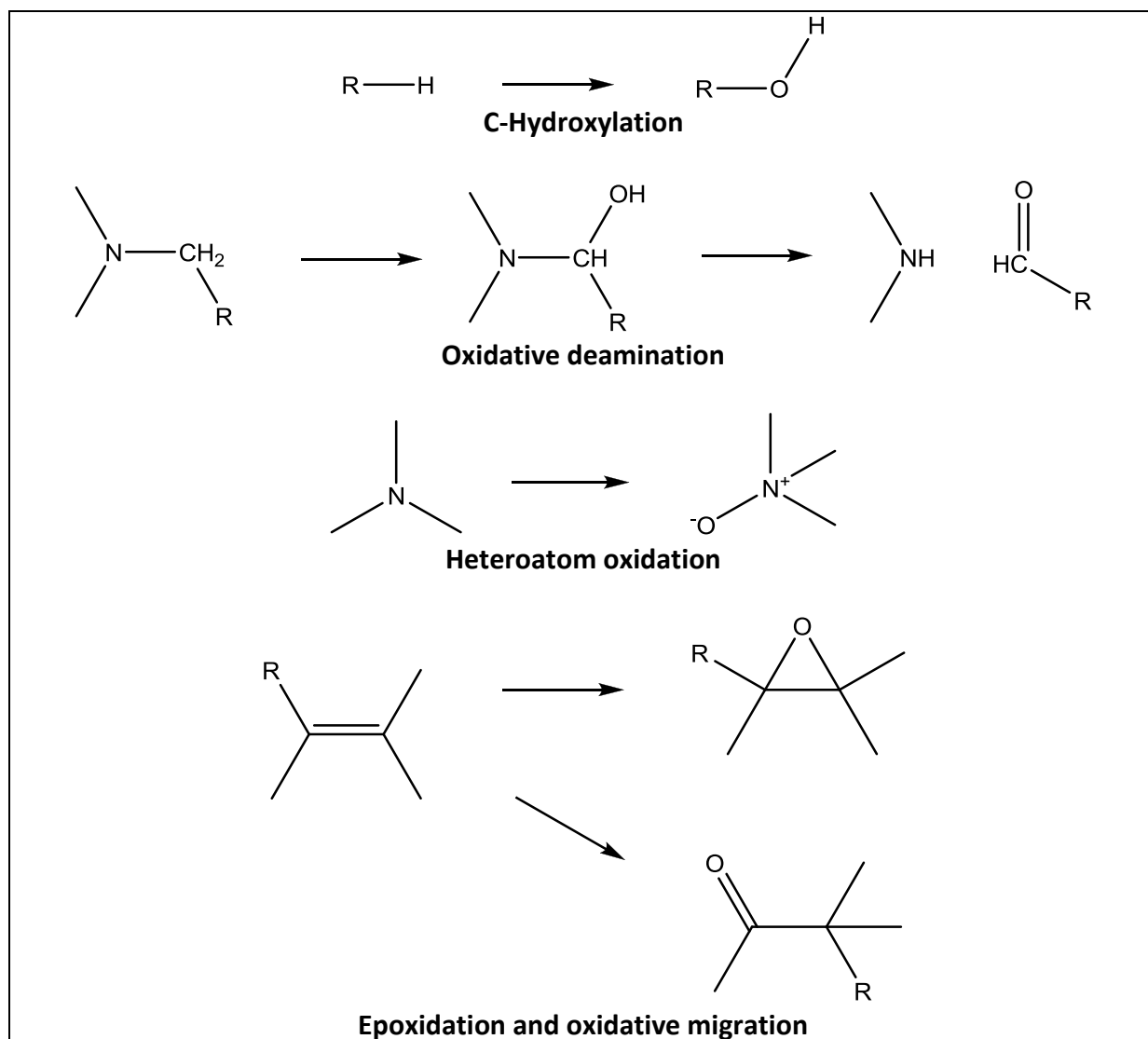


Figure 21: Commonly observed CYP450 mediated metabolic transformations.

The active site of cytochromes P450 contains a heme group (Figure 22). The catalytic cycle of CYP450 metabolism involves the binding of an oxygen molecule to the iron atom at the centre of the heme group. Reduction of the oxygen molecule, with the release of water, and a single electron transfer lead to the formation of an oxygen radical. This radical can react with the enzyme's substrate in a number of ways, leading to a variety of potential products. An overview of the catalytic cycle and more detailed mechanisms for a number of transformations are shown in Figure 23 and Figure 24.

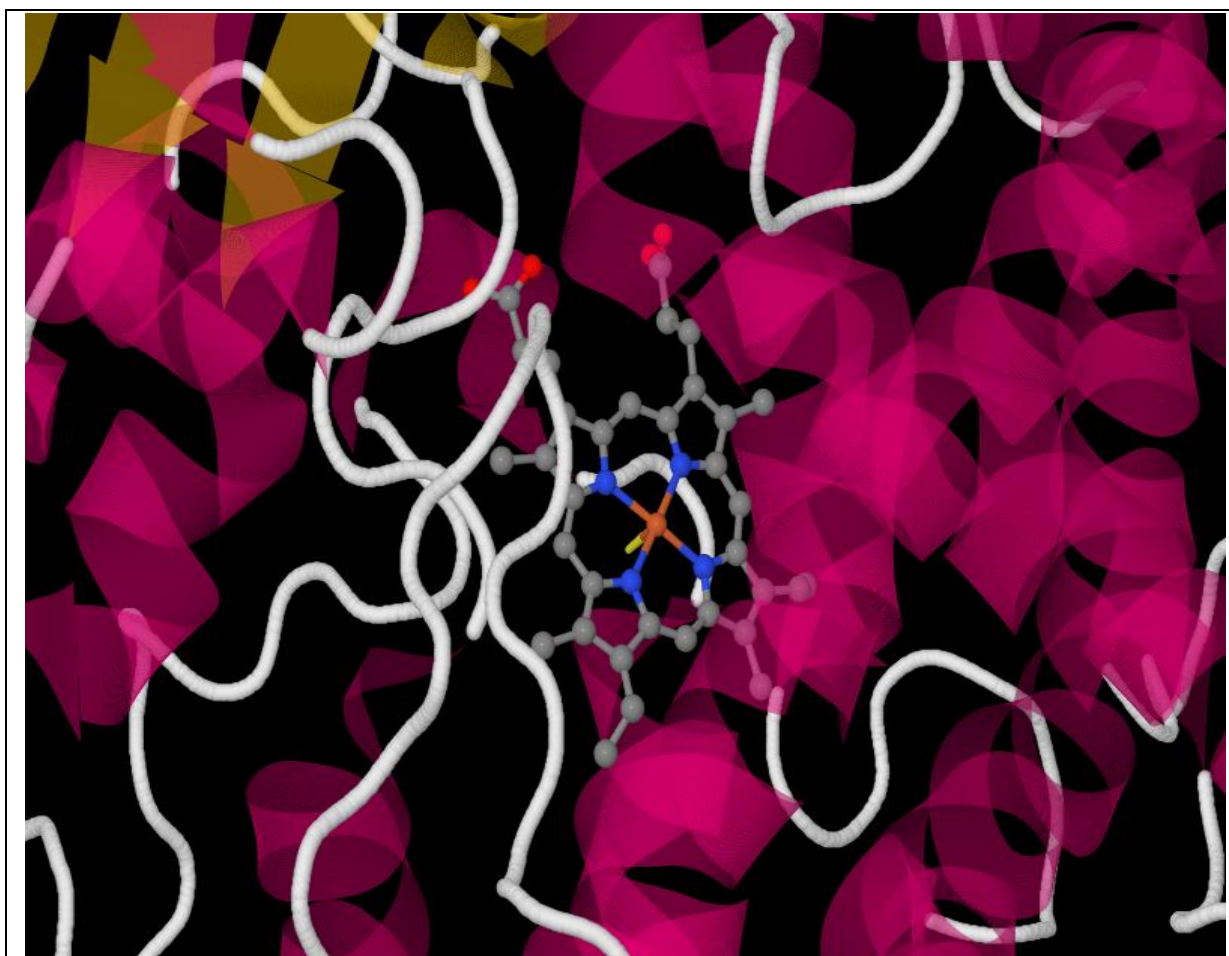


Figure 22: The heme group at the catalytic site of cytochrome P450 2C8; PDB ID: 2VN0 (211).

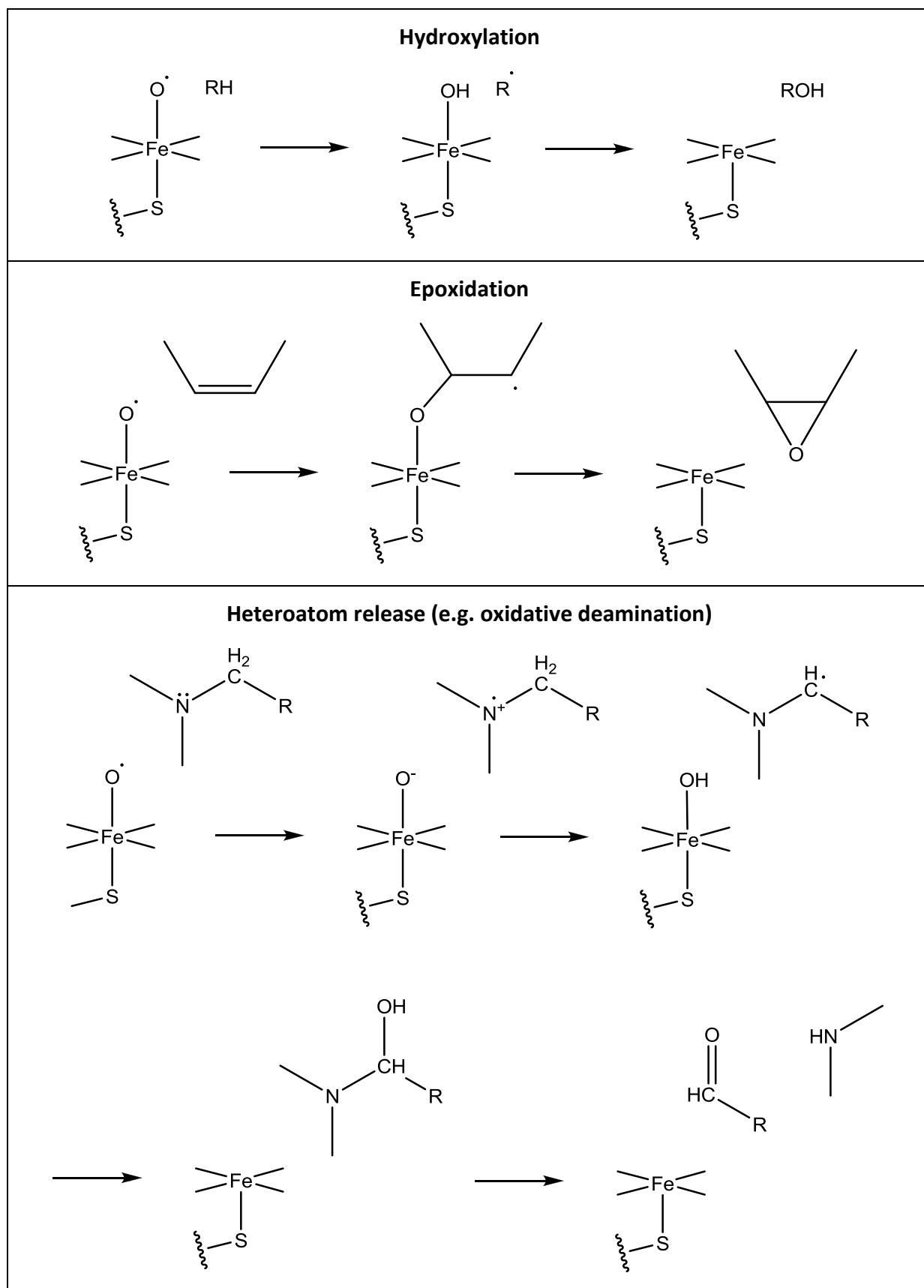


Figure 23: Mechanisms of some cytochrome P450 catalysed transformations (212).

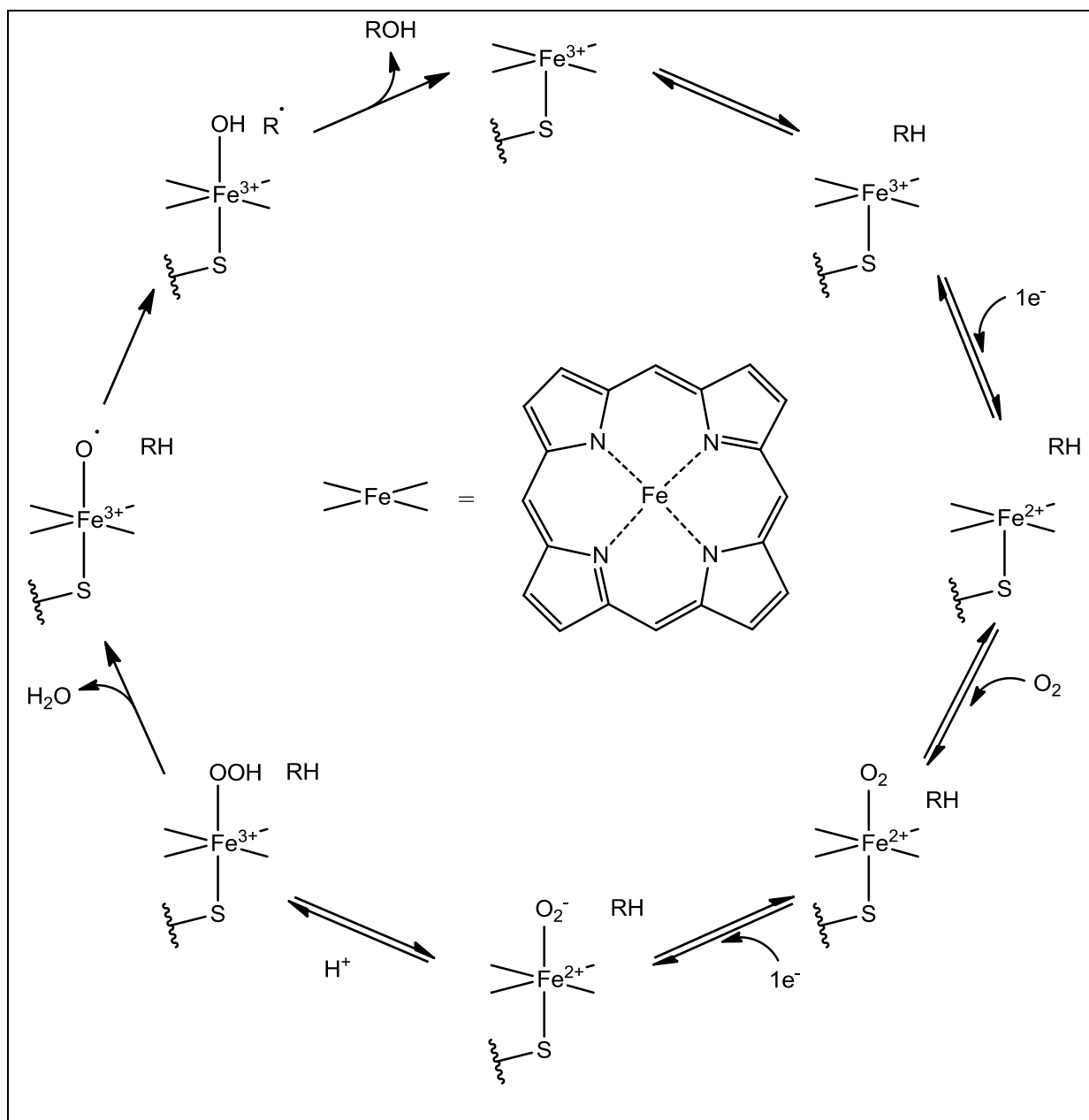


Figure 24: The basic catalytic cycle of CYP450 oxidation (209). First (top right), the substrate (RH in the figure above) binds close to the heme group in the CYP450 active site, displacing water. Oxygen binds to the heme group, and is activated by a single electron-transfer, resulting in the generation of a highly reactive Fe(V) oxo species. This reacts with the substrate, inserting an oxygen atom into the R-H bond, leading to the addition of a hydroxyl group.

The regioselectivity of CYP450 catalysed transformations – the atom or atoms where the reaction occurs – is determined by a number of factors: the energy required to remove a hydrogen atom from the substrate (hydrogen abstraction energy) and the stability of the resulting carbon radical, and also the structure and shape of the binding pocket of the

specific CYP450 variant catalysing the transformation. Each CYP450 isoform has a different sized and shaped binding pocket, which exhibits a different binding motif, favouring affinity to substrates displaying particular structural features.

While cytochromes P450 catalyse many metabolic transformations, a number of other enzymes are also involved in phase I metabolism (213). A number of enzymes facilitate oxidation reactions; Flavin-Containing Monooxygenase (FMO), Monoamine Oxidase (MAO) and Aldehyde Oxidase are among those commonly found. Although less common than oxidations, reductions often occur in metabolic pathways, and are catalysed by a number of enzymes including the cytochromes P450. Epoxide Hydrolases, Esterases and Amidases catalyse the hydrolysis of epoxides, esters and amides, respectively.

Phase II transformations are also catalysed by a number of different enzymes. Common phase II conjugates are glucuronic acid, glutathione and sulphate, though conjugation to many other molecules, including macromolecules such as proteins, DNA and RNA is also possible.

2.4 Predicting xenobiotic metabolism

Due to the interest in understanding the metabolism of xenobiotics, a considerable amount of effort has been put into the development of predictive tools. There are a number of different goals when predicting xenobiotic metabolism: identifying sites of metabolism, predicting the metabolites formed, predicting rates of metabolism and metabolite formation, and predicting the cytochrome P450 specificity of substrates. As might be expected, with a range of aims, a number of diverse approaches have been developed, with different techniques better suited to certain types of prediction.

The enzymes involved in xenobiotic metabolism are fairly nonspecific towards their substrates. If this were not the case then a vast array of different metabolic enzymes would be needed by an organism. The upshot of this is that enzyme specificity has less effect on the enzyme activity than is the case in many other receptor mediated systems, so ligand-based tools are often found to perform well. Since cytochrome P450 mediated oxidation is the most common route for xenobiotic metabolism many tools consider only this system.

2.4.1 Site of metabolism prediction

The atoms of a xenobiotic at which metabolic transformations are centred are termed its 'sites of metabolism'. Identification of the likely sites of metabolism of a drug enables medicinal chemists to design modifications to the drug's molecular structure in order to prevent its metabolism at that site. Blocking the major sites of metabolism of a drug can reduce the rate of first-pass metabolism, allowing it to enter systemic circulation at high enough concentrations to exhibit a therapeutic effect and can reduce the clearance rate of the drug, extending the time required between doses.

As trans-membrane proteins, it is difficult to produce crystal structures of mammalian cytochromes P450. Indeed it is only in the last decade that X-ray structures for CYP450s from humans and other mammals have been published (214,215,216). Prior to this, mammalian CYP450 structures could only be approximated through homology modelling based on crystal structures of soluble bacterial CYP450s e.g. (217).

Quantum mechanical methods/reactivity calculations

The mechanism of many cytochrome P450 catalysed biotransformations include a hydrogen abstraction step, where the heme-bound oxygen radical of the CYP450 active site removes a hydrogen atom from the substrate, creating a carbon radical. It is believed that this is the rate determining step of the transformation (218), and that the likely sites of metabolism can be predicted from the ease with which hydrogen abstraction can take place at each atom of the substrate.

Calculations using the AM1 semi-empirical quantum mechanical method have been performed on a number of radicals formed by hydrogen abstraction, and their parent drug compounds (219,220). In most cases the calculated radical stabilities showed good agreement with experimental bond dissociation energies. However, such quantum mechanical calculations are time-consuming, even if only a single energy minimized conformation of the drug molecule and each of the hydrogen abstracted radicals is considered.

Olsen *et al.* (221) have investigated the hydrogen abstraction energies of 24 substrates in a model CYP450 system using state of the art Density Functional Theory (DFT) calculations. They have also used this approach to study specific classes of transformations in detail:

aromatic oxidations (222) and sulfoxide, sulphur and nitrogen oxidation and dealkylation (223), and the structure of the CYP450 heme complex (224). These calculations can require days or weeks of CPU time for molecules the size of a typical drug, making them unsuitable as a tool for regular screening. However, these results have been used to establish a hierarchy of methods, from visual inspection of functional groups through semi-empirical calculations to DFT, which can be selected depending on the accuracy required and the complexity of the molecule under consideration. A rule-based method, derived from the high-level DFT calculations, estimates activation energies at different sites in a molecule relatively well (225), but with currently only eleven rules, this approach is not able to provide any discrimination between similar sites.

In a recent comparison of site of metabolism prediction tools (218) Afzelius *et al.* explored the possibility that the CYP450 catalysed biotransformations proceed via an alternative mechanism. It has been proposed that a single electron could be transferred from the substrate to the heme of the CYP450, creating a positively charged radical that reacts with either the heme/iron/oxygen complex or a neighbouring water molecule. According to this mechanism metabolism will be centred on the location of the spin 'hole' on the radical substrate, and the spin distribution can be estimated through quantum mechanical calculation.

Pharmacophores

Some cytochrome P450 families have been found to exhibit a pharmacophore that determines the orientation of the substrate in the active site (226). Through alignment of substrate molecules with this pharmacophore it can be predicted which atoms will be positioned near the heme group in the active site, and hence undergo metabolic transformation.

Docking methods

Various docking methods have been used to predict sites of CYP450 mediated oxidation. Afzelius *et al.* (218) made predictions using the Dock (227) and Glide (124) programmes (techniques they termed MetaDock and MetaGlide, respectively). Vasanthanathan *et al.* (228) have predicted sites of metabolism for cytochrome P450 1A2 ligands using GOLD (121).

These methods have all taken a common approach. The docking algorithms generate an ensemble of docked poses with varying substrate conformations and orientations. These poses are filtered, identifying any in which atoms are sufficiently close to the heme reactive centre for metabolism to be possible. The atoms are then scored on the basis of their distance from the catalytic site, and the energy of the binding pose or other scoring function of the docking algorithm.

Although only a small number of CYP450 isoforms have had their structures solved by X-ray crystallography, structures for some additional isoforms can be obtained through homology modelling, due to their degree of sequence similarity to solved structures, and these can be used for docking purposes (229).

QSAR

In an attempt to make faster predictions, a number of QSARs for metabolism prediction have been developed. Because even semi-empirical calculations take a considerable length of time, Singh *et al.* (230) have used the results of AM1 calculations on 50 known CYP450-3A4 substrates to generate a PLS QSAR model for the hydrogen abstraction energy. This enables fast estimates of the hydrogen abstraction energy, based on the local chemical environment of the hydrogen atom. They also added a sterically accessible solvent surface area requirement for substrate binding to the active site. While this model showed some predictivity, they found that it was “unable to predict the major site of metabolism in an appreciable number of cases”, and showed some systematic errors, notably the calculated dehydrogenation energy always suggesting that the piperidine ring carbons adjacent to nitrogen of N-methylpiperidines is the likely site of metabolism while CYP3A4 has almost always been observed to oxidise the methyl groups (231).

Besides predicting sites of metabolism, QSARs have also been developed to predict other aspects, such as rates of clearance (232).

Enzyme/substrate interactions

With the availability of structures of cytochromes P450, various techniques of predicting metabolism by examination of enzyme/substrate interactions have been developed. Both Molecular Interaction Field (MIF) (80) and Receptor Interaction Surface (RIS) methods have been investigated. In the MIF approach, a probe is positioned at regular intervals in a box

surrounding the active site and its interactions with the protein are calculated. In the RIS approach, rather than being positioned at grid points throughout the space containing the receptor, the probe is placed at regular points across the receptor's surface.

MetaSite

MetaSite bases its predictions on Molecular Interaction Field analysis. MIFs (80) are pre-calculated for the enzyme active site using the GRID force field and four types of probe – hydrophobic, hydrogen-bond acceptor, hydrogen-bond donor and charged. The distance of each probe position from the catalytic site – the oxygen bound to the heme group – is calculated, and the distribution stored (233). Each atom in a potential substrate molecule is assigned to one or more of the probe classes and an ensemble of conformers are generated and minimized. The distance distribution of probe types is calculated around each atom, and the complementarity between the active site and each atom determined. This provides a score for the fit of the substrate into the active site, with that atom at the catalytic site.

A reactivity score can optionally be used to weight the results of the interaction calculations. This is based on *ab initio* calculations of the hydrogen abstraction energy, but rather than computing this for each substrate compound, abstraction energies for small fragments, common to many drug-like molecules, have been pre-computed, and reactivity scores are generated by matching the most relevant fragment. Zhou *et al.* (234) reported that the inclusion of the reactivity weighting increased the accuracy of the predictions considerably: from an average of 30% to 60% that the highest ranked atom was a site of metabolism, and from 40% to 70% that one or more of the three highest ranked atoms was a site of metabolism.

Data mining

Boyer and Zamora (33) proposed a method of data-mining to the prediction of sites of xenobiotic metabolism. The Symyx® Metabolite database (at the time, the MDL Metabolite database) is widely used by chemists investigating whether a substructure is involved in any sorts of metabolic transformation. Boyer and Zamora generated small atom-centred fragments including the neighbouring 3-4 atoms and ring systems, and searched for transformations involving these fragments within the Metabolite database. Counting the number of occurrences of these transformations within the Metabolite database, along with

the total number of occurrences of the substrate fragment within the database, allows the calculation of an occurrence ratio, giving a probabilistic score for the likelihood of the transformation taking place. Boyer *et al.* (235) automated this process through the generation of fragments centred on each atom in a compound under consideration using circular atom environment fingerprints (60).

2.4.2 Metabolite prediction

A number of tools for the purpose of predicting the metabolites formed, rather than just the sites of metabolism of molecules, have been published. A number of these are now described; all follow a fairly similar approach, describing potential transformations using rules, and searching a molecule for sites where each rule matches.

META

META (236,237,238) has two dictionaries of transformations – one of CYP450 transformations and a second of spontaneous transformations. Each transformation consists of a target fragment and a product fragment. A prediction is made by identifying any occurrences of the target fragments in an input molecule, and substituting them with the corresponding product fragment. An example CYP450 fragment pair would be “replace occurrences of ‘N-CH₂’ with ‘N-CH-OH’” – meaning hydroxylate aliphatic carbons α to a nitrogen atom. Each CYP450 product is then processed with the spontaneous reaction transformations, until no further target fragment matches are found. In cases where tautomers are formed, a quantum mechanical calculation is performed to identify which is the most stable tautomeric form.

Experts have assigned each transform a priority value, according to the prevalence of the observed metabolites. This is based on a combination of data from “any mammalian source”, so the model represents an “average mammal” (237). If the rules were not prioritized then a combinatorial explosion of metabolites could be generated. As the number of transformation rules increased, accurately deciding on this prioritization was found to be challenging, and a genetic algorithm was utilised to optimise the priorities.

As of 2002, META contained over 750 transformation rules, developed from pharmacological data on around 150 xenobiotics. This included 43 transformation rules for

CYP450 aliphatic hydroxylation, 28 transformations modelling dealkylation of aliphatic and aromatic ethers and 30 dehalogenations (237).

Meteor

Meteor (239,240), developed by Lhasa Ltd., operates on a similar principle to META. Meteor's prediction engine contains two sets of rules. In order to predict a compound's metabolites, Meteor first applies a set of biotransformation rules, encoded by human experts, describing possible transformations, and the features of a molecule required to make that transformation permissible. Meteor allows the expression of sophisticated biotransformation rules, or 'biophores', such as "a single or double bond in a five- or six-membered ring, but not fused to another ring", rather than simply encoding functional groups.

Where many competing transformations could apply to a molecule, a system of reasoning rules (such as "benzylic oxidation is more likely than ring oxidation") is applied in order to determine the most likely transformations. The reasoning rules were developed through computational analysis of experimental data, in order to determine priority of the different biotransformation rules. As of 2002, Meteor contained 217 biotransformations, together with 841 reasoning rules (240) and by 2005 the knowledgebase had grown to more than 300 biotransformations and over 1000 reasoning rules (241).

Once potential metabolites have been identified, Meteor assigns likelihoods to each, using rules associated with each biotransformation, depending on the logP value of the substrate molecule. In order for a biotransformation to take place a substrate must have a logP value that allows it to enter and leave lipid membranes, and enough hydrophobic regions to facilitate enzyme binding.

Syigma

Another rule-based tool for the prediction of metabolites is Syigma (242) (Systematic Generation of potential Metabolites), developed and used in-house at Organon (now Schering-Plough). Syigma's rule-base was developed through the refinement of an initial set of very broad rules, such as 'oxidation of primary alcohol' and 'O-glucuronidation'. These rules were refined through a series of iterations, in each step of which more general rules were split on the basis of their performance; for example the general rule for the oxidation

of primary alcohols was divided into two separate rules for aliphatic and aromatic primary alcohols. Rather than the small number of likelihood classes employed by Meteor, Sygma assigns more finely grained likelihoods to its predictions. Associated with each rule is an empirical probability score calculated from the performance of that rule against a set of 6187 known metabolic reactions in humans.

MetaDrug

A further rule-based tool for predicting metabolism is GeneCo's MetaDrug (243). Rules describing 65 metabolic pathways were developed. QSAR models were constructed (for the 23 reactions with sufficient data) through kernel-partial least squares (K-PLS) analysis of 317 molecules randomly extracted from the MetaDrug database (244), and these are utilised to filter and prioritize the generated metabolites.

Microbial catabolism

A rule-based approach, very similar to that of Meteor, has been applied by Hou *et al.* (245) to the prediction of the biodegradation of chemicals in the environment by microbes. Over 1000 curated biotransformations from almost 200 metabolic pathways recorded in the University of Minnesota Biocatalysis/Biodegradation Database (UM-BBD) (246,247) have been used for the construction of a set of biotransformation rules. These rules, compiled by biotransformation experts analysing the UM-BBD, each consist of a SMARTS pattern matching a functional group, and the biotransformation which that group undergoes (e.g. aldehyde to carboxylic acid).

With around 200 rules, there are many possible transformations that could be applied to most molecules, leading to a combinatorial explosion in the number of predicted products. This leads to a high rate of false positives – predicted transformations that are not observed in nature. In order to overcome this, each of the rules is assigned to a likelihood group (very likely, likely, neutral, unlikely, very unlikely). A system of relative reasoning rules was also developed through analysis of all pair-wise occurrences of rule hits within the training data, and whether compounds triggered one or both of the rules (248). Together these are used to prioritize and filter rule hits.

2.5 Conclusion

In this chapter we have reviewed the importance of understanding the metabolism of xenobiotic compounds. Metabolic transformations affect a drug's efficacy and toxicology, so it is vital that any metabolic liabilities are identified. If metabolic transformations can be anticipated then modifications can be made to overcome any liabilities, and the body's metabolic systems can even be exploited, with the development of prodrugs.

A wide variety of methods have been employed for the prediction of xenobiotic metabolism, though many focus only on cytochrome P450 catalysed transformations, ignoring other mechanisms. Of the methods described, only a small number are publically available and straightforward for a chemist to use. Some are in-house tools, only accessible to workers in the company that developed the tool. Others, particularly the docking and quantum mechanical approaches, require complex calculations to be performed, and analysis of the results is complex, and these are typically only useable by experts. Of the tools that are generally available, most are commercial offerings.

The remainder of this thesis describes the development and evaluation of MetaPrint2D, a new tool for predicting sites of xenobiotic metabolism, and its extension to the prediction of types of transformation and metabolites likely to be formed.

3. Development of MetaPrint2D: a tool for predicting sites of xenobiotic metabolism

This chapter describes the development of MetaPrint2D – a tool for predicting sites of xenobiotic metabolism, based on the previously published Substrate/Product Occurrence Ratio Calculator (SPORCalc) (33,235). The initial goal of this work was to perform a more extensive evaluation of the SPORCalc program than had previously been carried out, and to remove the non-free dependencies (OEChem (249) and CORINA (85)), making the tool more readily distributable.

As work progressed, it became apparent that being based on fingerprint similarity techniques, which generally offer very high performance, the SPORCalc method had the potential to form the basis of a site of metabolism prediction tool that was fast enough for a chemist to work with it in an interactive manner. Unfortunately limitations in the SPORCalc software's architecture meant that this performance could not be realised, so the decision was taken to develop a new tool – MetaPrint2D. Additionally, a number of modifications to the method have been developed that could potentially improve accuracy, and the effects of these have been investigated.

In this chapter the SPORCalc approach to metabolic site prediction is reviewed, and the available data on metabolic transformations from the Symyx® Metabolite (250) database examined. The development of MetaPrint2D and the software distribution available are then described.

The next chapter presents the method and results of the evaluation of MetaPrint2D, and Chapter 5 describes the extension of MetaPrint2D to the prediction of types of transformation and metabolites formed.

3.1 Substrate/Product Occurrence Ratio Calculator

As was briefly discussed in the previous chapter, the Substrate/Product Occurrence Ratio Calculator (SPORCalc) is a data-mining tool, designed to exploit the biotransformation data recorded in the Symyx® Metabolite database, in order to generate structure-metabolism

relationships. SPORCalc introduced the use of knowledge-based statistical modelling to site of metabolism prediction. An overview of the SPORCalc procedure is given in Figure 25.

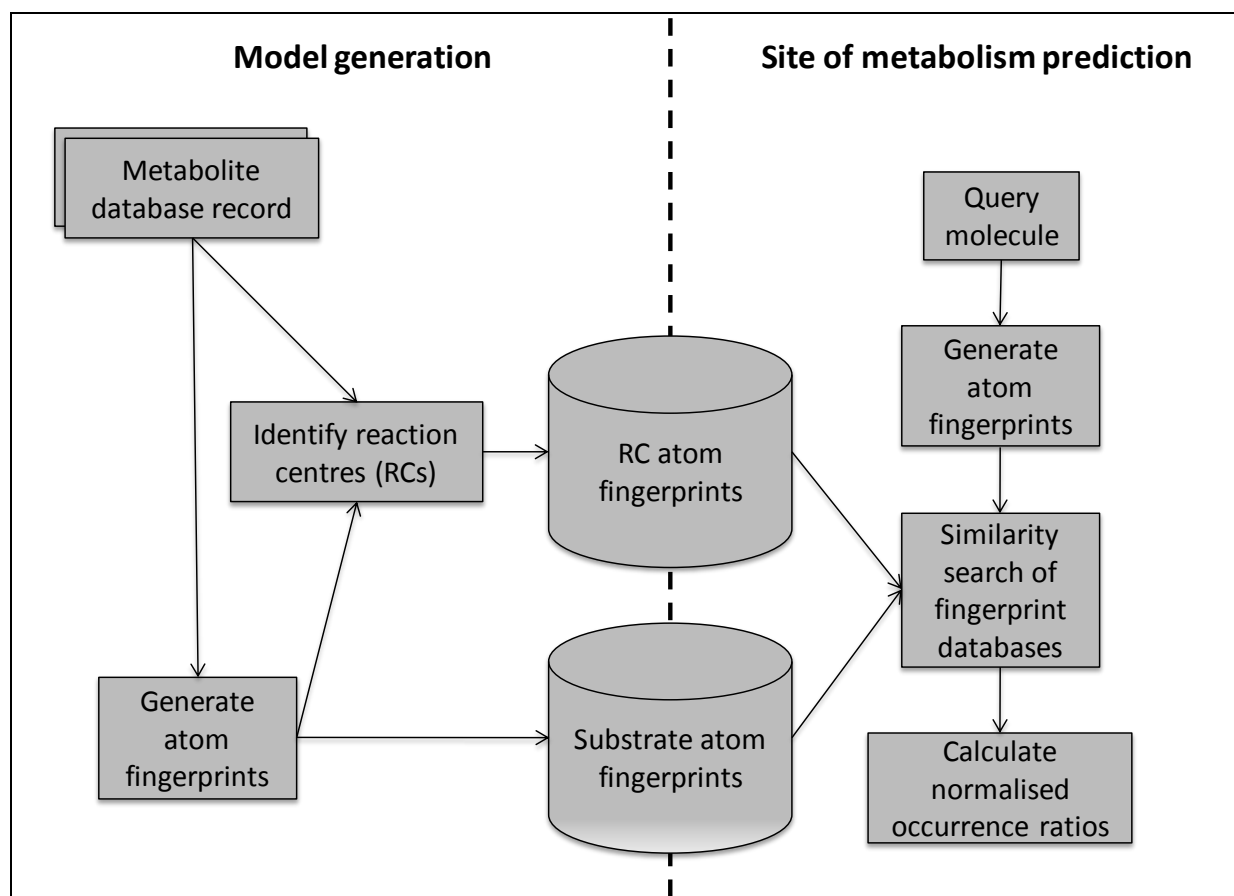


Figure 25: Overview of the SPORCalc procedure for site of metabolism prediction.

SPORCalc contains two databases of atom fingerprints: one listing the fingerprints of all the substrate atoms found in transformations contained in the Symyx® Metabolite database (the Metabolite database is described in more detail on page 75), and a second listing the fingerprints of only those atoms occurring at a reaction centre (site of metabolism). In order to investigate the sites of metabolism of a novel compound, SPORCalc generates fingerprints describing each of the atoms, and performs a similarity search against these two databases. This enables calculation of occurrence ratios – the ratio between the number of reaction centre atoms in the Symyx® Metabolite database that occupy a similar chemical environment to each atom in the query structure, and the total number of occurrences of atoms in a similar chemical environment in the entire database.

While the SPORCalc program was envisaged as a data mining tool, its output is effectively a prediction of the likely sites of metabolism of a molecule, since the calculated occurrence ratio is equivalent to the calculation of a conditional probability:

$$\begin{aligned}
 P(S|E) &= \frac{P(S \cap E)}{P(E)} \\
 &= \frac{N(S \cap E)/N_{tot}}{N(E)/N_{tot}} \\
 &= \frac{N(S \cap E)}{N(E)} \\
 &= \text{SPORCalc Occurrence Ratio}
 \end{aligned}$$

where:

$P(S|E)$ is the conditional probability that an atom is a site of metabolism, given the environment it occupies.

$P(S \cap E)$ is the probability that an atom is a site of metabolism and occupies the specified environment.

$P(E)$ is the probability that an atom occupies the specified environment.

$N(\dots)$ is the count of the number of atoms meeting the specified condition.

N_{tot} is the total number of atoms in the database.

Once calculated, the occurrence ratios are normalised, so that the highest scoring atom always has a score of one. This *normalized occurrence ratio* indicates the relative likelihood of each atomic site in a molecule being a centre of metabolism, while making no prediction as to the absolute likelihood of the molecule undergoing metabolic transformation.

Apart from the normalization step, this is a similar calculation to that performed by a Naïve Bayesian classifier. The major difference is that a Bayesian classification would consider both the likelihood that an atom is at a site of metabolism, given its environment, and the likelihood that it is not. A Bayesian classifier would usually report the likelihood ratio (LR):

$$LR = \frac{P(S|E)}{P(!S|E)}$$

Where a likelihood ratio greater than 1.0 would indicate that, given its environment, it is more likely that the atom is at a site of metabolism than it is not, and a likelihood ratio of less than 1.0 would indicate the opposite.

SPORCalc represents the chemical environments occupied by atoms using circular atom environments fingerprints (described on page 15, in Chapter 1) with Tripos' SYBYL® (46) atom types (251). Fingerprints of depth six – the central atom, and topological neighbours up to five bonds distant – are generated. Each layer of the fingerprint contains 33 bins, one for each SYBYL® atom type. Each bin holds a count of the number of occurrences of the respective atom type in that layer. This leads to fingerprints with a total of 198 bins.

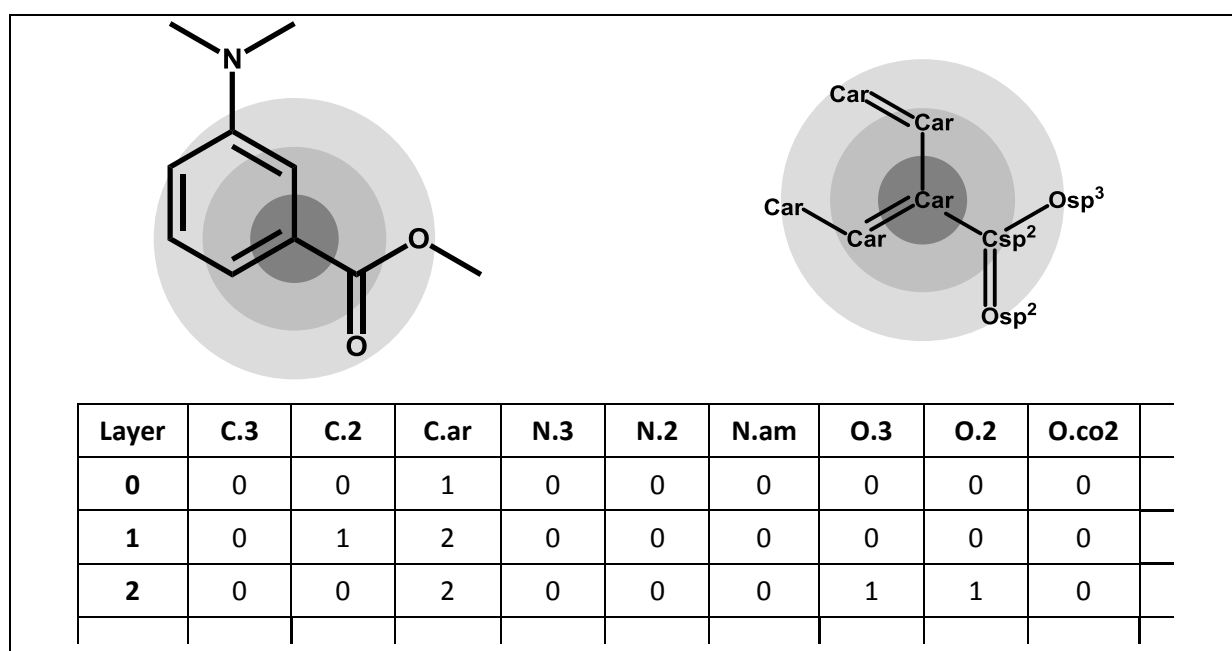


Figure 26: Illustration of the region of a structure forming an atom environment fingerprint (first three levels highlighted), and the contents of a subset of fingerprint's bins.

The SPORCalc package* consists of two separate programs: the database builder, which constructs the fingerprint databases with which SPORCalc performs its calculations, and the calculator which accepts a query molecule, input through a web interface, and generates a web page displaying the predicted sites of metabolism of the molecule.

* We gratefully acknowledge Lars Carlsson and AstraZeneca for sharing with us the latest version of SPORCalc.

3.1.1 SPORCalc databases

A SPORCalc model requires two databases of atom fingerprints; one containing the fingerprints of all the reaction centre atoms in the training data, and a second containing the fingerprints of all the atoms in the training data. The fingerprints are stored in text files, as a space separated list of integers – with one value for the occupancy of each of the 198 bins making up the six-level fingerprint. The reaction centre fingerprints are stored in a single file. Due to their greater number, the substrate fingerprints are split into separate files, one for each type of central atom.

[illegible]

Figure 27: A section of a SPORCalc reaction centre fingerprint file.

3.1.2 Site of metabolism calculator

The site of metabolism calculator is written mostly in Python, with a C++ program used to carry out the computationally expensive fingerprint similarity searching. The Python code is designed to run as CGI scripts on a web server. C++ was used in place of Python for the fingerprint searching since being a compiled rather than an interpreted language it is often much faster.

The workflow of the SPORCalc calculator is shown in Figure 28, below. The calculator takes an input molecule, using the SMILES (252) representation, and runs the CORINA (85) program to generate a PDB file containing a 3D structure of the molecule. This step acts to check that a valid SMILES has been specified. The calculator then uses the OEChem (249) library to load the SMILES, remove any hydrogen atoms, since they are not used in the calculation, and generate the tree-structure of the fingerprints. The molecule, with hydrogen atoms removed, is written to an MDL molfile, and OpenBabel (253) is used to convert this to a MOL2 file, from which the atom type assignments are read. CORINA is run again to generate a PDB file of the structure without hydrogen atoms, which is used to display the results. The fingerprints, the selected database and parameters for the

each of the atoms in the same manner as the calculator program. The fingerprints are written to the reaction centre and substrate data files, as appropriate. An overview of the database builder's workflow is shown in Figure 29.

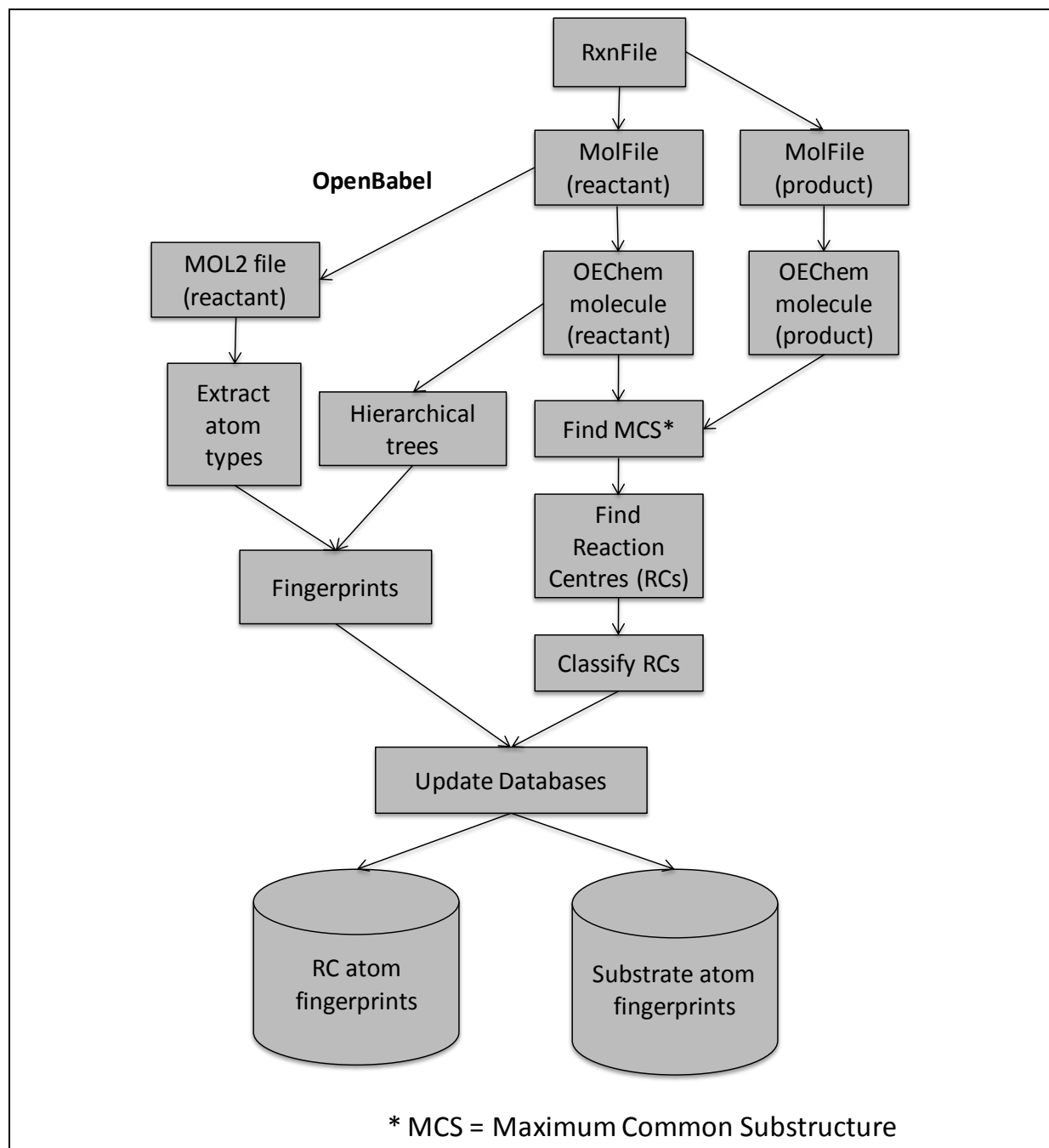


Figure 29: Overview of the SPORCalc database builder's workflow.

Reaction centre identification and classification

Reaction centres are identified through the comparison of the substrate and metabolite structures, detecting added and eliminated atoms and bonds, and changes to bond order.

SPORCalc

SPORCalc uses a process of up to four stages to identify the correct mappings between the atoms and bonds of the substrate and metabolite structures. In each of the first three stages of searching, putative mappings generated by maximum common substructure (MCS) searches carried out using the OEChem toolkit (249) are compared with the annotated mappings from the Symyx® Metabolite database. If none of the mappings is found to match the database's annotations then the search is repeated with increasingly strict matching criteria. If none of the search configurations generates an MCS mapping in agreement with the annotations from the Metabolite database then mappings from an MCS generated using the intermediate strictness are utilised.

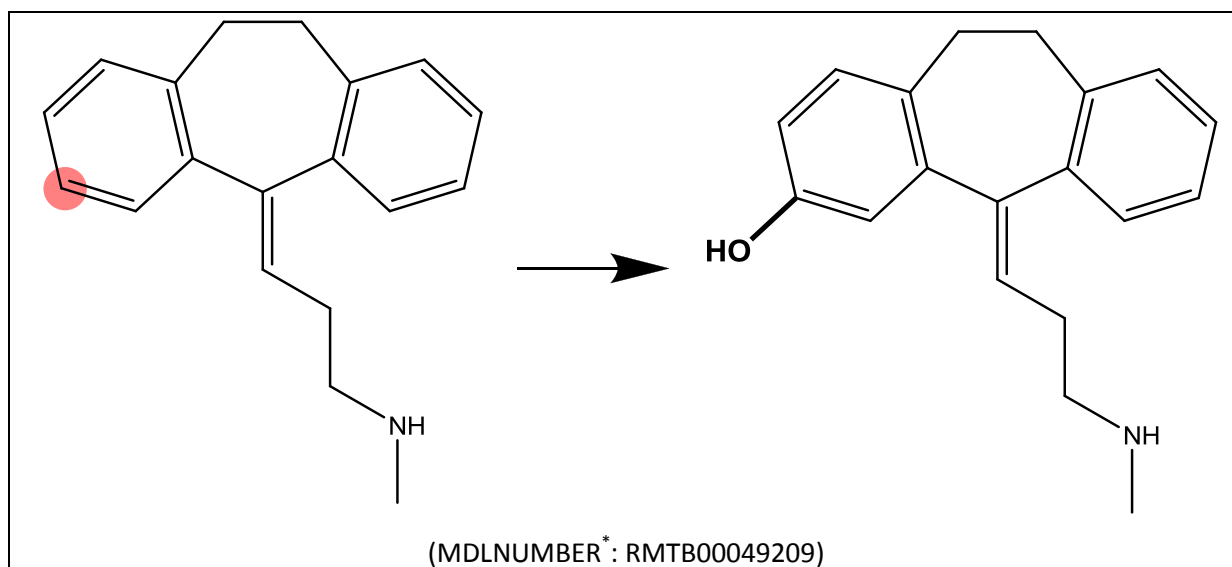
Step	Matching criteria
1	Atoms: atomic number Bonds: order (single and double can interchange)
2	Atoms: atomic number, charge, aromaticity Bonds: order, aromaticity
3	Atoms: atomic number, charge, hydrogen count, mass, ring membership, chirality Bonds: order, aromaticity, ring membership, chirality

Table 1: The SPORCalc database builder's MCS matching criteria.

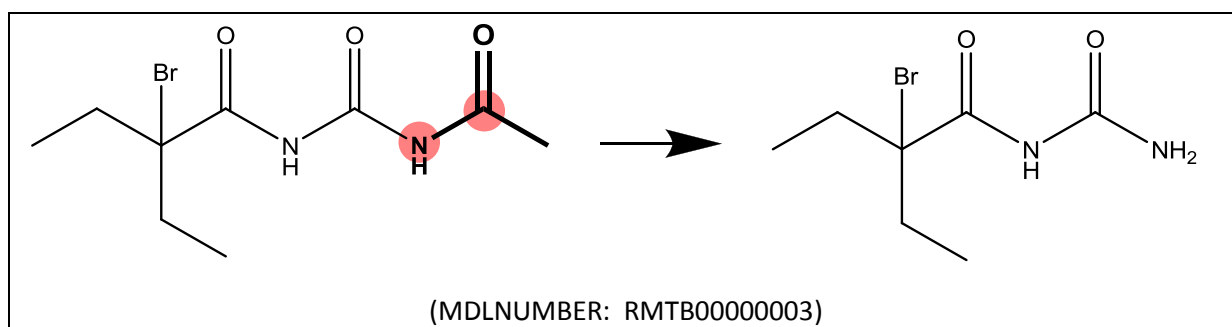
Once the MCS has been determined, the database builder identifies and classifies the reaction centres in the structure. SPORCalc classifies reaction centre atoms as being involved in one or more of phase I addition (defined as the addition of a single oxygen atom – i.e. hydroxylation, oxidation or epoxidation), phase II addition (addition of any group other than a single oxygen atom), elimination, bond order change, bond broken and bond created. In addition, any atoms flagged as both addition and elimination reactions are also flagged as substitutions. Being concerned primarily with phase I transformations, by default SPORCalc discarded labelling other than phase I addition and/or elimination.

Examples of each of the classes of transformation are shown below. Added, eliminated and changed portions of the structures are highlighted, as are the assigned reaction centres.

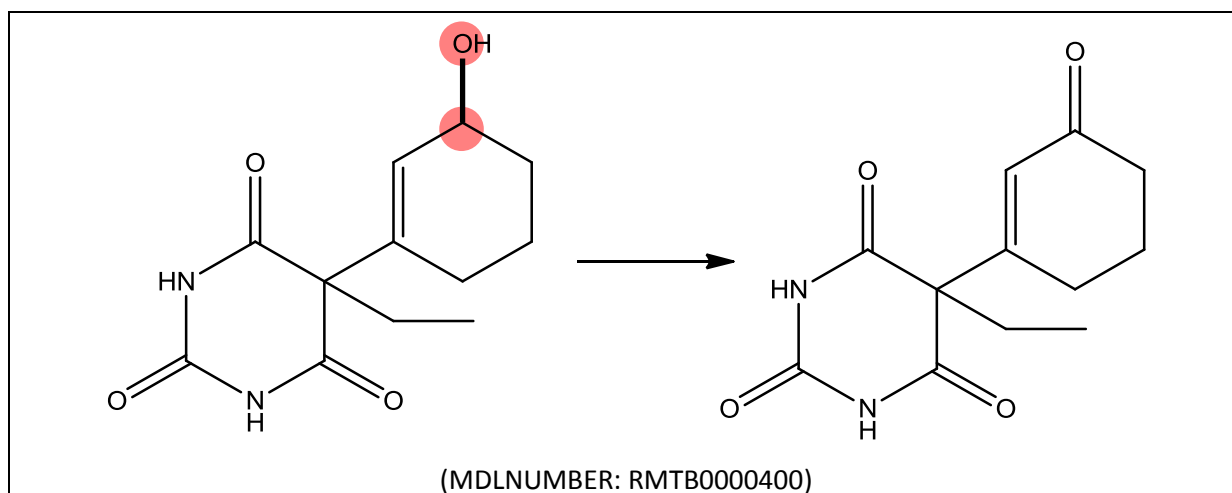
Addition (Phase I)



Elimination

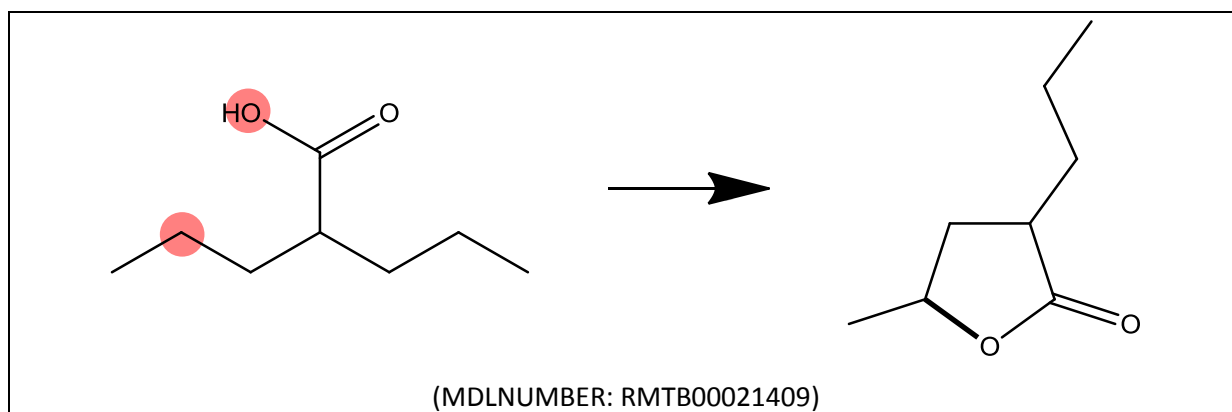


Bond order change

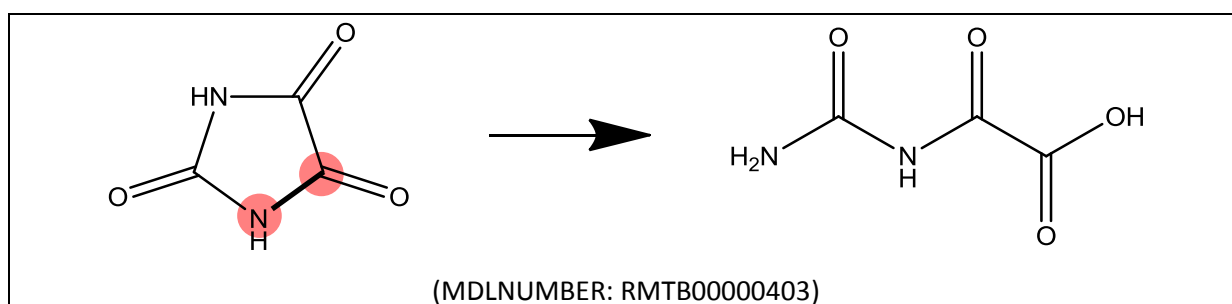


* The MDLNUMBER is the record identifier from the Symyx® Metabolite database.

Bond made



Bond broken



3.2 Development of MetaPrint2D

Early investigations of SPORCalc found that the software's architecture would make it difficult to perform the large-scale evaluation that was envisioned. The processing time of several minutes per compound made any evaluation quite time-consuming to perform, and being accessible only through a website made the process difficult to automate. In addition, since the OEChem library was integral to much of SPORCalc's processing, it would be difficult to substitute and evaluate alternative software approaches – one of the goals of this work.

Given this, it was decided to develop MetaPrint2D, a completely new piece of software, based on the SPORCalc approach to site of metabolism prediction. MetaPrint2D has been designed in an extensible manner, enabling it to be integrated with other software, and its use automated. This has enabled the introduction of a number of optimizations and other improvements to be evaluated.

Since SPORCalc had already undergone several evaluations, and was being used both within AstraZeneca and through a website run by the Unilever Centre for Molecular Science

Informatics, it was considered to be important to, so far as possible, maintain compatibility and ensure that a fair comparison was possible between the tools. To this end, the initial development of MetaPrint2D aimed to replicate SPORCalc's results as closely as possible. It was decided that, by default, MetaPrint2D should use the same training data as SPORCalc (the Symyx® Metabolite database) and the same type of models.

3.3 The Symyx® Metabolite database

3.3.1 Overview of the Symyx® Metabolite database

The Symyx® (formerly MDL) Metabolite database (250), provides information on the metabolic fate of xenobiotics, abstracted from primary literature, conference proceedings and New Drug Applications. The 2008.1 release of the database contained 87446 transformations, with around 5000 new transformations being added each year. Transformations are annotated with a variety of information including references to the literature reporting the transformation, details of the species and systems in which the transformation has been detected and classification of the types of reaction the transformation involves. Each individual transformation does not necessarily record all of these details; indeed different transformations report varying subsets of this information.

Database Version	2005.1	2006.1	2007.1	2008.1
Transformations	72599	78009	82671	87446
Single step	58757	62147	65732	69402
Product not reported	811	831	834	882
Newly added		5410	4662	4775

Table 2: Overview of the contents of the Symyx® Metabolite database

A transformation in the Symyx® Metabolite database consists of a single reactant (substrate) molecule, and a single product (metabolite). One substrate compound may undergo a number of competing metabolic transformations, leading to a variety of different products. Each of these is recorded in a separate record in the database.

Database version	2005.1	2006.1	2007.1	2008.1
Molecules	45486	47855	50515	53247
as substrate	22654	23934	25272	26588
as product	37677	39569	41786	44136
Schemes	11280	11923	12502	13052

Table 3: Contents of the 2005.1-2008.1 Symyx® Metabolite database releases.

Transformations are collated into metabolic schemes, with each scheme containing the collection of pathways originating from a distinct *parent compound*. The database contains one record for each single step transformation (e.g. P→1A; 1A→2D; P→1J – in Figure 30 below) and an additional record for the overall transformation achieved in each multi-step pathway (in addition to the records for the individual transformations such as P→1A and 1A→2A, there will be records for overall transformations like P→2A).

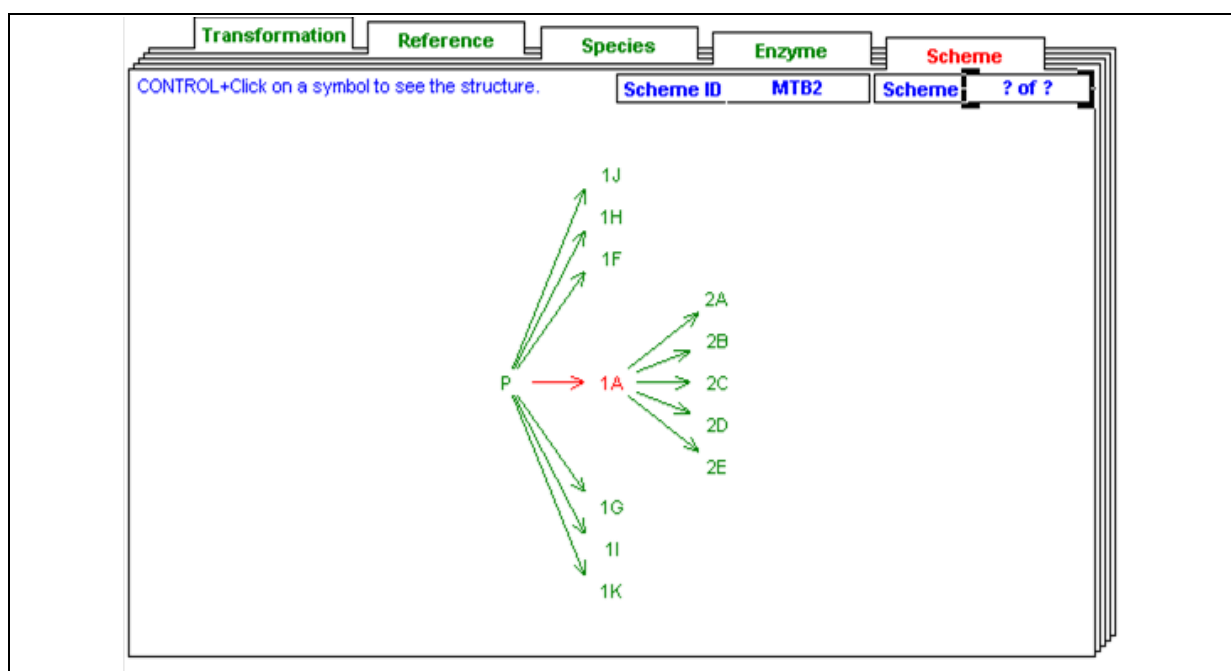


Figure 30: Screenshot of a metabolic scheme from the Symyx® Metabolite database in the ISIS/Base Metabolite Browser. The parent compound is designated 'P', and the 1st and 2nd generation metabolites 1X and 2X, respectively. The database contains a record for each transformation (designated by a reaction arrow), and an additional record for the overall transformation from the parent compound to each non-first generation product.

3.3.2 Consistency of the Symyx® Metabolite database

As will be discussed in more detail later in the chapter, this work required the simultaneous use of several different releases of the Metabolite database. This made it important that a method was identified to track molecules and transformations between database releases, and to ensure that the data was consistent between the different versions of the database. Examination of the records in different releases of the Metabolite database shows that the transformation indexes and scheme identifiers change between database versions; however there is a hidden field – MDLNUMBER, which contains a unique identification reference for each molecule and transformation. This identifier should remain the same with every release of the Metabolite database (254).

In order to check that the data was consistent between releases of the Metabolite database and the MDLNUMBER identifier preserved, as expected, InChI™ (255,256) (IUPAC International Chemical Identifier) canonical identifiers were generated for each molecule – substrate and product – for every transformation in each Metabolite database release that was being used. The InChIs were recorded along with the MDLNUMBERS of the molecules and of the transformation. This information was used to check whether the MDLNUMBERS are preserved, and whether or not they too are canonical.

During the InChI generation process one problem was encountered: the Symyx® Metabolite database contains a number of entries with generic R-groups representing parts of the structure (e.g. covalently bound proteins or DNA), but the InChI algorithm and software do not currently support the concept of ‘wildcard’ atoms – the connection table of molecules must be completely specified, and all atoms assigned a valid chemical element. To overcome this limitation, any R-groups encountered were substituted with iodine atoms, selected because iodine has the same valence, but is relatively rare within the Metabolite database (a search of the 2008.1 database found only 140 iodine containing molecules), so unlikely to cause a clash. This substitution was only carried out to facilitate the generation of InChI canonical identifiers for the molecules, and was not applied to any other analyses.

Consistency of molecule and reaction identifiers

The molecule’s MDLNUMBERS were found to be consistent between the 2006.1 and subsequent releases of the Metabolite database – the identifier always described the same

structure (the corresponding InChI is consistent). Between 2005.1 and 2006.1, however, the structure of a small number of molecules was changed. Examination of the altered molecules suggests that the changes are the result of a remediation process – fixing incorrect structures and stereochemistry:

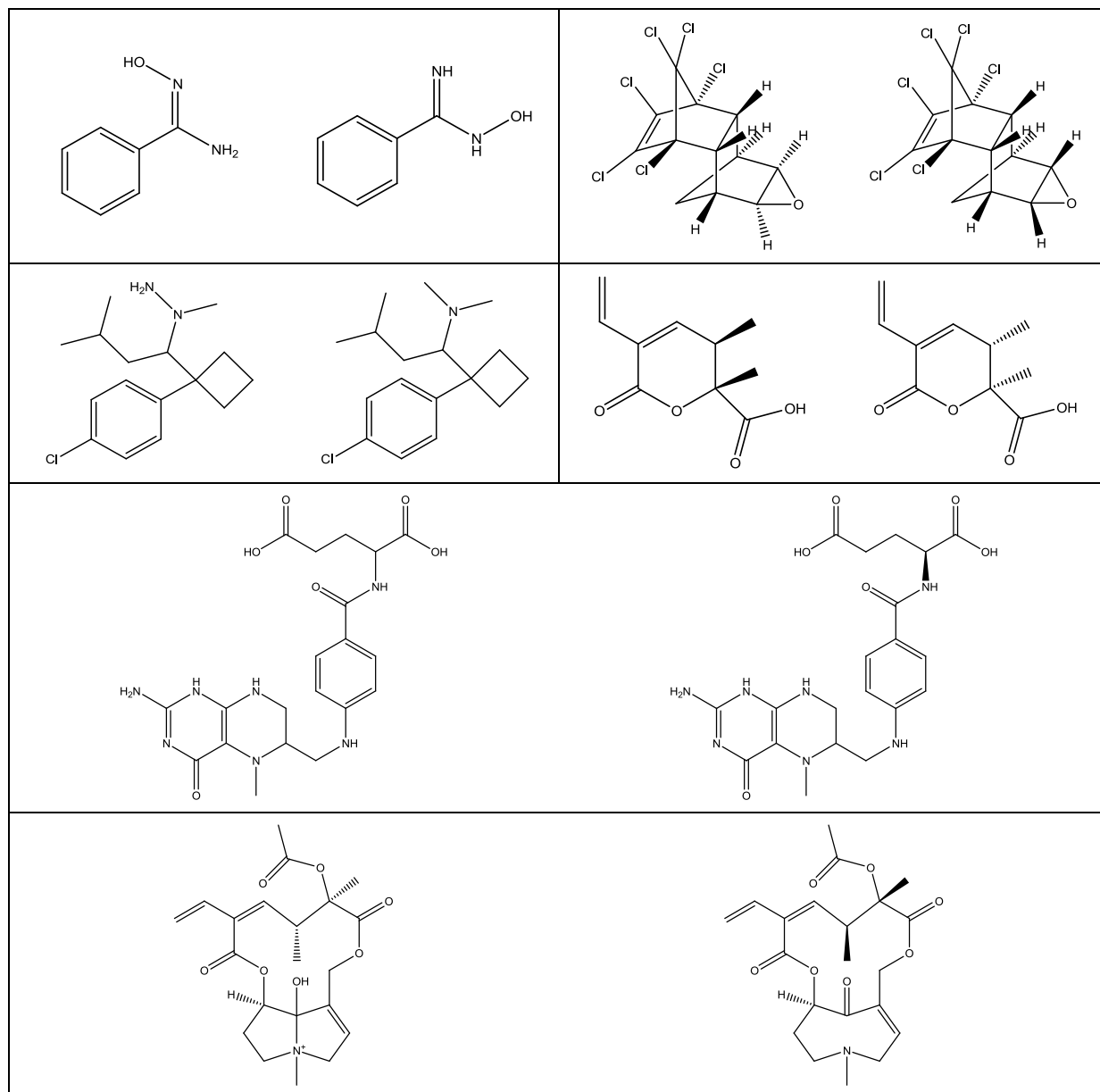


Figure 31: Examples of structures changed between 2005.1 and 2006.1 database releases (left: 2005.1; right: 2006.1 and later). In a few instances alterations are made to elements and connectivity, but in the majority of cases the only changes are the addition or correction of stereochemistry.

Other than these changes through remediation, the transformations' MDLNUMBERS are also consistent (always describe a transformation between molecules having the same

InChIs) between database versions. With respect to molecule MDLNUMBERS the transformation identifiers are consistent throughout the remediation process too.

Canonicity of molecule identifiers

An analysis was also carried out to determine whether the MDLNUMBER is a canonical identifier of molecules – i.e. whether all occurrences of molecules with the same structure are assigned the same MDLNUMBER. In order to perform this analysis the MDLNUMBERS associated with each InChI were determined, and any InChI having more than one MDLNUMBER associated was identified.

Initially it did appear that some molecules were assigned multiple MDLNUMBERS, but on further investigation it was found that in the majority of cases this could be explained by the recording of relative versus absolute stereochemistry. Many transformations have been reported by a number of sources, some of which have specified the absolute stereochemistry of the molecules, and others the relative stereochemistry. The metabolite database records these separately. When the input structures' chiral flag (indicating whether the structure represented a specific enantiomer) was taken into consideration this apparent duplication of molecules was eliminated.

The remainder of problem cases were due to 'indeterminate metabolites'. In a number of metabolic schemes the presence of intermediate metabolites whose structure are unknown is reported. The Symyx® Metabolite database represents these cases as empty structures, so while the molecules' structures are in fact different (but unknown), identical (empty) InChIs were generated by this analysis.

Canonicity of reaction identifiers

The canonicity of reaction identifiers was also investigated. Each reaction was described in terms of the InChI of its substrate and metabolite compounds, and these were mapped against the MDLNUMBER of the transformation. Analysis of this data found that while the reaction identifiers were consistent across the database releases investigated, they were not canonical. Some transformations from the same substrate to metabolite molecules are recorded multiple times in the Symyx® Metabolite database. This is due to reaction schemes centred on different parent compounds converging on a common intermediate metabolite, and from that point on following identical reaction pathways.

3.3.3 Data formats

The RDfile format

The entire Symyx® Metabolite database can be exported as an RDfile. The RDfile format is one of a family of related formats, known as CTfile (chemical table file) formats (257,258), developed at MDL Information Systems. The first CTfile format is the molfile, which is used to specify the structure of a single molecule, and consists of a Header block and a Ctab block – providing a connection-table description of the molecule. Related to molfiles are rxnfiles, which specify reactions and consist of a RXN Header block, together with a number of embedded molfiles – one for each reactant and product molecule. The structures of one or more molecules, together with associated data and properties for each molecule, can be stored in an SDfile (structure-data file), which is again made up of a number of embedded molfiles, but also includes name-value data entries for each molecule. Similarly, RDfiles provide the option to store one or more reactions, together with associated data and annotations, in a single file.

\$RDFILE 1	Header line
\$DATM 7/22/2008 15:12:35	Timestamp
\$RFMT \$RIREG 1	Reaction record indicator, internal ID = 1
\$RXN	Embedded rxnfile, containing reactant and product molecules
...	
\$DTYPE RXN:RXNREGNO	Field title (RXN:RXNREGNO)
\$DATUM 1	Field value (1)
\$DTYPE RXN:VARIATION(1):RXNREF(1):PATH	Field title
\$DATUM MTB1-A	Field value
\$DTYPE RXN:VARIATION(1):RXNREF(1):STEP	Field title
\$DATUM 1 Step	Field value
...	
\$RFMT \$RIREG 2	Next reaction record indicator
...	

Figure 32: Left, a portion of an RDfile export from the 2008.1 release of the Symyx® Metabolite database; and right, a description of the contents of each line or section of the file.

Data fields in the Metabolite database form a hierarchical tree structure, with each node in the tree containing child data fields, a value, or a list of values, as illustrated in Figure 33. The Rdf file format, however, flattens this tree structure and stores reaction data as pairs consisting of a field name and the associated data.

Tree Structure	Corresponding Field Names
<pre> RNX ├── RXNREGNO ├── VARIATION() │ ├── RXNREF() │ │ ├── PATH │ │ └── STEP │ └── LITREF() │ ├── JOURNAL_JRNL │ └── AUTHOR ├── REACTANT_LINK() └── PRODUCT_LINK() </pre>	<pre> RXN:RXNREGNO RXN:VARIATION(1):RXNREF(1):PATH RXN:VARIATION(1):RXNREF(1):STEP RXN:VARIATION(1):LITREF(1):AUTHOR RXN:VARIATION(1):LITREF(2):AUTHOR RXN:REACTANT_LINK(1):... RXN:PRODUCT_LINK(2):... </pre>

Figure 33: A section of the tree structure holding the reaction data in the Symyx® Metabolite database, and the corresponding data field names from an Rdf file export of the database.

Rxnfile format

The structure of an rxnfile is shown in Figure 34. The file contains a short header block which identifies the file as an rxnfile, allows the reaction to be named, and a short comment (up to 80 characters long) to be included. The header can also contain the initials of the user who created the file, the identity and version of the software used to generate the file, the date and time the file was created, and an internal registry number for the reaction. Following the header, an rxnfile has a line indicating the number of reactant and product molecules contained in the file, followed by those molecules embedded using the molfile format. The molecules are ordered as reactants followed by products.

\$RXN	Header line
	Reaction name (blank in Metabolite)
ISIS 072220081512	Information on user/software
	Line for comments
1 1	The numbers of reactant and product molecules
\$MOL	Molecule delimiter
...	Embedded molfile
\$MOL	Molecule delimiter
...	Embedded molfile

Figure 34: Overview of the format of an rxnfile.

Molfile format

The molfile format is similar to that of the rxnfile. Molfiles contain a header allowing the molecule to be named and a comment added. As with rxnfiles information regarding the user who created the file, the software used and an internal registry number can be included, but in addition it can be specified whether the file contains 2D or 3D coordinates, together with scaling factors and if used with a modelling program, a steric energy value. Following the header is the connection table (Ctab) block. This starts with a counts line, specifying the number of atom, bond and property records in the block and a chiral flag. The counts line is followed by one line for each atom in the molecule, specifying the atom's coordinates, element type, charge, isotope number and various other properties. The atom records are followed by bond records, again with one line for each bond in the molecule, indicating the atoms making up the bond and the order of the bond, along with some

annotations. The final section of the Ctab block contains a list of additional properties, including information on the isotopic composition, atomic charges and radical centres. In earlier versions of the molfile format the number of properties lines was included in the Ctab block's counts line, however the properties block is now terminated by a line reading 'M END'. This structure is illustrated in Figure 35, below.

	Molecule name (blank in Metabolite)
-ISIS- 07220815122D	Information line
	Line for comments
15 14 0 0 0 0 0 0 0 0999 V2000	Counts line; first two figures are number of atoms and number of bonds in molecule
-8.5869 -2.2723 0.0000 C 0 0 3 0 0 0 0 0 0 1 0 0	Atoms block
...	
1 2 1 0 0 0 2	Bonds block
...	
M CHG 1 3	Properties block
...	
M END	Molecule terminator

Figure 35: The structure of a molfile.

3.3.4 Data fields in the Symyx® Metabolite database

SPORCalc databases were generated from a list of rxnfiles. This approach lost much of the information contained in the Symyx® Metabolite database, since the rxnfile format does not allow the inclusion of any additional data fields. In the development of MetaPrint2D it was felt that although they are a more complex format, it was better to work with RDfiles, since they provide access to the full content of the Metabolite database.

RXN:RXNREGNO	Internal ID; reactions indexed from 1
RXN:SCHEMEID	Reaction scheme ID (e.g.: MTB1)
RXN:VARIATION(1):RXNREF(1):PATH	Reaction path ID (e.g.: MTB1-A)
RXN:VARIATION(1):RXNREF(1):STEP	Reaction step (e.g.: 1 Step; 2 of 5; 3 Steps)
RXN:VARIATION(1):LITREF(1):ANIMAL(1):SPECIES	Species/systems in which transformation has been observed to occur (e.g.: <i>in vitro</i> (Rabbit Liver Homogenate))
RXN:VARIATION(1):LITREF(1):ANIMAL(2):SPECIES	
RXN:VARIATION(1):LITREF(2):ANIMAL(1):SPECIES	
RXN:VARIATION(1):RXNCLASS(1):RXNCLASS	Annotated reaction types (e.g.: Deacetylation)
RXN:VARIATION(1):RXNCLASS(2):RXNCLASS	
RXN:VARIATION(1):MDLNUMBER	Unique reaction ID (e.g.: RMTB000000005)
RXN:REACTANT_LINK(1):MOL(1):MDLNUMBER	Unique molecule ID for reactant (e.g.: MMTB000000001)
RXN:PRODUCT_LINK(1):MOL(1):MDLNUMBER	Unique molecule ID for product (e.g.: MMTB00002974)

Figure 36: Selected fields from the Symyx® Metabolite database, relevant to the development of MetaPrint2D, with field names, descriptions and example entries.

3.4 MetaPrint2D's implementation

There are two primary factors affecting the accuracy of SPORCalc and MetaPrint2D – the quality and breadth of the data in the Symyx® Metabolite database, and the correctness of the identification of sites of metabolism in the training data. The former is something over which users of the database have no control (other than reporting any problems identified to the database's publishers, to be fixed in subsequent releases), the latter, however, is open to investigation.

Among other uses, OEChem was required by SPORCalc in order to carry out the maximum common substructure searches performed for the identification of reaction centres. Since the goals of this work included the removal of commercial dependencies, such as OEChem,

from the software, and the implementation of dependable algorithms to improve the reliability of predictions, this area has been investigated. Analysis of the method through which SPORCalc identified sites of metabolism suggested a number of alternative approaches.

MetaPrint2D was written using the Java programming language. Java is widely used in the chemical computing community, and offers the advantage of being easily portable between computers running different operating systems, while not suffering from the performance problems of purely interpreted languages.

3.4.1 Reaction centre identification

Over the course of the development of MetaPrint2D a number of approaches to the identification of sites of metabolism in the transformations from the training data were considered:

- Bond annotations
- Atom-atom mappings
- Maximum common substructure search

Each of these approaches is discussed below.

Symyx® Metabolite database bond annotations

The first option examined was to make direct use of the annotations contained in the Symyx® Metabolite database. The CTfile formats provide support for bond annotations detailing their 'reacting centre status', and this has been used in the construction of the Metabolite database. The available annotations are listed in Table 4.

Value	Meaning
0	Unmarked
1	A centre
-1	Not a centre
2	No change
4	Bond made (if in product)/broken (if in reactant)
8	Bond order changes
12 (4+8)	Both made/broken and bond order changes

Table 4: CTfile reaction centre status annotations.

Unfortunately, the bond annotations were not found to map well to the substrate reaction centres. Figure 37 shows a small selection of transformations from the Metabolite database, with the annotated bonds and atoms considered to be reaction centres highlighted.

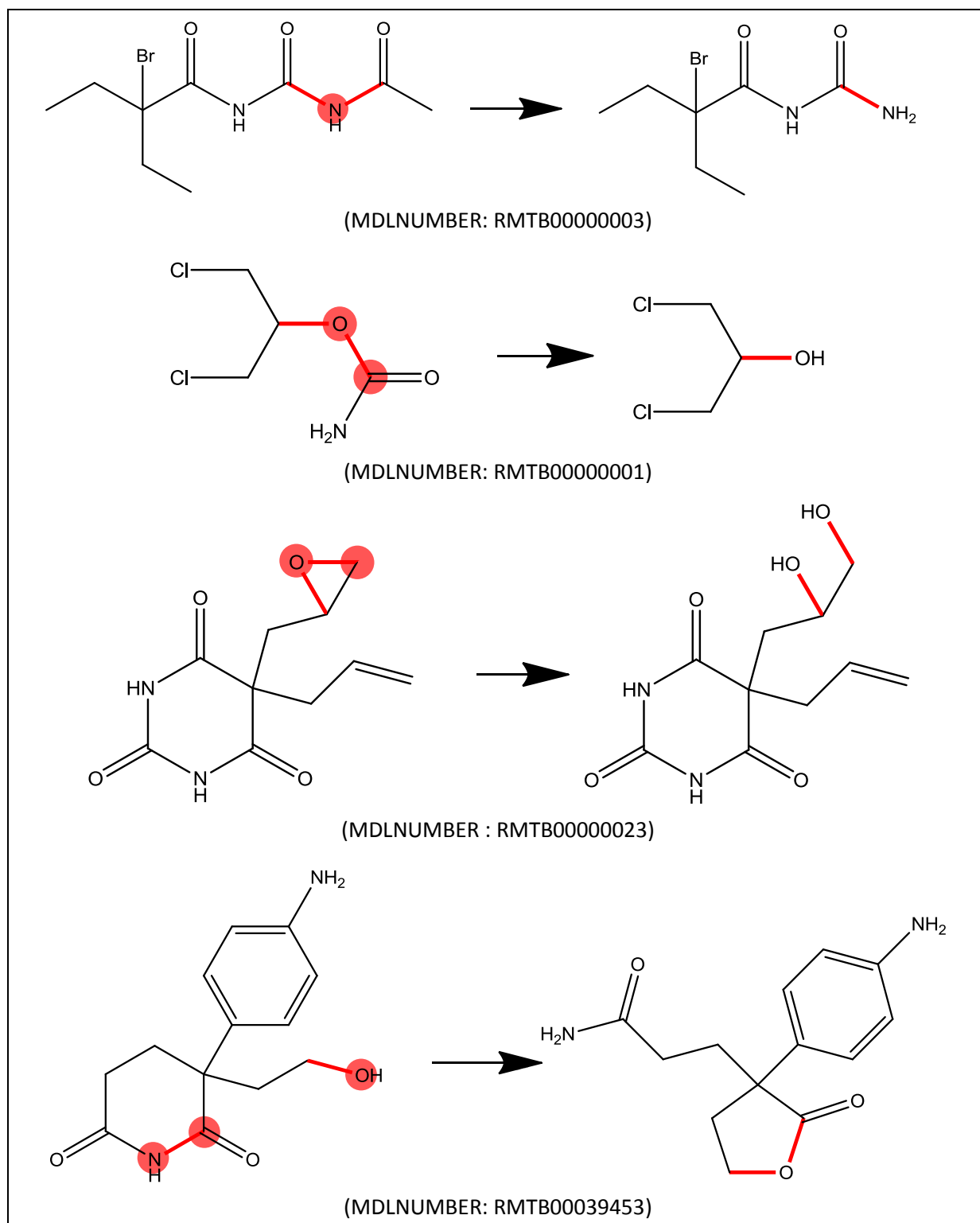


Figure 37: Symyx® Metabolite database bond annotations; the highlighted bonds are all marked 'bond made/broken', and the highlighted atoms are those that we would consider to be reaction centres.

It is clear that there is no simple correspondence between the bond annotations in the Symyx® Metabolite database and the reaction centre atoms; assigning as reaction centres

all those atoms belonging to an annotated bond would lead to a large number of extra atoms being labelled as reaction centres. The possibility that the bond annotations could be due to the mechanistic detail of the transformation has been considered, but given the likely mechanism of the hydrolysis (shown in Figure 38) taking place in the first transformation, this is unlikely to be the case.

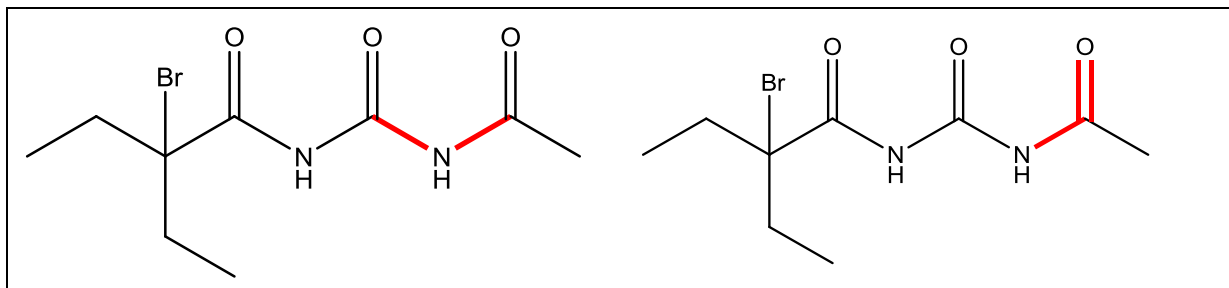


Figure 38: Left, substrate of the hydrolysis shown in Figure 37 with annotated bonds from the Metabolite database highlighted, and right, with the mechanistically important bonds highlighted. The bond annotations do not correspond to the mechanistically important bonds.

Symyx® Metabolite database atom-atom mappings

In addition to the 'reacting centre status' annotations of the bonds, transformations from the Symyx® Metabolite database are annotated with atom-atom mappings, indicating correspondence between atoms in the substrate and metabolite structures. Each atom that is conserved between the substrate and metabolite molecules is assigned a unique number, and annotated with that number in each of the structures. SPORCalc made use of these annotations in its determination of reaction centres.

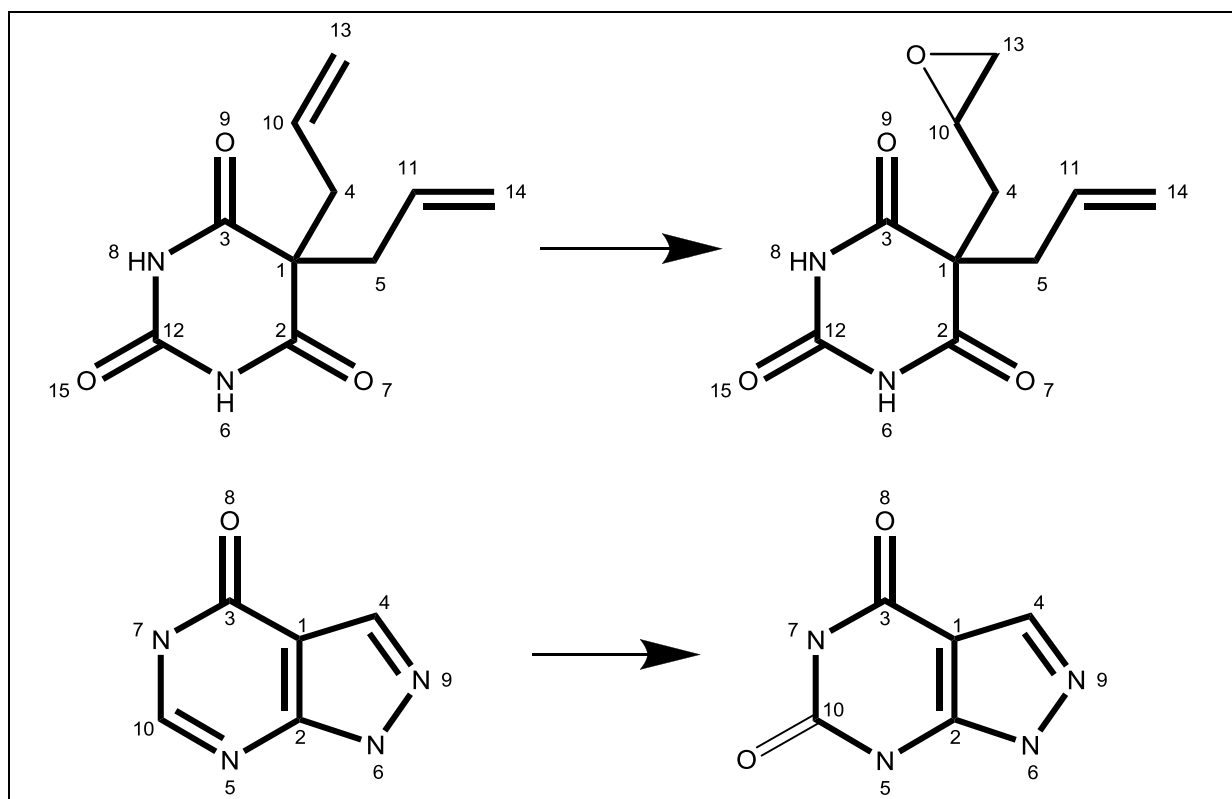


Figure 39: Examples of atom-atom mapping annotations from the Symyx® Metabolite database. The MCS is highlighted in bold, and the Metabolite database supplied atom mapping numbers are displayed. (Top – MDLNUMBER: RMTB00000022; Bottom – MDLNUMBER: RMTB00015481)

In many cases the atom-atom mapping annotations provided by the Metabolite database do give a good indication as to the atoms conserved between the substrate and metabolite compounds, and hence the locations of the reaction centres. However a number of problems were identified.

Missed annotations

Examination of the atom-atom mapping numbers in the Symyx® Metabolite database showed that they could not be used on their own to identify sites of metabolism. Some transformations are not annotated with any atom-atom mapping information, and in many cases some conserved atoms are missed out from the mapping. This is illustrated by the sulfuration shown in Figure 40, below.

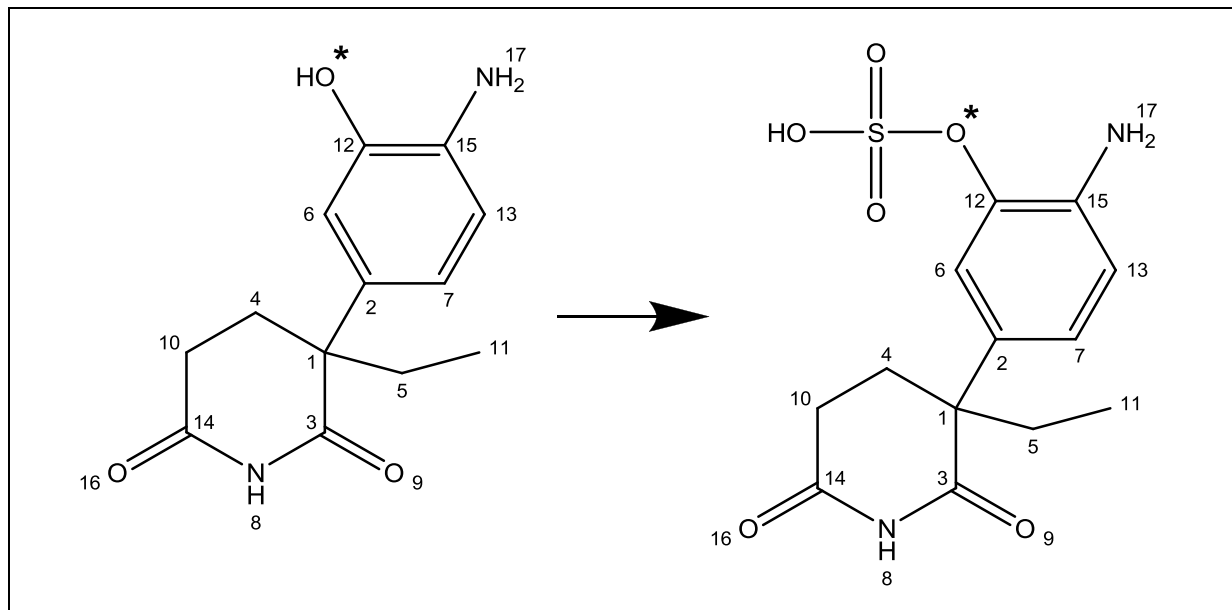


Figure 40: Atom-atom mapping numbers from the Symyx® Metabolite database (MDLNUMBER: RMTB00046671). The oxygen atom marked with an asterisk is not assigned a mapping in the database despite being conserved between the substrate and product.

Although the documentation for the Symyx® Metabolite database states that “atom-atom maps are usually assigned based on the apparent change in the transformation, rather than the actual transformation mechanism” (259), we have considered the possibility that the oxygen at which the reaction takes place (indicated with an asterisk in the figure) could have been excluded from the atom-atom mappings for mechanistic reasons. However, the mechanism through which sulfotransferases act (260) would conserve the atom over the course of the transformation, as shown in Figure 41. This suggests that the indicated atom is omitted from the mapping in error.

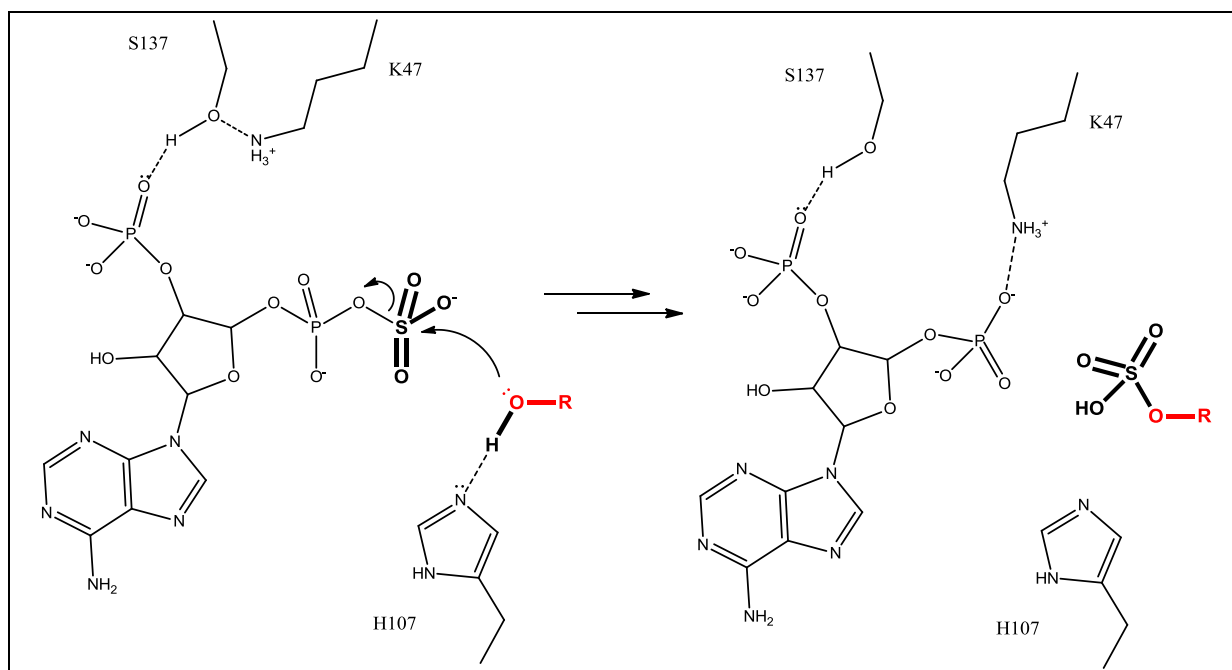


Figure 41: The proposed mechanism for sulfuration (260): Sulfuration takes place in the active site of a sulfotransferase enzyme, with cleavage of the sulphate group from 3'-phosphoadenosine 5'-phosphosulfate (PAPS) proceeding via an S_N2 -like mechanism, stabilized by surrounding charged and polar residues. The substrate, metabolite and sulfonate group are shown in bold, with the conserved substrate substructure highlighted in red. The substrate's hydroxyl oxygen is retained in the metabolite's structure.

The omission of mappings for atoms found at centres of addition such as this is not occasional, but rather seems to have been a systematic choice by the database's curators. The result of this is that the mapping numbers alone cannot be used to identify the conserved structure between a substrate and metabolite.

Mapping errors

While the majority of the atom-atom mappings provided in the Symyx® Metabolite database do appear to be accurate, aside from the missed mappings, there are a number of instances where they are incorrect, leading to strange apparent conserved structures. One such case is shown in Figure 42(a) below. The annotated atom-atom mapping numbers, and the conserved structure they imply, clearly do not correspond to the structure that is in actual fact conserved between the substrate and metabolite. Interestingly, in a similar transformation from the same metabolic scheme, shown in Figure 42(b), the annotations have been correctly assigned, although the exclusion of the oxygen adjacent to atom 4 from

the mapped structure, suggests that the mappings in this case could have been assigned on the basis of mechanism, rather than just the apparent change in the transformation, in spite of what is indicated in the database's documentation (259).

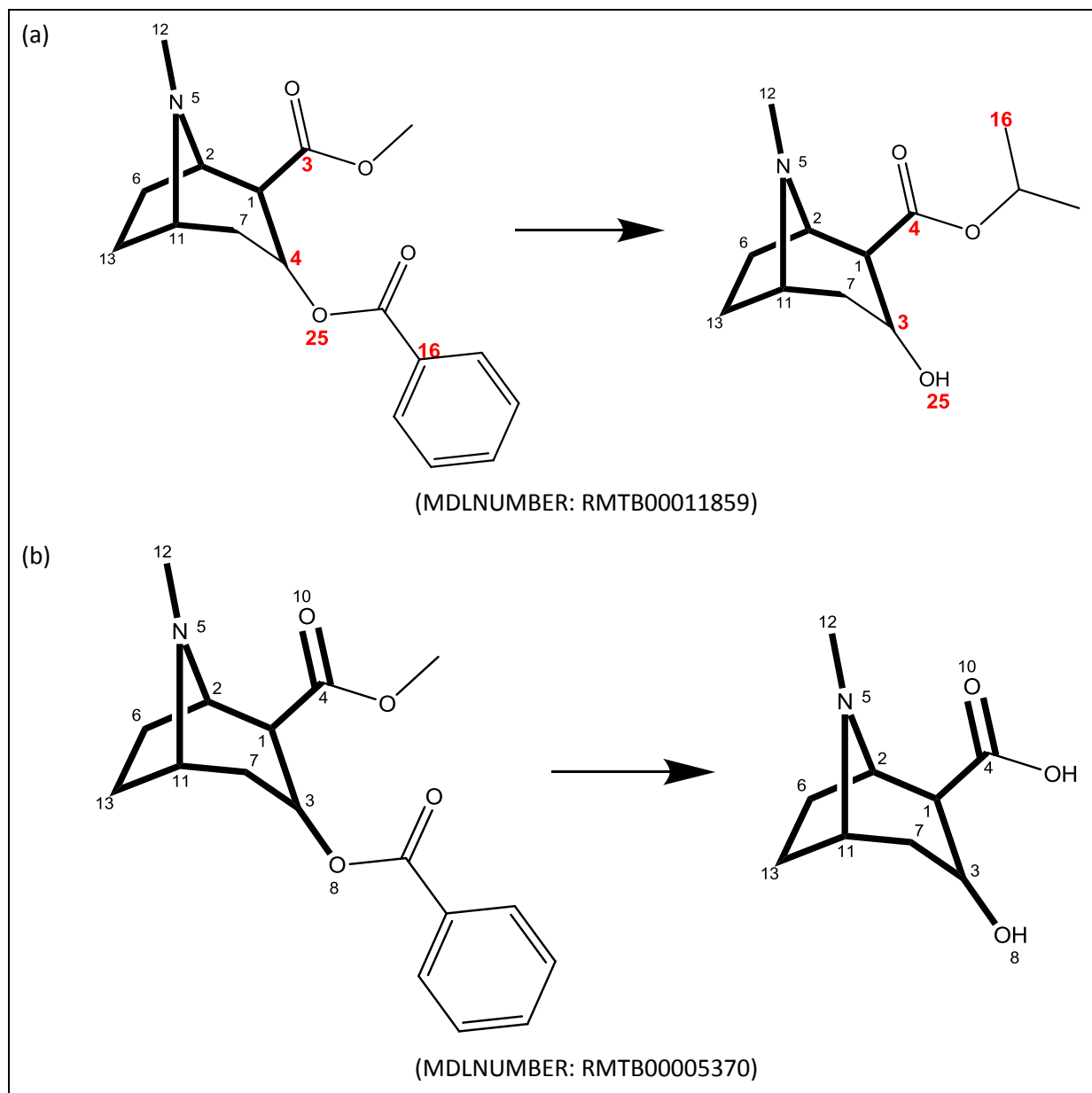


Figure 42: (a) An example transformation from the Symyx® Metabolite database, where the atom-atom mappings are incorrect; (b) A related transformation where the atom-atom mappings have been correctly assigned.

Maximum common substructure search

Well over two decades ago it was suggested that the changes occurring in the course of chemical reactions could be identified through determination of the maximum common subgraph of the reactant and product (261) molecules, and this formed the basis of the

approach taken by SPORCalc. The authors of SPORCalc were aware of the problem of incomplete mapping annotations, and, as described earlier, performed a series of maximum common substructure (MCS) searches with increasingly strict matching criteria until a result was found in which the mappings of all annotated atoms were in agreement with the mappings specified by the annotations. If no such mapping can be found then SPORCalc defaults to using the mappings produced by an MCS search with an intermediate strictness of matching criteria.

Since the annotated mappings are incomplete, an approach taking the Symyx® Metabolite database's atom-atom mapping annotations as a starting point for the maximum common substructure search, and 'growing' the MCS from that structure was considered. However, due to the identification of errors in the mappings such as that described above, it was decided not to pursue that method.

Instead, a scheme based on the MCS between the substrate and metabolite has been adopted. The difficulty in handling the data from the Symyx® Metabolite database is that each record contains only a single reactant and a single product – the main substrate and primary metabolite formed by the transformation. Additionally, some transformations represent the overall result of a number of elementary reaction steps, posing additional challenges. Rather than taking SPORCalc's approach of trying MCS generated using various configurations, until a match with the Metabolite database's annotations is found, MetaPrint2D identifies the 'best' conserved structure that it can, and only in the case of multiple equally good structures uses the Metabolite database's mappings to choose between them.

Exactly what is the best conserved substructure between a substrate and metabolite is not always easy to define. Figure 43, below, shows three possible MCS for a transformation from the Metabolite database, each of which is the outcome of an MCS search performed according to a different configuration of the search algorithm. If the search is performed with the requirement that bond orders must be conserved between the substrate and metabolite, then the MCS shown in (a) is found. If this requirement is relaxed, then the result shown in (b) is detected – with 9 atoms and 8 bonds conserved, compared to the 7 of each for the result of the first search. A third possibility, found if disconnected results are

permitted, is shown in (c); this also contains 9 conserved atoms and 8 conserved bonds, and unlike that shown in (b) the bond orders are conserved.

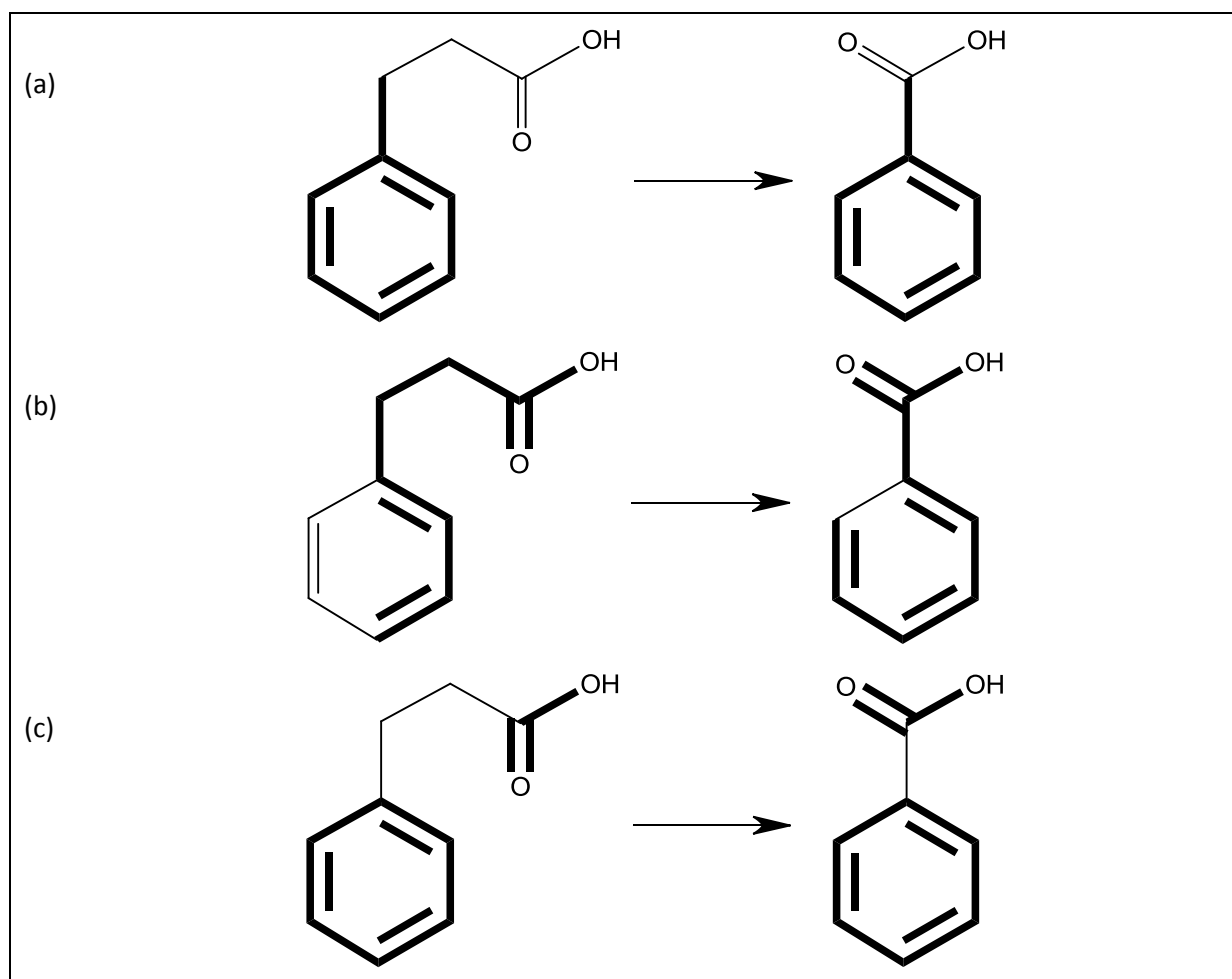


Figure 43: The output of maximum common substructure searches with various settings: (a) bond orders conserved; (b) bond orders not conserved; (c) bond orders conserved, disconnected structures permitted. (MDLNUMBER: RMTB00016651)

It is clear from this example that a MCS does not necessarily reflect the atoms and bonds conserved over the course of a reaction. It is important to note that mappings of bonds cannot be restricted on the basis of their order, as these often change over the course of a reaction. In order to determine the most appropriate 'reaction conserved substructure', MetaPrint2D relies on a set of heuristics to generate constraints on the permitted atom mappings, and then performs a search for the best maximum common substructure within the bounds of those constraints.

'Simple' transformations

In many cases either a simple addition such as hydroxylation or acetylation, or a simple elimination such as deacetylation or dealkylation, has occurred. In these instances there have either been additions or eliminations of atoms and bonds (other than hydrogen atoms), but not both. Examples of such transformations are illustrated in Figure 44. Whether a transformation potentially represents a simple addition or elimination can easily be ascertained by comparing the numbers of atoms in the reactant and product structures. If the product contains more atoms than the reactant then an addition may have taken place, and this can be determined by checking whether the reactant structure is completely contained within the product structure. Alternatively, if the reactant contains more atoms than the product then an elimination reaction may have taken place, in which case the product will be a substructure of the reactant. Testing whether one structure is completely contained within another – the so called 'subgraph isomorphism problem' is much quicker and simpler than maximum common subgraph-isomorphism, so this test is carried out at the start of the analysis of each transformation.

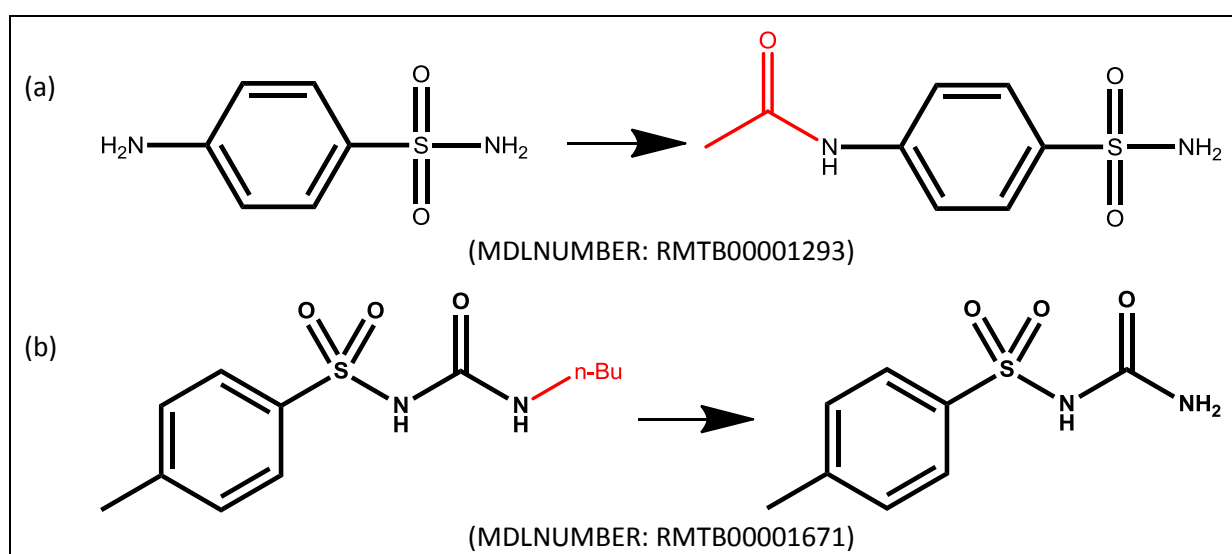


Figure 44: Examples of 'simple' transformations, where either the entire substrate or product structure is conserved. The added or eliminated portion of the structure is shown in red, and the conserved structure highlighted in bold. (a) Addition: acetylation (b) Elimination: dealkylation.

This test also identifies cases in which the Symyx® Metabolite database describes the result as 'optical resolution', where no transformation has occurred; rather a single enantiomeric

form is selected. Since MetaPrint2D currently disregards stereochemical information, these transformations are discarded.

Constrained MCS search

If there is no 'simple' mapping between the substrate and metabolite structures then MetaPrint2D performs a maximum common substructure search in order to determine the conserved substructure. However, in order to determine the most relevant MCS a series of heuristics imposing constraints on the permitted atom mappings have been developed.

These constraints are based around the principle of ring conservation – that if there are equivalent ring systems or single rings in both the substrate and metabolite structures then it is likely that they are conserved over the course of the transformation. MetaPrint2D first checks whether the Murcko framework (262), or scaffold, of either molecule is conserved, and then checks for any conserved ring systems (sets of simple rings sharing one or more atoms or bonds) and finally any remaining simple rings.

1. Scaffold constraints

Murcko frameworks consist of the set of ring atoms and bonds in a molecule, together with the atoms and bonds contained in linkers between rings. The first constraints on the MCS mappings that MetaPrint2D attempts to generate are based on the detection and conservation of this scaffold.

In order to do this the scaffold structures of both the substrate and metabolite molecule are identified, and a regular substructure search performed to determine whether one is completely contained within the other – i.e. whether a scaffold is conserved between the substrate and metabolite. If this is the case then the constraint that the conserved scaffold atoms must map to their equivalent atoms in the other compound is imposed.

If a conserved scaffold is identified then no further search for constraints is performed, since all of the ring atoms and bonds from the structure with the smaller scaffold will have had their potential mappings constrained.

An example illustrating the generation of scaffold-based atom mapping constraints is shown in Figure 45, below.

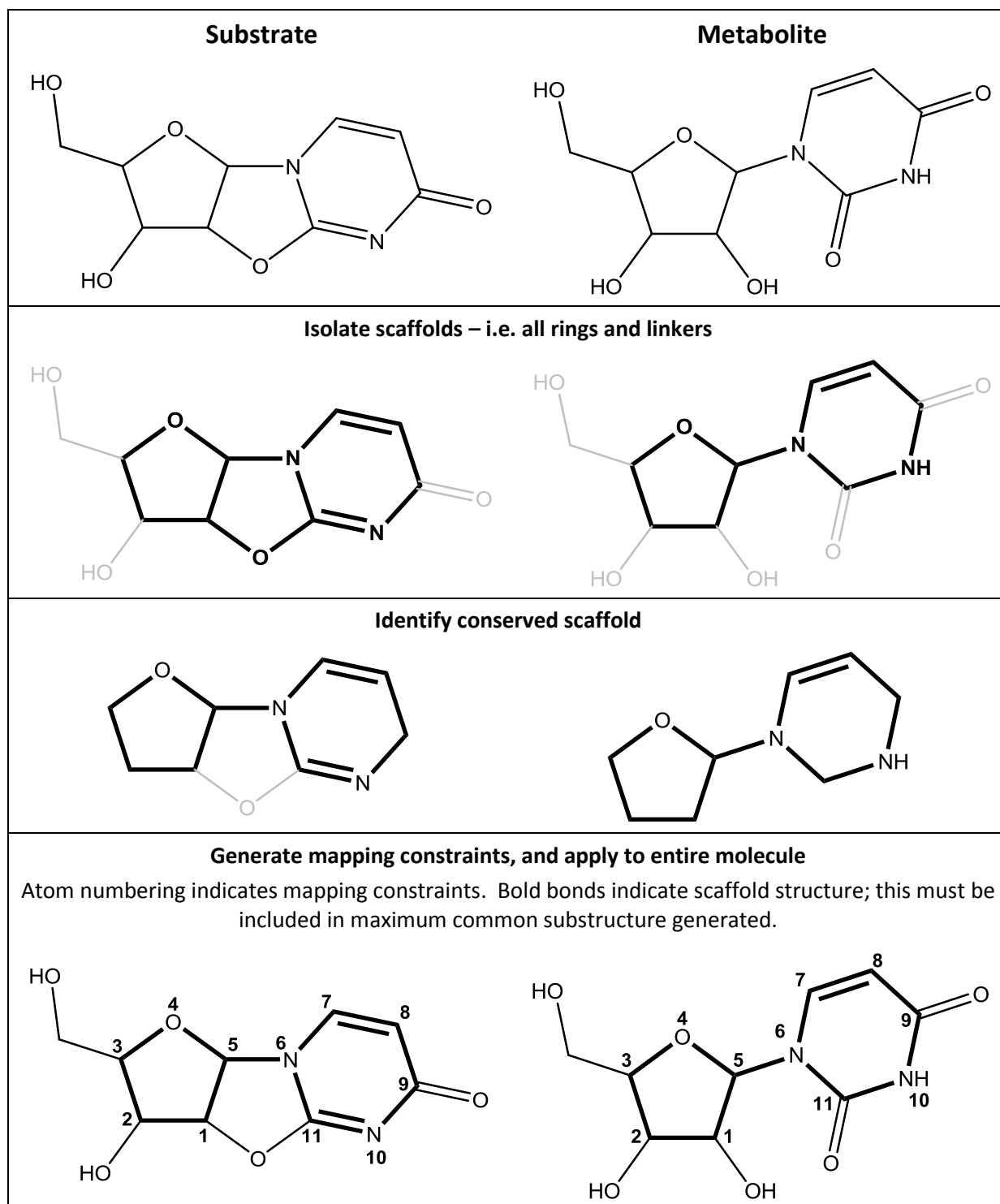


Figure 45: The generation of scaffold atom mapping constraints for MCS search illustrated for the metabolism of Ancitabine (MDLNUMBER: RMTB00036597). The scaffolds, or Murcko frameworks, of the substrate and metabolite structures are identified, and it is determined whether the smaller scaffold is completely contained within the larger. If this is the case then atom mappings between the two structures are generated. In this example each scaffold atom has a unique mapping to an atom in the other structure, but often groups or classes of equivalent atoms are detected.

2. *Ring constraints*

If the scaffold structure is not completely conserved, then a search for conserved ring systems and finally for conserved simple rings is carried out. Ring systems consist of either lone rings, or of sets of single rings having one or more atoms or bonds in common, resulting in bridged, fused or spiro systems. The ring systems in both the substrate and metabolite structures are detected, and any ring systems common to both molecules identified. If there are ring systems common to both structures then atom mapping constraints are generated for the atoms in these systems. In the case that a structure contains more than one identical ring system, mapping constraints are only generated if the other structure contains the same number of occurrences of a matching ring system.

Finally, after any conserved ring systems are identified, conserved structures between any rings that remain unmapped are explored.

An example illustrating the process is shown below in Figure 46.

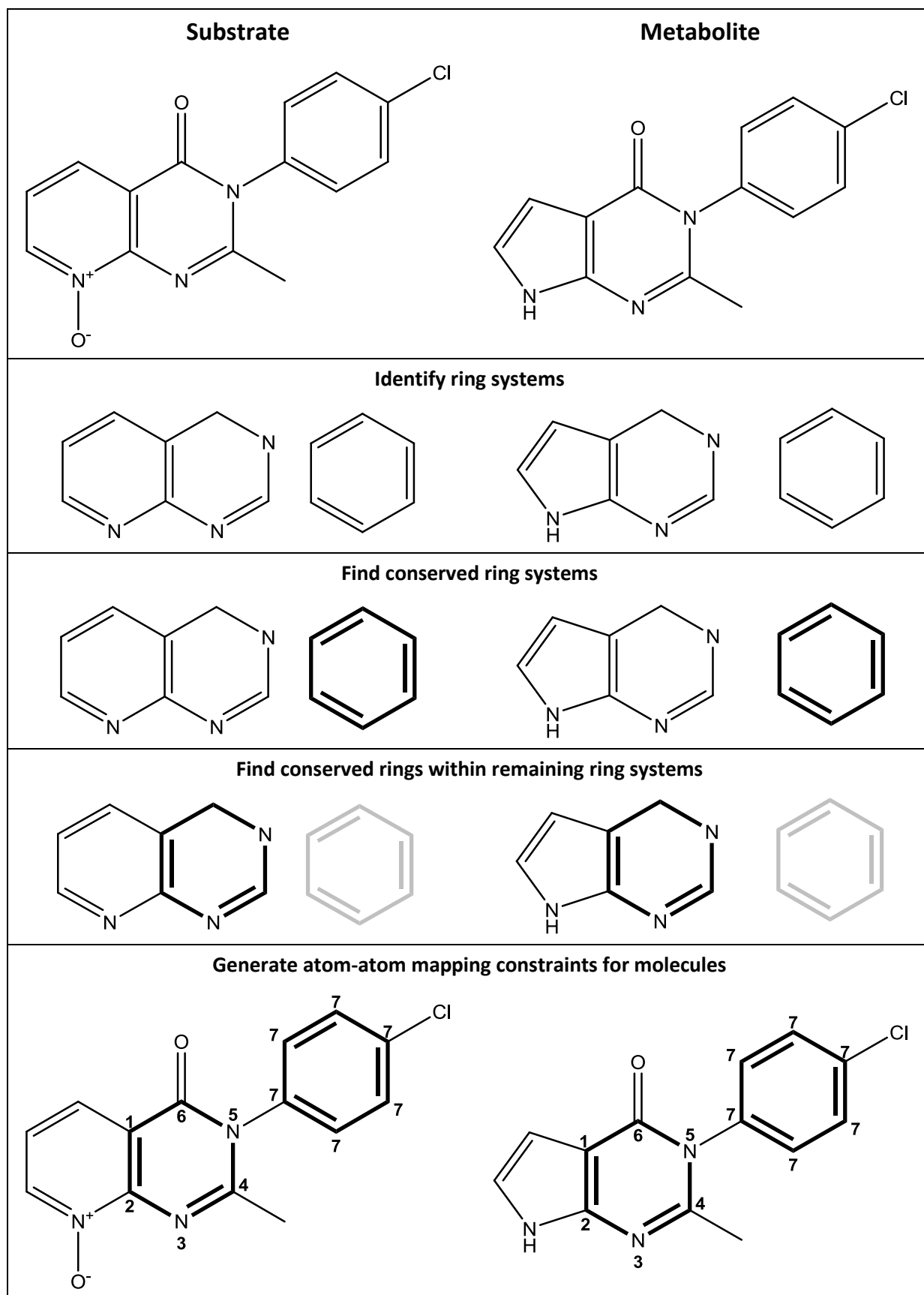


Figure 46: Generation of ring system constraints (MDLNUMBER: RMTB00000482).

MCS search algorithm

Once conserved scaffolds, ring systems and rings have been identified and constraints on permitted atom-atom mappings generated, a final heuristic is applied: any atoms that are not contained in any of the conserved structures may only map to another atom that is not contained in a conserved structure. A maximum common substructure search is then performed within the bounds of these constraints.

Maximum common substructure searching is an example of the 'maximum common subgraph isomorphism' problem, a computational task which is described as NP-complete (263), meaning that there is no efficient algorithm guaranteed to find the best solution. Between two structures having m and n atoms, respectively, there are $m^{n+1}n$ potential MCS solutions (264). However the performance of algorithms can be vastly increased in most cases through the application of appropriate heuristics; in the case of the searches performed by MetaPrint2D, both heuristics within the search algorithm itself, and additionally the constraints on the allowed atom-atom mappings that are generated.

Many algorithms for the determination of maximum common substructures have been proposed (265). MetaPrint2D relies on a modified version of the *Recursive backtracking algorithm*, developed by Krissinel and Henrick (264) to carry out its maximum common substructure searches. This algorithm is itself an enhancement of the well known *Ulmann algorithm* (266). The alternative approach would be to employ an algorithm based on maximal clique detection (267,268).

Each iteration of the recursive backtracking algorithm picks an unmapped atom from the query structure and identifies the set of atoms in the target structure to which it may be mapped without violating the constraints imposed by the previously mapped atoms. Each candidate mapping is picked in turn, with the search continuing until no more query atoms are available, in which case the algorithm backtracks to its previous state, and picks the next candidate mapping. The time required to perform the MCS search depends on the number of recursive calls made. Krissinel and Henrick have developed a strategy for efficiently pruning the search space, eliminating time consuming exploration of undesirable branches that cannot lead to a good solution to the search, at the expense of a small additional overhead per iteration.

At each iteration the algorithm checks whether the current search direction is worth continuing, or whether it should backtrack and pick a different path to search, by testing whether the number of nodes which could be mapped (the sum of the number of nodes currently mapped and the number of unmapped nodes with permitted mappings remaining) is at least as high as the number of nodes in the best result found so far. When picking the next node, the node with the fewest potential mappings is selected, narrowing the search space as rapidly as possible. Finally, when each mapping is made, the potential mappings of all remaining unmapped nodes are refined on the basis that nodes neighbouring the last mapped node in the query structure must neighbour the target structure node to which it was mapped, and similarly, nodes not neighbouring the query structure node may not map to a neighbour of the mapped node in the target structure.

The recursive backtracking algorithm utilised by MetaPrint2D is slightly modified from Krissinel and Henrick's published algorithm. The option to specify constraints on the permitted atom and bond mappings has been added, as has an option to ensure that only connected results are generated, by ensuring that at each step in the search the current structure can only be extended into neighbouring atoms.

Filtering suggested maximum common substructures

Often there are several potential MCSs, with different mappings between reactant and product atoms. If this is found to be the case then Occam's razor is applied, and it is assumed that the simplest explanation – i.e. the MCS with the fewest reaction centres and fewest added, removed or changed bonds – is the best.

The following examples illustrate how the problem is addressed.

1. Minimize the number of reaction centres

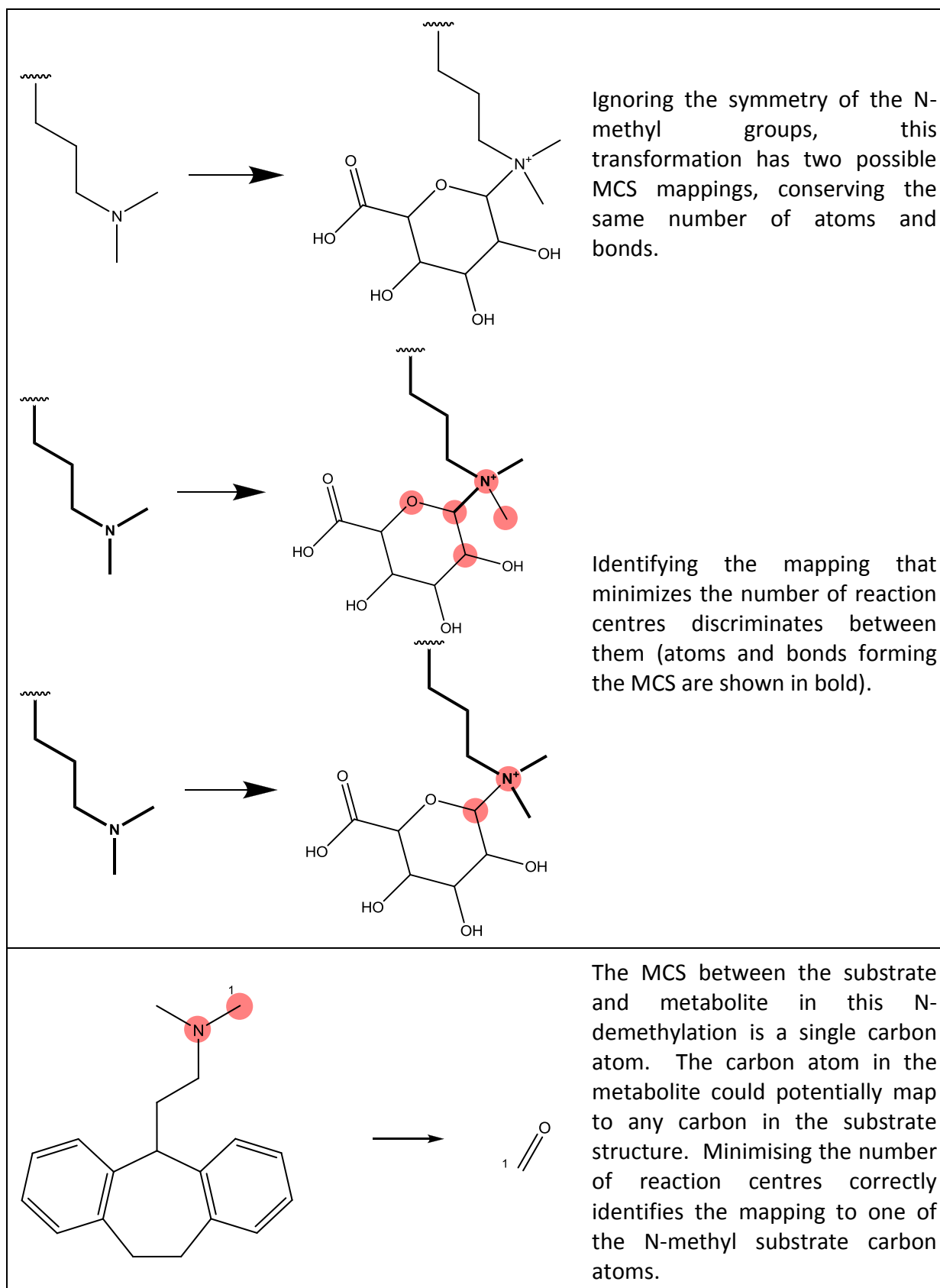


Figure 47: Illustration of the selection of 'best' MCS, by minimizing the number of reaction centres.

2. Pick the MCS with the greatest number of unchanged bonds

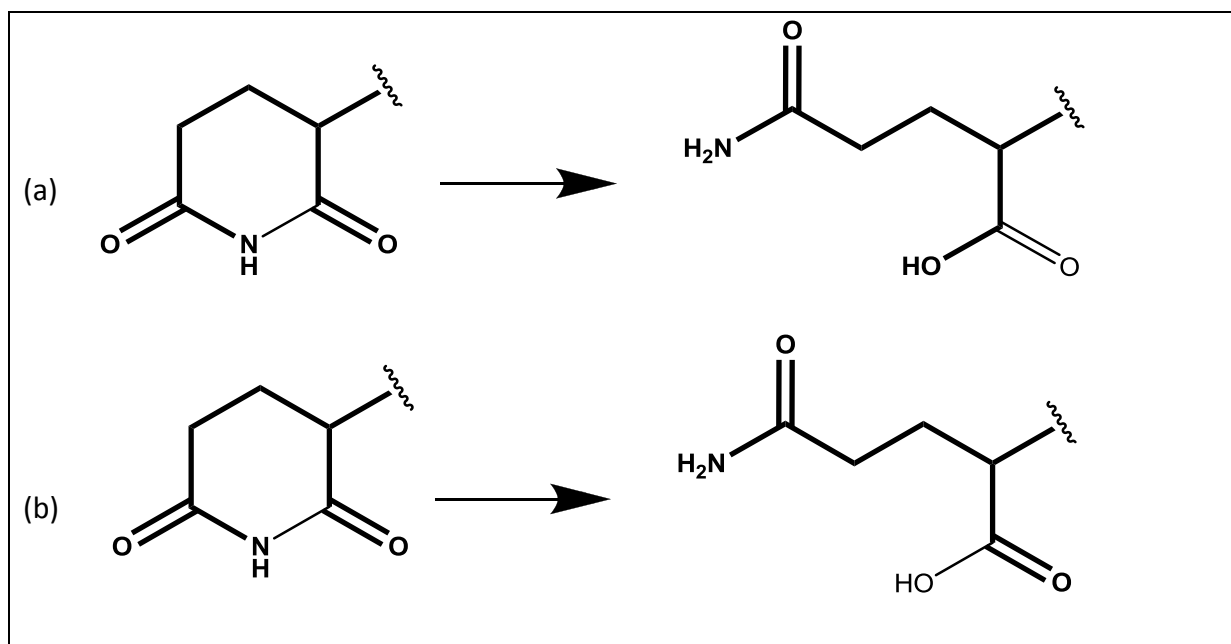


Figure 48: Alternative mappings for an amide hydrolysis; both have the same number of reaction centres; (b) is preferred since the carbonyl oxygen's bond order is conserved.

The performance of the reaction centre identification has been evaluated through manual inspection of 300 randomly selected transformations from the Symyx® Metabolite database. The results of this evaluation are presented on page 120, in Chapter 4.

Classifying reaction centres

Once the structure common to the reactant and product has been identified, reaction centres are detected and classified. In order to maintain compatibility with SPORCalc the same classification scheme is used. Reaction centres are identified through examination of bonds changed between the reactant and product structures. Bonds found in the reactant molecule but not in the product are listed, as are bonds in the product molecule but not in the reactant. In addition, bonds whose order is changed between the reactant and product are identified.

Atoms are marked as reaction centres and assigned to one or more of the same reaction classes used by SPORCalc. This classification is performed on the basis of the atom and bond changes, and in the case of additions, also on the group added:

- Phase I addition – defined as the addition of a single oxygen atom, which covers hydroxylation, oxidation and epoxidation
- Phase II addition – defined as the addition of any other group
- Elimination
- Bond breaking
- Bond formation
- Bond order change
- Substitution – defined as both an addition and an elimination centred on the same atom

Filters can be applied during the model construction process enabling generation of models for any combination of these reaction types. In order to facilitate comparison with SPORCalc, and other site of metabolism prediction tools, in the course of this work models have been restricted to the prediction of Phase I additions and eliminations.

3.4.2 Multi-component structures

There are a small number of transformations for which the Symyx® Metabolite database reports more than one component in either the substrate or metabolite structures. In all cases the additional component is due to the presence of a counter-ion, such as the acetate anion in Figure 49, below.

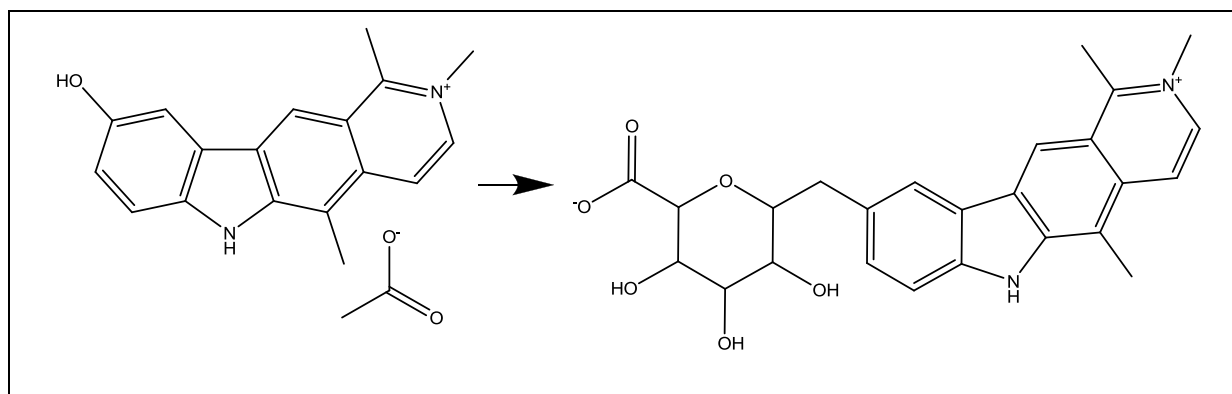


Figure 49: Example of a transformation in which the substrate structure contains an acetate counter-ion (MDLNUMBER: RMTB00042423).

In most cases the presence of the counter-ion is reported in either the substrate or metabolite, not both. For such transformations only the largest component is analysed, and any counter-ions are discarded

3.4.3 Aromaticity detection

When determining whether a region of a molecule has undergone a metabolic transformation it is important that different aromatic resonance forms are taken into consideration; the representation of aromaticity and other delocalised systems is a challenge for chemical information systems. In cases such as that illustrated in Figure 50, it can appear that bonds have changed order over the course of a transformation, when in actual fact the substrate and metabolite structures are different resonance forms of the same aromatic system.

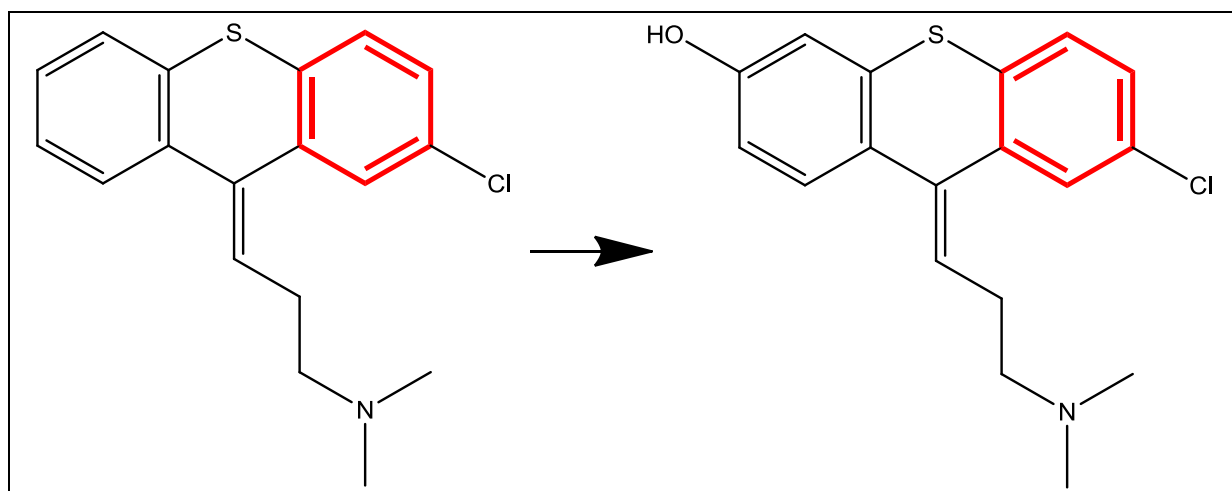


Figure 50: The bonds in the aromatic ring highlighted in red appear to change order between the substrate and metabolite structures. These, however, do not represent reaction centres: the two structures are equivalent resonance forms of the delocalised π -system. (MDLNUMBER: RMTB00007698)

In other cases, however, a transformation really has occurred:

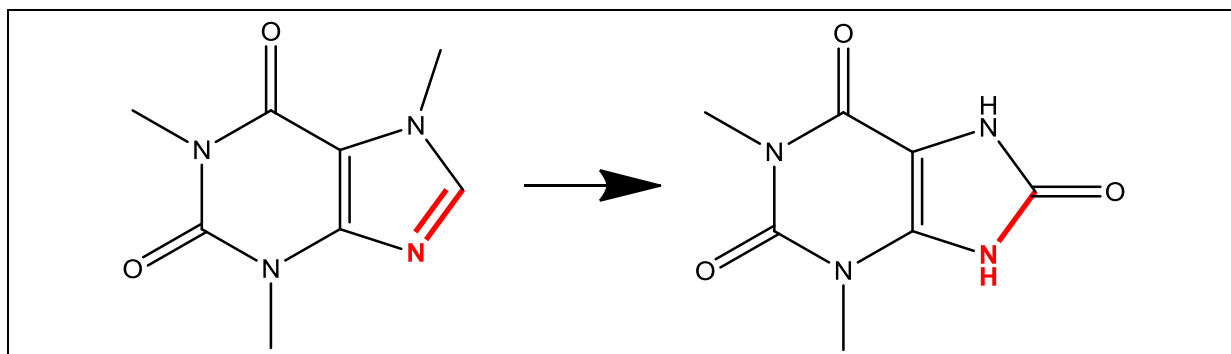


Figure 51: In this case the change in order of the aromatic bond highlighted is due to a metabolic transformation, and is not the result of an alternative resonance structure. (MDLNUMBER: RMTB00004389)

In Figure 51 both of the five-membered rings containing the highlighted bond can be considered aromatic by chemical information systems using Huckel's $4n+2$ electron rule. However, the change in bond order is not simply down to a change of resonance structure. This can be recognised in two ways: firstly, the change in bond order is not a concerted change of an alternating single/double bond system, and secondly a hydrogen atom has been added to the nitrogen atom, increasing its number of substituent atoms.

3.4.4 Fingerprint generation

MetaPrint2D utilises the same type of circular atom environment fingerprints as SPORCalc (see Page 68) to represent the chemical environment occupied by atoms. These atom-centred fingerprints consist of lists of the atom types encountered at successive topological distances from a central atom. MetaPrint2D first assigns SYBYL[®] atom types (251), listed in Table 5 below, to each atom in a structure. Fingerprints are then generated for each atom in turn, by means of a depth limited breadth-first search (BFS) encompassing the central atom and all atoms up to five bonds distant. At each depth within the search, a list of the atom types encountered at that depth, together with their frequency of occurrence, is recorded. Atoms are only recorded the first time they are encountered by the BFS, regardless of any cycles in the structure. These atom type lists form the basis of the fingerprints used in MetaPrint2Ds.

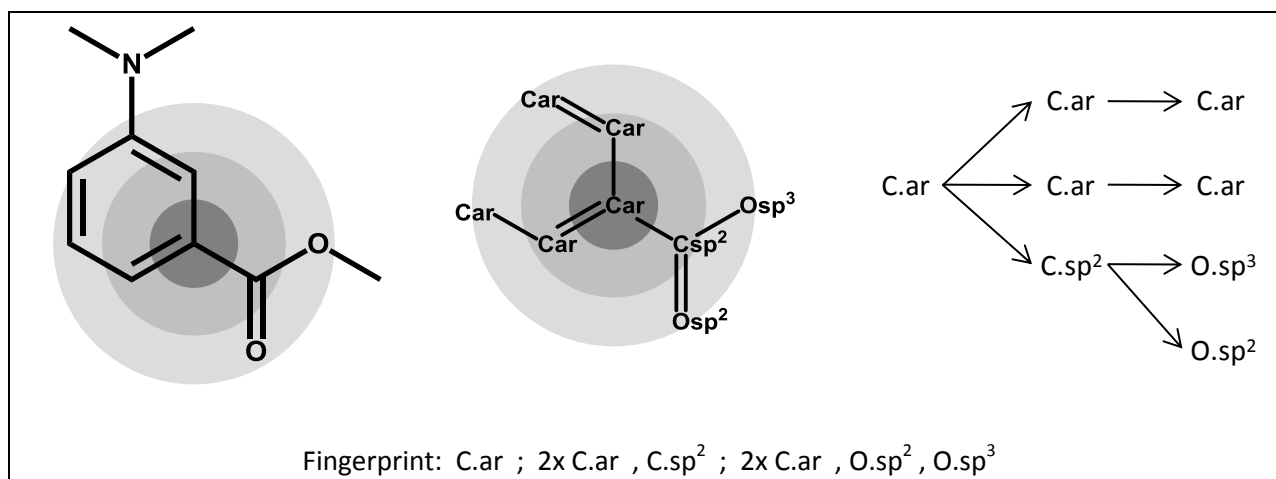


Figure 52: Atom environment fingerprint generation. Neighbouring atoms are visited via breadth-first search, and the number of occurrences of each atom type at each topological distance is recorded.

Atom type	Description	Atom type	Description
C.3	sp ³ hybridized carbon	N.3	sp ³ hybridized nitrogen
C.2	sp ² hybridized carbon	N.2	sp ² hybridized nitrogen
C.1	sp hybridized carbon	N.1	sp hybridized nitrogen
C.ar	aromatic carbon	N.ar	aromatic nitrogen
C.cat	carbocation (C ⁺)	N.am	amide nitrogen
O.3	sp ³ hybridized oxygen	N.pl3	planar 3-coordinate nitrogen
O.2	sp ² hybridized oxygen	N.4	positively charged sp ³ nitrogen
O.co2	carboxylate/phosphate oxygen	Li	lithium
S.3	sp ³ hybridized sulphur	Na	sodium
S.2	sp ² hybridized sulphur	K	potassium
S.O	sulphoxide sulphur	Ca	calcium
S.O2	sulphone sulphur	Al	aluminium
P.3	sp ³ hybridized phosphorus	Si	silicon
F	fluorine	H	hydrogen
Cl	chlorine	Du	dummy atom
Br	bromine	LP	lone pair
I	iodine		

Table 5: The SYBYL[®] atom types used by SPORCalc and MetaPrint2D.

3.4.5 Fast fingerprint searching

As with SPORCalc, the model generation process for MetaPrint2D can be carried out beforehand, and the data saved, so the efficiency of this step does not affect the experience of users when making a query. However, the speed of the searching of this data contained in the model is vital in determining the user's experience. Ideally predictions should be fast enough to be made in 'real-time' as far as a user is concerned, enabling immediate visualisation of the effects of alterations to a structure.

SPORCalc's database simply listed the environment of each atom in the Symyx® Metabolite database in a set of data files; one file contained the atom environments of all the reaction centres in the database, and other files contained the full list of atom environments, separated by the type of their central atom. Predictions were made by iterating through all these files, comparing each atom environment fingerprint to the fingerprints of the atoms in the query structure, and keeping count of the number of reaction centre and substrate hits.

To hold all this data, SPORCalc's data files were very large – in the region of 600MB for each model. Storing and searching this quantity of data led to SPORCalc taking several minutes per molecule to generate predictions. A number of alternative approaches to storing and searching the model have been investigated, with the aim of increasing the speed of prediction.

The 2008.1 release of the Symyx® Metabolite database contains 1352387 atoms, occupying 166766 distinct environments. Storing a list of atom environments, each with pre-computed reaction centre and substrate occurrence counts, rather than recording each atom individually leads to a reduction in both file size and computation time of almost 90%.

Input/Output operations (reading from/writing to a disk) are very time-consuming in comparison to equivalent operations on data stored in Random Access Memory (RAM). Calculations can be performed much more rapidly if MetaPrint2D's dataset can be held entirely in RAM, rather than being read from disk on every use. Each fingerprint consists of six levels of 33 bins, each of which contains a value from a small range (typically 0-5). Storing each bin of the fingerprint in a single byte of memory, a fingerprint would take up 198 bytes. This means that the fingerprints for the 165951 distinct environments found in the 2008.1 release of the database would require around 31MB of memory. Additional

memory is needed for the storage of occurrence counts, and overheads associated with the data structures, but that will not expand the memory requirements to the point where this quantity of data cannot easily be stored in the memory of modern computer systems, which typically have gigabytes of RAM available.

Since MetaPrint2D is intended to be used as a library, potentially integrated into larger systems, it is still desirable to minimize the storage requirements as much as possible as MetaPrint2D may not have access to the computer system's entire resources. Additionally, under its default settings the Java virtual machine only has access to a small proportion of the host computer's resources. Users may wish to work with models generated with various constraints (e.g. Human/Rat...), so it is beneficial to be able to hold several models in memory at the same time.

Memory usage has been further reduced through exploitation of the hierarchical structure of the fingerprints. At each level, the fingerprint contains a count of the number of occurrences of each atom type at that distance from the atom on which the fingerprint is centred. These single-level sub-fingerprints are often identical to a single-level of many other full six-level fingerprints. For example, all fingerprints centred on an aromatic carbon will have an identical first level (containing a single C.ar typed atom), and many will have an identical second level (containing two C.ar typed atoms, and nothing else). This structure can be exploited in two ways. Firstly, if the fingerprints are sorted on the basis of the hierarchy of single-levels fingerprints, only the sub-fingerprints for the levels differing from the previous fingerprint need be stored. Alternatively the memory requirements of the model can be reduced through caching these single-level sub-fingerprints in memory, and having the six-level fingerprints share instances of them.

These approaches are illustrated in Figure 53.

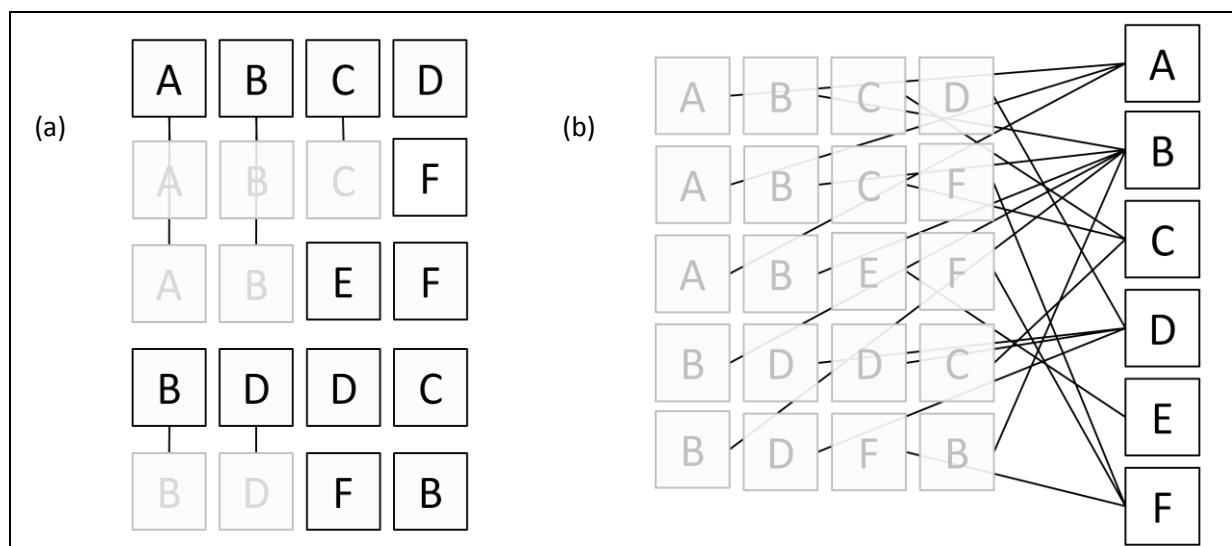


Figure 53: Illustration of the memory savings of the schemes described above. Each row of cells (e.g. ABCD) illustrates a full fingerprint, formed from a number of sub-fingerprints. The sub-fingerprints represented by cells in a dark colour must be stored in full, while those greyed out can reference a copy of the data stored elsewhere, thereby saving memory. (a) Hierarchical difference: if the fingerprints are sorted according to their hierarchy, only the differences from the previous fingerprint need be stored. (b) Single-level caching: alternatively, a single copy of each sub-fingerprint can be kept in a pool, with all the relevant fingerprints referencing that copy.

The memory savings afforded by these optimizations are shown in Table 6, below.

Storage Method	Memory Usage / MB
No optimization	77.05
Hierarchical difference	44.51
Single-level caching	27.02

Table 6: The memory requirements for storage of a MetaPrint2D model constructed from all the data in the 2008.1 Symyx® Metabolite database. No optimization: full fingerprints held in memory; Hierarchical difference: hierarchical structure of fingerprints exploited; Single-level caching: single-level sub-fingerprints cached, and shared between full six-level fingerprints. (Memory usage recorded on a Dell Inspiron 6400 laptop with Intel Core 2 T5300 @ 1.73GHz; 3.24GB RAM)

Indexing

In order to further reduce the time required to search the atom environment data it is indexed as it is loaded into memory. This means that when a search is performed, rather than having to iterate through the entire dataset, the relevant portions can be rapidly

retrieved and the number of more computationally demanding similarity calculations performed kept to a minimum.

In the standard parameterizations of SPORCalc the similarity searches were carried out with the constraint that the first one, two or three levels must exactly match the query fingerprint. This clearly lends the data to being indexed on the first three fingerprint levels. In order to facilitate flexibility in searches, constrained to any level of exact matches, each of the first three levels of fingerprint is indexed separately. The alternative would be to index level 1, level 1+2, level 1+2+3 – which would require much more memory, for only a small extra gain in performance.

Each index takes the form of a hash map (269). This is a data structure mapping identifiers (keys) to associated values in an efficient manner. The indexes in MetaPrint2D use single-level sub-fingerprints as keys, and the set of all fingerprints having that pattern at the indexed level as the corresponding value. When a search is performed, the sets of data having the required fingerprint at each level are looked up, and the conjunction of all such data (i.e. those data points with the correct sub-fingerprint at every level) is returned. Set operations are very fast, and in this instance their speed is further increased by considering the index search results in order of increasing size, which minimises the number of calculations to be carried out.

A hash map consists of a simple sequence of buckets, each of which can hold one or more data items (key/value pairs). A *hash function* is employed to calculate the index of the bucket in which a data item should be stored, from the item's key. When fetching a data item from the hash map only a single bucket has to be inspected, rather than searching all the data, allowing for very fast retrieval. Hash maps can offer “constant time performance” (269) – meaning that the retrieval time is independent of the number of items stored in the hash map, as opposed to storing data in a simple list where search time scales rapidly with the quantity of data. Once a certain capacity has been reached the number of buckets is increased, and the data items redistributed between them. MetaPrint2D makes use of the standard implementation of a hash map data structure provided by the Java language (`java.util.HashMap`).

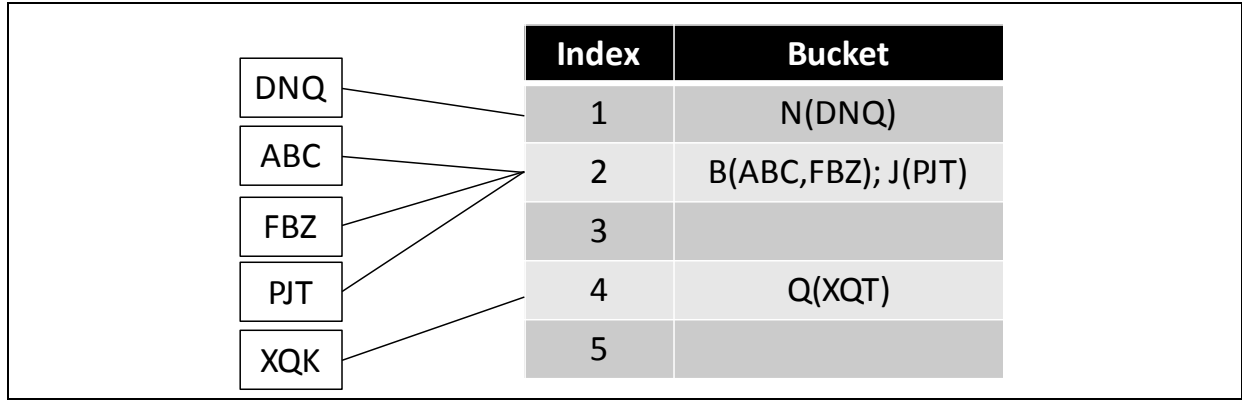


Figure 54: Illustration of the indexing of fingerprints by a hash map. The boxes represent three level fingerprints, with each letter representing a single-level sub-fingerprint. The fingerprints are stored in a hash map, indexed by their second level. Each fingerprint is mapped to a bucket based on its second level sub-fingerprint and a hash function. All of the fingerprints having a sub-fingerprint in common are stored in a set in the corresponding bucket.

The performance of a hash map is dependent on the quality of the hash function used. The hash function should be quick to calculate, and distribute the data items uniformly between the buckets, with few collisions (generating the same value for items with different keys). The hash function used by MetaPrint2D is based on the hash function for strings described in the Java language specification (270):

$$hash\ code = \sum c_i \times 31^i$$

In both cases a hash function is required for a similar data structure; a string consists of a sequence of characters (represented by their numeric ASCII/UNICODE values) and a MetaPrint2D fingerprint consists of a sequence of bin values. Through trial and error it was determined that in order to minimise hashing collisions, optimal parameters for the exponential term in the hash functions for MetaPrint2D fingerprints and sub-fingerprints are 63 and 15, respectively.

Since the fingerprints in MetaPrint2D are immutable (un-changeable) the output of the hash function for a particular fingerprint will never change. This means that performance can be further increased by calculating the hash of a fingerprint a single time, and caching the value with the fingerprint.

Data file size and loading times

A number of optimizations to the data file were investigated, in order to decrease its size and speed-up loading. As mentioned above, rather than storing the fingerprint for each atom of the Symyx® Metabolite database, the total occurrence counts for each distinct atom environment are stored.

Within each atom environment fingerprint, the majority of the atom types at each single-level have a count of zero, so, rather than storing the value of every bin in the full fingerprint a sparse fingerprint representation is employed, where only the indexes of the non-zero bins, together with their values, are stored. The hierarchical nature of the fingerprints is also exploited. The inner-layers of the fingerprints are much less variable than the outer layers, so rather than storing each fingerprint in full, when writing the data file the fingerprints are sorted in an ascending order, and only the single-level sub-fingerprints that differ from those of the previous fingerprint are stored, in a similar manner to that described on page 110.

SPORCalc stored fingerprints in an ASCII text format: a space delimited string of numbers. In order to store the fingerprints as numerical values in memory, as required by the similarity calculations, this string must be split into a list of numbers, and then the text representations of the numbers converted to their numeric equivalents. This computation can be reduced, decreasing the data's loading time, by storing the data in binary format, so the byte values can be read directly from the files, removing the need for conversion.

As shown in Table 7, together these optimizations have reduced the size of the data files for a model from over half a gigabyte to just over three megabytes, or well under one megabyte if GZip compression is applied to the file, and loading times have reduced from around two and a half minutes to well under one second, on commodity hardware (Dell Inspiron 6400 laptop Intel Core 2 T5300 @ 1.73GHz; 3.24GB RAM).

File format	File size/MB	t ₁ /s	t ₂ /s	t ₃ /s	t ₄ /s	t ₅ /s	t _{av} /s
6lvfp files	558.9	149.08	146.86	146.55	147.59	145.20	147.03
Text file	63.8	17.09	16.49	16.47	16.75	16.56	16.67
Bin file	32.7	4.70	4.13	4.31	4.42	4.28	4.37
Sparse diff file	3.35	0.84	0.68	0.63	0.61	0.61	0.67
Compressed file	0.839	0.92	0.74	0.68	0.67	0.68	0.73

Table 7: File sizes and load times (recorded on a Dell Inspiron 6400 laptop Intel Core 2 T5300 @ 1.73GHz; 3.24GB RAM). 6lvfp files = six level fingerprint files – full list of atom fingerprints as used by SPORCalc; text file = atom environment and occurrence counts stored in ASCII format; bin file = atom environment and occurrence counts stored in binary format; sparse diff file = binary file with sparse fingerprints exploiting hierarchical structure; compressed file = sparse diff file compressed using GZip compression. The data are for models generated from the 2008.1 Symyx® Metabolite database, containing 1,352,387 atoms occupying 166,766 distinct environments. Timings have been recorded on five independent runs, each of which is reported, together with their mean.

3.5 Software availability

The core of MetaPrint2D's calculation engine has been designed as a self-contained library, providing a straightforward Application Programming Interface (API). This enables a variety of different user-interfaces to be developed, and makes it straightforward for MetaPrint2D to be embedded into larger applications.

Three interfaces to the MetaPrint2D library, designed to facilitate a range of use cases, have currently been produced: a website, a command-line utility and a plug-in for the Bioclipse rich client platform. These are now described, and their relative merits and disadvantages, along with potential future applications are discussed.

3.5.1 Web site

The first interface provided is a website, hosted at the Unilever Centre for Molecular Science Informatics in the Cambridge University Chemical Laboratories (<http://www-metaprint2d.ch.cam.ac.uk/>). This interface is the most straightforward to use, requiring no set-up or configuration, just access to a graphical web browser which comes preconfigured on almost all modern computers. Users can input molecules using SMILES, or sketch a structure using the JME editor (271), and results are clearly presented in a form a chemist

will immediately recognise. Predicted sites of metabolism are highlighted using a traffic-light system. Full details of the results are accessible by moving the cursor over an atom. Any regions of the molecule that are not well covered by the database are highlighted in grey, giving the chemist insight into the reliability of the model's predictions (this will be discussed further in the next chapter).

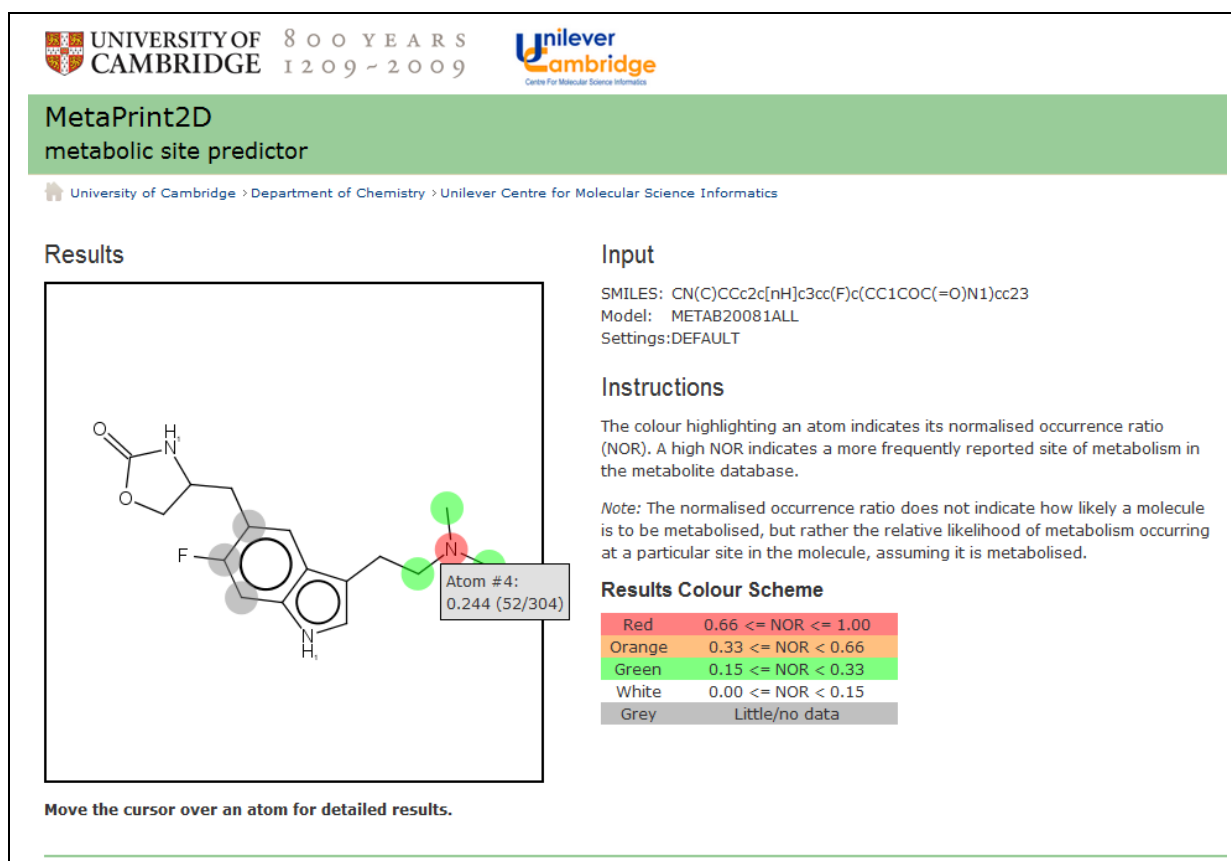


Figure 55: Results of a query carried out through the MetaPrint2D website.

Installing and maintaining the MetaPrint2D software on a central server, rather than requiring users to set-up and maintain the code on their own machines, makes it simple to keep the application up-to-date and means that users can be sure they are accessing the most recent version of the software and data files. There are, however, downsides to the server-client model. The software is only accessible when users are connected to the internet, and the server provides a single point of failure, vulnerable to heavy usage or malicious attacks, though this can be mediated through techniques such as queuing computationally expensive tasks, and limiting the frequency with which users can make requests. If demand were to grow, then flexible cloud-like compute resources, such as the Amazon Elastic Compute Cloud (EC2) (272) could be used to increase service availability at

times of high demand, without having to maintain the infrastructure of a server pool with a large amount of redundant capacity.

There can also be issues regarding security of intellectual property (IP) rights when working with remote services. Users may be wary of submitting confidential molecules across the internet; indeed many companies have absolute bans on doing so, though this position is becoming less prevalent with the growth in usage of Electronic Laboratory Notebook (ELN) systems, with the greater IP protection they afford (273).

The web-based version of MetaPrint2D can also be packaged up with a small web-server, producing a stand-alone version, ideal for demonstration purposes. This removes the need for a connection to the internet, and running the service locally also means that there can be no concerns concerning the transmission of confidential data across public networks. The web-based version of MetaPrint2D cannot, however, currently be distributed due to license restrictions prohibiting the distribution of the JME editor, which is currently used for structure input.

The current web interface could readily be adapted to provide a SOAP or RESTful 'webservice' interface facilitating integration with workflow tools and other remote applications.

3.5.2 Command-line Utility

A command-line based interface for MetaPrint2D has also been developed and released. This interface can take as its input a single SMILES string, a file containing a list of SMILES, or an SDF file containing one or more molecules, and generates site of metabolism predictions for each molecule, and optionally images displaying the likely sites of metabolism of the compounds, similar to those produced by the website (shown in Figure 55, above). The command-line MetaPrint2D application carries out computations locally on the user's computer, so does not require internet access to run, and removes the IP considerations surrounding the transmission of potentially sensitive data to remote services. Use of the command line tool does however require some degree of technical expertise, and as such is more appropriate for the power-user wishing to batch-process a large number of compounds, or integrate MetaPrint2D's site of metabolism predictions into a script or workflow.

The command-line application's batch processing mode is able to fully leverage the speed of the MetaPrint2D calculation library, enabling high-throughput virtual screening of large compound collections.

3.5.3 Bioclipse plug-in

The third interface to MetaPrint2D currently available is a plug-in for the Bioclipse, created in collaboration with the developers of Bioclipse*. Bioclipse (274) is an open source chemo- and bioinformatics platform, built on the Eclipse (275) rich client platform.

The integration with Bioclipse provides the most powerful interface to the MetaPrint2D library. Chemists are able to draw molecular structures into the editor, in a manner common to many other applications, and visualise how predicted sites of metabolism change as they modify the structure, in real time – the screenshot below shows that Bioclipse was able to capture the molecular structure from the editor, assign the required atom types, generate site of metabolism predictions with the MetaPrint2D library, and render the results of those predictions in the editor, all in 172 milliseconds. Bioclipse is also able to run MetaPrint2D over large files of structures, predicting sites of metabolism for each structure, and displaying the output in a scrollable table, and due to its speed, this is a fairly trivial task.

* The MetaPrint2D plug-in for Bioclipse was written by Ola Spjuth of Uppsala University, one of the developers of Bioclipse, with assistance from the author of MetaPrint2D. The plug-in provides an interface between Bioclipse's internal data-structures and user interface, and the MetaPrint2D library.

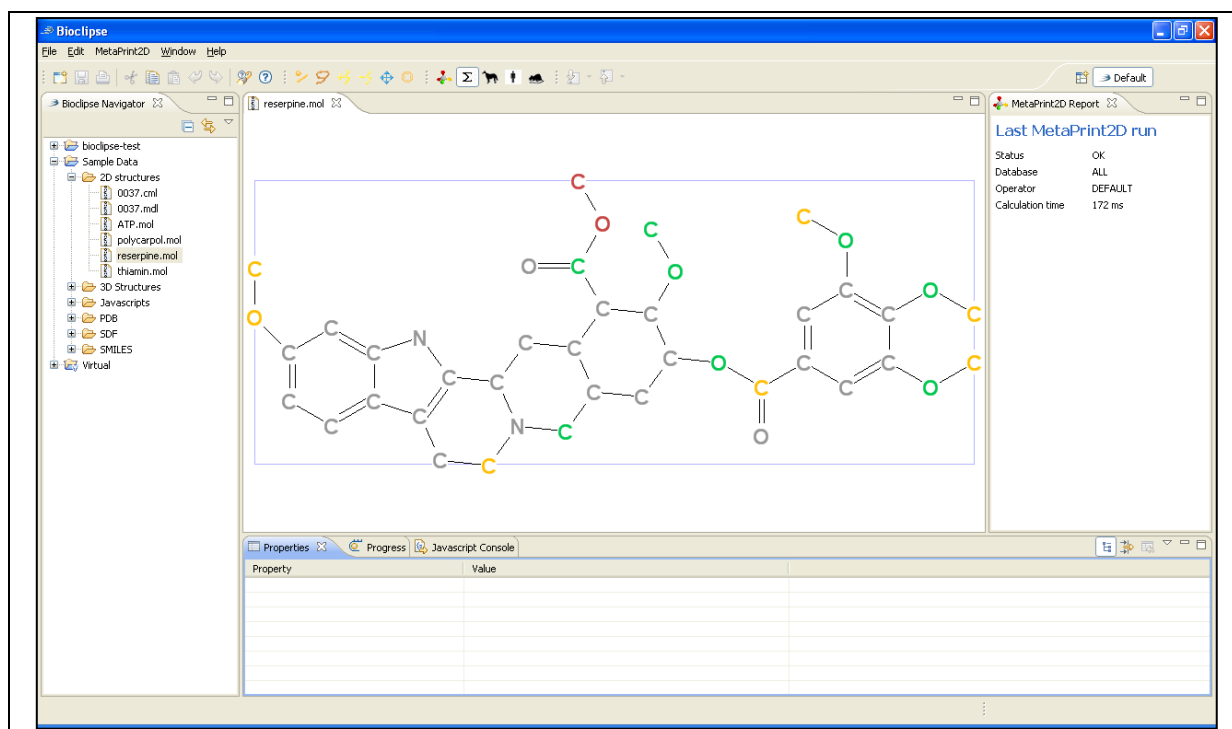


Figure 56: MetaPrint2D running in Bioclipse.

As with MetaPrint2D's command-line application, Bioclipse removes the considerations regarding intellectual property rights that are associated with use of the MetaPrint2D website, since all calculations are carried out locally on the user's computer, and no information is transmitted across the Internet. The downside of the application is, however, that a significant effort may be required for initial set-up and configuration – the Bioclipse application must be installed and configured, and the MetaPrint2D plug-in then added. In a corporate environment where use of computer systems is often governed by strict security policies, approval may be required before Bioclipse can be installed, and the system administrator's assistance may be required for installation. Users will also need to take time to familiarise themselves with the more complex interface than that presented by the website.

3.5.4 Other possible applications

The design of MetaPrint2D, placing the core calculation engine into a library independent of any user interface, means that many applications and interfaces beyond those described here can be developed. The work undertaken in collaboration with the Bioclipse project has shown how MetaPrint2D can be integrated with other applications. Similar work could be

carried out with chemical editors and electronic laboratory notebook systems, in order to make MetaPrint2D's predictions more readily available to chemists.

The potential to use MetaPrint2D for high-throughput virtual screening or as a component in a workflow has already been mentioned in connection with the command-line application. This could also be achieved through direct integration of the MetaPrint2D library with a workflow engine, possibly leading to the development of a molecular descriptor based on the likelihoods of sites in a molecule being metabolised. Anecdotal evidence from users of SPORCalc within AstraZeneca suggests that compounds with three or more highly likely sites of metabolism (scoring 'red' in the web interface's traffic-light system) are highly metabolically labile, and potentially toxic. Observations such as this could lead to MetaPrint2D's integration with some sort of structural alerts or warning system.

3.5.5 Licensing

The MetaPrint2D library has been released as an Open Source project, hosted on the SourceForge community site (<http://sourceforge.net/projects/metaprint2d/>). The code is published under the GNU Affero General Public License (AGPL). This ensures that the MetaPrint2D software will be made as widely available as possible, and permits any individuals and organisations to freely use and modify the code to suit their needs, with the proviso that anyone wishing to 'convey' (distribute, or make available to others through a web service) copies of MetaPrint2D, or any derivative works, must also make available the source code containing their modifications under the terms of the AGPL. This will ensure that future development of the MetaPrint2D library will benefit the whole community.

4. Evaluation and optimization of MetaPrint2D

This chapter describes the evaluation and optimization that have been performed on MetaPrint2D. The result of the assessment of MetaPrint2D's reaction centre identification algorithms is reported. A number of data pre-processing steps are proposed and their effects investigated, and the reliability of MetaPrint2D's predictions analysed. MetaPrint2D models generated from data on specific cytochrome P450 isoforms are also discussed, as is the quality of the available test data, and the results of MetaPrint2D's evaluation have been compared to that reported for other site of metabolism prediction tools.

4.1 Reaction centre identification

In order to evaluate the reaction centre detection the results of an analysis of 300 randomly sampled transformations from the 2008.1 release of the Symyx® Metabolite database were manually inspected. In order to do this a 'debug' application (shown in Figure 57) was created. This displays the substrate and metabolite structures in two adjacent panels. The MCS atoms and bonds are displayed in a solid colour, and the remainder of the structure is shown 'greyed out'. Atoms identified as reaction centres are also highlighted.

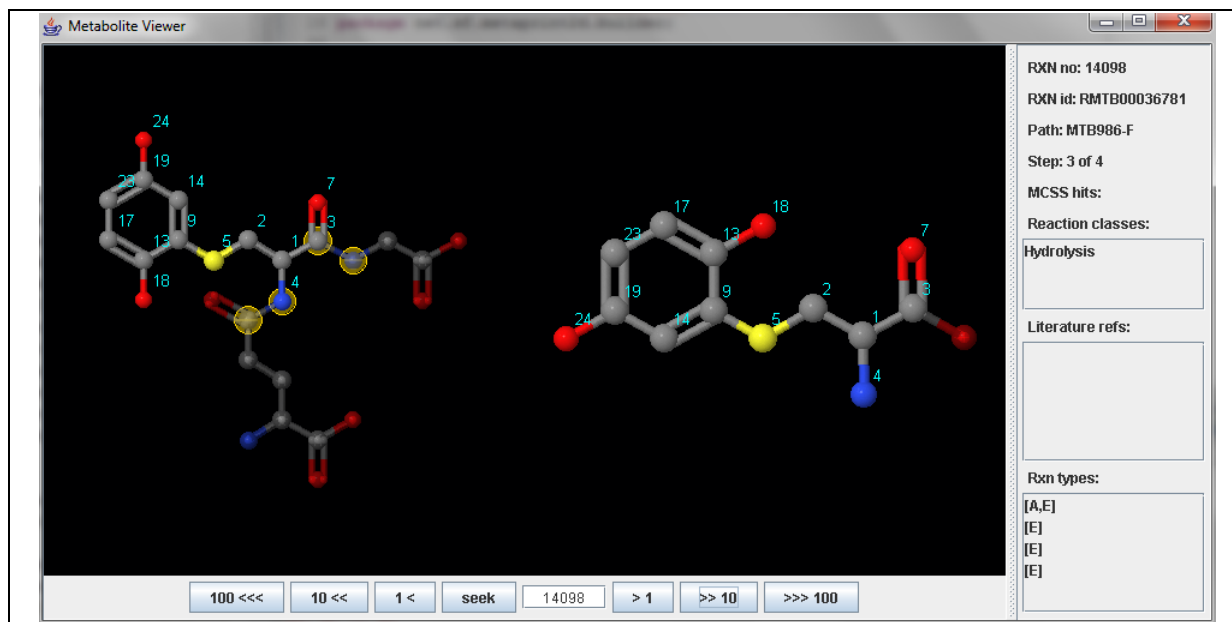


Figure 57: Metabolite analyser debug viewer used to evaluate reaction centre identification. The conserved structure and identified reaction centres are highlighted.

For eight of the selected transformations it was not possible to carry out any analysis as the metabolite structure was not specified in the Metabolite database. Of the remaining 292 transformations examined only two (0.7%) caused problems, with the conserved structure correctly identified in all the other cases. The problem transformations are shown in Figure 58.

In the case of transformation *RMTB00003542*, we believe it is likely that the detected MCS is correct; however no confirmation of this is available from the Symyx® Metabolite database, since this entry does not contain any mapping annotations. In the case of transformation *RMTB00043645*, however, the detected MCS is in error; the transformation is a demethylation, and the true MCS is shown in (b).

This error results from the structure of the Metabolite database: each transformation record contains only a single substrate and metabolite molecule. In the case of a reaction producing several products, there can be a number of records in the database, one recording for each product, though many reactions only record the largest/major metabolite formed. For the majority of demethylation reactions in the Metabolite database the fate of the methyl group is not reported.

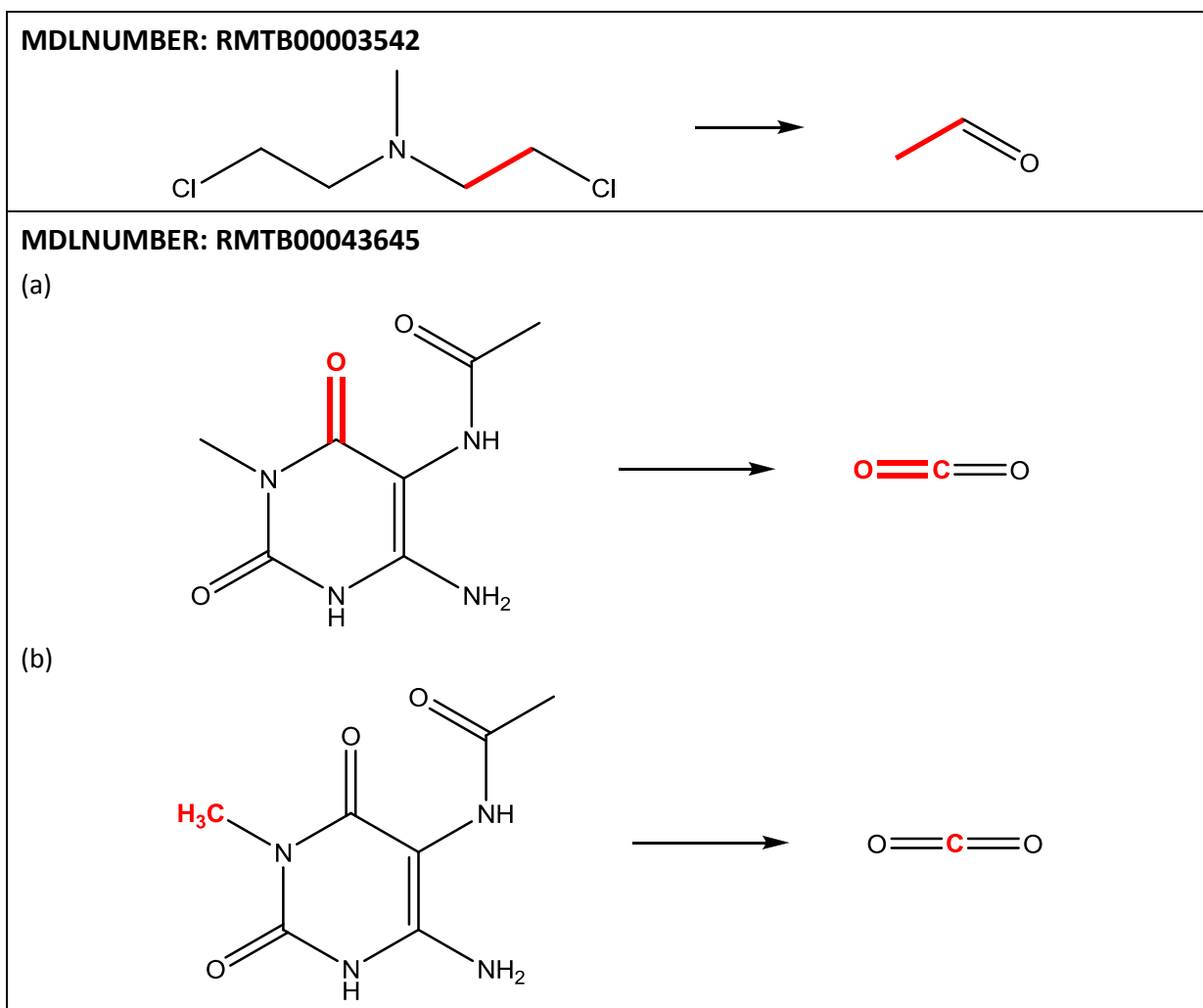


Figure 58: Problem transformations identified during the evaluation of the reaction centre identification algorithm. These are discussed in the paragraph above.

4.2 Pre-processing of Symyx® Metabolite data

Models for the SPORCalc metabolic site predictor were constructed using every transformation in the Symyx® Metabolite database. During the development of MetaPrint2D a number of possible data pre-processing steps were identified, some or all of which were thought to potentially improve the quality of the models generated. In order to determine whether any of the pre-processing steps should be used, models were trained applying each of these techniques and their performance evaluated and compared.

4.2.1 Multi-step transformations

Related transformations in the Symyx® Metabolite database are organised into 'schemes' organised around a parent compound, as illustrated in Figure 59.

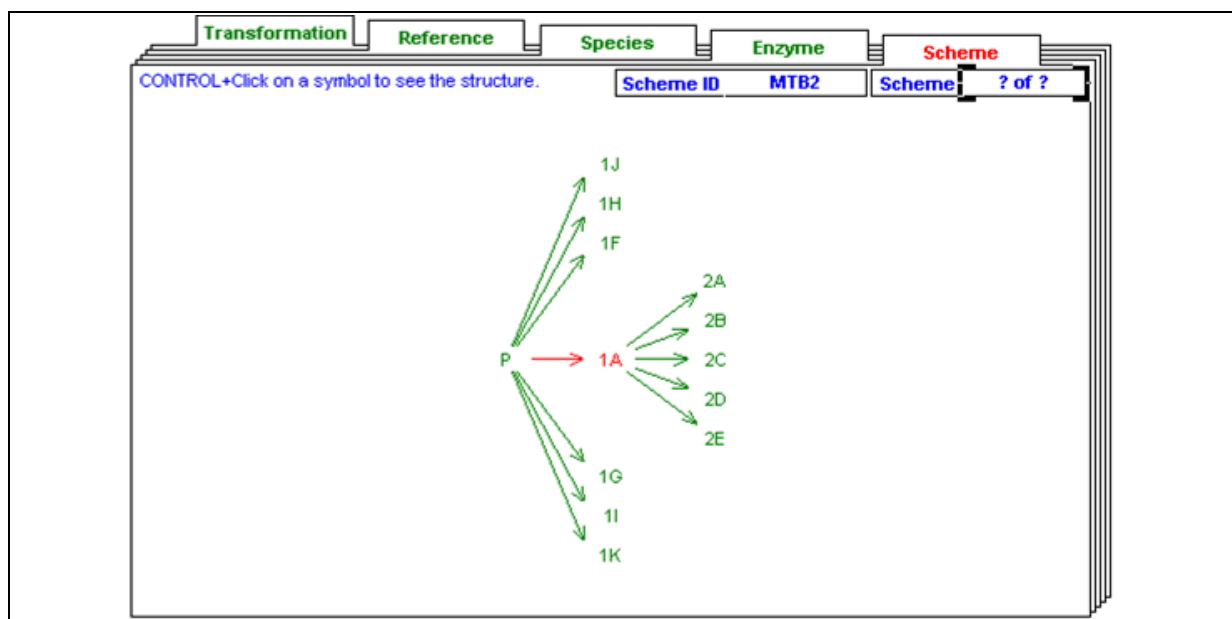


Figure 59: Metabolic scheme MTB2, from the Symyx® Metabolite database, 2008.1. The database contains 17 transformation records in this scheme – 12 corresponding to the transformations shown by arrows above, and a further five representing the overall transformation from the parent compound to each of the second generation metabolites 2A-2E.

The database contains a separate entry for each transformation in the scheme, e.g. $P \rightarrow 1A$ (representing the transformation from the parent compound P to the first generation metabolite $1A$), $P \rightarrow 1J$, $1A \rightarrow 2A$, and also an additional entry for each final product not formed in a single step from the parent compound e.g. $P \rightarrow 2A$, $P \rightarrow 2B$. When analysing the records representing the overall transformation of a multi-step reaction path, the changes can be so great that it becomes difficult to determine the conserved structure between the parent compound and the final metabolite, and hence determine the sites of metabolism. This is particularly true of longer pathways; the Metabolite database contains reaction schemes with pathways up to 13 steps deep.

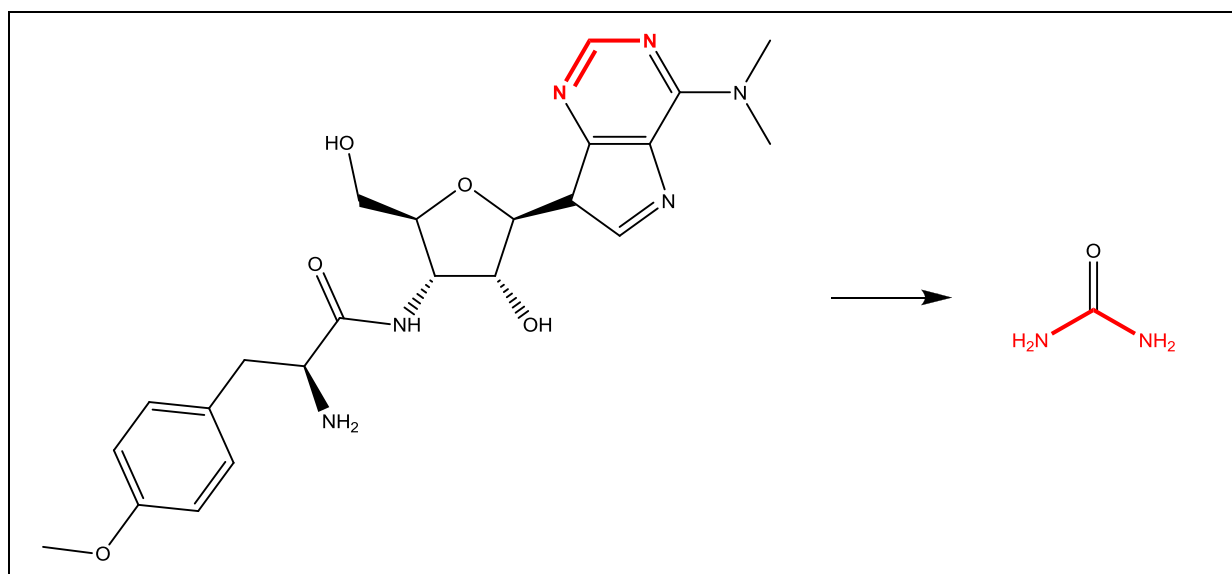


Figure 60: The overall transformation resulting from a seven step metabolic pathway (MDLNUMBER: RMTB00005520). It is not immediately obvious how the metabolite structure maps to that of the parent compound, or which atoms are sites of metabolism. Manual inspection of the individual steps of the pathway has identified the conserved substructure, and this is highlighted in red.

Inclusion of multi-step records in the construction of MetaPrint2D models could cause two further problems. As illustrated by the multi-step transformation shown in Figure 61, atom environments found to be sites of metabolism in the overall transformation may not correspond to sites of metabolism in any of the individual steps making up the pathway. Additionally, inclusion of the extra transformations leads to double (or higher) counting of some reaction centre and substrate atom environments, distorting the model.

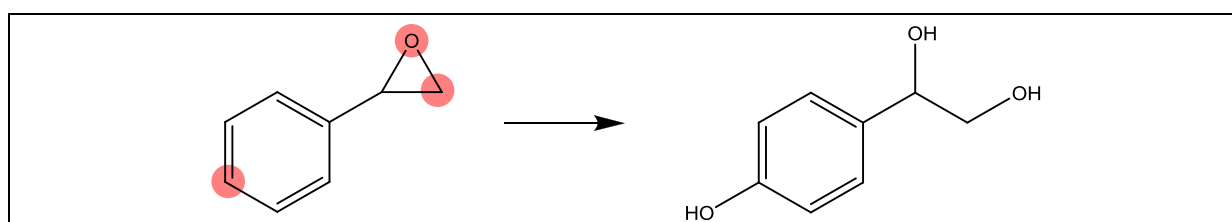


Figure 61: Analysis of this overall transformation (MDLNUMBER: RMTB00052516) suggests the presence of sites of metabolism (highlighted red) at both ends of the parent compound. However, inspection of the individual steps in the reaction scheme shows that hydrolysis of the reactive epoxide always occurs before the hydroxylation of the phenyl ring.

4.2.2 Per-transformation versus per-molecule

There is a second means by which the generation of MetaPrint2D models from the transformations in the Symyx® Metabolite database can lead to double (and higher) counting of the atom environments and reaction centres in certain compounds. This is through the repetition of substrate compounds for different transformations.

Considering the seven first generation metabolites of the parent compound 'P' in the reaction scheme shown previously (Figure 59, page 123), those atoms never occurring at a reaction centre will be recorded as such seven times – once for each transformation. The atom at the reaction centre for the transformation to '1K' will be recorded once as occurring at a site of metabolism and six times as not occurring at a site of metabolism. The atoms involved in the remaining 1-step transformations will each be recorded as occurring at a site of metabolism twice, and as not occurring at a site of metabolism five times. If the multistep transformations directly between 'P' and '2A-E' are included, as just discussed, then the picture becomes even more complicated. This is illustrated in Figure 62.

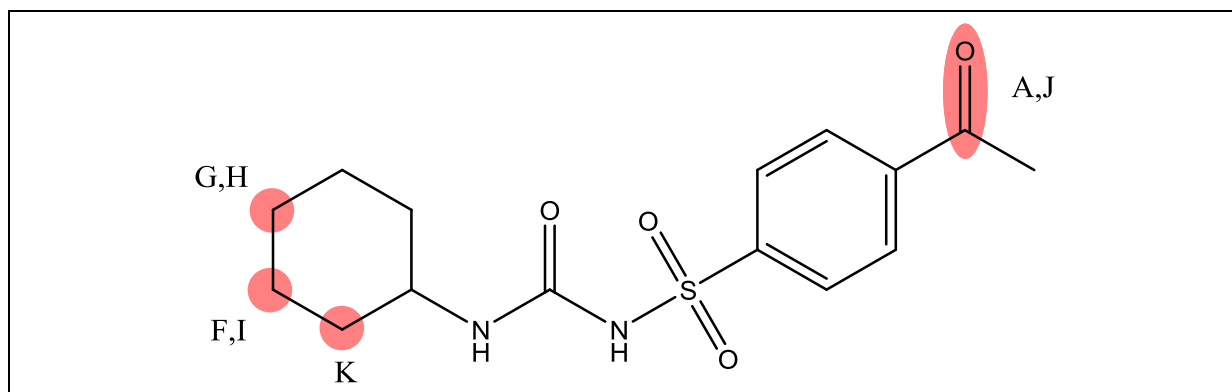


Figure 62: The sites of metabolism (highlighted red) from all the transformations of a compound (MDLNUMBER: MMTB00000002). The letters adjacent to sites of metabolism indicate the products resulting from metabolism at that site.

4.2.3 Symmetry

A further consideration is the metabolism of molecules exhibiting symmetry. In instances where a site of metabolism exists in a symmetrical region of a molecule, as illustrated in Figure 63, the occurrence counts for the environment of the atom at the site of metabolism are updated twice, once as occurring at a reaction centre, and once as not. In fact, the metabolic transformations are equally likely to occur at the equivalent atoms.

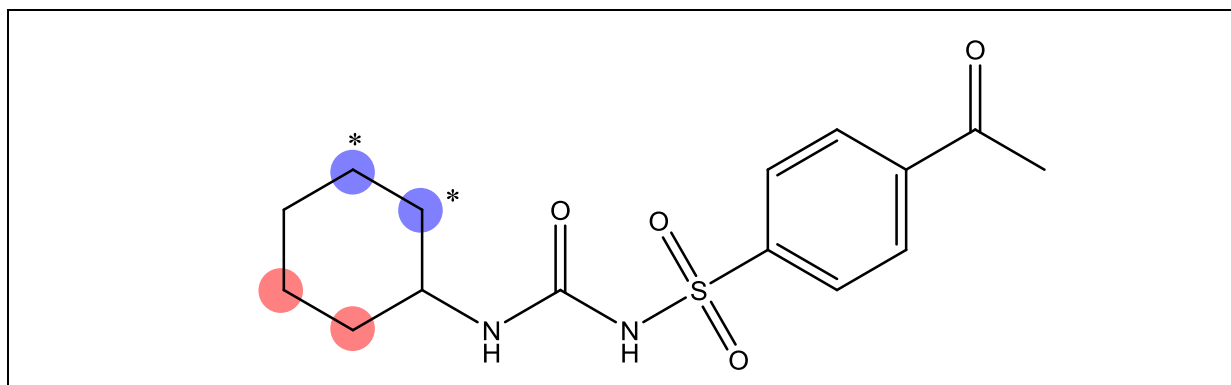


Figure 63: An example of a compound (MDLNUMBER: MMTB00000002) with sites of metabolism (highlighted in red) occurring in a symmetrical region of a molecule. The atoms marked with an asterisk are not recorded as sites of metabolism, despite being chemically identical to those from transformations on the opposite side of the ring.

4.2.4 Duplicate transformations

Since the metabolism of different parent compounds can produce a common metabolite, which may be metabolised further, and the Symyx® Metabolite database is organised into schemes structured around a parent compound, some transformations appear in the Metabolite database multiple times. This raises the question of whether recording all reports of such transformation biases the model, and whether such transformations should be identified and only recorded a single time.

It was considered that this repeated counting of some atom environments could distort the results, though it is possible that this duplication could actually improve the results, by increasing the weighting of regularly occurring environments.

4.3 Evaluating metabolic site predictions

4.3.1 Current approaches to evaluation

A number of recent studies have included evaluations of the ability of various software tools to predict sites of metabolism on molecules e.g. (218,235,234,276). Two main approaches to assessing performance have been followed: a qualitative analysis *via* the visual inspection of a tool's output compared to the known sites of metabolism of a molecule and a quantitative analysis. The quantitative analysis reports the percentage of molecules for which the highest ranked predicted site of metabolism is an experimentally observed site,

and the percentage of molecules for which at least one of the top three ranked sites is an experimentally observed site of metabolism.

In the evaluation of MetaPrint2D the percentage occurrence of experimentally observed sites of metabolism in the top one and top three ranked hits has been calculated, in order to enable some degree of comparison to previous studies; however these commonly used test metrics contain some intrinsic flaws. The expected values of the test metrics are dependent on both the sizes of the molecules under investigation, and the number of sites of metabolism each possesses. This leads to a bias towards higher test scores for studies on smaller molecules or molecules with a greater number of sites of metabolism.

Figure 64(a) shows a box plot of the distribution of sizes of substrate molecules in the Metabolite database, and the wide variation in size of metabolised compounds is clearly visible. Considering a molecule with ten atoms, metabolised at a single site; ranking the atoms at random there is a 10% chance of metabolism occurring at the highest ranked site, and a 30% chance of it occurring at one of the top three ranked sites. In comparison, for a molecule with 20 atoms, also metabolised at a single site, the chances of metabolism occurring in the top one or top three randomly ranked site are 5% and 15%, respectively – half that of the molecule with ten atoms.

The number of sites at which metabolism has been reported to occur can also vary widely between molecules, as illustrated by the compounds in Figure 64(b). In benzopyrene metabolic transformations have been reported to occur at 60% of the non-hydrogen sites in the molecule, while for 5-chloro-2-mercaptobenzothiazole metabolism has only been reported at a single site – meaning there would be only a 9% chance of selecting the correct site at random.

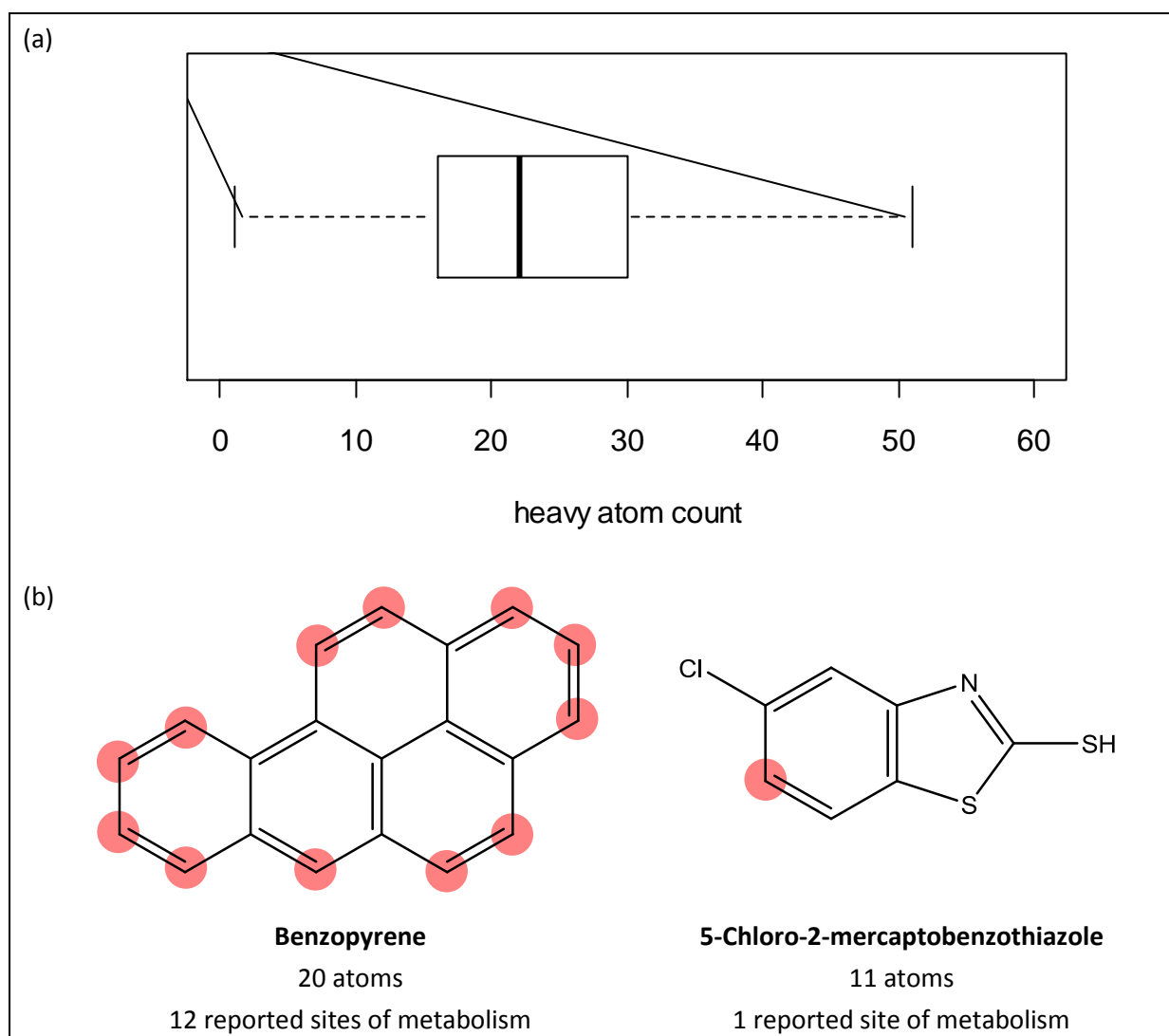


Figure 64: (a) Box plot showing the distribution of sizes (heavy atom count – i.e. the number of non-hydrogen atoms in the molecule) of the substrate molecule on version 2008.1 of the Symyx® Metabolite database. Outliers, as calculated using the `boxplot.stats` method from the statistical package R (108), with default parameters, are not shown. (b) Both the size of molecules and the number of sites at which they are metabolised (highlighted in red) can vary widely.

4.3.2 Area under the ROC curve-based performance measure

In order to overcome these biases, an alternative method of evaluating the performance of metabolic site prediction tools was proposed. This approach, based on the receiver operating characteristic (ROC) curve (277), is independent of both the size of molecules and their numbers of metabolic sites.

Receiver operating characteristic (ROC) curves

Receiver operating characteristic curves provide a technique for visualising a classifier's performance, depicting the trade-off between hit rates, and rates of false positives. ROC curves were first developed during the Second World War for the analysis of radar signals, where it was important to determine whether a signal was from an enemy plane, or due to noise, and have long been used in signal detection theory. In recent years ROC curves have been applied in a wide range of fields such as medical diagnostics and machine learning.

A ROC curve consists of a plot of the True Positive Rate versus the False Positive Rate as the threshold at which the classifier discriminates between positives and negatives is varied.

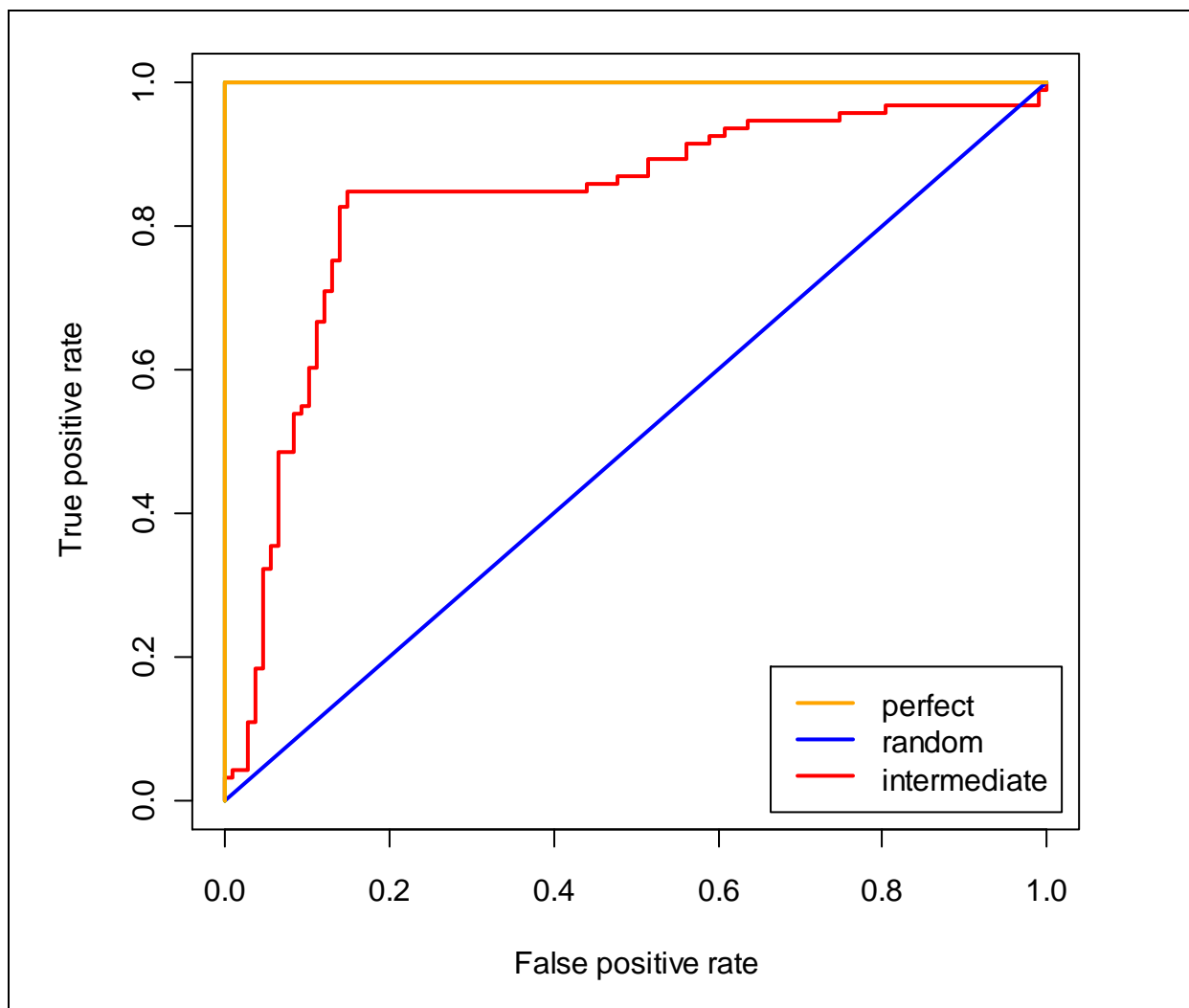


Figure 65: Example ROC curves for three different cases. The curve in orange is that for a classifier that performs perfectly, and that in blue for one that is completely random. In red is the more usual outcome – predictions that are better than random, but not perfect.

There are two major advantages of using ROC curves over common measures of classifier performance, such as accuracy, sensitivity, specificity, precision and recall. Firstly, through use of the relative scores produced by the classifier, ROC curves can measure the ability of the classifier to distinguish between positive and negative instances without having to be calibrated to produce good probability estimates, and secondly, they are insensitive to the relative number of positive and negative instances.

The area under the ROC curve (AUC) is a commonly used summary statistic, used to represent the 2-dimensional curve in a single number, and enable simple comparisons between classifiers. The AUC varies from 0.0 to 1.0, with 1.0 indicating that the classifier discriminates between positive and negative instances perfectly, 0.5 indicating that the performance is equivalent to randomly assigning classes, and a value of less than 0.5 indicating that the classifier is generating negative classifications (so multiplying by -1 gives a classification). The AUC has been shown to be equal to the probability that a randomly selected positive instance will be ranked higher than a randomly selected negative instance, which is the equivalent to the nonparametric Wilcoxon rank test (278,279).

Evaluating site of metabolism prediction using ROC curves

Prediction of the sites of metabolism in a molecule can be treated as a binary classification problem: each atomic position in a molecule either is or is not a site of metabolism. Each individual molecule from a test set presents an independent classification problem, and the ability of the prediction tool to discriminate between its sites of metabolism and atomic sites that are not metabolised can be assessed.

If an AUC value is determined for each molecule in the test set (giving a measure of the performance of the method under evaluation when identifying the sites of metabolism in that molecule) then the overall performance of the tool can then be evaluated by examining the distribution of AUC values generated using standard statistical techniques such as averages and variance (278).

4.3.3 Generation of test data

Selection of test data

In order to carry out an unbiased evaluation of a prediction tool it is important that the evaluation is carried out using data not used in the development of the method. This is often achieved through techniques such as cross-validation. Since access to a series of annual releases of the Symyx® Metabolite database was available while this work was being carried out, an obvious alternative to cross-validation presented itself: training MetaPrint2D using data from one release of the Metabolite database, and testing using the data added to subsequent releases. This has the added advantage of simulating a likely usage scenario, whereby MetaPrint2D is trained using all available data and then used by chemists to investigate the new compounds they are working with. This evaluation scheme also facilitates investigation of the robustness of MetaPrint2D to updates of the Metabolite database.

Many compounds within a single metabolic scheme exhibit only relatively minor variation. In order to ensure that the test data contained a diverse selection of compounds, and was not biased by clusters of very similar compounds, the selection of test compounds was restricted to new parent molecules. This means that new molecules present either in later generations of new metabolic schemes, or newly identified metabolites in previously known schemes are excluded from the evaluation.

The number of compounds in each test set is shown in Table 8.

Release	Novel parent compounds	Containing A/E
2006.1	601	498
2007.1	546	461
2008.1	509	408

Table 8: The number of novel parent compounds identified in each release of the Metabolite database, and the number of those which contain phase I additions and/or eliminations (labelled 'Containing A/E' in the table above).

Processing of test data

In order to identify all the sites of metabolism of a test molecule, all of the single step transformations in which it appears as the substrate were identified, and their reaction centres mapped onto a single copy of the molecule. Symmetrical points in the molecule were also identified, and reaction centres mapped between equivalent atom positions, producing a complete set of reaction centres for the compound. This process is illustrated for the metabolic sites of Flavanone in Figure 66.

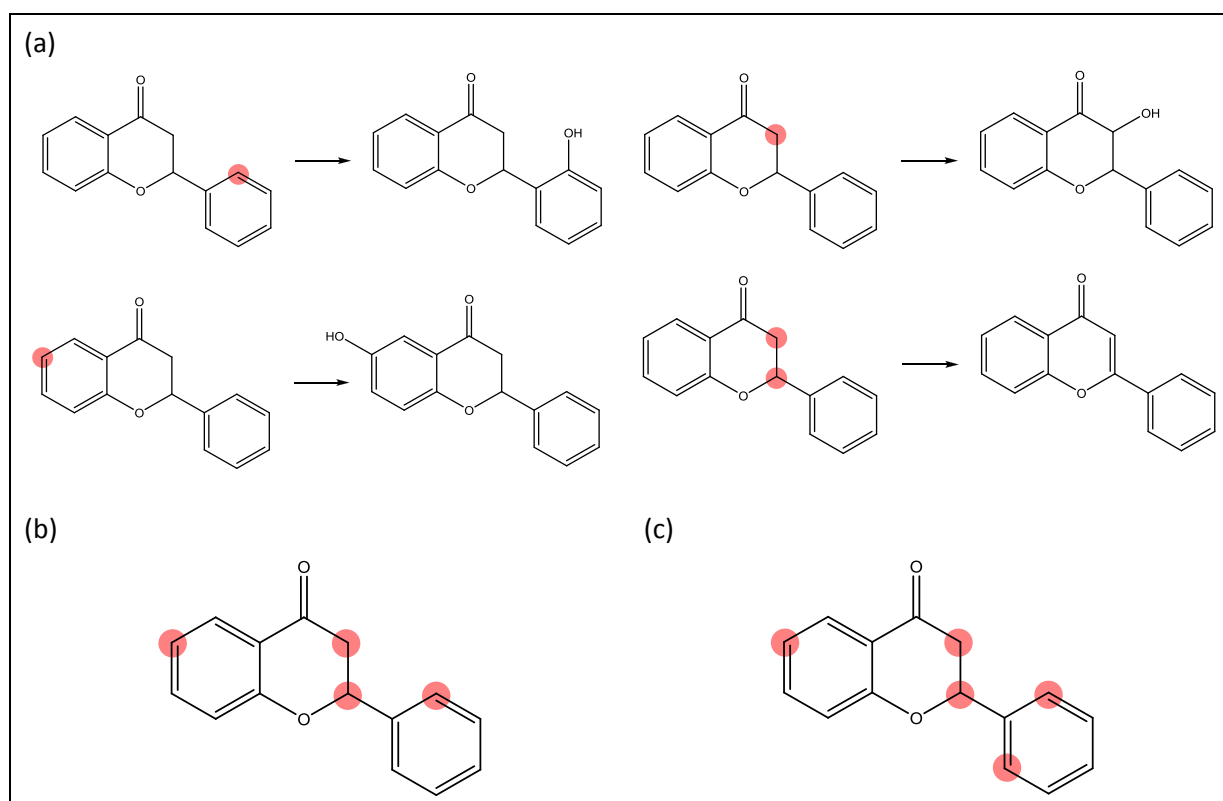


Figure 66: (a) Four of the metabolic transformations of Flavanone (280), with the reaction centres marked on the parent compound; (b) Flavanone with reaction centres from all four transformations merged; and (c) with symmetry mappings applied.

4.4 Evaluation of MetaPrint2D and the effects of data pre-processing options

In order to evaluate the performance of MetaPrint2D, and the effects of the various pre-processing options proposed, a number of MetaPrint2D models have been generated and their performance evaluated. Models have been generated from the data contained in each of the 2005.1, 2006.1 and 2007.1 releases of the Symyx® Metabolite database. The

predictions by each model have been evaluated against the transformations of novel parent compounds from subsequent releases of the Metabolite database. This process has been repeated for each of the pre-processing options discussed above.

Details of the models generated and the evaluation metrics calculated are given in Table 9 and Table 10, and the results of this evaluation are presented below.

Model	Description
all	Model constructed using all transformations
nomulti	Multi-step transformation records were excluded
nodup	Duplicate transformations were excluded
sym	Symmetry mappings were applied
merge	All transformations for each compound were merged
hasrc	Molecules with no phase I reaction centres were excluded

Table 9: The MetaPrint2D models constructed.

Name	Description
Top 1	% of molecules for which highest rank atom is a site of metabolism
Top 3	% of molecules for which at least one of the three highest ranked atoms is a site of metabolism
Mean AUC	Mean area under the ROC curve
Median AUC	Median area under the ROC curve

Table 10: The evaluation metrics calculated.

4.4.1 Results

Results are presented to three significant figures in order to illustrate the small degree of variation. The significance, or otherwise, of this variation is discussed on page 135.

Training data: 2005.1

Model	Top 1			Top 3			Mean AUC			Median AUC		
	2006	2007	2008	2006	2007	2008	2006	2007	2008	2006	2007	2008
all	60.2%	59.7%	58.8%	76.1%	77.0%	75.5%	0.766	0.780	0.792	0.903	0.895	0.882
nomulti	61.0%	60.3%	60.3%	76.5%	77.9%	75.7%	0.770	0.785	0.794	0.920	0.889	0.892
nodup	60.6%	59.7%	59.8%	76.5%	77.7%	75.7%	0.767	0.781	0.792	0.917	0.903	0.880
sym	60.0%	58.4%	59.3%	75.5%	75.5%	76.0%	0.764	0.779	0.791	0.899	0.885	0.887
merge	57.0%	57.5%	59.3%	75.3%	77.9%	76.5%	0.764	0.784	0.791	0.891	0.901	0.893
hasrc	59.4%	61.2%	59.8%	76.3%	77.2%	74.3%	0.767	0.781	0.787	0.918	0.895	0.875

Table 11: These models were trained using data from the 2005.1 release of the Metabolite database, and evaluated using novel parent compounds from the 2006.1, 2007.1 and 2008.1 releases.

Training data: 2006.1

Model	Top 1		Top 3		Mean AUC		Median AUC	
	2007	2008	2007	2008	2007	2008	2007	2008
all	59.0%	59.6%	78.1%	77.0%	0.781	0.799	0.896	0.889
nomulti	59.9%	59.6%	77.4%	76.0%	0.785	0.802	0.889	0.891
nodup	59.2%	60.0%	78.3%	77.0%	0.781	0.800	0.901	0.897
sym	58.1%	59.3%	75.5%	76.5%	0.779	0.799	0.885	0.893
merge	58.8%	59.8%	77.7%	76.0%	0.784	0.799	0.904	0.897
hasrc	60.7%	59.8%	77.7%	76.0%	0.781	0.794	0.900	0.881

Table 12: These models were trained using data from the 2006.1 release of the Metabolite database, and evaluated using novel parent compounds from the 2007.1 and 2008.1 releases.

Training data: 2007.1

Model	Top 1	Top 3	Mean AUC	Median AUC
	2008	2008	2008	2008
all	59.6%	77.2%	0.804	0.900
nomulti	59.3%	76.5%	0.805	0.902
nodup	60.3%	77.2%	0.803	0.900
sym	59.6%	76.7%	0.803	0.900
merge	60.0%	75.7%	0.803	0.913
hasrc	60.5%	76.2%	0.799	0.892

Table 13: These models were trained using data from the 2007.1 release of the Metabolite database, and evaluated using novel parent compounds from the 2008.1 release.

4.5 Analysis of MetaPrint2D's performance

The performance of MetaPrint2D's predictions changes very little with the pre-processing options discussed. Wilcoxon signed rank tests (281,282) have been carried out to determine whether there is any significant variation between the distributions of AUC scores generated from models constructed using all available data, and those constructed with each of the pre-processing options. Wilcoxon's signed rank test is used in place of the paired Student's t-test since the distribution of AUC scores is not normally distributed, as can clearly be seen in Figure 67 on page 137. The results of these tests are presented in Table 14 below. The only pre-processing option to consistently improve the performance of the model (p-values much lower than 0.05) is the exclusion of multi-step transformations from the training data, and this only produces a very small improvement in the AUC (~0.01).

There is also little variation between predictions using models generated from different releases of the Metabolite database; the quality of MetaPrint2D's predictions on test data from the 2008.1 Metabolite database shows little variation between models trained using the 2005.1, 2006.1 or 2007.1 releases of the database.

training	testing	nomulti	nodup	sym	merge	hasrc
2005	2006	2.25E-08 $\Delta = 0.011$	0.00274 $\Delta = 0.0083$	0.187	0.204	0.470
	2007	1.37E-07 $\Delta = 0.011$	0.741	0.0620	0.236	0.553
	2008	0.00463 $\Delta = 0.0079$	0.371	0.192	0.827	0.0069 $\Delta = -0.0081$
2006	2007	1.69E-06 $\Delta = 0.0099$	0.392	0.00138 $\Delta = 0.0068$	0.282	0.545
	2008	0.00030 $\Delta = 0.0093$	0.0872	0.352	0.967	0.0910
2007	2008	0.00325 $\Delta = 0.0079$	0.446	0.0178 $\Delta = 0.0064$	0.831	0.00579 $\Delta = 0.0080$

Table 14: p-values for Wilcoxon signed rank tests comparing the distributions of AUC scores generated from models constructed using all data, and models constructed using each of the pre-processing options. Variations between distributions that are significant at the 95% confidence level (p-value < 0.05) are highlighted in bold, and the shift in the distribution's median (Δ) given.

4.5.1 Distribution of MetaPrint2D's performance scores

The distribution of area under the ROC curve (AUC) scores for the novel compounds from the 2006.1 database tested on a model constructed using the 2005.1 database, are shown in Figure 67 below. This is representative of the distributions of the other combinations of data/models. Examination of the distribution of AUC scores shows that in the majority of cases predictions are very accurate, with AUC scores in the range 0.95-1.00, however there is a long tail to the distribution, with MetaPrint2D performing much worse than random (AUC < 0.5) in a small number of instances.

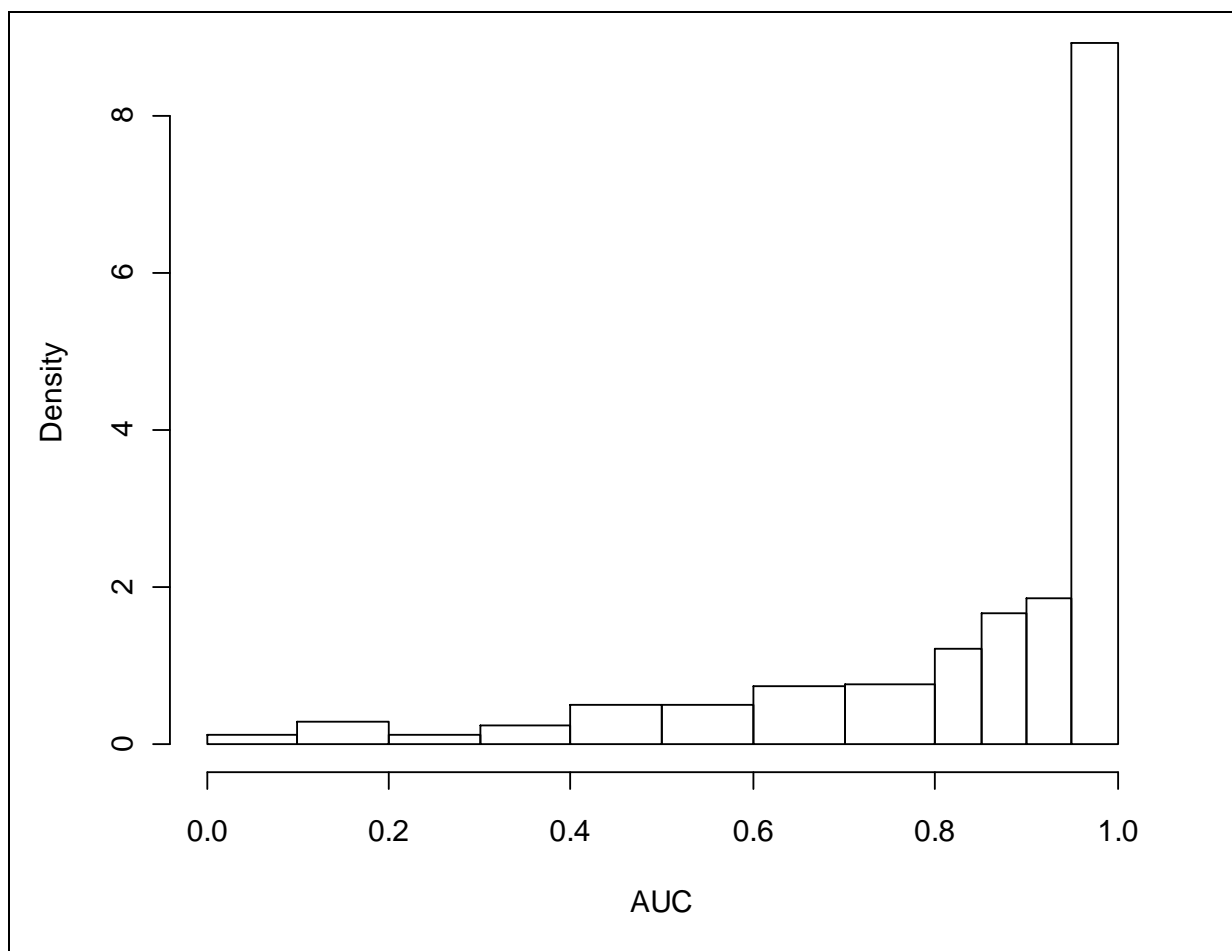
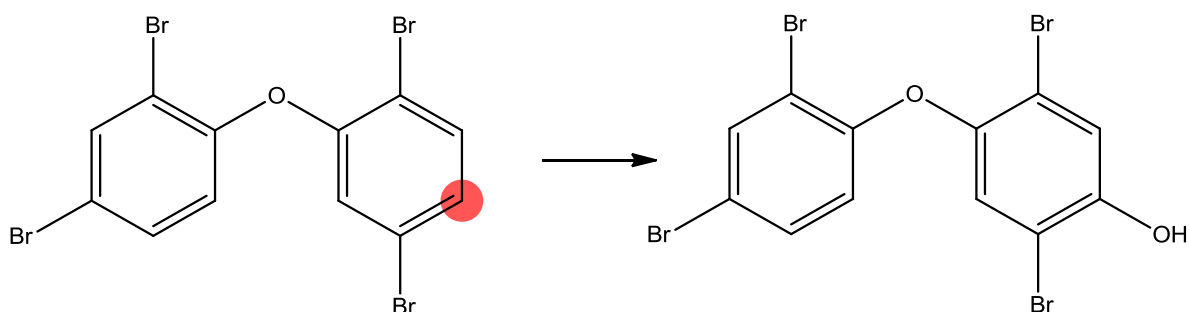


Figure 67: Histogram showing the distribution of AUC values generated using a model build on all compounds from Metabolite database 2005.1, and test compounds from 2006.1. The distribution is highly skewed towards higher values, indicating very good predictions were made for the majority of test compounds.

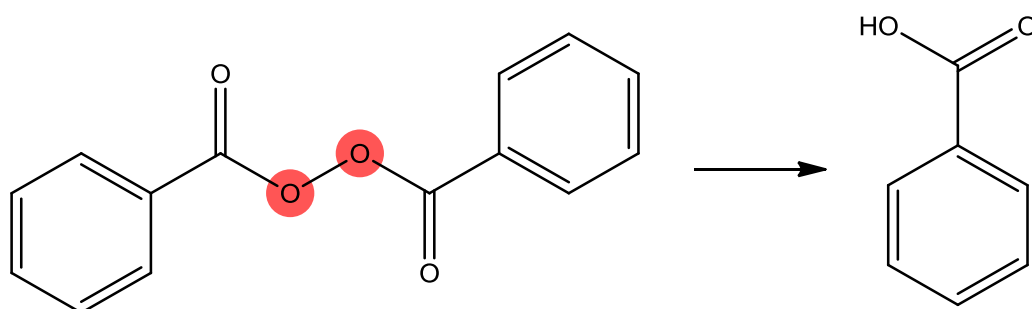
The molecules with the very lowest AUC scores have been identified and a selection of these is shown below. On examination of the molecules' transformations, a trend is apparent: the sites of metabolism occur at atoms occupying novel environments (atom environments not represented in the training data), so are assigned a normalized occurrence ratio of zero, but there are other environments within the molecule that have been found to occur at reaction centres in the training data, so receive a non-zero normalized occurrence ratio.

MDLNUMBER: MMTB00051019; AUC = 0.0625



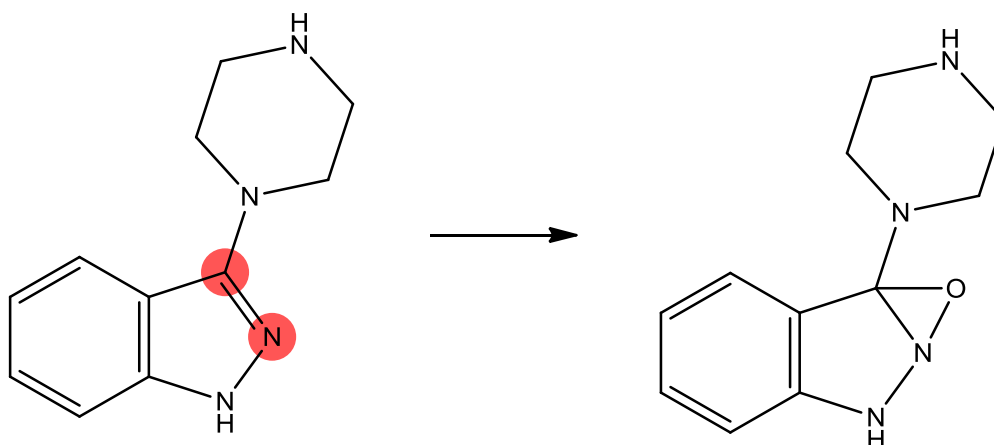
Searching the training data for the atom environment at the reaction centre in this transformation did not produce any hits, indicating that the environment is completely novel to the model

MDLNUMBER: MMTB00052513; AUC = 0.125



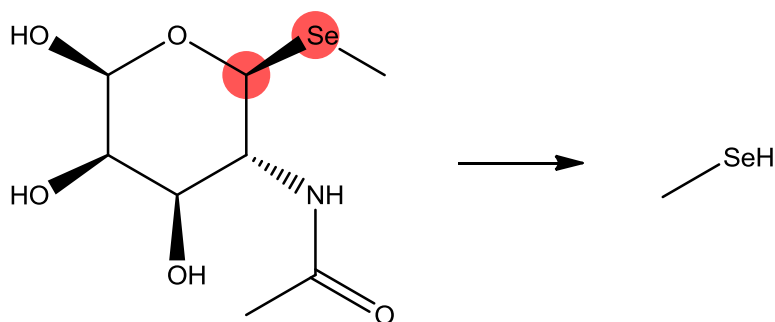
Searching the training data for the reaction centre fragment did not produce any hits; the phenyl ring, however, is found in many records in the Symyx(R) Metabolite database, and is observed to undergo a variety of transformations.

MDLNUMBER: MMTB00050992; AUC = 0.1538



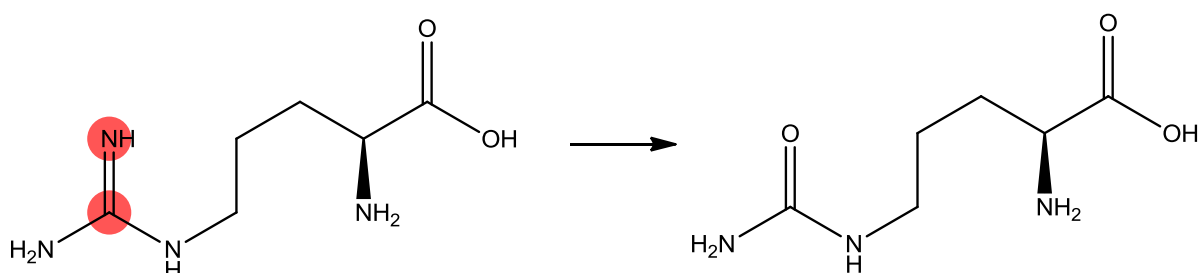
Searching the training data for the reaction centre fragment produced only three hits, none of which underwent a metabolic transformation centred anywhere within this region of the compound.

MDLNUMBER: MMTB00044960; AUC = 0.167



Searching the training data for any molecules containing selenium produced 51 hits, none of which occupied a remotely similar atom environment.

MDLNUMBER: MMTB00050971; AUTOC = 0.167



Searching the training data for other molecules containing the reaction centre fragment revealed only a single hit, which undergoes a completely different transformation, centred on an atom in a different atom environment:

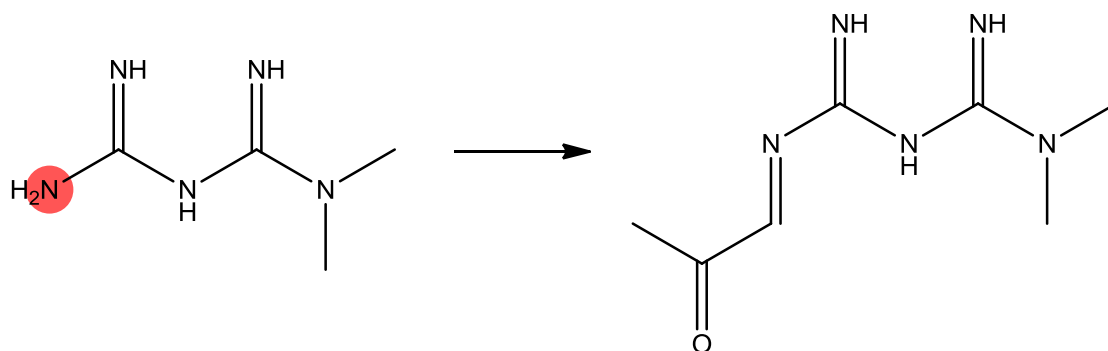


Figure 68: Examples of molecules for which MetaPrint2D's predictions had the lowest AUC scores.

4.5.2 Novel atom environments

Given that the molecules for which MetaPrint2D generated the worst predictions all had some atoms occupying novel atom environments the relationship between the proportion of novel atom environments in a molecule, and the reliability of MetaPrint2D's site of metabolism predictions has been investigated.

In order to generate as detailed an analysis as possible, all 1367 novel parent compounds from the 2006.1, 2007.1 and 2008.1 Symyx® Metabolite databases have been collated, and

their sites of metabolism predicted using a MetaPrint2D model trained on the 2005.1 release of the Metabolite database. For each molecule the proportion of the atoms that occupy novel atom environments has been recorded. The area under the ROC curve (AUC) statistics calculated for the test molecules have been binned according to the proportion of novel atom environments in the molecule. The results of this are shown in Figure 69, below.

Analysis of Figure 69, below, shows that the sites of metabolism of molecules in which no atoms occupy novel atom environments are predicted well. The quality of predictions clearly decreases as the proportion of atoms with a novel environment increases. When all atoms in a molecule are in novel atom environments performance of the classifier is essentially random (AUC=0.5). This is to be expected, since atoms in a novel environment will all be assigned the same normalised occurrence ratio (nOR=0.0).

The greatest uncertainty in the quality of prediction is found when between a third and a half of the molecule's atoms are occupying novel atom environments. This is again due to the novel environments being assigned a normalised occurrence ratio of 0.0. In instances where the novel environment is not found at a reaction centre accurate predictions for that molecule are still possible. When the novel environment is found at a reaction centre, however, the low score assigned means that the accuracy of prediction for that molecule will be very low, since the majority of the other atoms in the molecule will have a higher (than zero) normalised occurrence ratio, even when the likelihood of their being a site of metabolism is very low.

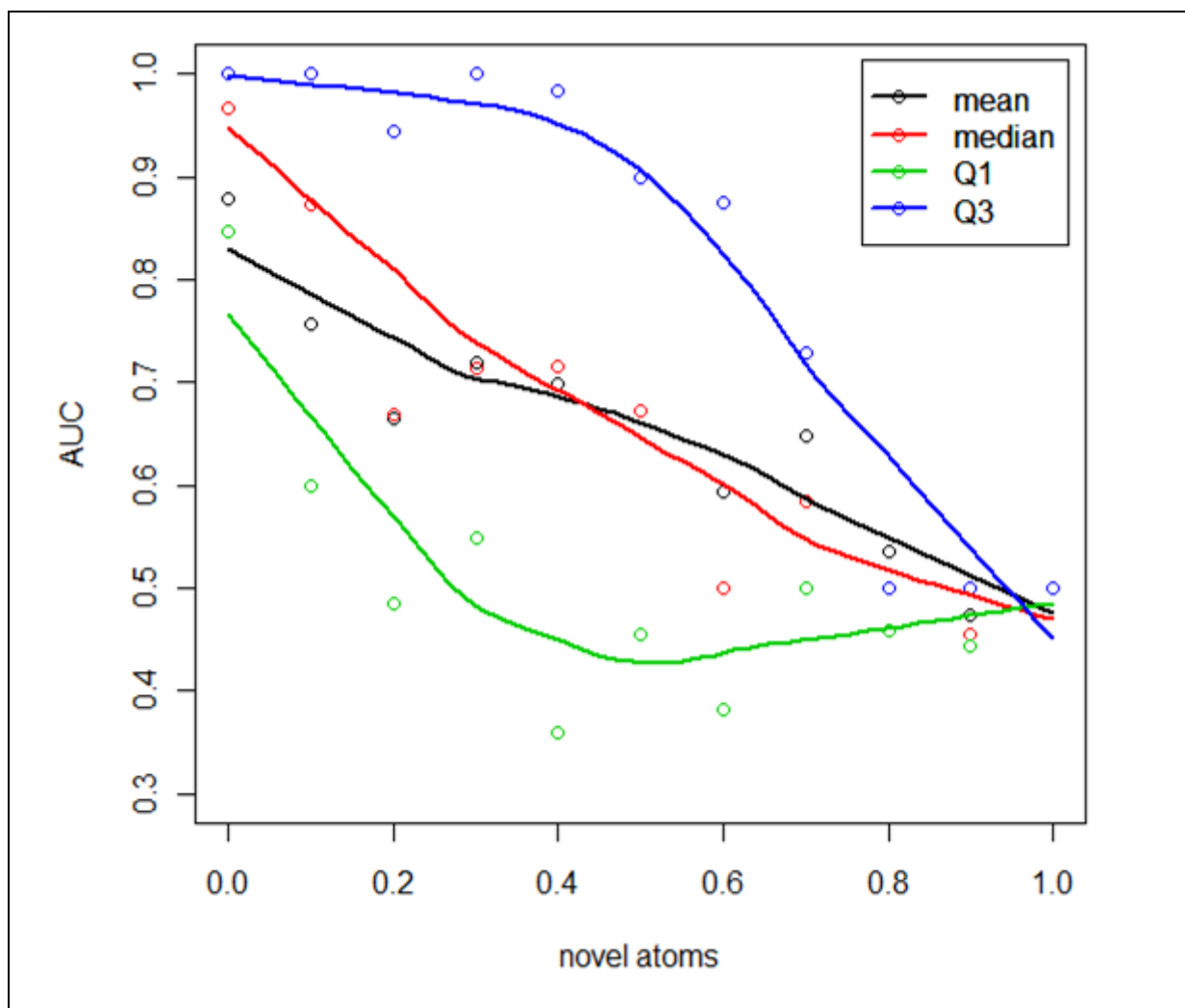


Figure 69: Graph showing the variation in performance of MetaPrint2D with the proportion of atoms in query structures occupying novel environments. Predictions on 1367 novel parent molecules from the 2006.1, 2007.1 and 2008.1 releases of the Metabolite databases were made using a model trained on data from the 2005.1 release. AUC statistics and the proportion of novel atoms in the query compound were calculated for each molecule (novel atoms = 0.0 means that no atoms in the compound are in novel atom environments, while novel atoms = 1.0 indicated that all atoms in the compound occupy novel atom environments). The data was binned according to the novel atom proportion, split at boundaries 0.05, 0.15, 0.25...0.85, 0.95. For each bin the mean, median, first and third quartile (Q1 and Q3 in the figure above) values are plotted, and locally weighted scatter plot smoothing (LOESS) lines have been fitted using the `loess.smooth` function from the statistical package R (108), with default parameters.

Ideally, atom environments that have been observed in the training data and found to occur only very rarely, or never, at reaction centres should receive lower scores than atom

environments for which no information is available. To facilitate this MetaPrint2D was been modified so that novel atom environments are assigned the mean normalised occurrence ratio of all the non-novel atoms from a sample set of test data (nOR=0.159), and the above analysis repeated. The result of this change is shown in Figure 70, below.

As a result of this modification the position of the 75-percentile line (Q3) barely changed, however the mean, median and 25-percentile (Q1) AUC all increased, indicating that this modification has resulted in fewer badly predicted molecules.

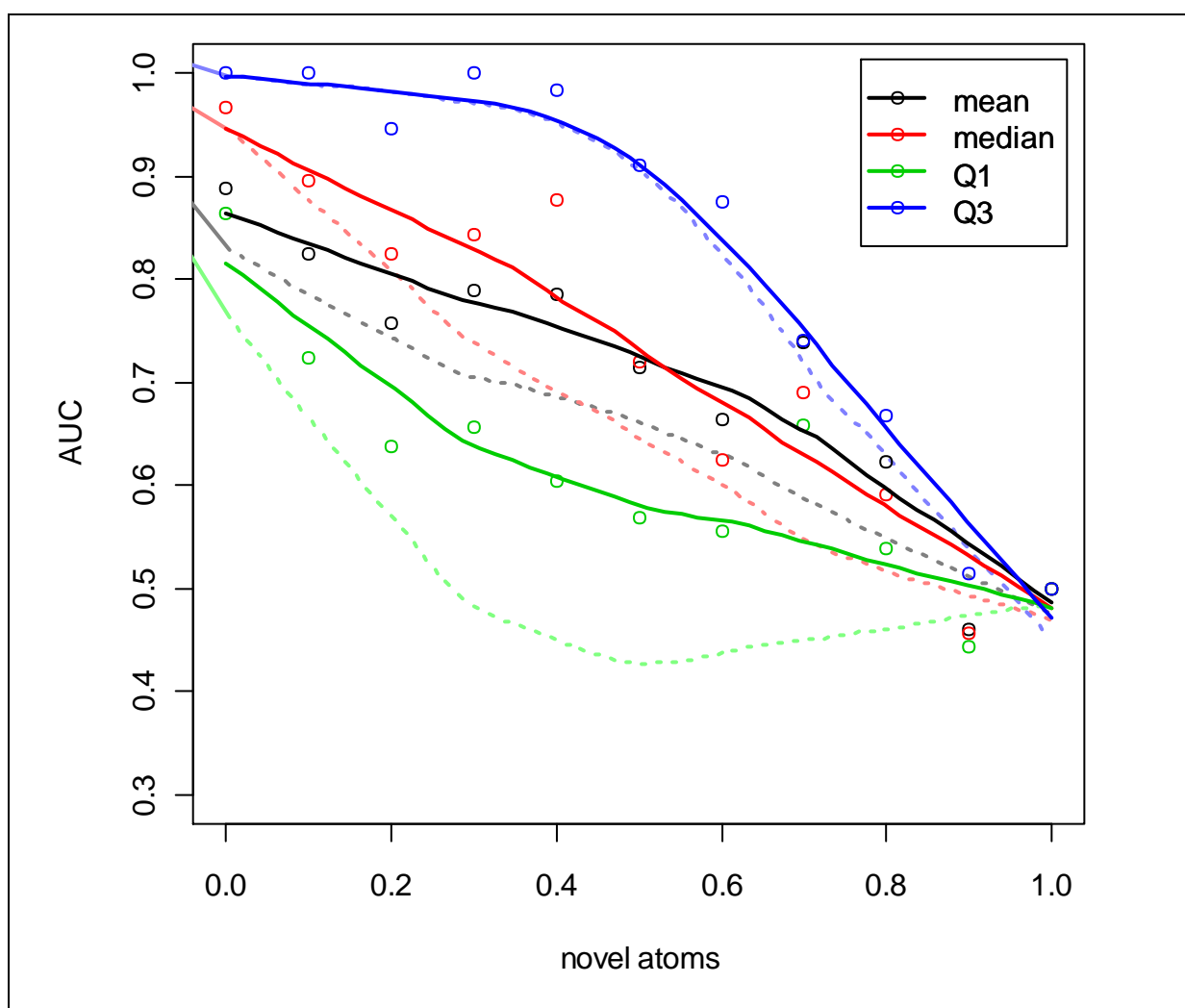


Figure 70: Graph showing the variation in performance of MetaPrint2D with the proportion of atoms in query structures occupying novel environments. Dotted lines taken from Figure 69, above, generated when novel atom environments are assigned a normalised occurrence ratio of 0.0; solid lines have been generated in the same manner, but with novel atom environments assigned the mean normalised occurrence ratio of the data set: 0.159.

The distribution showing how the performance of MetaPrint2D varies with the degree to which a molecule fits the descriptor space described by MetaPrint2D's model can be used to estimate the reliability of a prediction generated by MetaPrint2D. Figure 71, below, shows the distribution of novel atoms in the molecules from the test data. Almost all of the atom environments in the majority of compounds are well characterised by the model.

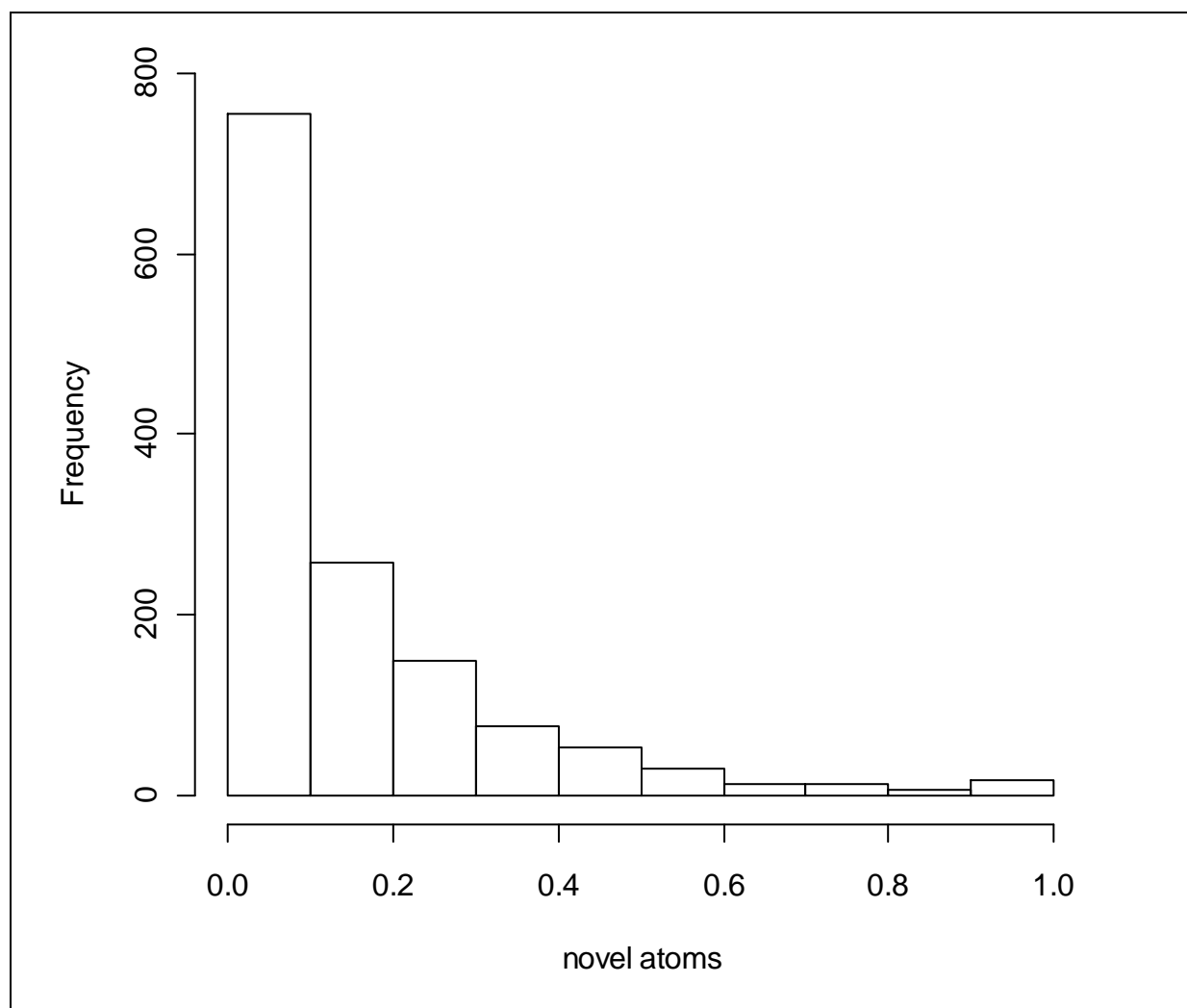


Figure 71: The distribution of proportions of novel atoms among test molecules.

4.6 Speed of predictions

The speed of MetaPrint2D has been assessed, through predictions of the sites of metabolism of 232 common drug molecules (mean heavy atom count: 20.9 atoms). The total time taken to perform the calculations (including loading the model and assigning SYBYL® atom types) averaged 6.81 seconds over five runs, which is equivalent to less than 30ms per molecule.

Run times					Mean time (seconds)	Mean time per molecule (milliseconds)
t ₁ /s	t ₂ /s	t ₃ /s	t ₄ /s	t ₅ /s		
6.81	6.86	6.80	6.81	6.79	6.81	29.37

Figure 72: Time taken for MetaPrint2D to generate site of metabolism predictions for 232 common drug molecules (recorded on a Dell Inspiron 6400 laptop Intel Core 2 T5300 @ 1.73GHz; 3.24GB RAM). Timings recorded on five independent runs, and averaged.

4.7 Parameterization of MetaPrint2D

SPORCalc provided the following pre-configured parameterizations, together with the option to set a custom parameterization:

Setting	Similarity threshold	Exact levels	Level weightings					
			1	2	3	4	5	6
Loose	1.0	2	-	-	1.0	0.75	0.50	0.25
Default	0.5	3	-	-	-	0.75	0.50	0.25
Strict	0.1	4	-	-	-	-	0.5	0.25

Figure 73: The pre-configured parameterization of SPORCalc.

The initial evaluations of MetaPrint2D were performed using this parameterization, and the effects of varying the parameterisation have subsequently been explored. Early investigations found that there were very few atom environments being discarded due to their exceeding the similarity threshold; the major source of variation between parameterizations lies in the number of number of levels to which exact fingerprint matches are required.

Figure 74 shows how the distribution of AUC scores changes as the number of fingerprint levels to exactly match is varied.

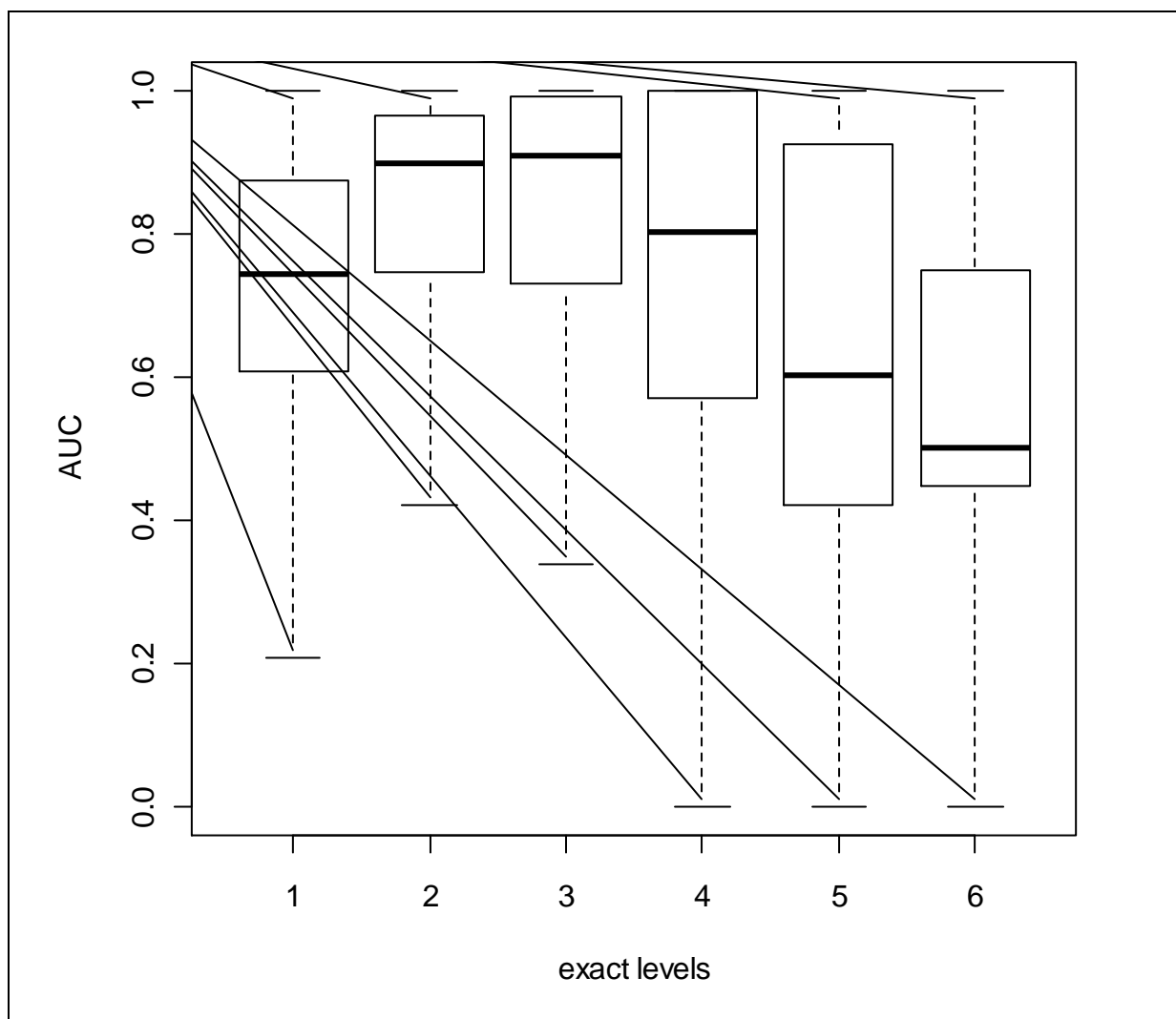


Figure 74: Box plots showing the changes in distribution of AUC scores for MetaPrint2D predictions as the number of fingerprint levels to match exactly is varied. Outliers, as calculated using the `boxplot.stats` method from the statistical package R (108), with default parameters, are not shown.

In general the best results (highest AUC scores) are found when the first two or three fingerprint levels are exactly matched. Exact matching to three levels produces slightly more very well predicted molecules than exact matching to two levels (the 4th quartile is higher), but also has more badly predicted molecules (the 1st quartile extends lower).

The kernel density plots in Figure 75 show the AUC distributions in more detail.

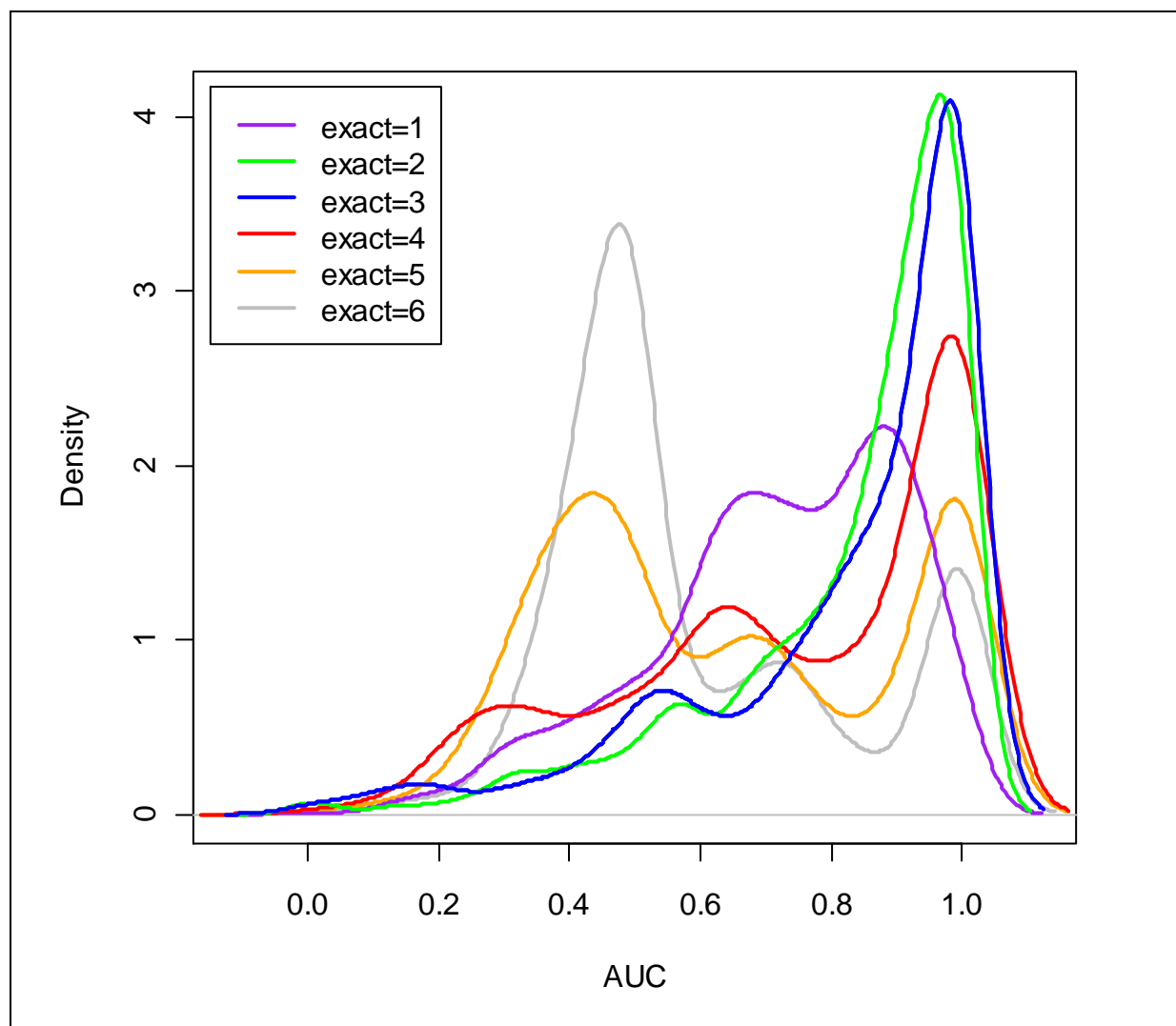


Figure 75: Kernel density plots showing the changes in distribution of AUC scores for MetaPrint2D predictions as the number of fingerprint levels to match exactly is varied.

When a high number of levels of exact matching are required ($n=5$ and $n=6$) the distribution is predominantly bimodal, with peaks around $AUC=0.5$ and 1.0 . The peak at $AUC=1.0$ is due to molecules well represented by the model; due to the specific nature of searches carried out using these settings, very good results are generated for such molecules. However, for many of the atom environments in the test data such a search requiring exact matches to so many fingerprint levels returns little or no data, and the classifier is unable to discriminate between such atoms. This leads to an AUC of around 0.5 . There is also a third, smaller, peak in the $AUC=0.6-0.7$ region of the distribution. This is due to atoms occupying environments that sometimes occur at sites of metabolism; typically this arises from regions of a molecule which remain unaltered during some steps of a metabolic scheme, while containing a centre of metabolism in other steps.

As the number of levels that must be matched exactly is reduced the number of molecules for which MetaPrint2D makes good predictions increases, and reaches a maximum when two or three levels of exact match are required.

4.8 Isoform specific models

While there are a large number of cytochrome P450 isoforms, only a relatively small number account for the majority of known CYP450 catalysed xenobiotic transformations. A small proportion of the transformations in the Symyx® Metabolite database have annotations indicating which, if any, cytochrome P450 isoforms catalyse the transformation. The 2008.1 release of the database contains 29602 metabolic transformations observed to occur in humans. Of these only 3839 (13.0%) are annotated with a specific cytochrome P450 isoform. The reported CYP450 substrates are shared between over 150 isoforms and their variants. However, for the majority of CYP450 isoforms only a few substrates are reported. In the 2008.1 Metabolite database only eleven isoforms have one hundred or more reported substrates. These are listed in Table 15, below.

Isoform	Substrate count
CYP3A4	1019
CYP1A2	607
CYP2D6	559
CYP2C9	470
CYP2E1	412
CYP2C19	401
CYP2B6	314
CYP1A1	295
CYP2A6	287
CYP2C8	265
CYP3A5	201

Table 15: The eleven CYP450 isoforms with more than one hundred substrates reported in the 2008.1 release of the Symyx® Metabolite database.

Separate MetaPrint2D models have been constructed for each of these isoforms, and their performance assessed. Due to the much smaller numbers of substrate molecules available than for the models described earlier, a different assessment strategy was adopted.

Cytochrome P450 3A4, which has the largest number of reported substrates, has around one hundred new substrates reported with each database release; however, the numbers of the other cytochromes P450 are smaller. This means that the previous approach of constructing models using all the data available in a particular release of the Symyx® Metabolite database and assessing the performance of that model using data added in a subsequent release of the database would leave very small quantities of data available for testing.

Instead, the models have been generated and evaluated using Monte Carlo cross-validation (283). For each CYP450 isoform 20 modelling runs were performed. In each run 80% of the isoform's substrates were randomly selected to be used to construct the model, the performance of which was tested using the remaining 20% of the substrates. The same statistics as used previously – percentage correct in top one and top three hits, and the mean and median areas under the ROC curve – were generated for each set of test data, and the values averaged over the 20 runs.

Cytochrome P450 Isoform	Number of Substrates	% Top 1	% Top 3	Mean AUC	Median AUC	% Novel
CYP3A4	1019	56.0 (3.3)	71.5 (3.2)	0.816 (0.016)	0.904 (0.020)	16.4
CYP1A2	607	56.8 (4.8)	75.1 (3.5)	0.795 (0.024)	0.859 (0.033)	22.9
CYP2D6	559	64.5 (4.0)	78.8 (2.9)	0.836 (0.018)	0.938 (0.019)	22.8
CYP2C9	470	59.4 (4.7)	74.2 (4.5)	0.802 (0.026)	0.899 (0.038)	23.5
CYP2E1	412	57.5 (5.3)	78.0 (3.4)	0.772 (0.028)	0.840 (0.045)	27.5
CYP2C19	401	59.1 (4.5)	74.8 (4.4)	0.813 (0.028)	0.906 (0.043)	24.5
CYP2B6	314	59.5 (7.1)	73.9 (6.6)	0.790 (0.034)	0.869 (0.066)	27.5
CYP1A1	295	49.2 (6.3)	66.9 (5.9)	0.764 (0.022)	0.805 (0.041)	26.3
CYP2A6	287	54.7 (6.9)	70.4 (5.3)	0.758 (0.028)	0.820 (0.050)	30.8
CYP2C8	265	52.7 (6.4)	68.8 (5.6)	0.772 (0.030)	0.854 (0.041)	29.2
CYP3A5	201	49.1 (8.2)	67.1 (7.3)	0.778 (0.036)	0.815 (0.080)	25.9

Table 16: Performance of cytochrome P450 Isoform specific models. The results are the mean of 20 Monte Carlo cross-validation runs, with standard deviations of each value given in parentheses.

The performance of these models (presented in Table 16) shows little variation from that of the global models reported earlier. The small standard deviations of the performance

scores, indicates that the results showed little variation between cross-validation runs. As a result of the considerably smaller data sets a much greater number of atom environments have little or no data in the training set. In the earlier models only 3.5% of atoms in test compounds occupied novel atom environments, but for the isoform specific models 16-30% of the atoms occupy environments with little or no data.

In order to explore the specificity of the isoform specific models the models generated for each CYP450 isoform have been used to predict the sites of metabolism of the substrates of each of the other isoforms. Table 17 contains the mean area under the ROC curve results for each of these experiments. The isoforms are listed in order of decreasing data quantity. Initial inspection appears to show a trend in the performance of each test set decreasing with the model size. This is confirmed by statistical testing; Pearson's correlation tests on the performance of each test set against the natural logarithm of the number of substrates used to generate the models give correlation coefficients in the range 0.750 – 0.898, indicating a positive correlation between the two, at the 95% confidence level, for all isoforms. This correlation is clearly visible in Figure 76, below.

		Isoform of Test Data										
		CYP3A4	CYP1A2	CYP2D6	CYP2C9	CYP2E1	CYP2C19	CYP2B6	CYP1A1	CYP2A6	CYP2C8	CYP3A5
Isoform of Training Data	CYP3A4	0.816	0.882	0.900	0.906	0.859	0.910	0.925	0.874	0.914	0.918	0.962
	CYP1A2	0.783	0.795	0.871	0.867	0.847	0.870	0.898	0.899	0.897	0.898	0.810
	CYP2D6	0.767	0.820	0.836	0.866	0.821	0.867	0.860	0.808	0.854	0.852	0.789
	CYP2C9	0.779	0.807	0.872	0.802	0.822	0.887	0.846	0.791	0.843	0.894	0.793
	CYP2E1	0.734	0.807	0.829	0.813	0.772	0.810	0.865	0.822	0.882	0.841	0.742
	CYP2C19	0.753	0.792	0.854	0.846	0.790	0.813	0.823	0.760	0.831	0.871	0.764
	CYP2B6	0.741	0.800	0.810	0.799	0.813	0.820	0.790	0.806	0.853	0.845	0.785
	CYP1A1	0.704	0.783	0.780	0.755	0.761	0.750	0.800	0.764	0.791	0.789	0.778
	CYP2A6	0.737	0.804	0.814	0.808	0.832	0.820	0.850	0.821	0.758	0.850	0.762
	CYP2C8	0.726	0.778	0.800	0.811	0.771	0.831	0.820	0.778	0.815	0.772	0.761
CYP3A5	0.695	0.684	0.701	0.680	0.670	0.694	0.718	0.731	0.704	0.733	0.778	

Table 17: The mean AUC performance of models trained on substrates metabolised by one CYP450 isoform on predicting sites of metabolism of substrates of other isoforms. The diagonal cells, highlighted, show the cross-validated performance of each model, from Table 16.

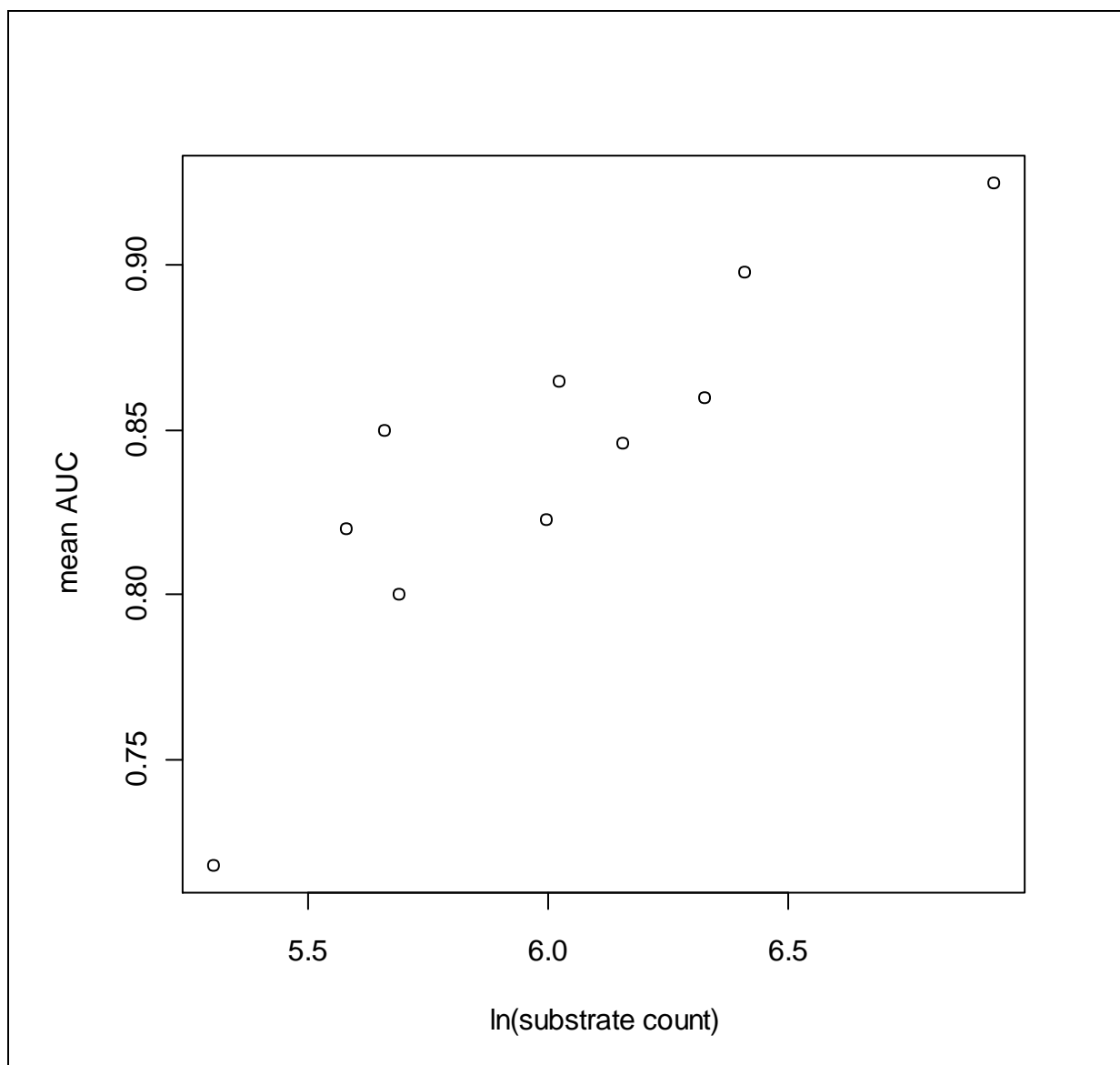


Figure 76: Performance predicting sites of metabolism of CYP2B6 substrates by models trained on the substrates of other isoforms plotted against the natural logarithm of the numbers of substrates used to build the model. The data has a correlation coefficient of 0.817. Analysis of substrates of other CYP450 isoforms shows a similar correlation.

These results show that MetaPrint2D models trained using the substrates of one CYP450 isoform can predict the sites of metabolism of substrates of other isoforms as well as, or in some cases better than, other substrates of the isoform on which the model was developed. This shows that there is a fairly small variation in the specificity of the different CYP450 isoforms, and the effects of this variation are small enough to be masked by the uncertainty in the model. Given how small a quantity of data is available for each CYP450 isoform, and that the various isoforms catalyse the same types of reactions, differing mainly in the size

and shape of their binding pocket and hence which compounds they are able to metabolise, it is not too surprising that the inclusion of additional data improves the discrimination between potential sites of metabolism, even if it is from molecules metabolised by a different isoform.

4.9 Comparison with other tools

Table 18 shows the reported performances of several site of metabolism predictions methods.

Isoform	SPORCalc	MetaSite	MetaGlide	QMBO	Scientist
CYP2C9	55% / 81% ^a	70% / 91% ^a	33% / 67% ^a	58% / 84% ^a	49% / 81% ^a
	49% / 87% ^b	53% / 60% ^b	42% / 79% ^b	53% / 82% ^b	70% / 88% ^b
		83%, 84% ^c			
CYP3A4		61% / 87% ^a			
	55% / 81% ^a	49% / 74% ^b	39% / 65% ^a	58% / 84% ^a	49% / 81% ^a
	49% / 87% ^b	90%, 86% ^c	45% / 71% ^b	51% / 87% ^b	70% / 88% ^b
		41% / 72% ^d			
		21% / 40% ^e			
CYP2D6		62%, 85% ^c			

Table 18: Performance of selected site of metabolism prediction methods reported in other studies. SPORCalc is the in-house version running at AstraZeneca. MetaSite is a commercial offering. MetaGlide makes predictions based on docking using Glide. QMBO is a quantum mechanical method based on hydrogen abstraction energy. Finally the predictions of a biotransformation scientist were included in one study. Sources of data: (a) Afzelius *et al.* (218) public data set, % top 1 and top 3 hits contain site of metabolism; (b) Afzelius *et al.* (218) in-house data set, predictions are centred on functional groups rather than atoms, % top 1 and top 3 hits contain site of metabolism; (c) Cruciani *et al.* (276) % correct prediction in top 2 hits from two in-house data sets reported; (d) Zhou *et al.* (234) % top 1 and top 3 hits with reactivity on; (e) Zhou *et al.* (234) % top 1 and top 3 hits with reactivity off.

The results of MetaPrint2D's evaluation, reported above, are similar to these results. Direct comparison is difficult since it is apparent that the reported performance of a method varies considerably with the dataset used for the evaluation.

4.10 Accuracy of the test data

The reliability of the analysis of site of metabolism predictions described here depends on the assumption that all of the sites of metabolism for a molecule are reported. This is also an issue with regards to the quality of the models constructed. Analysis of the relationship between the proportion of sites in a molecule at which metabolism is found to occur, and the year in which metabolic studies on the compound were first reported (Figure 77) shows a weak but clear trend towards a smaller proportion of more recently reported molecules being found to be sites of metabolism. This suggests that the metabolic profiles of more recently studied compounds may not yet be fully characterised, though it could be the case that there has been a tendency for those substrates undergoing a greater variety of metabolic transformations to have been identified earlier.

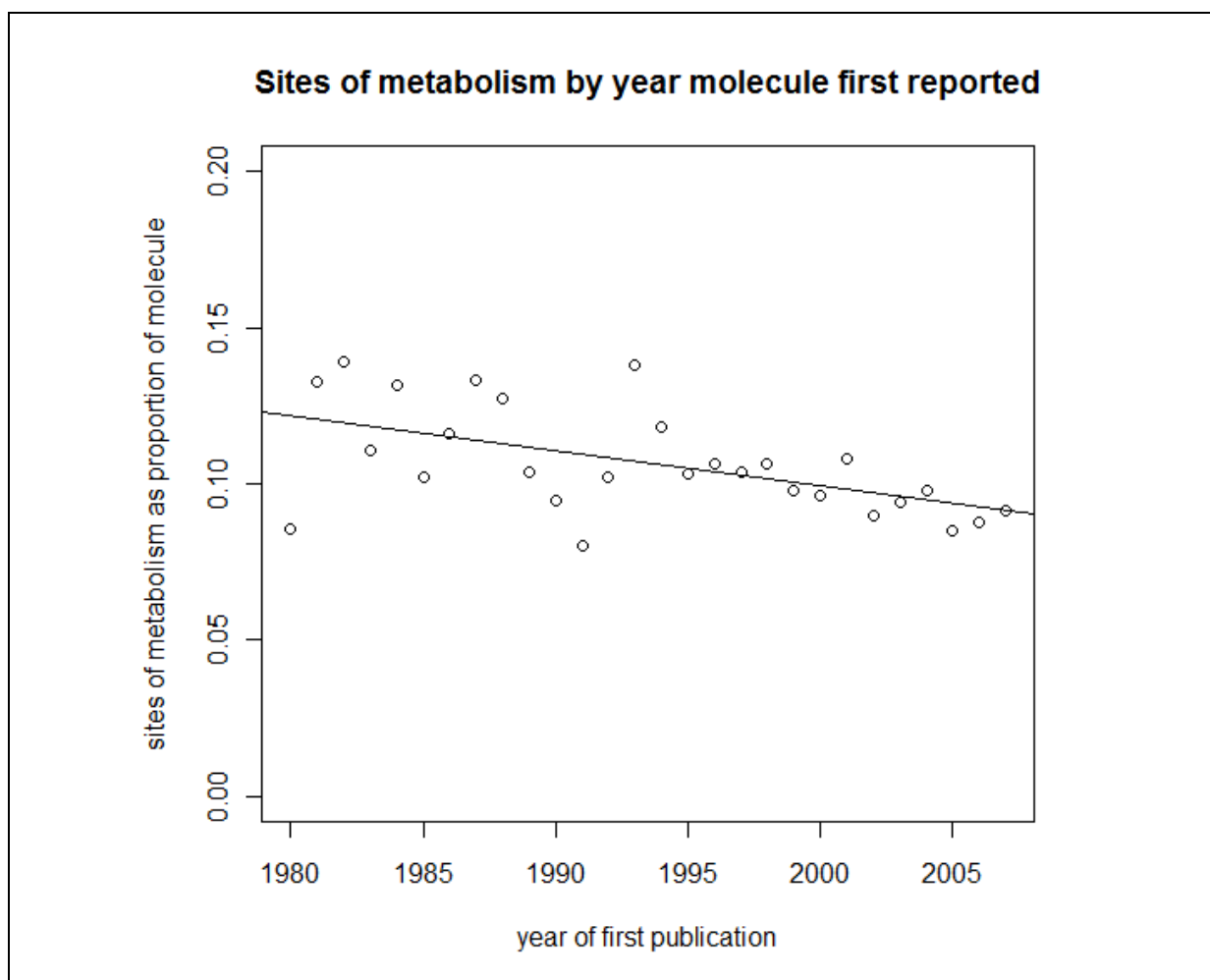


Figure 77: There is a weak but clear trend for more recently studied molecules to have a lower proportion of their structure reported to be sites of metabolism.

This trend has a Pearson's product-moment correlation coefficient of -0.543, which hypothesis testing ($H_0: \rho = 0$, $p = 0.003$) indicates is statistically significant.

Figure 78 shows a plot of the mean number of sites of metabolism identified in a molecule against the number of heavy (non-hydrogen) atoms it contains. At the extremes of high and low heavy atom counts there is considerable variation in the mean number of sites of metabolism, which can be accounted for by the low number of compounds of those sizes. For the molecule sizes with a considerable amount of data available (the data from around 10-40 heavy atoms) there is very little variation in the mean number of sites of metabolism per molecule, in spite of an almost quadrupling in molecule size.

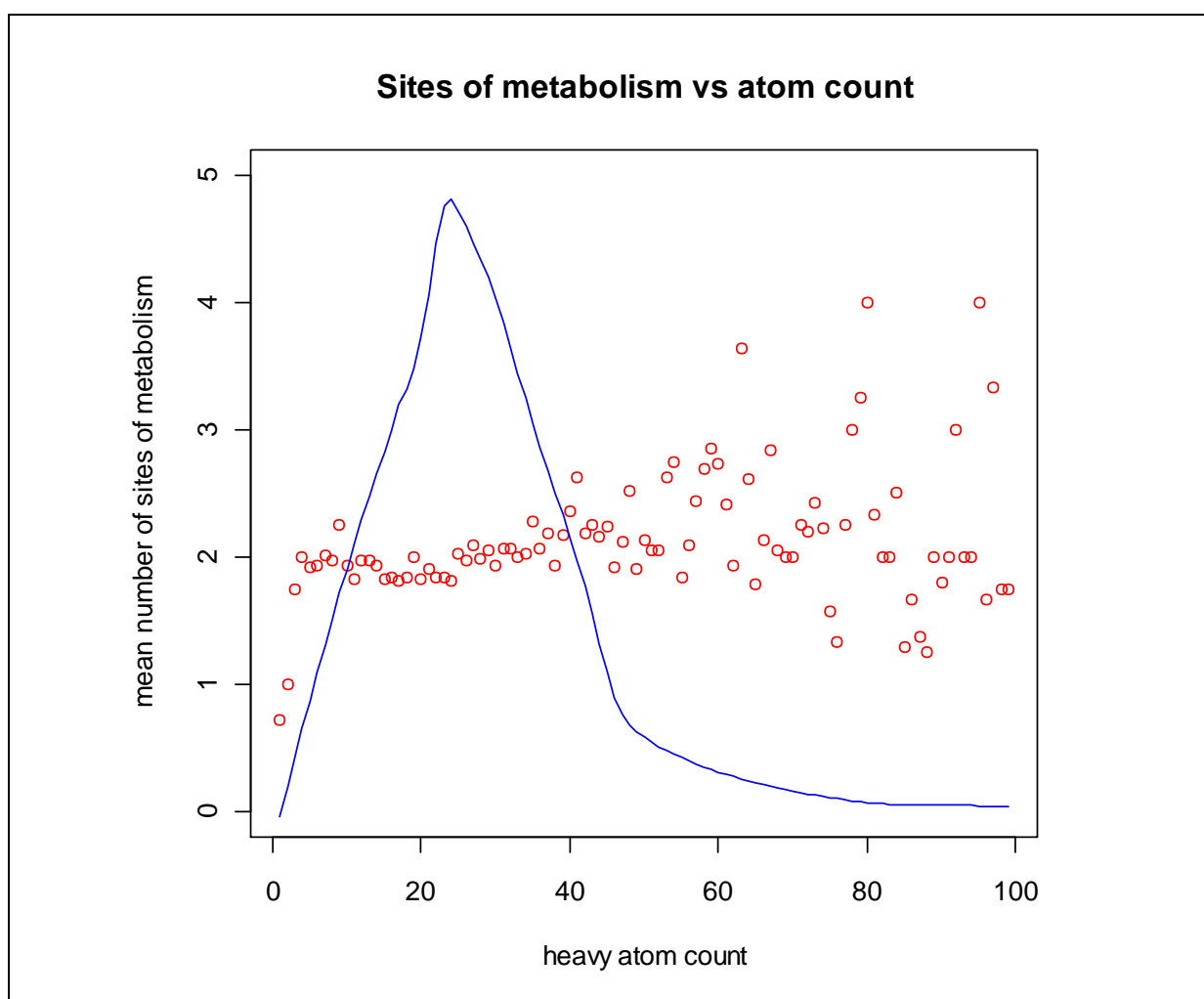


Figure 78: Red: The mean number of sites of metabolism of molecules of each size. Blue (scale not shown): Regression line indicating the number of molecules of each size. The number of sites of metabolism identified remains fairly constant, irrespective of the size of the molecule; for molecule with more than around 40 heavy atoms there is too little data to draw any conclusions.

This suggests that the sites of metabolism of larger compounds could be under-reported. It seems likely that the majority of studies are reporting only the two most major metabolites detected; it is possible that once two metabolites have been identified investigation of a compound ceases.

4.11 Conclusions

Chapters 3 and 4 have reported the development and evaluation of MetaPrint2D – a new tool for the prediction of sites of xenobiotic metabolism. MetaPrint2D has been made freely available as an Open Source library, meaning anyone can develop applications making use of it, and future improvements should feed back into the project. Three interfaces to the MetaPrint2D library have been provided, suiting a variety of use cases; MetaPrint2D is accessible through a website run from the Chemistry Department at the University of Cambridge, through integration with the Bioclipse platform, and as a simple command-line application.

The data-mining approach to site of metabolism prediction used by MetaPrint2D allows predictions to be made on tens of structures per second, on a regular desktop PC. This is much faster than other comparable tools, and means that for the first time chemists using a tool such as Bioclipse will be able to investigate how the likely sites of metabolism of a molecule change as they make modifications to its structure in real time. MetaPrint2D's speed will also enable the inclusion of site of metabolism predictions in high-throughput virtual screening programmes.

MetaPrint2D has undergone one of the most extensive evaluations reported for any site of metabolism prediction tools, with predictions tested on around 1200 substrates. In the course of this evaluation a novel ROC curve-based method for evaluating the performance of site of metabolism predictions has been proposed, which overcomes the biases inherent to the evaluation metrics currently used.

The evaluation of MetaPrint2D demonstrated the stability of the model to updates of the training data. It has also demonstrated how the model's applicability to a query compound can be straightforwardly estimated from the proportion of sites within the molecule for

which little or no data is available from the training database, and that this can be used to assign a degree of confidence to the model's predictions.

There are some limitations to MetaPrint2D's current approach to site of metabolism prediction: stereochemistry is ignored, as are 3-dimensional effects. Despite this, MetaPrint2D's performance has been found to be comparable to that reported for other tools. Finally, MetaPrint2D only predicts the relative likelihood of metabolism occurring at each site within a molecule, ignoring rates and yields.

The next chapter discusses the extension of MetaPrint2D in order to predict specific types of metabolic transformation, and the structures of potential metabolites.

5. Extension of MetaPrint2D to the prediction of transformation types and the generation of metabolites

As discussed in Chapter 2, current methods for predicting the metabolites formed through xenobiotic metabolism are rule based, often only provide a coarse-grained discrimination between the different products' likelihood of formation, and are prone to over-prediction of the number of metabolites. This chapter reports an extension of the data-mining approach to site of metabolism used by MetaPrint2D. Introduction of a list of reaction type definitions enabled identification of the transformations that each predicted site of metabolism is likely to undergo, and the metabolites generated.

5.1 Introduction

MetaPrint2D and its predecessor SPORCalc predict sites of xenobiotic metabolism, but make no prediction of the metabolites likely to be formed. MetaPrint2D-React is an extension of MetaPrint2D which includes predictions of the types of transformation that occur, and generates the structures of the metabolites formed. Like the other approaches to metabolite prediction described in Chapter 2, a set of reaction patterns are used to define the possible transformations, but MetaPrint2D-React provides much finer grained differentiation between the likelihoods of various metabolites being formed than the other rule-based tools.

As discussed previously, tools for the prediction of metabolites, such as Meteor (239), use a set of rules to define 'biophores' (descriptions of functional groups and other molecular properties) that are used to determine where in a molecule metabolic transformations may take place. In order to predict the metabolites of a compound, its structure is searched for the presence of each biophore, identifying all the sites in the structure where each transformation could occur. When different biophores indicate that several competing transformations could occur, relative reasoning rules can be used to assert one transformation's precedence over another. Each biophore has associated with it a likelihood score; this may be one of a small number of categories such as *very likely*, *likely*,

unlikely or may be a more finely grained score such as the empirical probabilities assigned to each rule by Sygma (242).

MetaPrint2D-React extends the data model used by MetaPrint2D in order to include sub-models for each class of reaction. In addition to recording the reaction centre and substrate occurrence counts, each atom environment fingerprint in MetaPrint2D-React also records occurrence counts for each type of transformation that has been observed to occur in the training data at an atom occupying that environment. Since this has been achieved through extension of the data structures described in Chapter 3, it adds very little overhead to the search performance.

In order predict the metabolism of a compound MetaPrint2D-React performs a search of the model's data for similar atom environments to each atom in the query structure in exactly the same manner as MetaPrint2D does. However, in addition to calculating the overall occurrence ratio for each atom in the structure, a separate occurrence ratio for each type of transformation reported in the training data is calculated. Structures of predicted metabolites are then generated through application of the reaction rules associated with each predicted transformation type.

This approach enables much finer grained differentiation between the relative likelihoods of predicted metabolites than any of the methods described in Chapter 2 since the occurrence ratio of each metabolite is based on a data mining search of the environments occupied by atoms in the molecule, rather than a match against one of a list of pre-defined transformation patterns, or biophores.

There are also a number of other benefits. Since pattern matching is not used to determine the sites at which transformations occur, a much smaller and simpler set of biotransformations can be defined, making maintenance of the rule base much simpler. For instance, rather than requiring many different rules for hydroxylation, defining precisely which substructures at which hydroxylation may occur, together with a separate likelihood for each rule, MetaPrint2D-React needs only a single definition of hydroxylation: the addition of an -OH group, and can determine the appropriate locations at which to apply the transformation, and the likelihood of it taking place, through a statistical analysis of the data in the Symyx® Metabolite database.

In addition, MetaPrint2D-React's metabolite prediction has the potential to be much faster than a purely rule-based approach. Substructure searches are a relatively computationally expensive procedure; in the case of a method relying on a list of several hundred biotransformation rules, the query structure must be searched for occurrences of the substrate pattern of each of the hundreds of rules. MetaPrint2D-React can determine the sites in a molecule at which each type of transformation can occur using the fast fingerprint search described previously. The most time consuming part of the MetaPrint2D search is the lookup of similar atom environments and the addition of occurrence counts for specific transformation types makes little difference to the calculation time.

Some methods, such as Meteor, base their estimate of a metabolite's likelihood of formation on an assessment of the structure of the metabolite formed. This means that even if the user only wishes to examine a subset of the predicted metabolites, say the 10 most likely, all of the metabolite structures must be generated and assessed. By generating likelihoods of formation from the results of atom environment fingerprint searching, MetaPrint2D-React allows small subsets of the metabolites to be rapidly selected for further analysis.

5.2 Identifying transformations

5.2.1 Metabolite database annotations

The 87,446 biotransformations recorded in the 2008.1 release of the Symyx® Metabolite database encompass a wide range of types of reaction. 68,900 of the records have reaction class annotations, assigning one or more of 286 reaction class labels such as *Hydroxylation* and *Hydrolysis* to the transformation. Of these reaction class labels, 115 are assigned to twenty or fewer biotransformations and only 95 are assigned to more than one hundred transformations. The most common reaction class labels are listed in Table 19.

C-Hydroxylation	8386	N-Oxidation	811	Oxidative Dealkylation	277
Hydrolysis	7658	N-Deacylation	768	Glycination	254
C-Oxidation	6033	Covalent Binding	757	S-Alkylation	230
Aromatic Hydroxylation	4063	Isomerization	751	N-Acetylcysteination	227
Aliphatic Hydroxylation	3855	O-Conjugation	746	Condensation	221
O-Glucuronidation	3666	Tautomerization	724	Cleavage	214
N-Dealkylation	3537	O-Methylation	678	Deglycosidation	208
Reduction	3108	N-Acylation	670	Epimerization	203
Ring Opening	2515	N-Reduction	632	Dealkylation	201
Oxidation	2232	Dehydration	604	O-Deacetylation	196
O-Dealkylation	1702	Decarboxylation	604	Dehydroxylation	189
N-Demethylation	1673	Phosphorylation	587	N-Deacetylation	185
Hydrogenation	1579	Ring Closure	573	Desulfuration	178
Conjugation	1464	Chain Shortening	550	O-Phosphorylation	173
Glutathionation	1322	Rearrangement	533	Lipid Binding	161
O-Sulfation	1315	O-Alkylation	531	C-Dealkylation	157
Dehydrogenation	1254	Deamination	519	Dimerization	155
Epoxidation	1232	O-Deacylation	478	Dephosphorylation	154
Dehalogenation	1187	Nucleophilic Addition	467	Chain Elongation	152
Aromatization	1166	Cyclization	459	O-Dephosphorylation	152
O-Demethylation	1156	Elimination	442	N-Methylation	152
Dearomatization	1072	Esterification	405	Inversion	150
S-Oxidation	1033	Oxidative N-Dealkylation	376	Radical Formation	147
Protein Binding	1027	N-Glucuronidation	337	Lactonization	147
Optical Resolution	1013	Hydration	333	Glycosidation	144
N-Acetylation	1013	Nucleophilic Substitution	306	O-Deglycosidation	140
DNA Binding	1010	Sulfation	290	Sulfuration	135
Glucuronidation	1005	S-Methylation	289	Amidation	133
Oxidative Deamination	875	S-Dealkylation	286	Oxidative Desulfuration	132
Hydroxylation	856	N-Hydroxylation	285	N-Alkylation	118

Table 19: The 90 most common reaction class labels from the 2008.1 release of the Symyx® Metabolite database, together with their occurrence counts.

Initially the use of these database annotations as the basis for determining the types of reaction to have occurred was considered, however this approach was not found to be feasible. Unfortunately the reaction type annotations in the Symyx® Metabolite database are inconsistent, and often incomplete. For instance, hydroxylation reactions occurring at a carbon atom are variously labelled with one or more of *Hydroxylation*, *C-hydroxylation*, *Aromatic Hydroxylation* and *Aliphatic Hydroxylation*. The hydroxylation transformations shown in Figure 79, illustrate this variability.

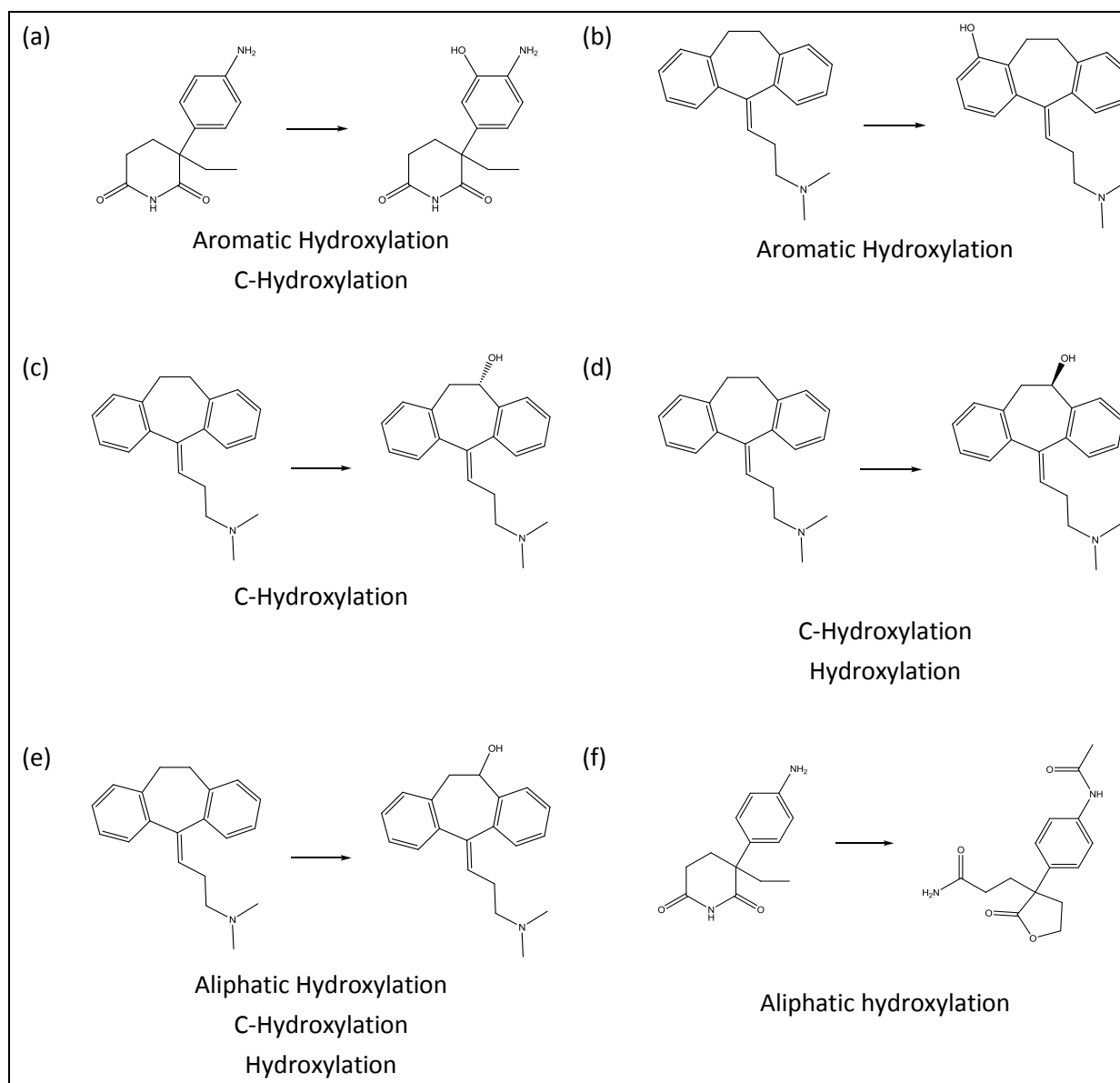


Figure 79: Examples of hydroxylation transformations from the Symyx® Metabolite database with their annotated reaction classes.

Transformations (b)-(e) from Figure 79 are all from the same reaction scheme, and even here there is little consistency in the annotations, while for transformation (f) no hydroxylation is immediately obvious. This single record in the database represents the combined result of a series of elementary reactions (shown in Figure 80), and the annotations describe the reactions occurring in the separate steps, rather than the overall transformation.

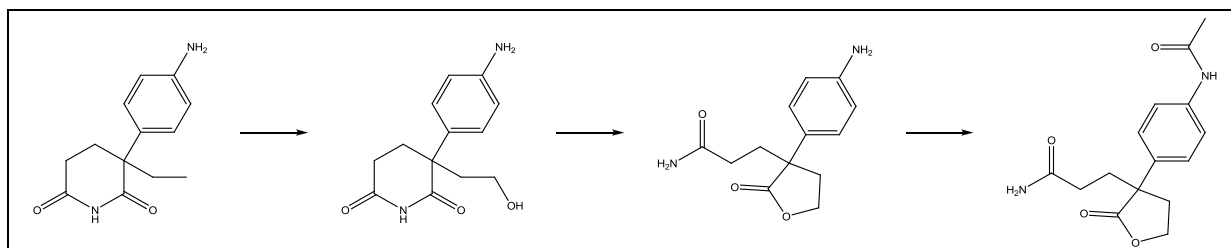


Figure 80: The series of elementary reactions making up the transformation shown in Figure 79(f). The record is annotated *Aliphatic Hydroxylation*, *N-Acetylation*, *N-Deacylation* and *Ring Closure*.

There is little consistency, however. In some similar cases where records represent the result of more than one elementary reaction the classes of some steps are omitted from the annotations, and in other cases the records are only annotated with the apparent reaction shown by the overall transformation. Alternatively, a record may represent the product of several transformations in different regions of the molecule, but only one of these is annotated.

The annotations also vary between database releases. The transformation shown in Figure 81 is described as an *Epoxidation* and *Hydrolysis* in the 2007.1 release of the database, but in the 2008.1 release is additionally annotated as *Aliphatic Hydroxylation*, *C-Hydroxylation* and *Hydrogenation*. The *Hydroxylation* annotations can be accounted for as the result of the overall process, but the last annotation does not appear to be related to the reaction scheme.

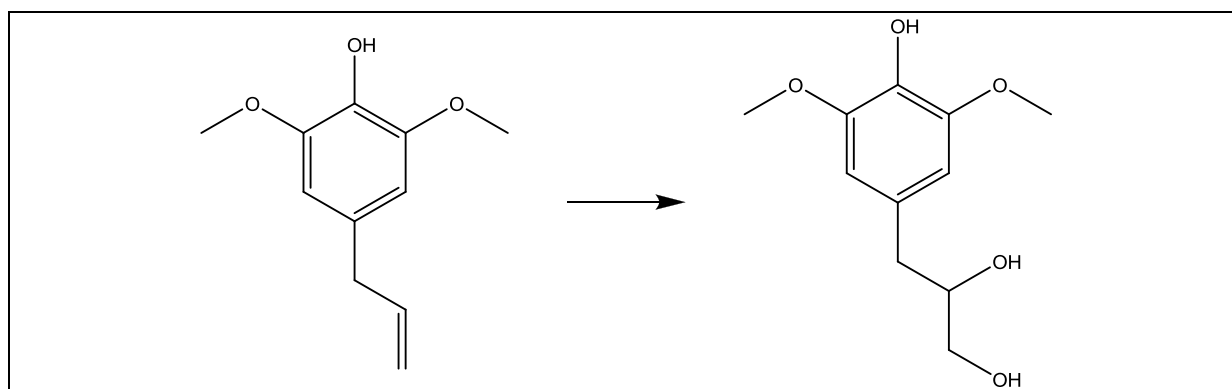


Figure 81: This transformation (MDLNUMBER: RMTB00078230) is annotated *Epoxidation* and *Hydrolysis* in the 2007.1 release of the Metabolite database. In the 2008.1 release it is additionally annotated *Aliphatic Hydroxylation*, *C-Hydroxylation* and *Hydrogenation*.

Other annotations are simply incorrect:

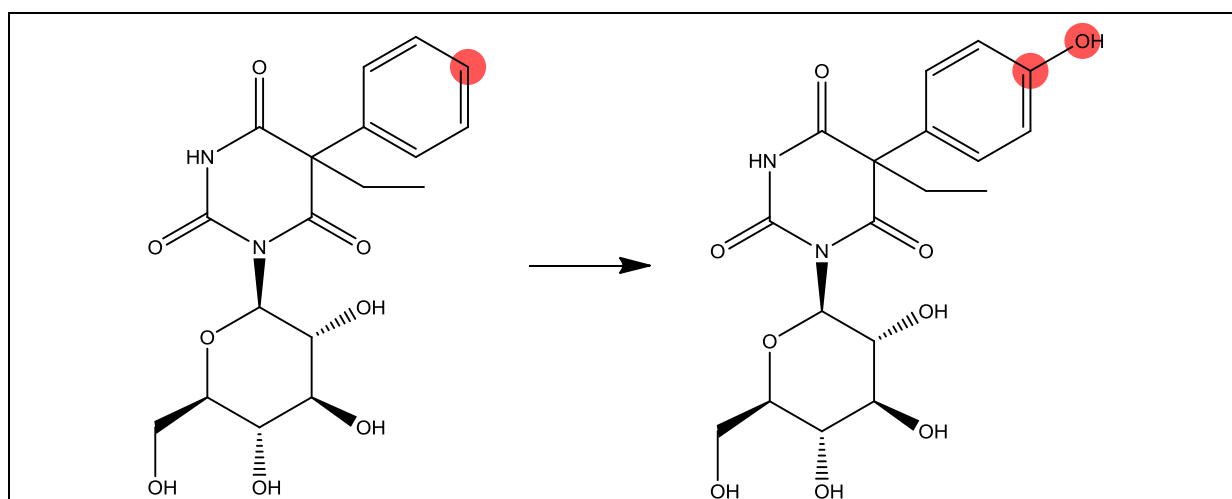


Figure 82: This transformation (MDLNUMBER: RMTB00010291) is annotated as an *N-deglucosidation* in the 2008.1 release of the Symyx® Metabolite database, but the transformation it represents is a hydroxylation.

A further challenge in working with the reaction class annotations from the Symyx® Metabolite database arises from the fact that each record in the database only represents a single product of the transformation. Reactions that generate more than one product are recorded in a series of records, one for each product. In these cases the assigned reaction classes can vary, depending on which product is under consideration. The reaction shown in Figure 83(a) is recorded as two separate transformations, originating from the same substrate, one leading to each of the product compounds. Transformation (b) is described as an '*Oxidative Deamination*' and an '*Oxidative N-Dealkylation*', while transformation (c) is

described as a '*C-Oxidation*' and an '*N-Dealkylation*' – despite both records being separate views onto the same metabolic transformation.

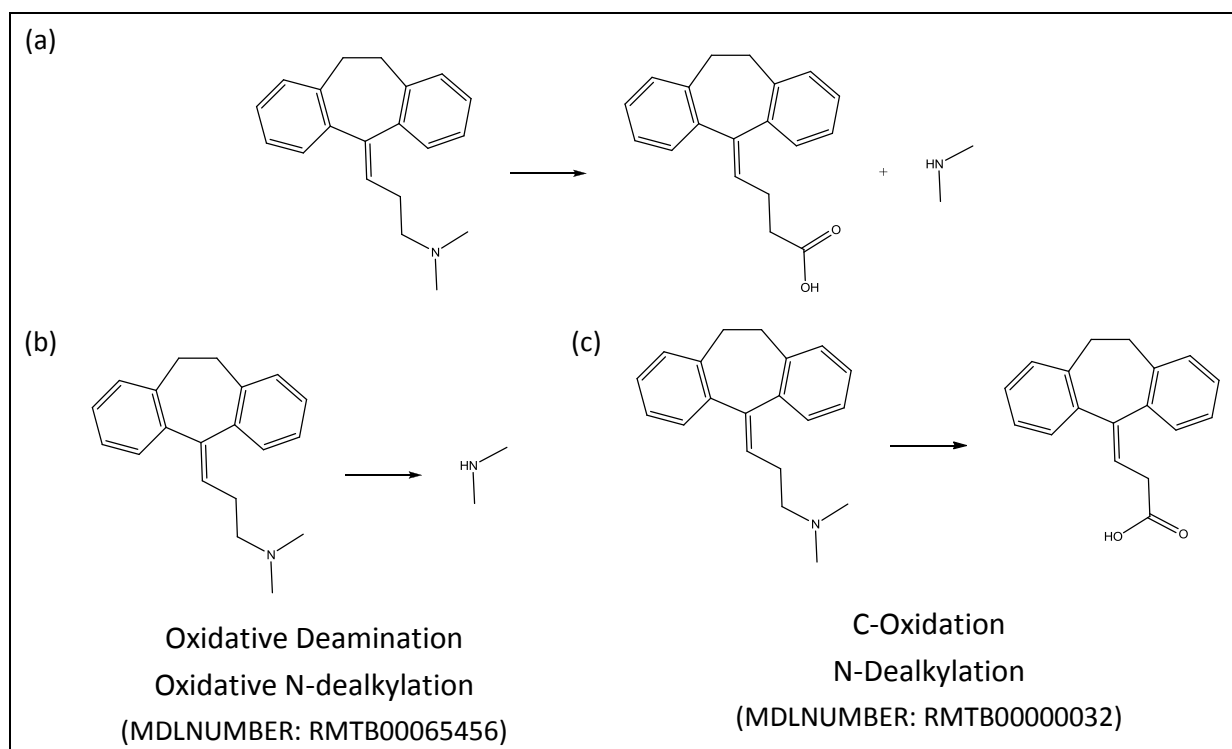


Figure 83: The Symyx® Metabolite database reports the overall reaction (a) in two separate records (b) and (c), one for each of the products. The reaction class annotations differ between these records, depending on which product is under consideration.

5.2.2 SMARTS patterns

Since the reaction class annotations in the Symyx® Metabolite database were unsuitable for use in assigning transformation types to sites of metabolism, a method for classifying the transformations using reaction SMARTS (55) patterns was developed. SMARTS is a language for describing molecular patterns, based on the widely used SMILES representation of molecular structure. Reaction SMARTS is an extension of the SMARTS language enabling description of reactions. Reaction SMARTS describe structures required to be present in reactant and product molecules, and can specify mappings between atoms in these structures. An example reaction SMARTS is described in Figure 84, below.

<chem>[OH:1]>>[O:1]C(=O)C</chem>
--

Figure 84: Reaction SMARTS pattern representing an esterification reaction. The reactant (to the left of the '>>') must contain an -OH group, and this maps to an oxygen in the product with an acetyl group added.

MetaPrint2D-React identifies reaction centres through a constrained maximum common substructure search, in exactly the same way as described in Chapter 3 for MetaPrint2D. MetaPrint2D assigned broad classifications (phase I addition, phase II addition, elimination, bond breaking, bond formed, bond order changed) to the reaction centre atoms. MetaPrint2D-React instead uses a list of reaction SMARTS patterns to classify the transformations on the basis of the structural changes between the substrate and metabolite compounds.

In order to determine which reaction types should be included in MetaPrint2D-React the most common reaction classes in the Symyx® Metabolite database were identified, along with common types of transformations reported in the literature. Reaction SMARTS patterns were then written to describe these transformations. MetaPrint2D-React stores these reaction type definitions in a configuration file, making it straightforward to make changes to the reaction types that are supported by the software.

Since the fingerprinting/data-mining approach of MetaPrint2D is used to determine the sites in a molecule where transformations should be applied, very general reaction rules can be used. Rather than the highly specific rules such as '4-Hydroxylation of 1,3-Disubstituted Benzenes' required by tools like Meteor, only a single rule for hydroxylation (using a wildcard to represent the atom to which the -OH group is added) is necessary. The same is true for many other types of reaction. If users wish to discriminate between transformation types at a more fine-grained level than these generic rules permit, for example between aliphatic C-hydroxylations, aromatic C-hydroxylations and N-hydroxylations, then this can easily be achieved with the addition of extra rules – appropriate reaction SMARTS patterns can be added to the reaction type definitions and the training data reprocessed to generate a new model.

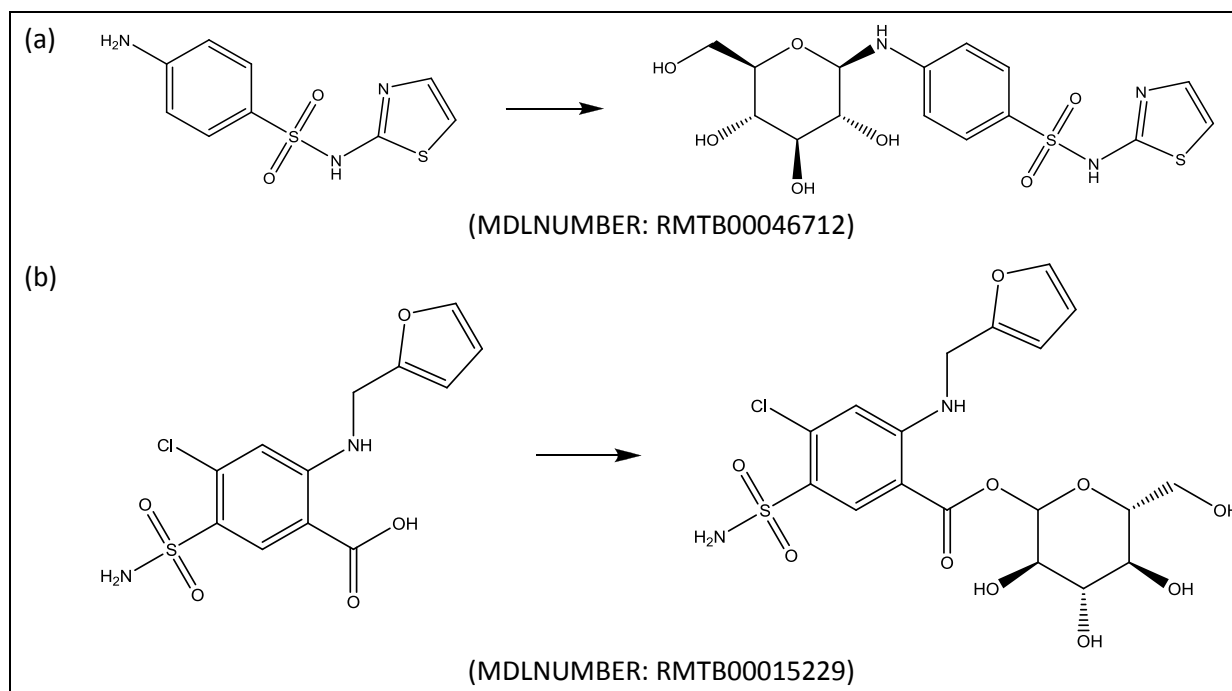


Figure 85: Both the N-glucosidation shown in (a) and the O-glucosidation shown in (b) can be described using the same reaction SMARTS pattern by representing the atom to which the glucose is conjugated by a wildcard which can match any chemical element.

The reaction types defined in MetaPrint2D-React mostly correspond to reaction classes in the Symyx® Metabolite database. Some of the Metabolite database's reaction classes are very broad, and have been assigned to a number of quite different reactions. The annotation '*Hydrolysis*', for instance, can describe an ester, amide or epoxide hydrolysis reaction. MetaPrint2D-React has separate reaction SMARTS patterns to handle these various cases.

The Symyx® Metabolite database also contains a number of quite generic reaction classes, such as acylation. In such cases MetaPrint2D-React can define several reaction patterns covering both specific commonly occurring cases and broader generic reaction types. In the case of acylation (illustrated in Figure 86), MetaPrint2D-React contains two rules; the first covering the specific case of acetylation, the most common type of acylation, and a broader rule to catch the remainder of cases. Similarly in the case of dealkylation reactions MetaPrint2D-React defines two rules: one covering the most common case – demethylation, and a broader rule for the generic case.

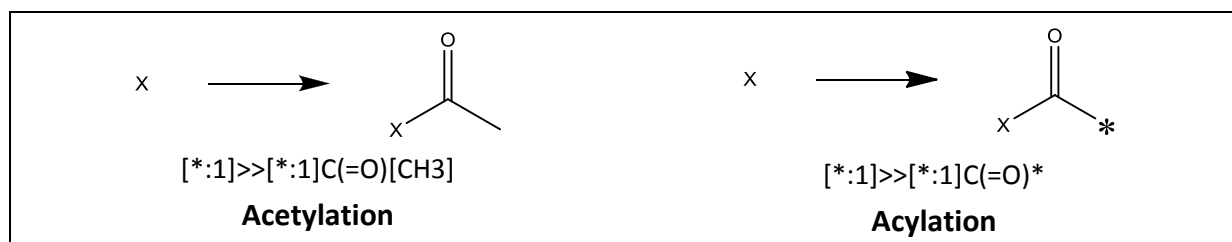


Figure 86: MetaPrint2D–React includes generic reaction type rules such as *acylation*, and also more specific rules to cover the most common instances, such as *acetylation*.

In other cases multiple rules may be required in order to represent variants of a group that is added:

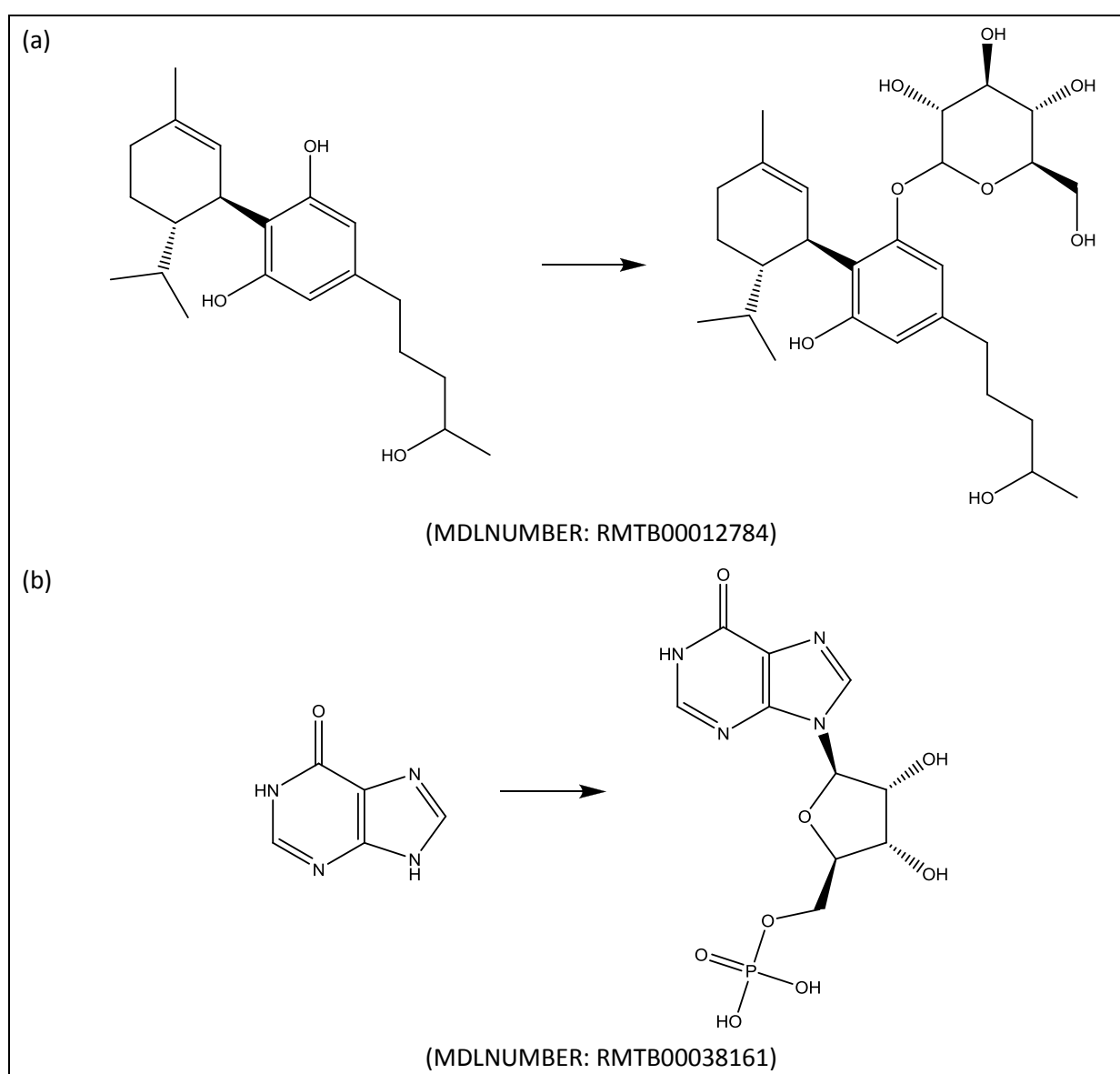


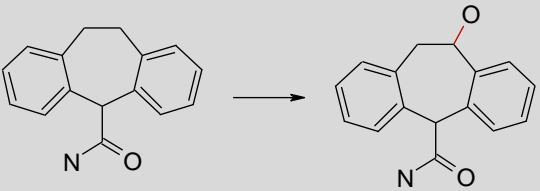
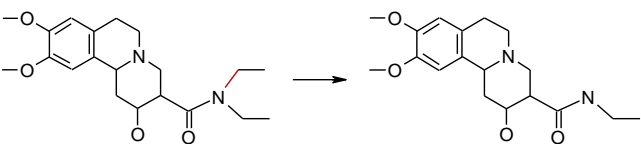
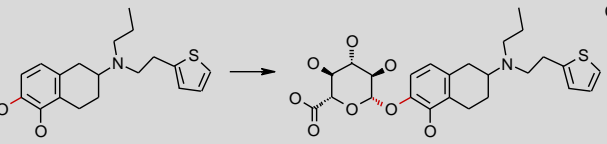
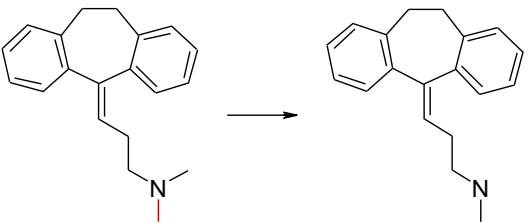
Figure 87: Separate reaction SMARTS patterns are required to describe the glycosidation reactions shown in (a), and with and additional phosphate group in (b).

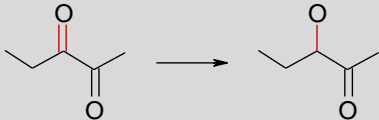
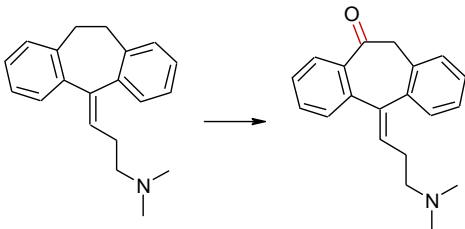
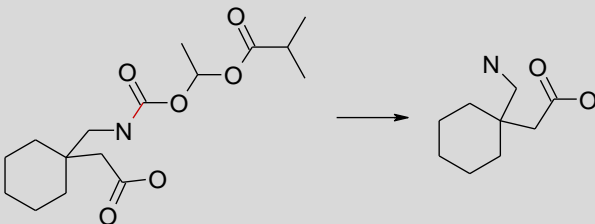
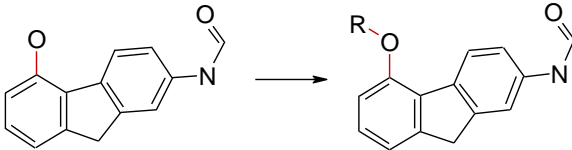
Finally, in order to avoid enumerating the large number of possible substitution reactions, MetaPrint2D-React has the capability to recognise substitutions as the combination of an addition pattern and an elimination pattern.


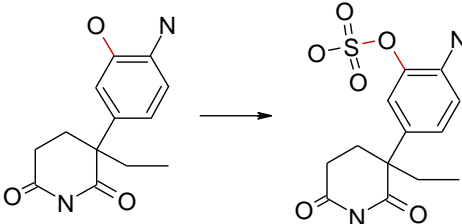
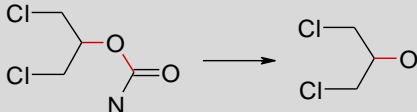
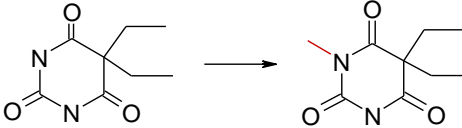
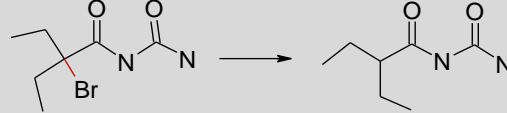
The browser tool used to evaluate the reaction centre identification algorithms used by MetaPrint2D (described on page 120) has been adapted to assist with the evaluation and refinement of the reaction type rules. The browser enables each record from the Symyx® Metabolite database to be inspected, highlighting the regions of the molecule undergoing a transformation. The reaction classes detected by the rules are listed and the atoms at which the reaction occurred highlighted. Any regions undergoing transformation that are not described by a transformation rule are also highlighted. The initial reaction type rules were refined through manual inspection of records sampled at random from the Symyx® Metabolite database using this browser. Common types of transformation that were not adequately described by the initial rule set were identified and the rules adapted accordingly.

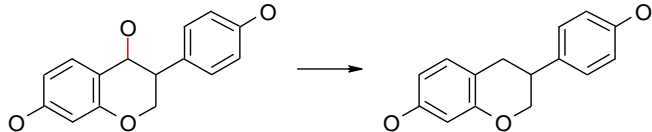
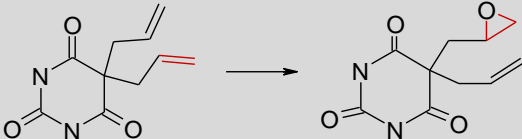
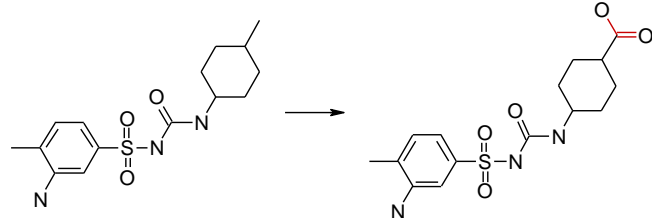
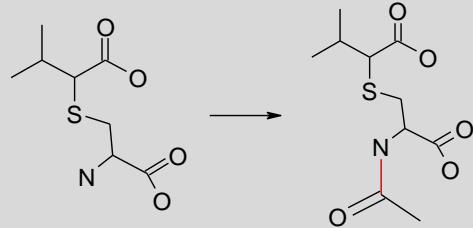
The reaction type rules developed through this process characterize 81.2% of reaction centre atoms identified in the 2008.1 release of the Symyx® Metabolite database. Examples of the reaction SMARTS patterns describing the most common reaction types, together with cases of the transformations they represent, are presented below.

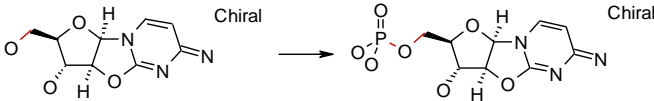

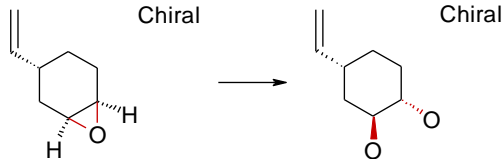
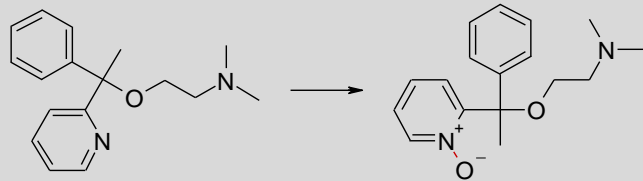
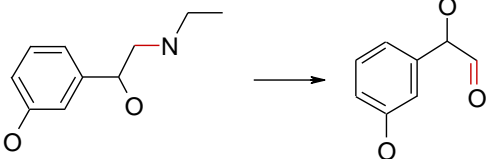
5.2.3 Reaction type definitions

Description	Frequency	Example Transformations	SMARTS Patterns
Hydroxylation	12732	 MDLNUMBER: RMTB00000015	<chem>[*:1]>>[*:1]-[OH]</chem>
Dealkylation	6996	 MDLNUMBER: RMTB00000099	<chem>[*:1]-C>>[*:1]</chem>
Glucuronidation	5741	 MDLNUMBER: RMTB00015763	<chem>[*:1]>>[*:1]C1C(O)C(O)C(O)C(C(=O)O)O1</chem>
Demethylation	3560	 MDLNUMBER: RMTB00000024	<chem>[*:1]-[CH3]>>[*:1]</chem>

Description	Frequency	Example Transformations	SMARTS Patterns
Reduction (double to single bond)	3466	 MDLNUMBER: RMTB00017724	<chem>[*:1]=[*:2]>>[*:1]-[*:2]</chem>
Oxidation (+O)	2905	 MDLNUMBER: RMTB00001469	<chem>[*:1]>>[*:1]=O</chem>
Amide hydrolysis	2707	 MDLNUMBER: RMTB00074205	<chem>[N:1]-[C\$(*=O):2]>>[*:1].[OH]-[*:2]</chem> <chem>[N:1]-[C\$(*=O):2]>>[OH]-[*:2]</chem> <chem>[N:1]-[C\$(*=O):2]>>[*:1]</chem>
Conjugation	2400	 MDLNUMBER: RMTB00066487	<chem>[*:1]>>[*:1]-[#0]</chem>

Description	Frequency	Example Transformations	SMARTS Patterns
Oxidation (single to double bond)	2218	 MDLNUMBER: RMTB00000313	<chem>[*:1]-![:*2]>>[*:1]=![:*2]</chem>
Sulfation	1988	 MDLNUMBER: RMTB00046671	<chem>[*:1]>> [*:1]-S(=O)(=O)-O</chem>
Ester hydrolysis	1971	 MDLNUMBER: RMTB00000001	<chem>[C:1](=[O:2])-[O:3]-[*:4]</chem> <chem>>>[C:1](=[O:2])-[OH].[OH:3]-[*:4]</chem> <chem>[C:1](=[O:2])-[O-]*</chem> <chem>>>[C:1](=[O:2])-[OH]</chem> <chem>C(=O)-[O:1]-[*:2]>>[OH:1]-[*:2]</chem>
Methylation	1154	 MDLNUMBER: RMTB00000079	<chem>[*:1]>>[*:1]-[CH3]</chem>
Dehalogenation	1115	 MDLNUMBER: RMTB00000221	<chem>[*:1]-[I,Br,Cl,F]>>[*:1]</chem>

Description	Frequency	Example Transformations	SMARTS Patterns
Dehydroxylation	1088	 MDLNUMBER: RMTB00070716	<chem>[*:1]-[O;H,-]>>[*:1]</chem>
Epoxidation	1035	 MDLNUMBER: RMTB00000022	<chem>[*:1]=[*:2]>>[*:1](-O1)-[*:2]-1</chem>
Oxidation (=O,-OH)	1004	 MDLNUMBER: RMTB00000113	<chem>[*:1]>>[*:1](=O)-[OH]</chem>
Acetylation	935	 MDLNUMBER: RMTB00011676	<chem>[*:1]>>[*:1]-C(=O)-[CH3]</chem>

Description	Frequency	Example Transformations	SMARTS Patterns
Phosphorylation	821	 <p>MDLNUMBER: RMTB00003618</p>	<chem>[*:1]>>[*:1]-P(=O)(-O)-O</chem>
Tautomerization	720	 <p>MDLNUMBER: RMTB00086505</p>	<chem>[*:1]=[*:2]-[*:3]>>[*:1]-[*:2]=[*:3]</chem>
Epoxide hydrolysis	701	 <p>MDLNUMBER: RMTB00059271</p>	<chem>[r:1]1-[r:1]-[Or:2]-1>>[*:1](-[OH])-[*:1]-[OH:2]</chem>
Hydroxidation	681	 <p>MDLNUMBER: RMTB00006966</p>	<chem>[*:1]>>[*:1]-[O-]</chem>
Oxidative deamination (=O)	678	 <p>MDLNUMBER: RMTB00060711</p>	<chem>[*:1]-N>>[*:1]=O</chem>

5.3 Predicting transformations

MetaPrint2D-React predictions are based on a circular atom environment fingerprint similarity search, performed analogously to that used by MetaPrint2D for site of metabolism predictions. In order to enable prediction of the types of transformation that can occur at each site, every fingerprint is associated with a count of how many times each reaction type has been observed at an atom occupying that environment in the training data. When predictions are made, each atom in the query compound has a score (normalized occurrence ratio) generated for each type of reaction, in addition to the overall score for the likelihood of metabolism occurring at that site. The overall score is equal to the sum of the scores for each reaction type at that site.

The fingerprint search determines the possible reactions at each site in the molecule, and these predictions are refined through checks that the reactant pattern for each reaction type matches the sites at which they are predicted to occur. For many reaction types, such as hydroxylation, this is a trivial process that can be omitted since the reactant pattern is simply the wildcard 'any atom', but for others, particularly those transformations involving several atoms in their pattern, this is an important step. In the case of reaction types with several atoms in their reactant pattern, such as hydrolysis or epoxidation, it is possible that the reaction is predicted to occur at some of the required atoms but not others:

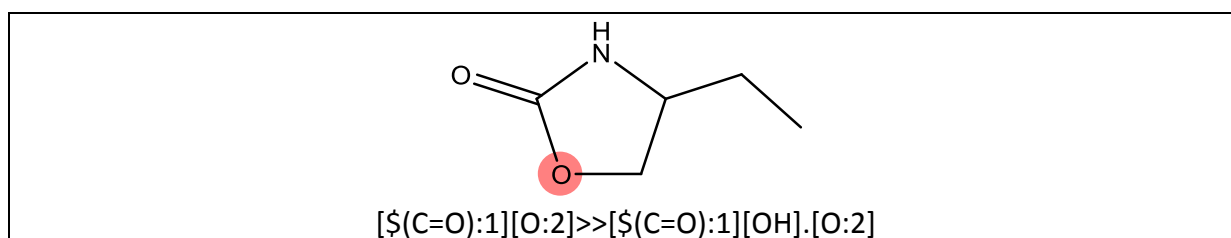


Figure 88: Hydrolysis has been predicted to occur at the highlighted position, but not at the adjacent carbonyl carbon, required by the reaction pattern. Since there is not a full match of the reaction pattern to atoms at which hydrolysis is predicted, the partial prediction is discarded.

The atoms' overall scores are analogous to the normalized occurrence ratios calculated by MetaPrint2D when making site of metabolism predictions. The values differ from the site of metabolism scores computed by MetaPrint2D however, as MetaPrint2D-React captures a different subset of the reaction centres. The site of metabolism prediction models used by

MetaPrint2D were trained using reaction centres classified as either phase I addition, or elimination. MetaPrint2D-React makes predictions on a greater range of additions, including phase II conjugations, but only includes the subset of elimination reactions specifically described by reaction SMARTS patterns.

5.4 Generating product structures

In order to generate the structures of metabolites the reaction SMARTS patterns are treated as SMIRKS (55) transformation patterns. SMIRKS is a language for defining transformations, derived from SMILES and SMARTS.

In order to apply a SMIRKS transformation to a structure the reactant component of the SMIRKS pattern is matched against the sites of the query structure predicted to undergo the transformation. The SMIRKS pattern is analysed in order to determine the atoms conserved between the reactant and product patterns (using mapping IDs incorporated in the SMIRKS pattern), and the added, deleted and altered atoms and bonds are identified. The metabolite structure is then generated through duplication of the query structure and application of the changes from the transformation pattern.

5.5 User interface

Like the site of metabolism prediction code, MetaPrint2D-React has been designed as a library, enabling different user interfaces to be developed independently of the prediction engine, and facilitating the embedding of the tool in larger applications. Currently only one user interface is available: a website, hosted at the Unilever Centre for Molecular Science Informatics in the Cambridge University Chemical Laboratories (<http://www-metaprint2d.ch.cam.ac.uk/metaprint2d-react>).

Query molecules can either be input using the SMILES format, or sketched using the JME editor. The output initially appears in a very similar form to that of the MetaPrint2D website: a structure diagram with atoms highlighted using a traffic-light system indicating the relative likelihood of metabolic transformations being centred on that site. Moving the cursor over an atom, however, reveals the list of reaction types predicted to occur at that site, with their relative scores, as shown in Figure 89.

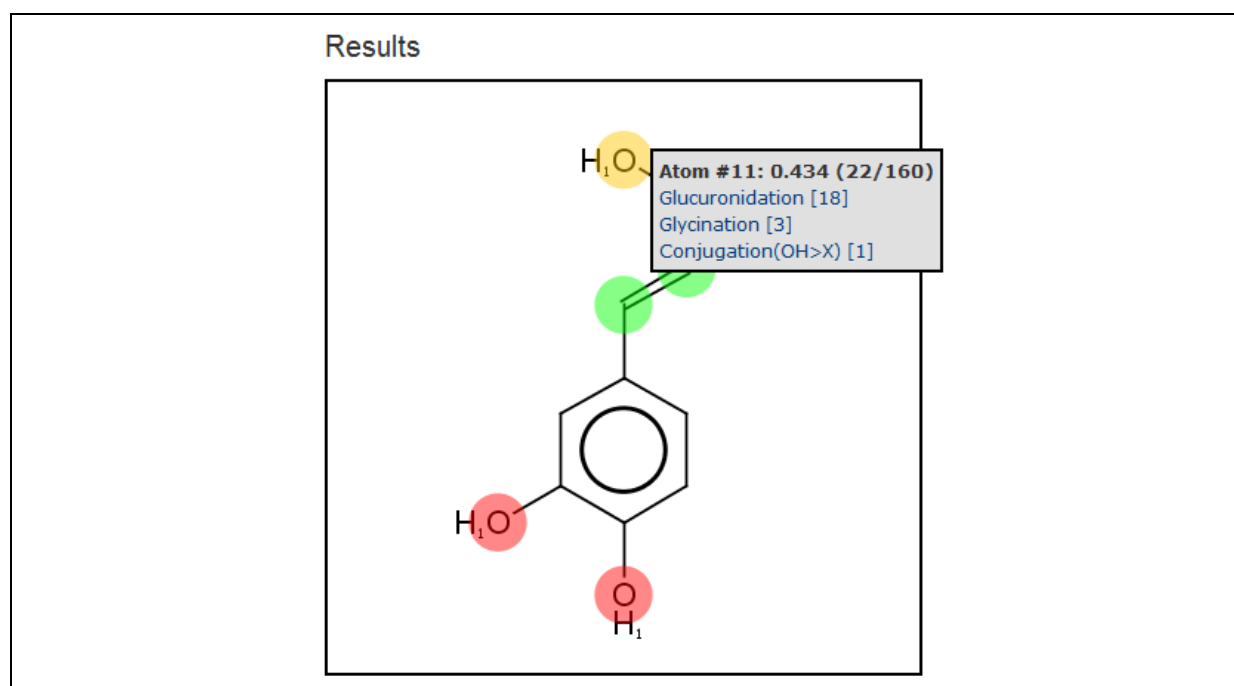


Figure 89: MetaPrint2D-React predictions for a compound. The relative likelihood of metabolism occurring at different sites in the structure is shown, and moving the cursor over an atom reveals more details, including the types of reaction predicted to occur at that site.

Clicking on one of the predicted reaction types generates the structure of the metabolite resulting from that transformation:

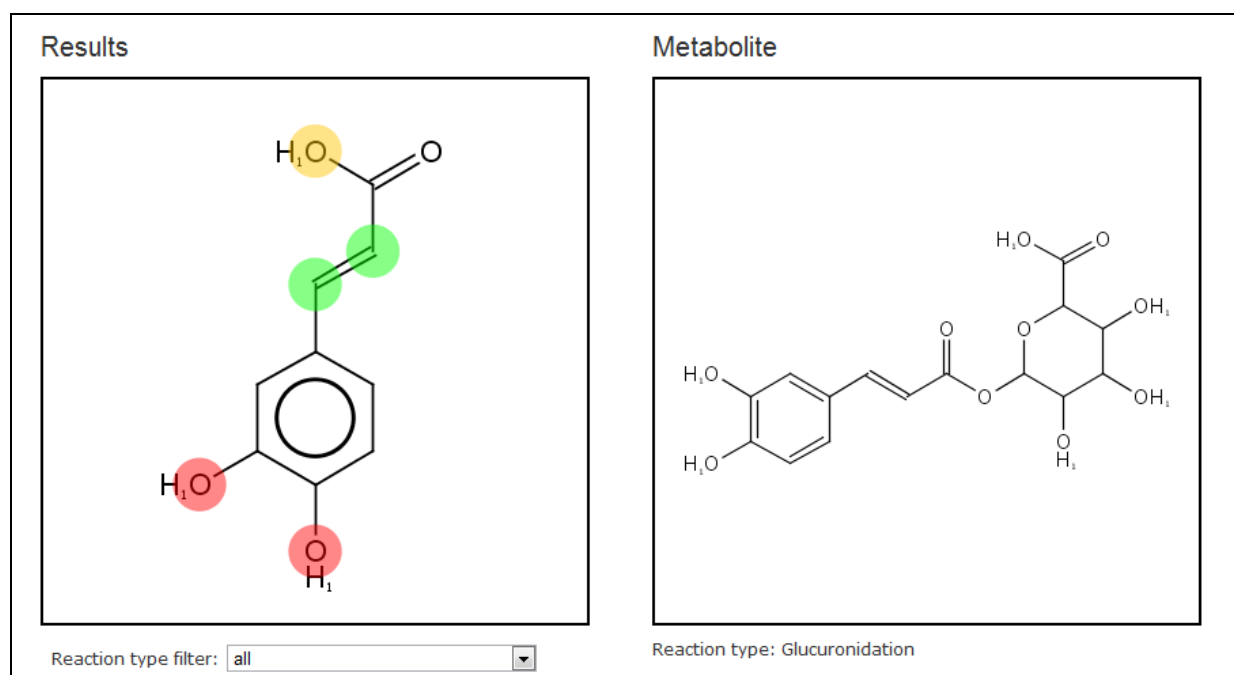


Figure 90: Clicking on one of the predicted reaction types generates the structure of the metabolite resulting from that transformation.

It is also possible to apply a filter to the results, presenting only the sites predicted for one particular class of reaction:

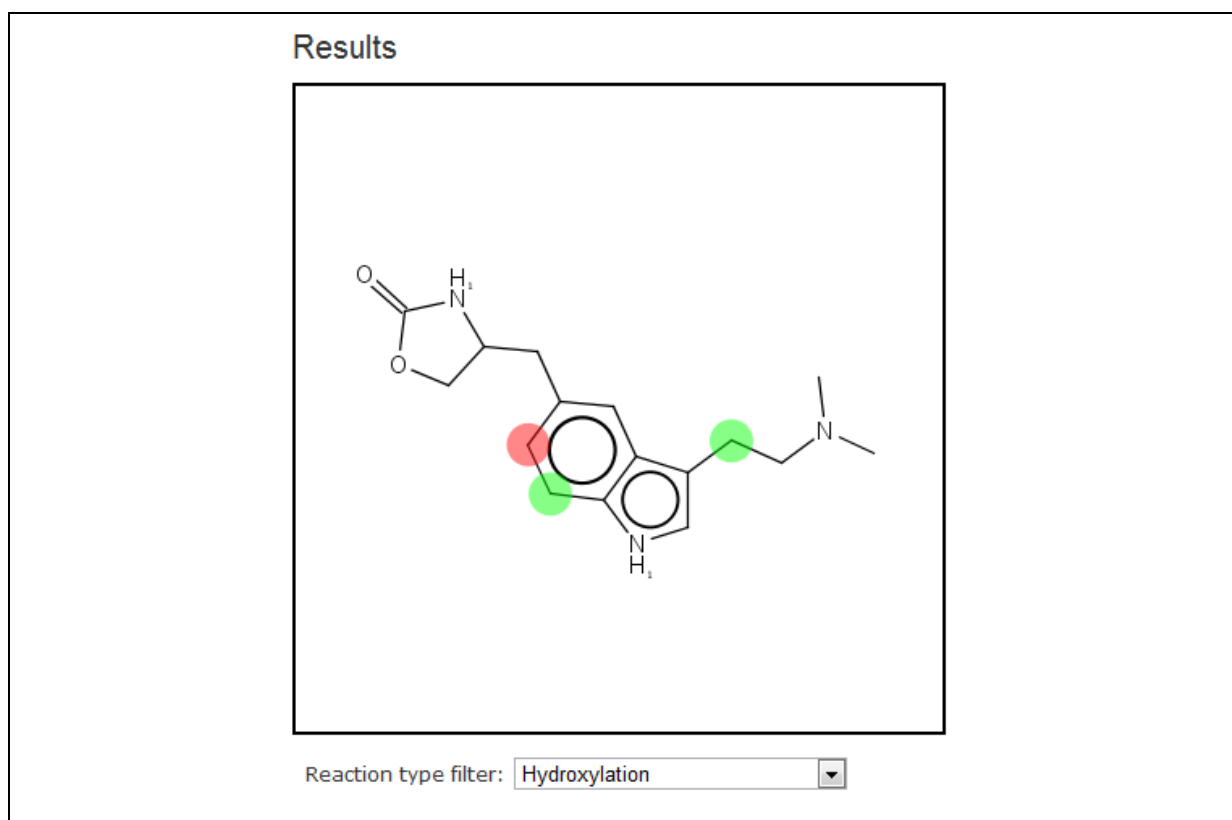


Figure 91: The reaction type filter can be used to limit predictions to one class of reaction; here it is being used to show the relative likelihood of hydroxylation occurring at different sites in a molecule.

5.6 Evaluation

Two evaluations of MetaPrint2D-React have been performed. The first evaluated the identification of sites of metabolism by MetaPrint2D-React and the second evaluated the predictions of specific reaction types.

The evaluation of site of metabolism predictions using MetaPrint2D-React was performed in the same manner as the assessment of MetaPrint2D, described in Chapter 4. The 2006.1 release of the Symyx® Metabolite database was used to train the MetaPrint2D-React model and the quality of predictions was evaluated using novel data from the 2008.1 release. The reaction schemes added to the Metabolite database between the 2006.1 and 2008.1 releases were identified, and one test compound was randomly selected from each scheme. The results of this evaluation are presented in Table 20, below.

Molecule Count	% Top 1	% Top 3	Mean AUC	Median AUC
922	58.9%	78.7%	0.812	0.918

Table 20: Results of the evaluation of site of metabolism predictions made using MetaPrint2D-React.

During this exercise the ability of MetaPrint2D-React to predict types of transformation was also assessed. At each reaction centre in the test compounds the types of transformation that MetaPrint2D-React predicted were compared to the transformation types reported at that site. Of the 2889 transformations reported to occur in the training data, 2257 (78.1%) were predicted by MetaPrint2D-React.

A separate assessment has been performed for each reaction type predicted by MetaPrint2D-React. Due to the variation in the number of times that transformation types are reported in the Metabolite database, training and test data were selected through Monte Carlo cross-validation (283), in a similar manner to the approach used for the training and evaluation of cytochrome P450 isotope specific models described previously in Chapter 4.

For each reaction type, 20% of the molecules from the 2008.1 release of the Symyx® Metabolite database identified as undergoing that reaction were randomly selected to form a test set, and the remainder of the data, excluding molecules occurring in the same metabolic scheme as any of the test compounds, used to generate a MetaPrint2D-React model. This model's predictions for the molecules in the test set were assessed using the same metrics as described previously – the percentage of molecules having a site of metabolism among the top one or three predicted sites of metabolism, and the mean and median area under the ROC curve (AUC) statistics. The selection of test and training data, model construction and evaluation of predictions were repeated ten times for each reaction type, and the results averaged. These are reported in Table 21, below.

Reaction Type	Molecule 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
Amide hydrolysis	1817	79.70%	90.30%	0.934	1.000
Conjugation to unknown structure	1777	55.80%	78.20%	0.857	0.985
Reduction (double to single bond)	1571	76.20%	87.80%	0.887	0.992
Ester hydrolysis	1495	91.50%	97.40%	0.963	1.000
Oxidation (single to double bond)	1450	72.80%	85.70%	0.851	0.966
Sulfation	1395	73.40%	87.60%	0.927	1.000
Dehydroxylation	694	75.30%	90.60%	0.859	0.974
Acetylation	672	79.30%	85.40%	0.892	1.000
Methylation	670	61.30%	80.00%	0.841	0.990
Oxidation (=O,-OH)	607	54.80%	75.10%	0.829	0.947
Epoxidation	582	62.90%	77.00%	0.812	0.915
Dehalogenation	570	74.40%	90.60%	0.798	0.852
Oxidative deamination (=O)	552	81.50%	87.50%	0.877	1.000
Hydroxylation	539	60.60%	72.30%	0.839	0.995
Phosphorylation	467	81.30%	85.40%	0.920	1.000
Epoxide hydrolysis	431	94.50%	96.40%	0.941	1.000
Glutathionation (+SX)	430	38.60%	62.90%	0.724	0.763
Oxidative deamination (=O,-OH)	335	75.00%	85.00%	0.874	0.993
Dehydration	334	79.80%	88.30%	0.818	0.846
Tautomerization	320	65.00%	80.20%	0.806	0.890
Hydroxylation/ Tautomerization(=O)	296	59.40%	74.30%	0.771	0.831
Dephosphorylation	270	74.50%	89.20%	0.929	0.993
Oxidative deamination (-OH)	268	57.40%	69.90%	0.754	0.761
Epoxide opening (+X)	267	80.90%	89.80%	0.886	1.000
Epoxidation/Hydrolysis	266	50.00%	75.50%	0.784	0.819
Reduction (=O)	250	72.00%	80.90%	0.804	0.906
Aromatization	220	57.70%	72.30%	0.795	0.864
Oxidative Elimination	220	73.10%	86.00%	0.791	0.831

Reaction Type	Molecule Count	% Top 1	% Top 3	Mean AUC	Median AUC
Acylation	196	45.10%	65.20%	0.738	0.808
Hydration	173	51.30%	61.60%	0.720	0.711
Glycination	171	79.40%	87.20%	0.866	0.995
Reduction (=O,-O)	170	90.00%	94.50%	0.925	1.000
Methoxylation	156	39.00%	60.80%	0.778	0.873
Dealkylation (x2)	153	54.90%	70.60%	0.768	0.813
Elimination	152	61.40%	75.90%	0.820	0.895
Alkylation	147	50.70%	70.20%	0.747	0.820
Epoxide Hydrolysis/Aromatization	144	88.40%	91.70%	0.929	1.000
Glutathionation (=)	144	53.00%	71.50%	0.725	0.737
Elimination (XX)	140	61.30%	76.30%	0.755	0.755
Conjugation (substituting OH)	133	62.00%	76.50%	0.790	0.882
Acetylcysteination	131	42.50%	61.40%	0.710	0.723
N-dealkylation	111	42.30%	55.50%	0.697	0.666
Dealkylation (3)	111	83.60%	96.00%	0.930	1.000
Desulfuration	105	77.50%	90.00%	0.818	0.907
Glutathionation (O>SX)	102	58.70%	78.30%	0.770	0.829
Glucosidation (+X)	97	42.50%	64.00%	0.778	0.916
CoA Binding	90	73.00%	85.00%	0.911	1.000
Dealkylation (2)	89	51.30%	60.60%	0.763	0.828
Oxidation(=O=O)	87	54.00%	70.70%	0.763	0.857
Cysteamination	80	38.80%	54.90%	0.685	0.629
Demethylation (x2)	76	79.90%	90.00%	0.863	0.976
N-Dearylation	70	53.50%	58.90%	0.777	0.855
Denitration	65	91.50%	96.50%	0.892	0.938
Amination	64	26.70%	61.30%	0.708	0.685
Protein Binding	62	41.90%	63.30%	0.752	0.783
Conjugation (+SX)	62	53.00%	62.70%	0.740	0.755
Glutamation	61	21.70%	32.40%	0.731	0.788
Esterification	58	65.50%	84.20%	0.872	0.989
Azo_cleavage	57	83.80%	92.50%	0.924	0.994
Ring_opening	57	26.30%	45.30%	0.655	0.635
Elimination (XH)	55	71.90%	80.30%	0.834	0.918
Chlorination	54	54.10%	71.20%	0.809	0.898

Reaction Type	Molecule Count	% Top 1	% Top 3	Mean AUC	Median AUC
Dehydrohalogenation	51	88.60%	92.80%	0.831	0.894
Conjugation (=)	48	45.10%	64.20%	0.695	0.724
Sulfuration	46	19.20%	44.70%	0.566	0.496
Methiolation	46	30.80%	38.10%	0.609	0.539
Epoxide Hydrolysis/Dehydration	38	76.90%	80.00%	0.886	0.962
DNA Binding	37	70.90%	81.00%	0.824	0.945
Cyanidation	35	32.90%	62.90%	0.708	0.693
Formylation	34	46.70%	52.00%	0.714	0.751
Nitrosation	33	52.00%	65.30%	0.714	0.711
Deamination (NH ₂)	29	58.00%	69.00%	0.648	0.584
Disulphide Reduction	28	30.00%	66.00%	0.788	0.803
Epoxide opening (3)	27	64.00%	72.00%	0.813	0.884
Glycosidation (+XP)	27	55.20%	83.80%	0.840	0.903
Aromatization/ Elimination	26	36.50%	69.50%	0.663	0.668
Sulfonation	25	58.50%	62.50%	0.769	0.825
Glycosidation (+X)	19	48.30%	80.00%	0.811	0.869
Oxidation/Dehalogenation	15	76.70%	76.70%	0.737	0.765
Hydroxylation/ Tautomerization(=O=O)	14	30.00%	60.00%	0.706	0.706
Thioester hydrolysis	13	85.00%	100.00%	0.962	0.962
Condensation	12	40.00%	55.00%	0.649	0.649
Oxidation (=O,-[O-])	11	75.00%	85.00%	0.713	0.713
Epoxide dehydration	9	50.00%	50.00%	0.743	0.743
N ₂ -elimination	9	80.00%	100.00%	0.900	0.900
Fluorination	8	30.00%	60.00%	0.641	0.641
Glucosidation (+OX)	8	0.00%	0.00%	0.444	0.444
Bromination	7	50.00%	70.00%	0.810	0.810
Peroxidation	6	10.00%	10.00%	0.475	0.475
Deamination (NHNH ₂)	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 21: MetaPrint2D-React's performance in predicting each type of reaction.

For the majority of reaction types MetaPrint2D-React's predictions are as accurate as, or better than, the site of metabolism predictions generated by MetaPrint2D. The best performing reaction types are those that can only occur at a specific substructure, such as ester hydrolysis, as opposed to transformations such hydroxylation which using a wild-card match can potentially be applied to any chemically relevant site.

The least accurately predicted transformations are all amongst those with the lowest numbers of occurrences, and those occurrences are in several different atom environments. This means that when the data is split into training and test sets in the course of cross-validation runs many of the atom environments occupied by reaction centre atoms in the test set do not appear in the training data.

The reaction types with the least occurrences have been removed from MetaPrint2D-React.

5.7 Conclusions

There are some limitations to the methods used by MetaPrint2D-React. Reaction types must be expressed as SMARTS patterns. This means that transformations must be defined in terms of exact substructures – SMARTS patterns cannot represent a query such as 'a chain of 3-5 carbon atoms'. SMARTS patterns cannot represent the formation of radicals, though extensions to address this issue have been proposed (253). These restrictions make it unfeasible for MetaPrint2D-React to capture certain types of reaction, such as dimerization and ring contractions, since to do so would require enumerating every possible structure. This is not a major limitation of the approach since such reactions occur quite infrequently.

A related limitation is in the handling of reactions involving transformations at different sites in a molecule, linked by a conjugated system. Examples of these are shown in Figure 92. Such transformations can occur with a varying number of bonds separating the main reaction sites, and again cannot currently be handled by MetaPrint2D-React unless all possible arrangements are enumerated.

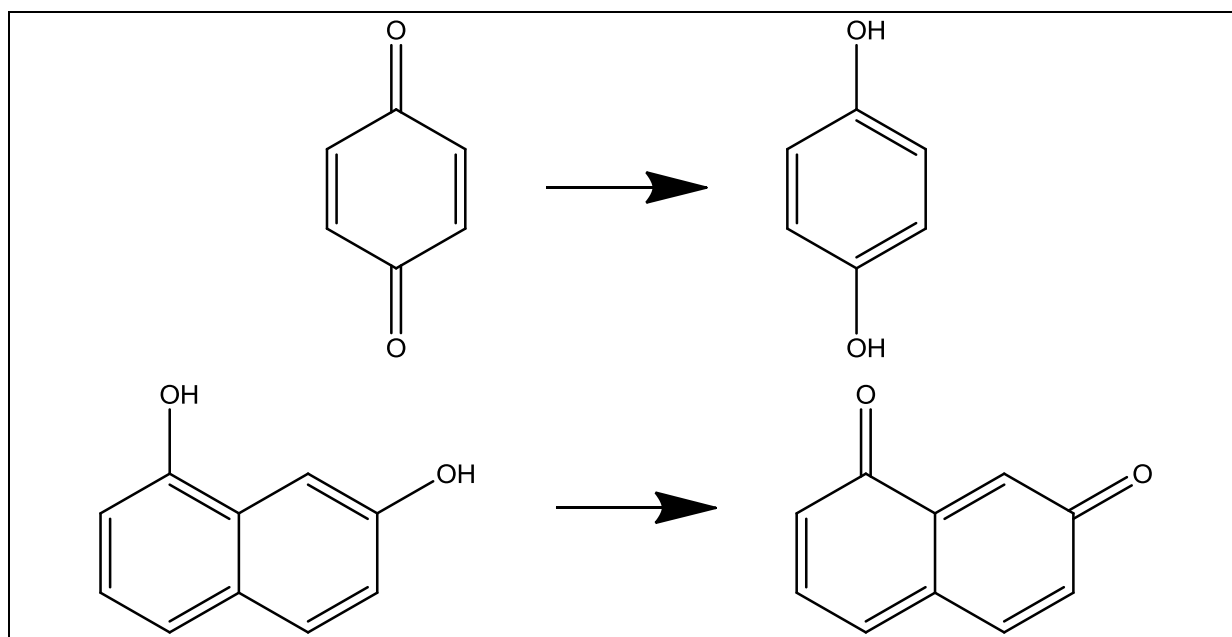


Figure 92: In order to represent reactions consisting of transformations at different sites in a molecule connected by a conjugated system using SMARTS patterns separate SMARTS are required for different lengths of conjugation.

There are several cases in which MetaPrint2D-React requires a number of separate reaction type definitions in order to describe closely related reaction types. An example where this arises is acylation – there is currently a specific rule for acetylation (the acetyl group is most common acyl group to be added) and a generic rule that captures the remaining cases of acylation. A similar situation occurs with alkylation where there is a specific rule describing methylation, and a generic rule describing other alkylation reactions. It may be useful, and improve the quality of predictions, if it were possible to establish hierarchies of reaction types. This would mean that methylation reactions could count towards the occurrences of alkylations, and the overall likelihood of alkylation could be predicted more accurately, while still enabling the likelihood of specific types of alkylation to be calculated. Similarly, in the case of oxidative deamination there are three different rules covering the hydroxyl, aldehyde and carboxy metabolites.

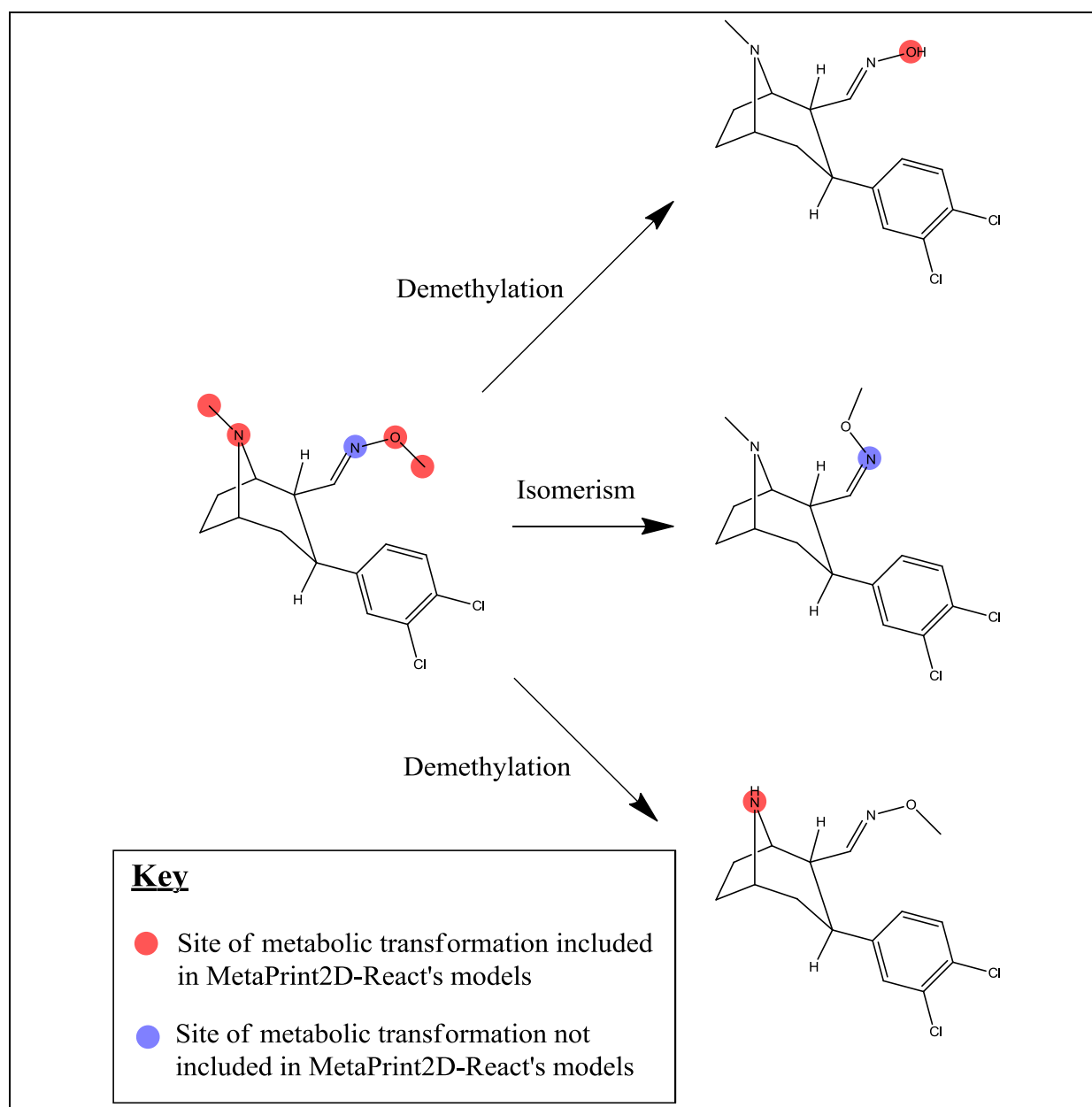
6. Retrospective prediction of recently published metabolic schemes

This chapter reports further work carried out to evaluate the performance of MetaPrint2D-React. Fifteen studies reporting the metabolic disposition of a novel xenobiotic compound were identified from the January to October 2008 issues of the journal *Drug Metabolism and Disposition*. Site of metabolism and type of transformation predictions for the parent compounds in each of these studies were generated by MetaPrint2D-React. These are compared to the compounds' reported metabolic dispositions below.

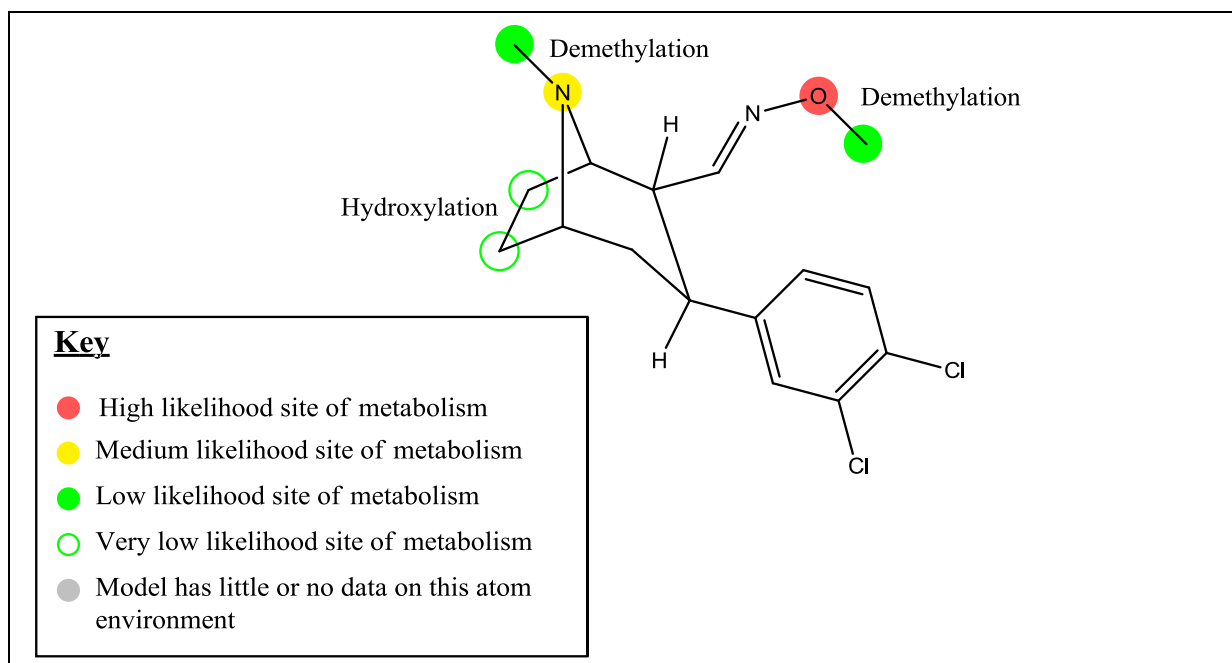
The metabolic studies used in this evaluation were selected on the basis of two criteria. Firstly, that the parent compounds in the studies were not included in the data used to train the MetaPrint2D-React model, and secondly that the paper reporting the study included a clear summary of the proposed metabolites of the parent compound.

6.1 [14C]Brasofensine (284)

6.1.1 Reported metabolites



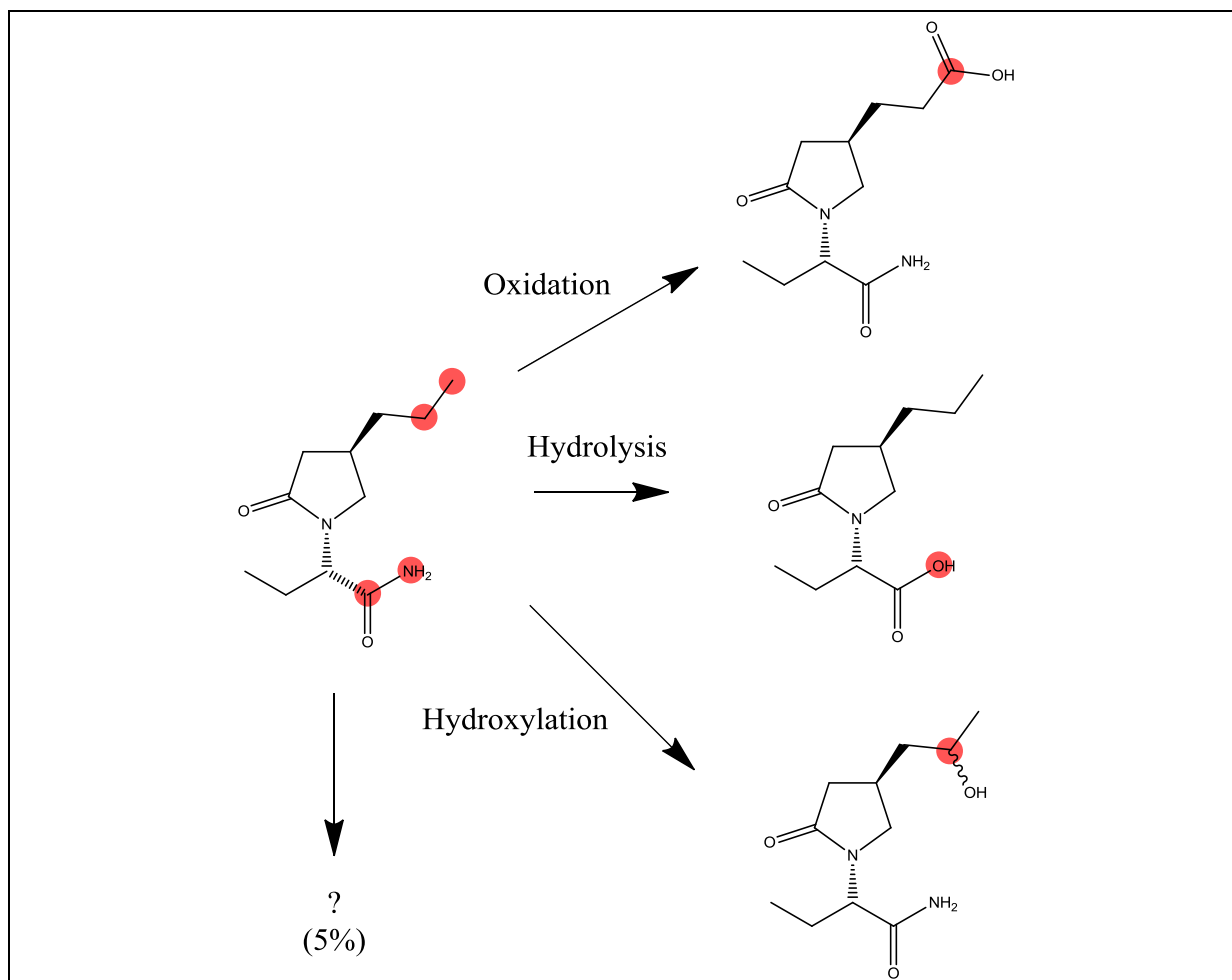
6.1.2 MetaPrint2D-React predicted transformations



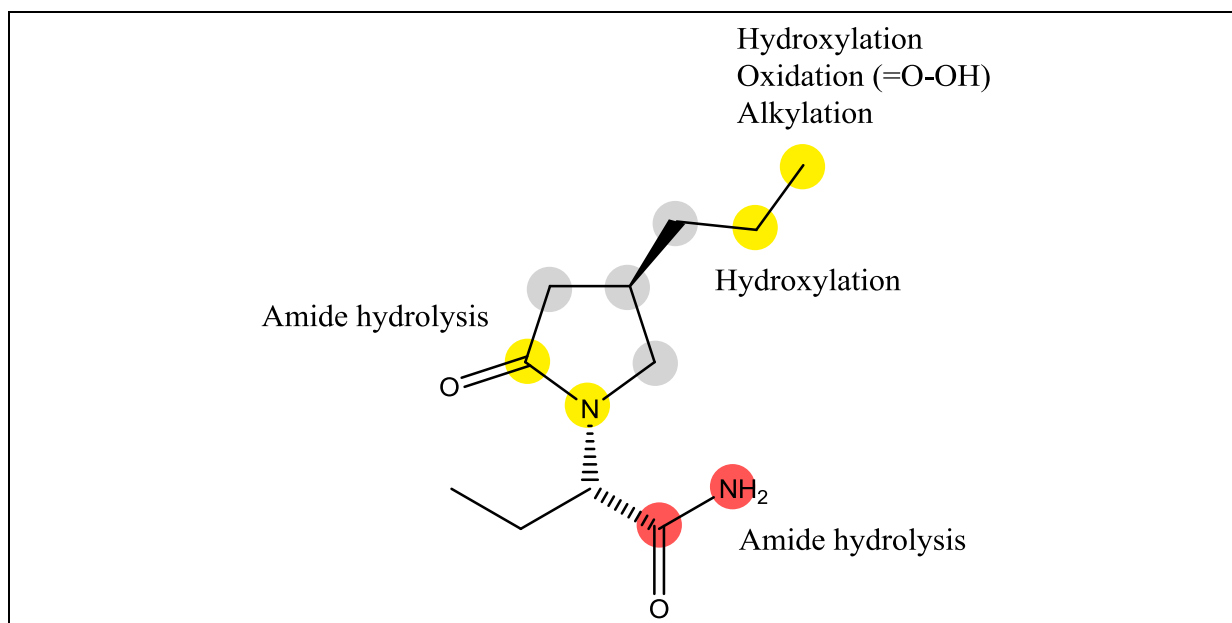
MetaPrint2D-React correctly identified the two possible demethylation reactions as the most likely transformations for the query molecule. The stereoisomerism was not predicted since this type of transformation is not included in MetaPrint2D-React's models since the descriptors currently used by MetaPrint2D-React do not include any information on stereochemistry. In addition to the reported transformations, MetaPrint2D-React predicted (with a low normalised occurrence ratio) two possible sites of hydroxylation. Overall, all four sites of metabolism (for transformations supported by MetaPrint2D-React) were identified.

6.2 ¹⁴C-Brivaracetam (285)

6.2.1 Reported metabolites



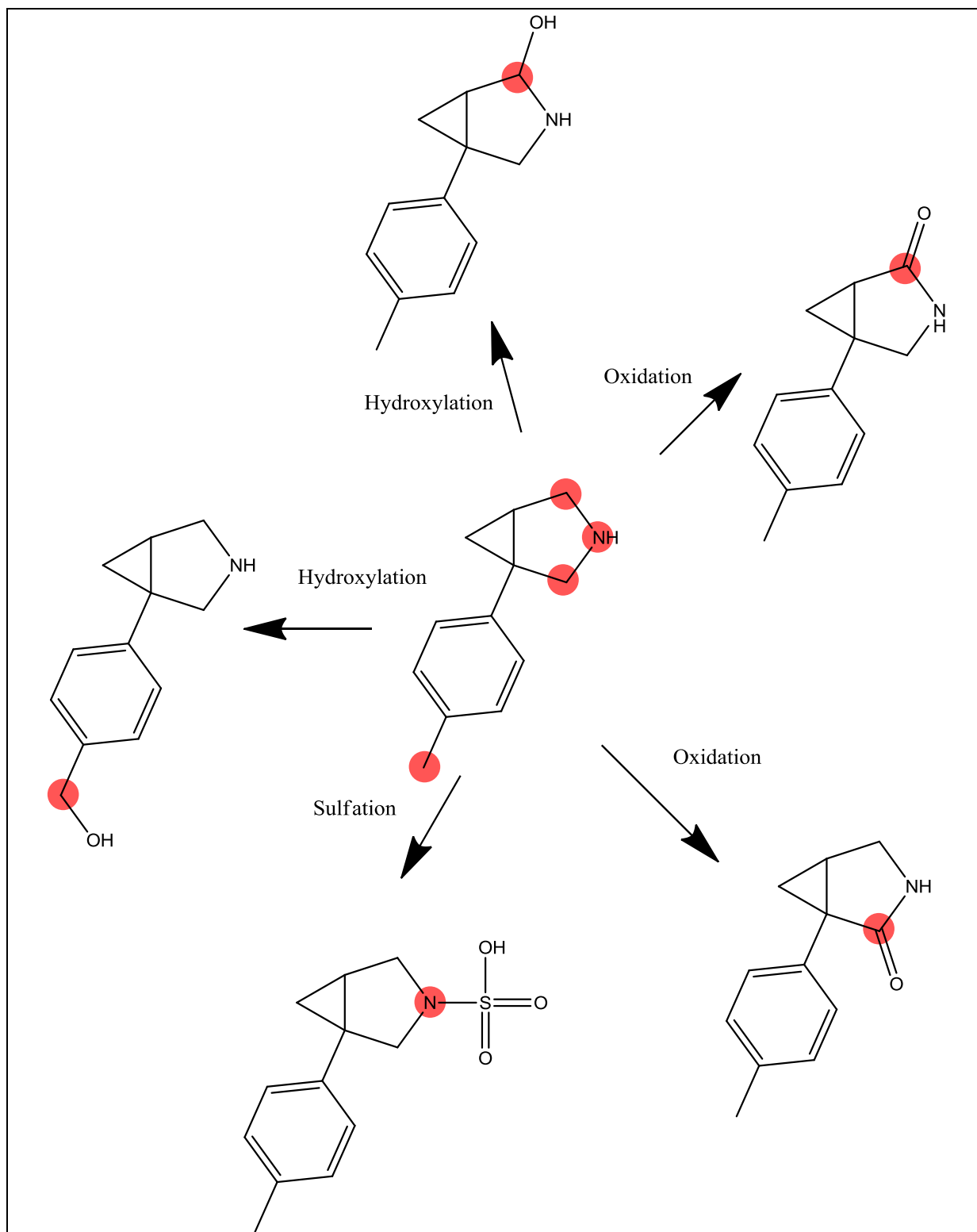
6.2.2 MetaPrint2D-React predicted transformations



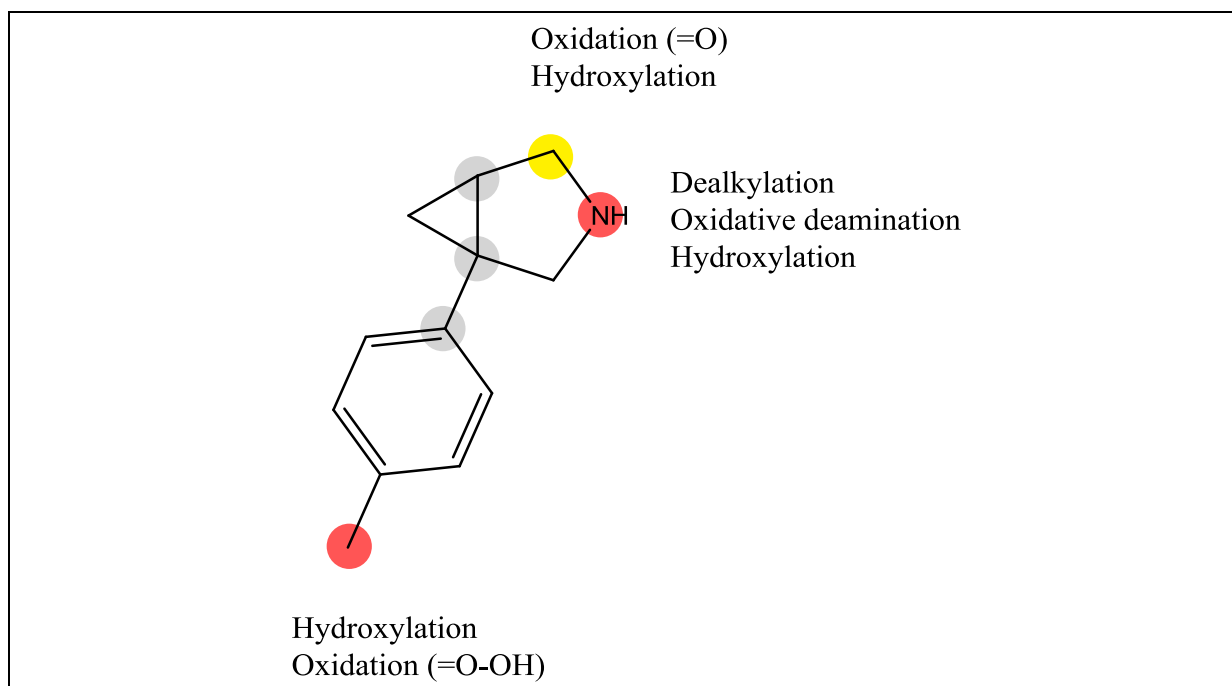
MetaPrint2D-React's model does not contain very much data on several of the atom environments occupied by atoms in this molecule, but correctly predicted the amide hydrolysis, hydroxylation and oxidation reactions. Hydrolysis of the cyclic amide and alkylation were also predicted, though not reported. Given that 5% of the metabolites identified in the study were not characterised, it is possible that these transformations could be occurring. All four sites of metabolism were identified.

6.3 Bicifadine (286)

6.3.1 Reported metabolites



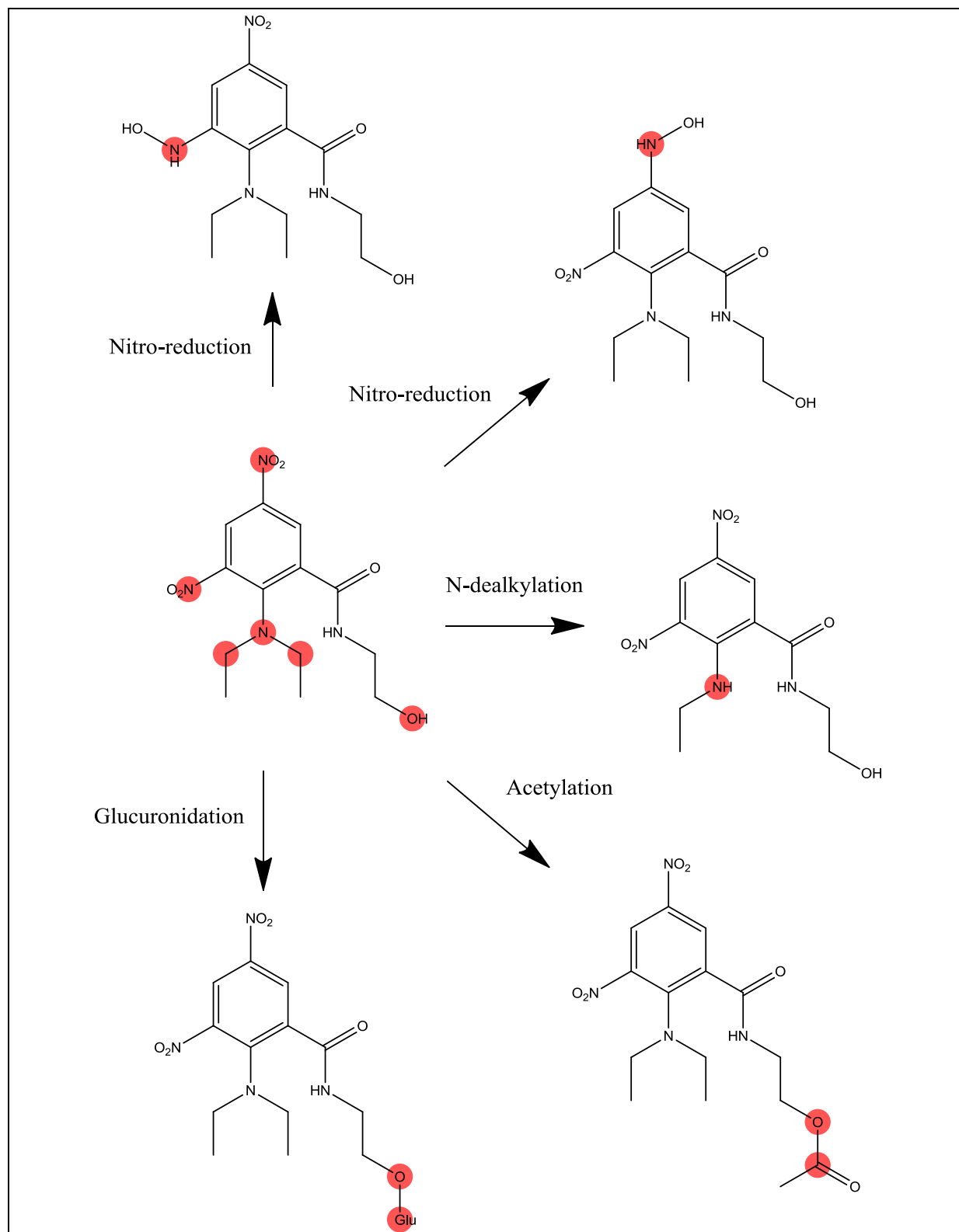
6.3.2 MetaPrint2D-React predicted transformations



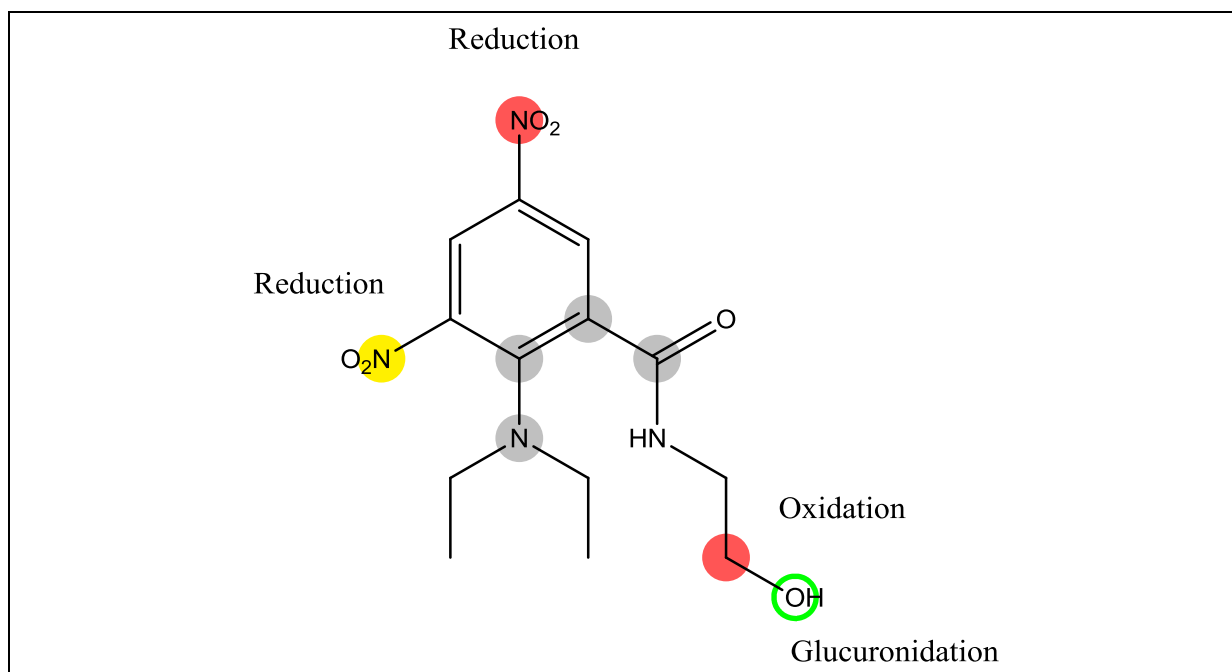
MetaPrint2D-React correctly predicted hydroxylation of the 5-membered ring, and one of the sites of oxidation. Hydroxylation of the methyl group was also correctly predicted. Sulfation of the nitrogen atom was not predicted, though hydroxylation, which is likely to be the first step of this process, was predicted. Three of the four reported sites of metabolism were identified.

6.4 N-(2-Hydroxyethyl)-3,5-dinitrobenzamide 2-mustard prodrug (287)

6.4.1 Reported metabolites



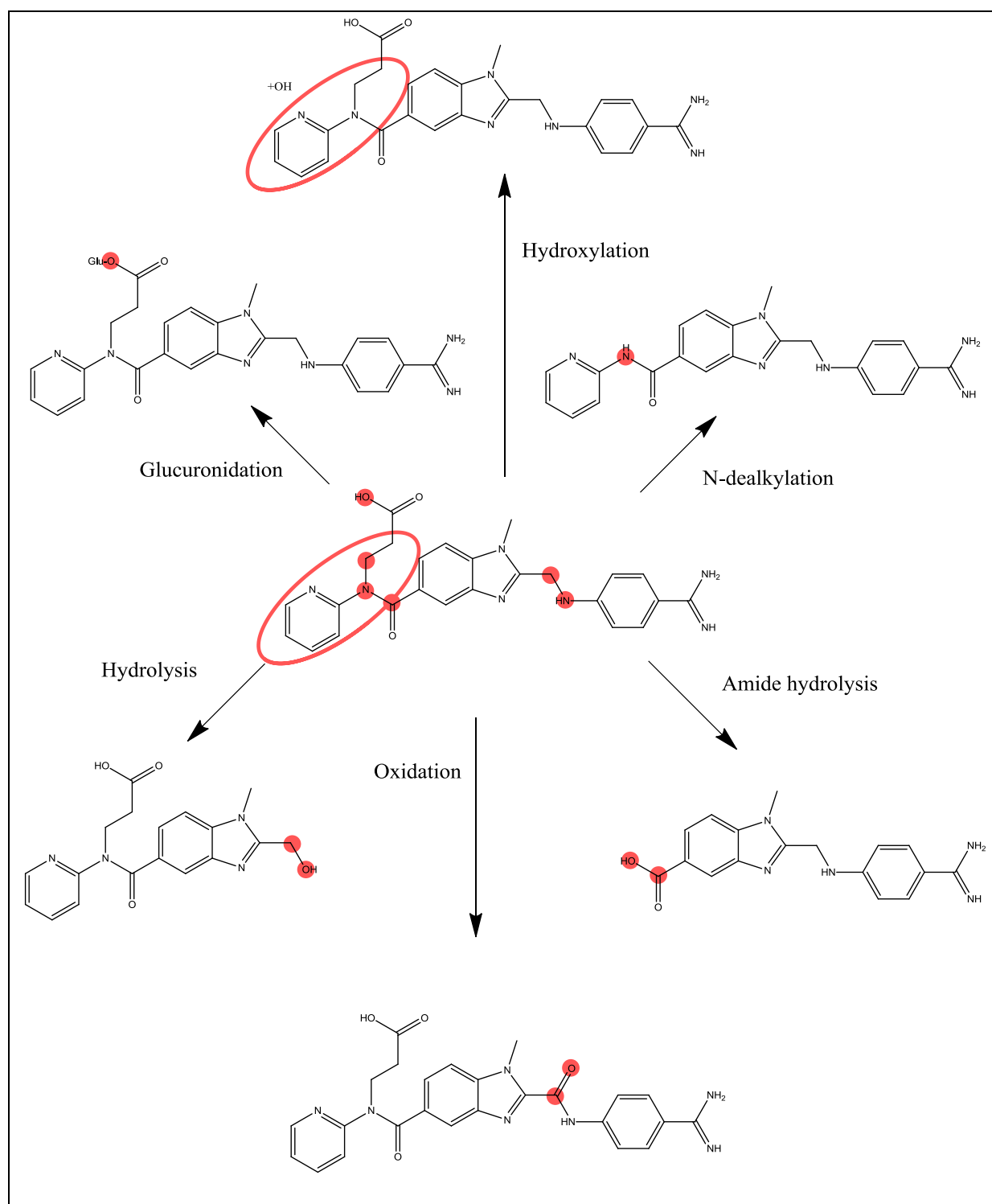
6.4.2 MetaPrint2D-React predicted transformations



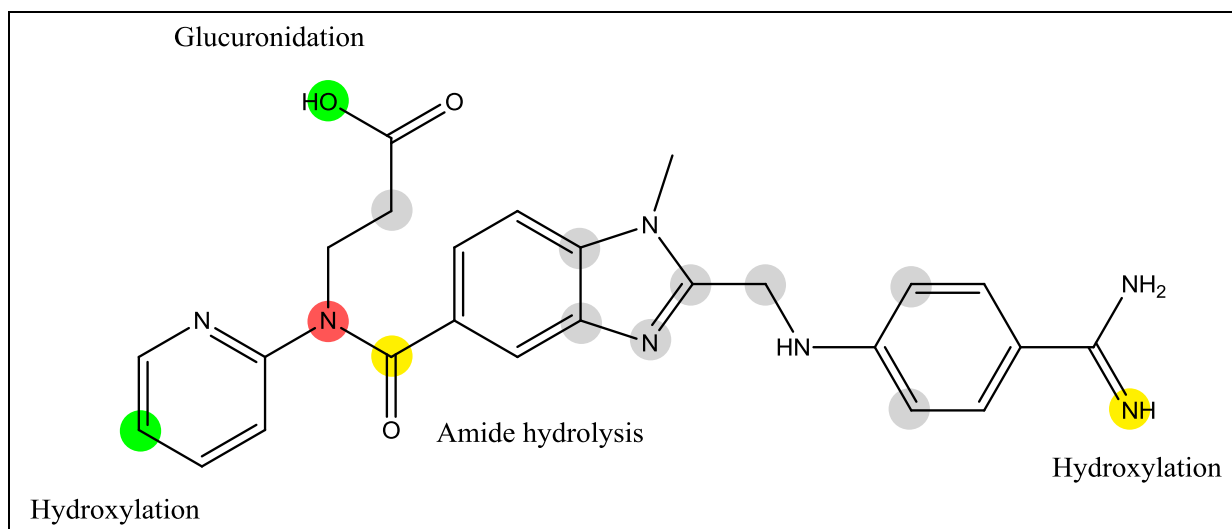
MetaPrint2D-React correctly predicted reduction of the nitro groups, and the glucuronidation reaction was also identified, but with a low likelihood ratio. The reported acetylation was not predicted, and neither was the N-dealkylation. In the latter case this was due to the nitrogen atom occupying a novel atom environment. In all, only three of the reported sites of metabolism were identified.

6.5 Dabigatran (288)

6.5.1 Reported metabolites



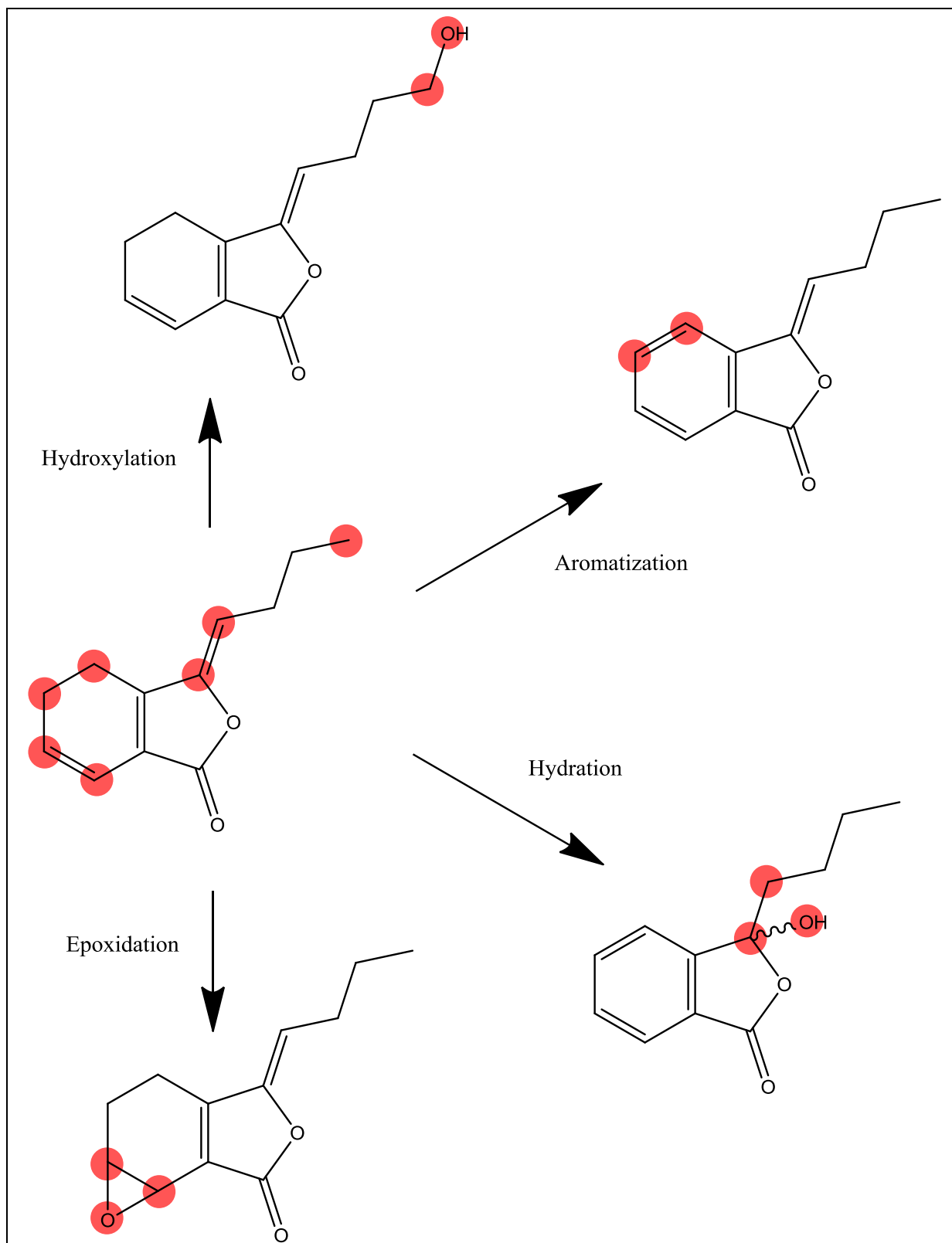
6.5.2 MetaPrint2D-React predicted transformations



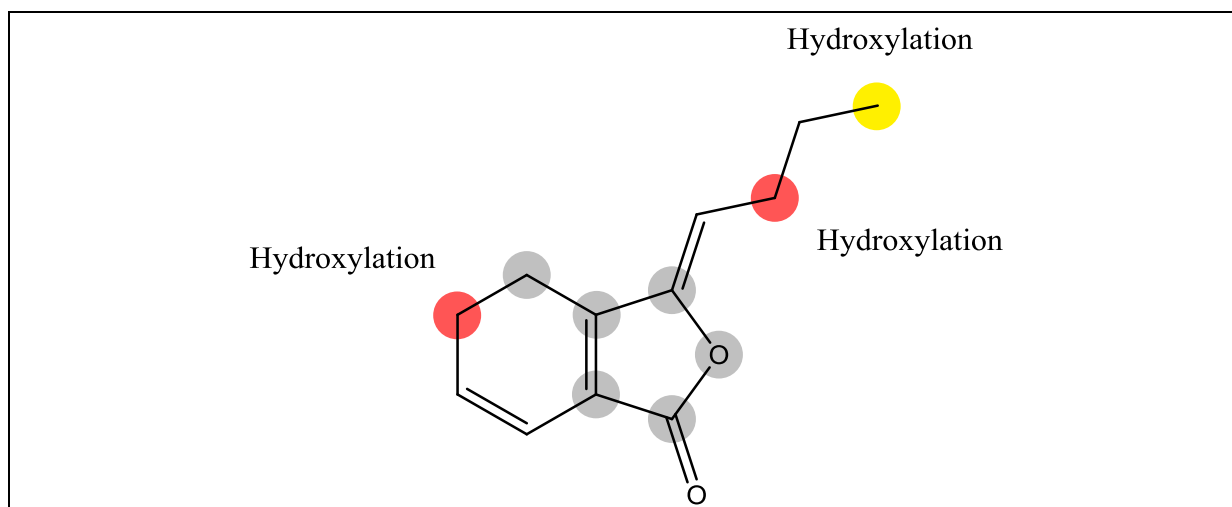
MetaPrint2D-React correctly predicted the amide hydrolysis and glucuronidation, and gave a likely location for the reported hydroxylation of the pyridine ring. The model contained little information on a number of atom environments found in the molecule, and possibly as a result of this failed to predict the amine hydrolysis or oxidation. The model also failed to predict the N-dealkylation, and suggested an additional hydroxylation that has not been reported to be observed. Overall, half of the reported sites of metabolism were identified.

6.6 Ligustilide (289)

6.6.1 Reported metabolites



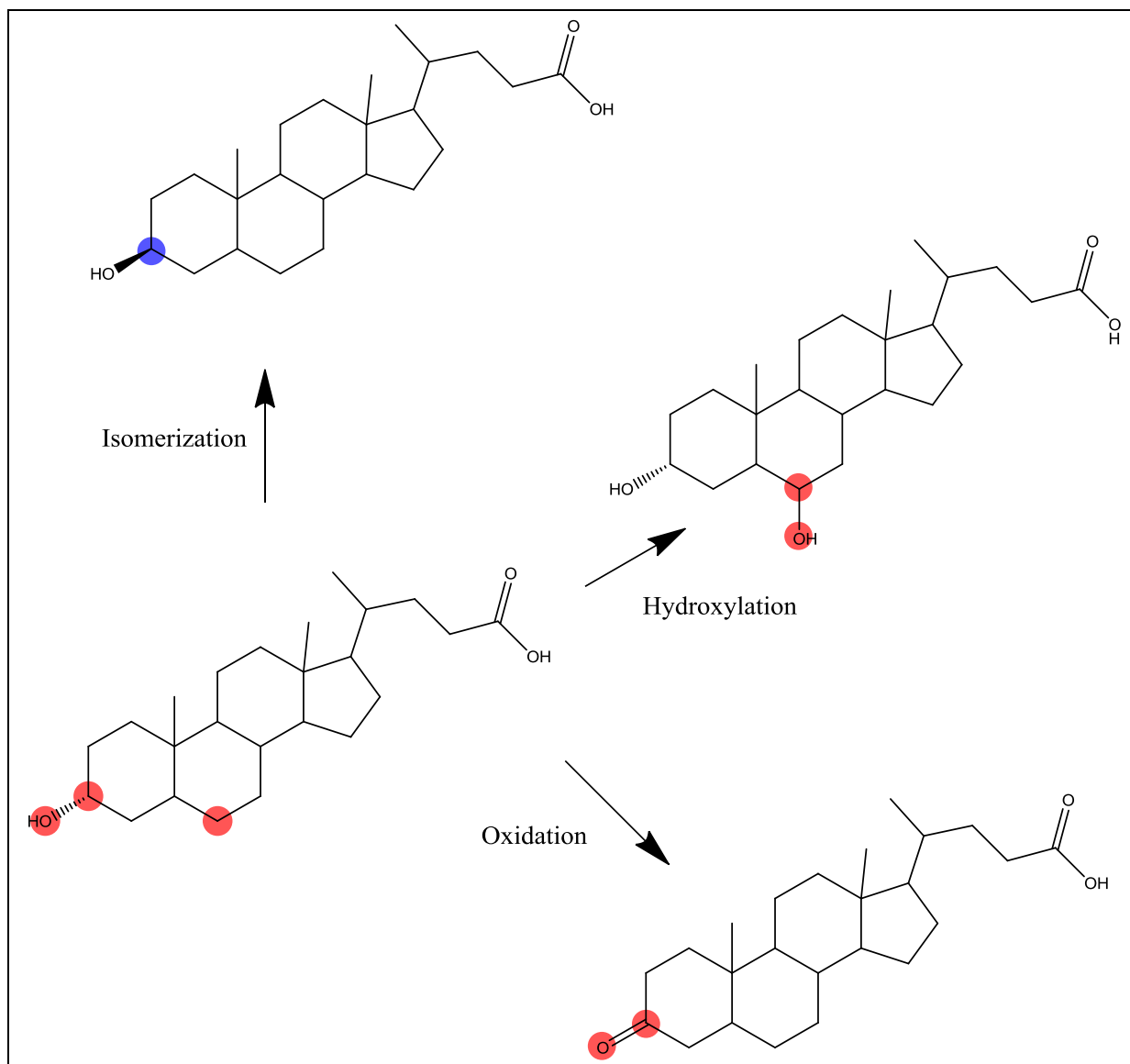
6.6.2 MetaPrint2D-React predicted transformations



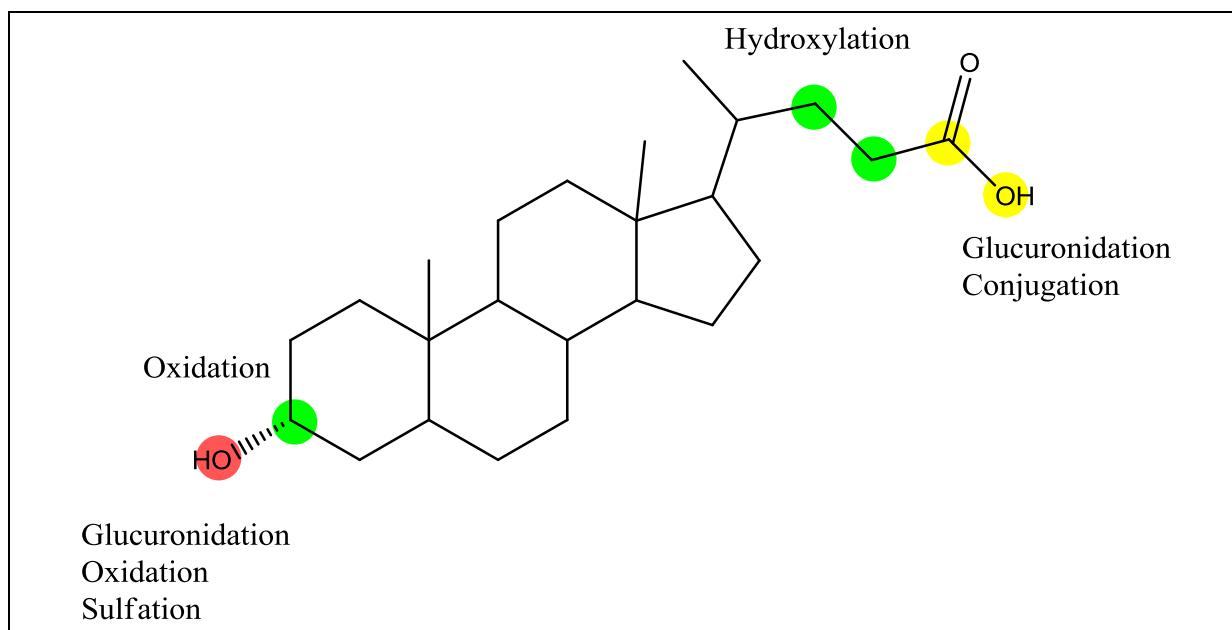
MetaPrint2D-React performed poorly on this molecule, only one of the predicted hydroxylation reactions was correctly located, and all the other reported transformations missed. This is not surprising given that almost half of the atoms in the structure are occupying novel atom environments, so lie outside of the model's domain of applicability.

6.7 Lithocholic acid (290)

6.7.1 Reported metabolites



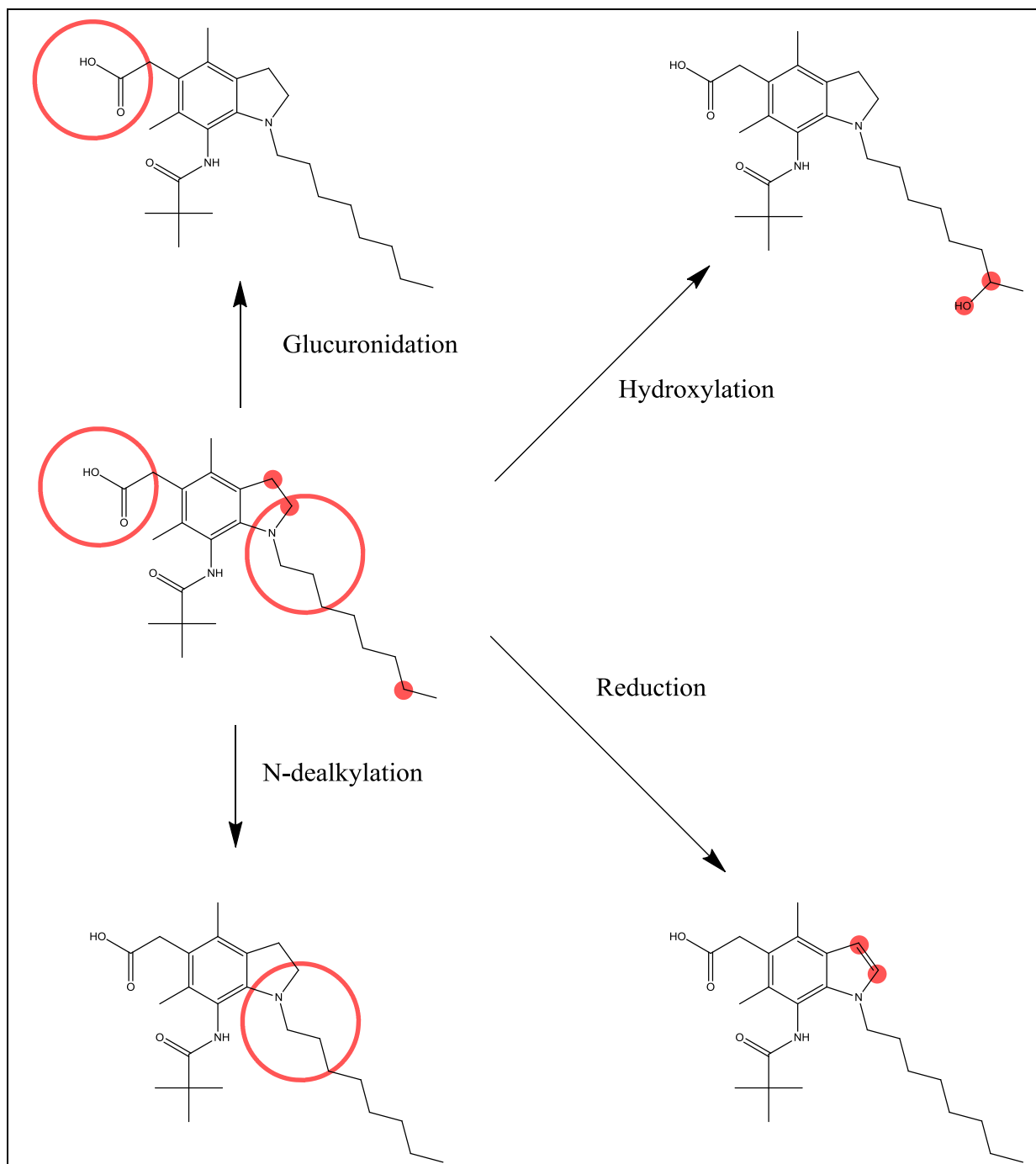
6.7.2 MetaPrint2D-React predicted transformations



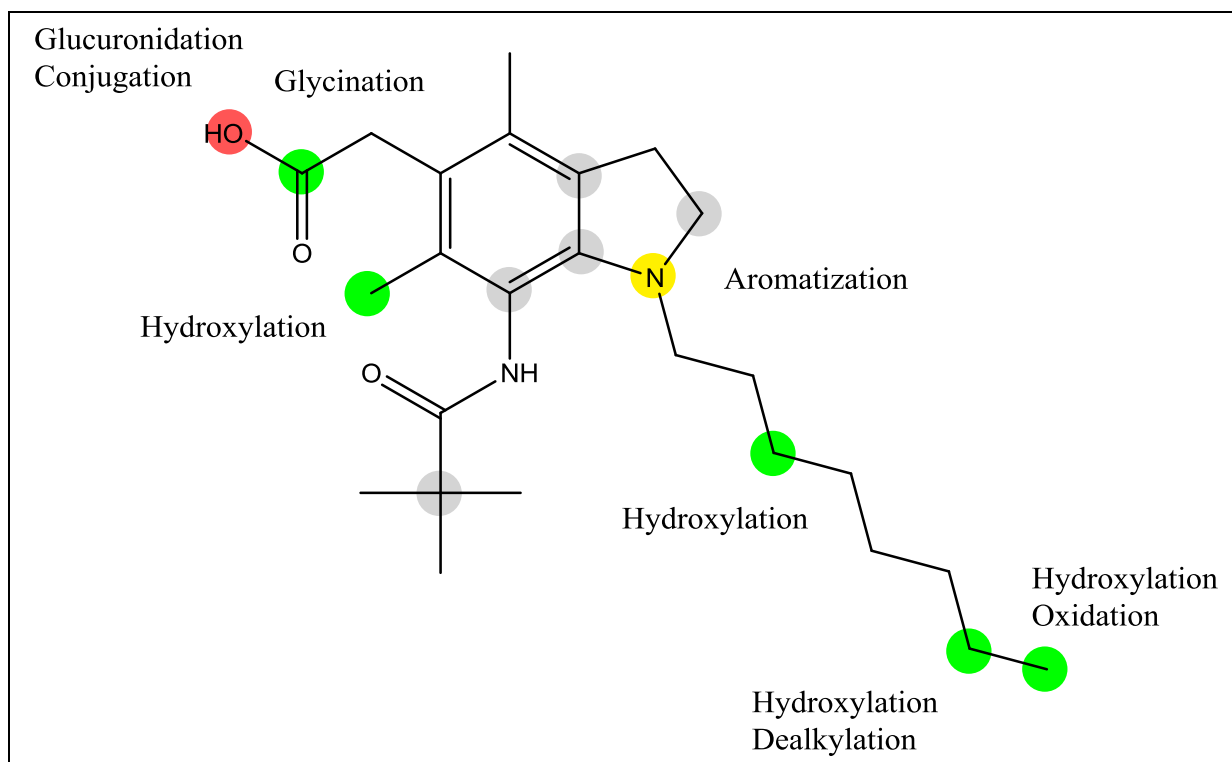
MetaPrint2D-React correctly predicted the oxidation transformation but failed to predict the hydroxylation reaction, and the stereoisomerism is beyond the scope of the model. The model also predicted a number of additional transformations which have not been reported.

6.8 Pactimibe (291)

6.8.1 Reported metabolites



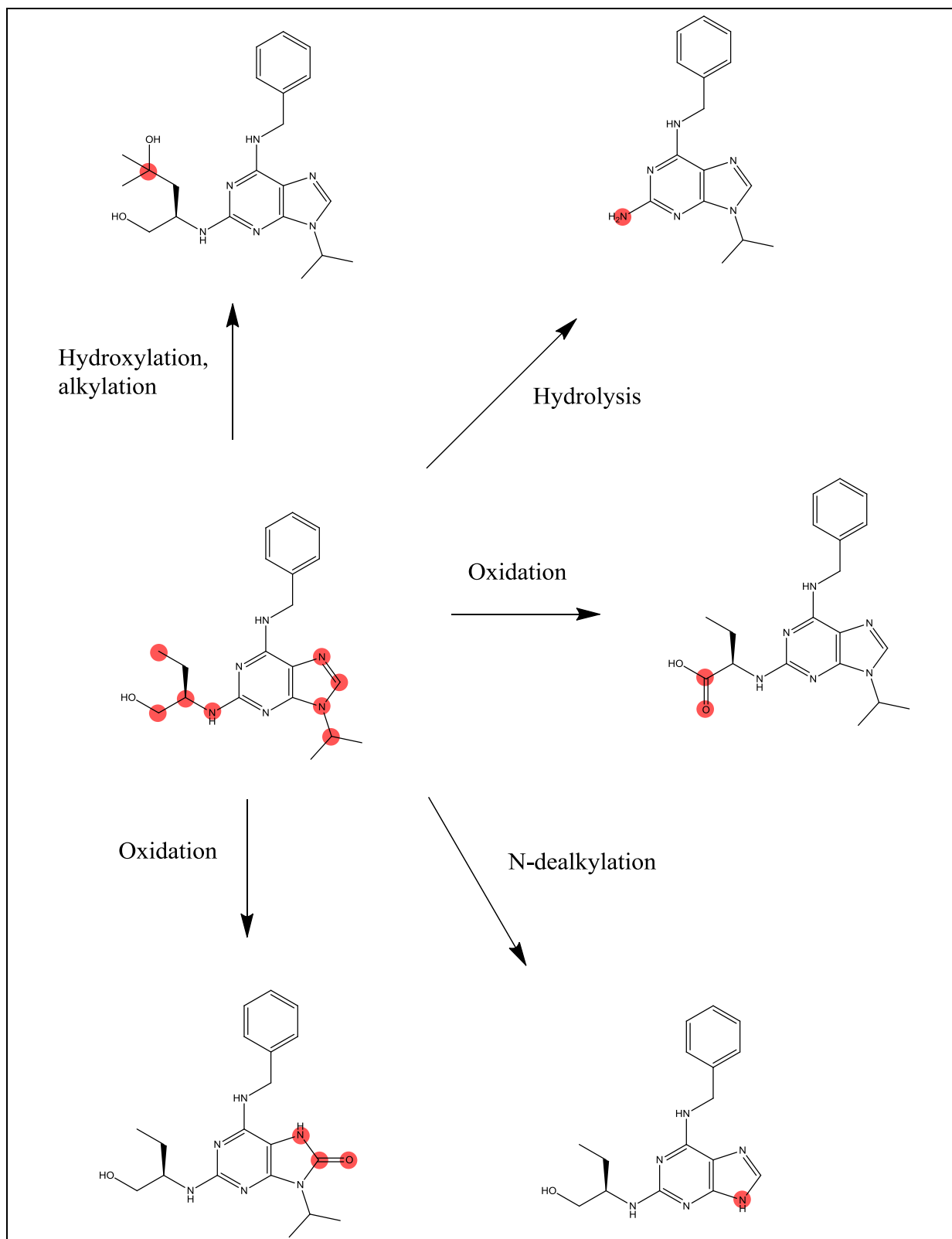
6.8.2 MetaPrint2D-React predicted transformations



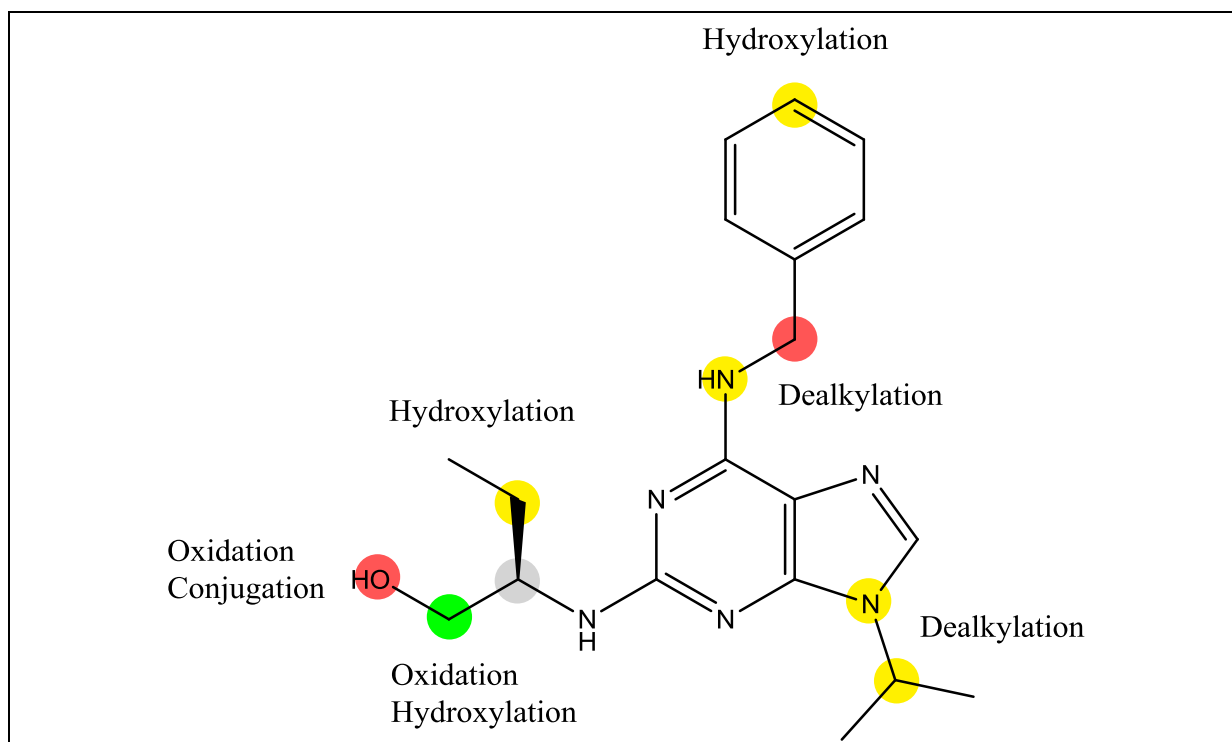
MetaPrint2D-React correctly predicted the glucuronidation reaction, and suggested that other types of conjugation could occur. The hydroxylation was also correctly predicted, though a number of additional potential sites of hydroxylation that have not been reported were suggested. Neither the reduction or dealkylation reactions were predicted; in both cases there are atoms occupying novel environments in the vicinity of the metabolic sites.

6.9 Seliciclib (292)

6.9.1 Reported metabolites



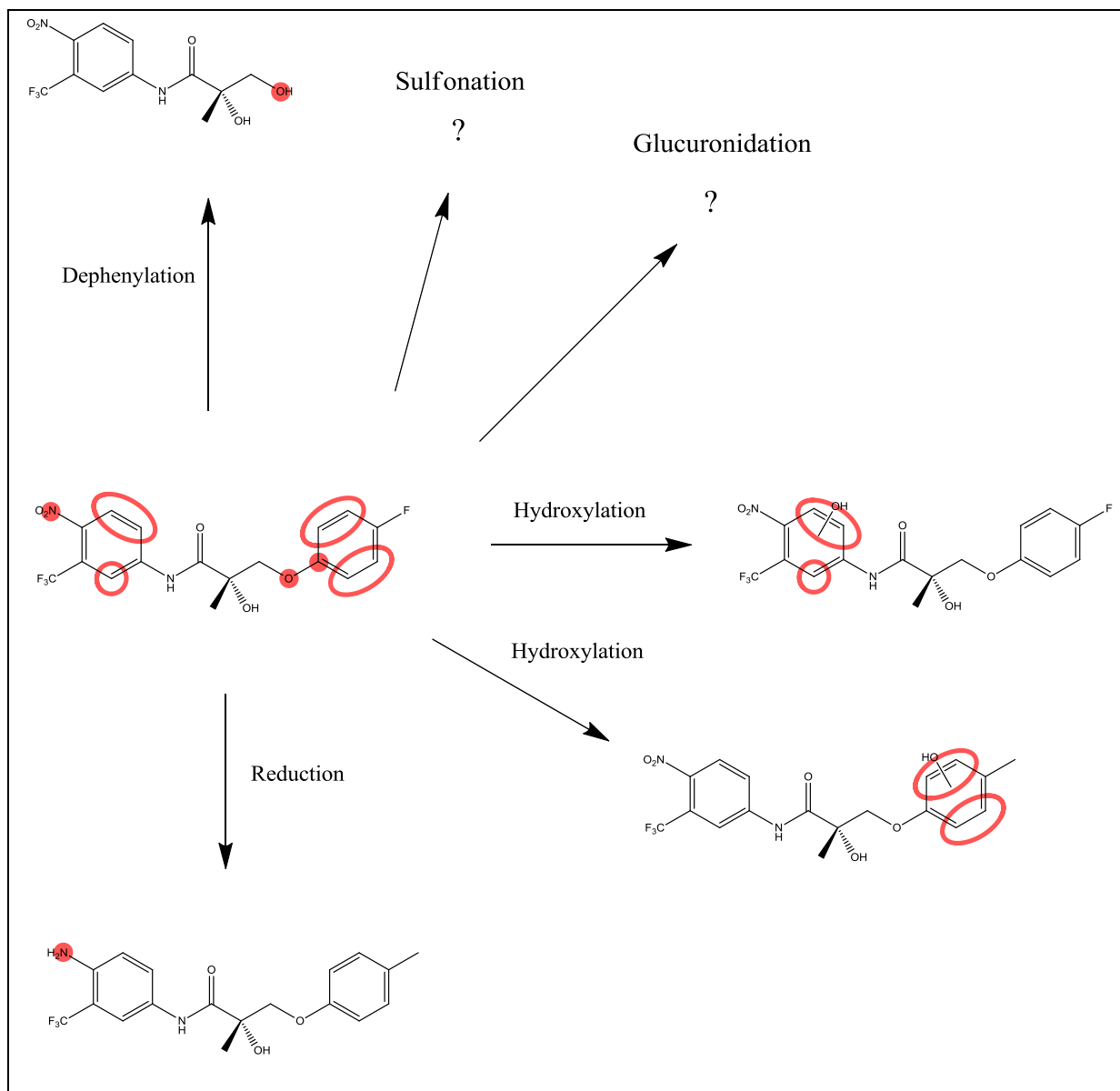
6.9.2 MetaPrint2D-React predicted transformations



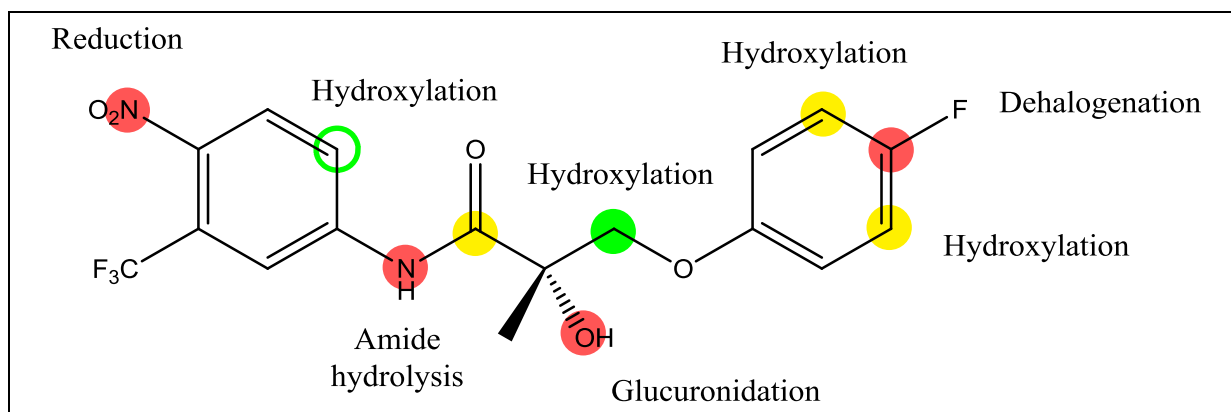
MetaPrint2D-React predicted half of the reported metabolic transformations of this compound. The hydroxylation, dealkylation and one of the sites of oxidation were identified; however the alkylation, hydrolysis and second site of oxidation were not. Several transformations that have not been reported were also suggested.

6.10 Aryl-propionamide derived selective androgen receptor modulator (293)

6.10.1 Reported metabolites



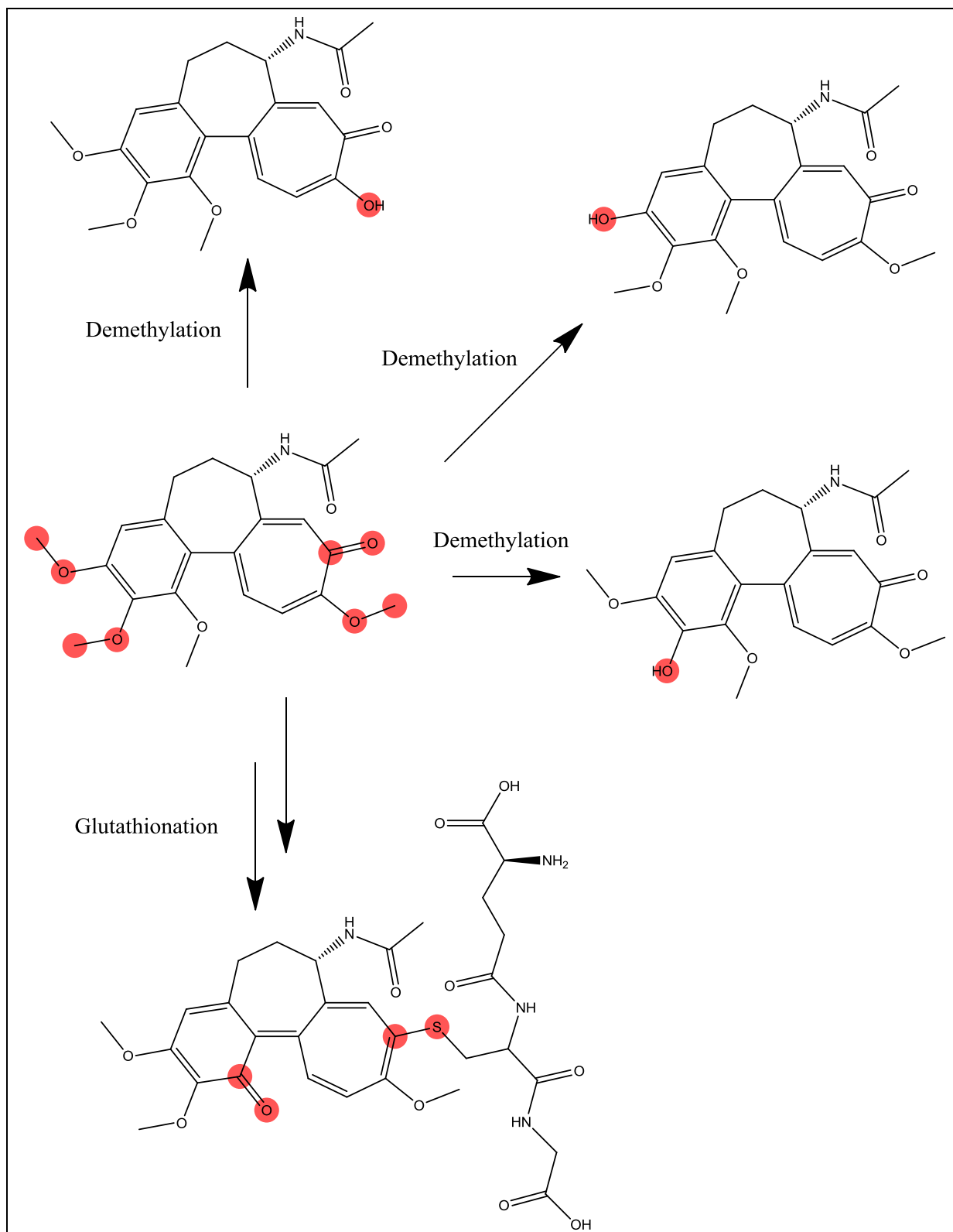
6.10.2 MetaPrint2D-React predicted transformations



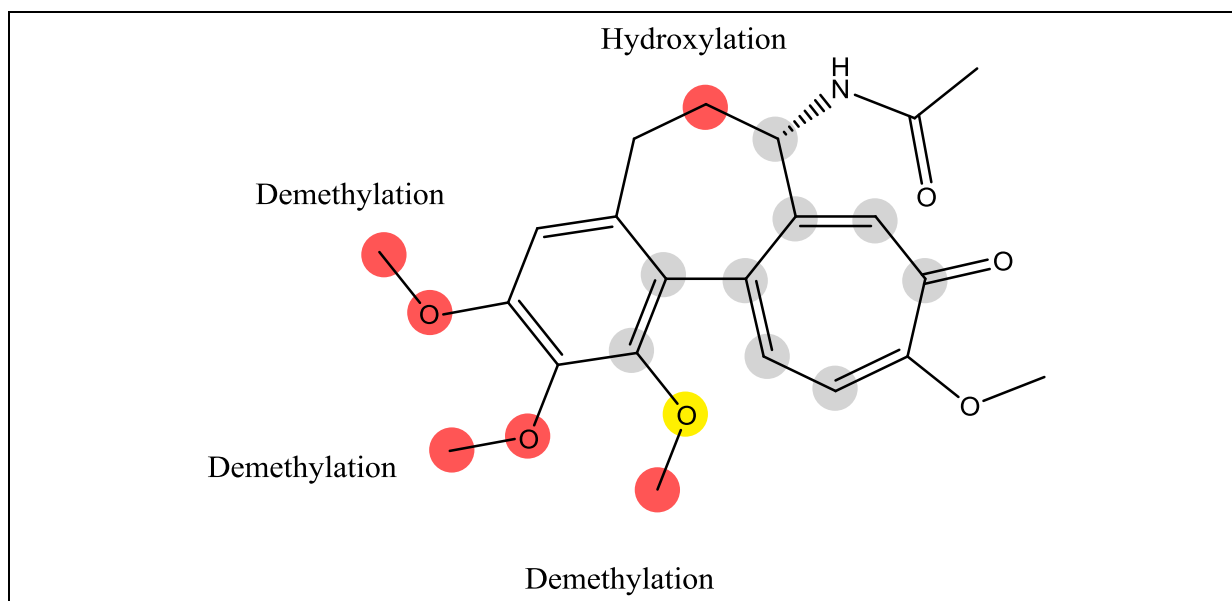
Quite a large number of metabolites have been reported for this compound. MetaPrint2D-React suggested locations for both of the reported aromatic hydroxylation reactions, and the glucuronidation, and predicted reduction of the nitro group. The model failed to predict the dephenylation reaction or the sulfonation, but did predict a number of transformations that have not been reported: dehalogenations, amide hydrolysis and an additional site of hydroxylation. In the case of both the dephenylation and sulfonation transformations it is likely that the first step in these processes would be a hydroxylation reaction and these were predicted at the appropriate sites.

6.11 Colchicine (294)

6.11.1 Reported metabolites



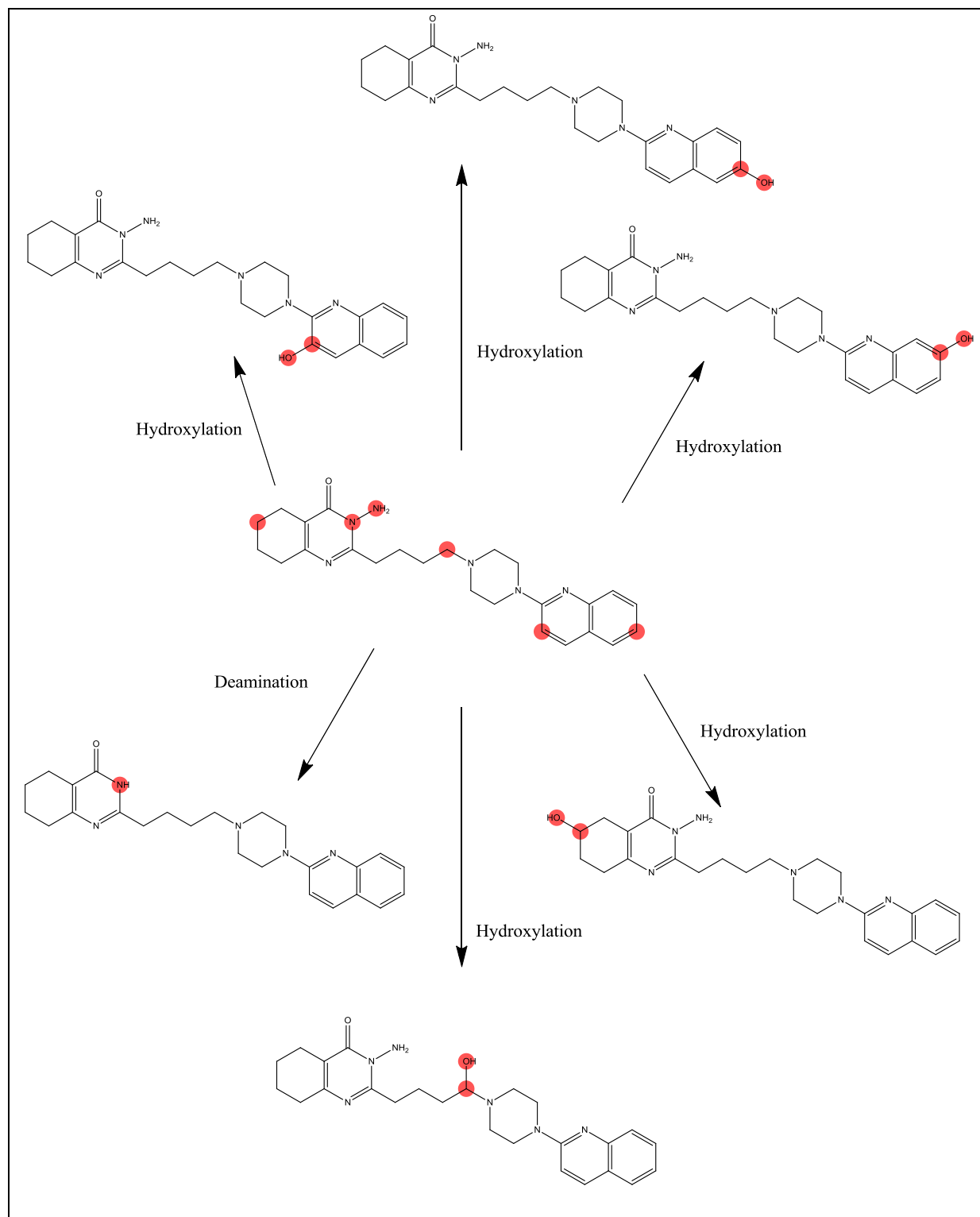
6.11.2 MetaPrint2D-React predicted transformations



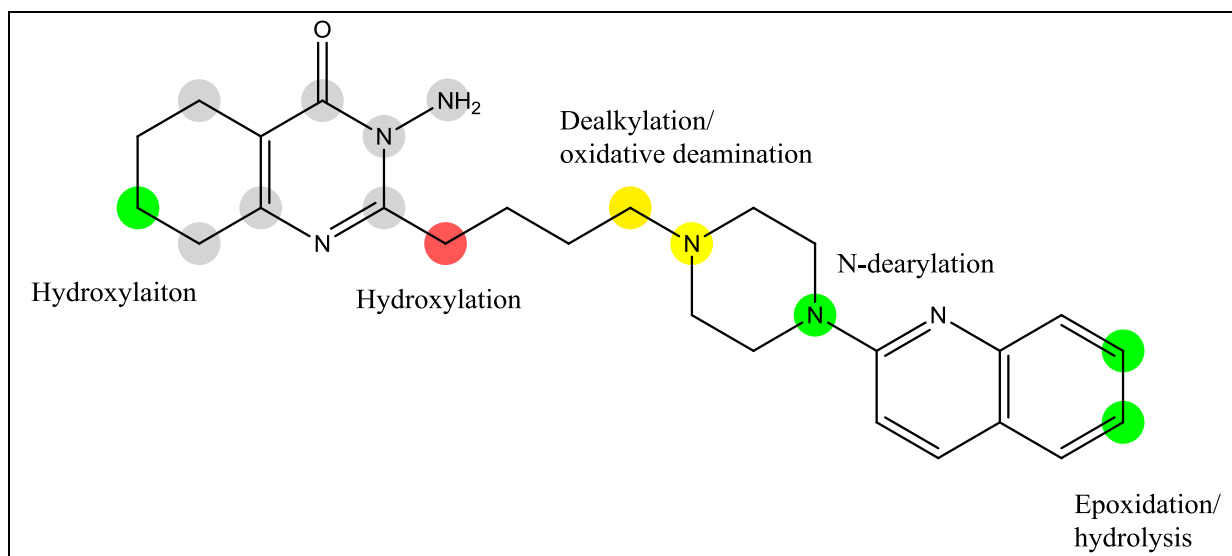
MetaPrint2D-React successfully predicted demethylation of three of the four methoxy groups, but failed to predict formation of the glutathione conjugate, though this was reported to be the result of a multi-step process. No predictions were possible for a large region of this compound due to the number of atoms occupying novel environments.

6.12 3-Amino-5,6,7,8-tetrahydro-2-{4-[4-(quinolin-2-yl)piperazin-1-yl]butyl}quinazolin-4(3H)-one (295)

6.12.1 Reported metabolites



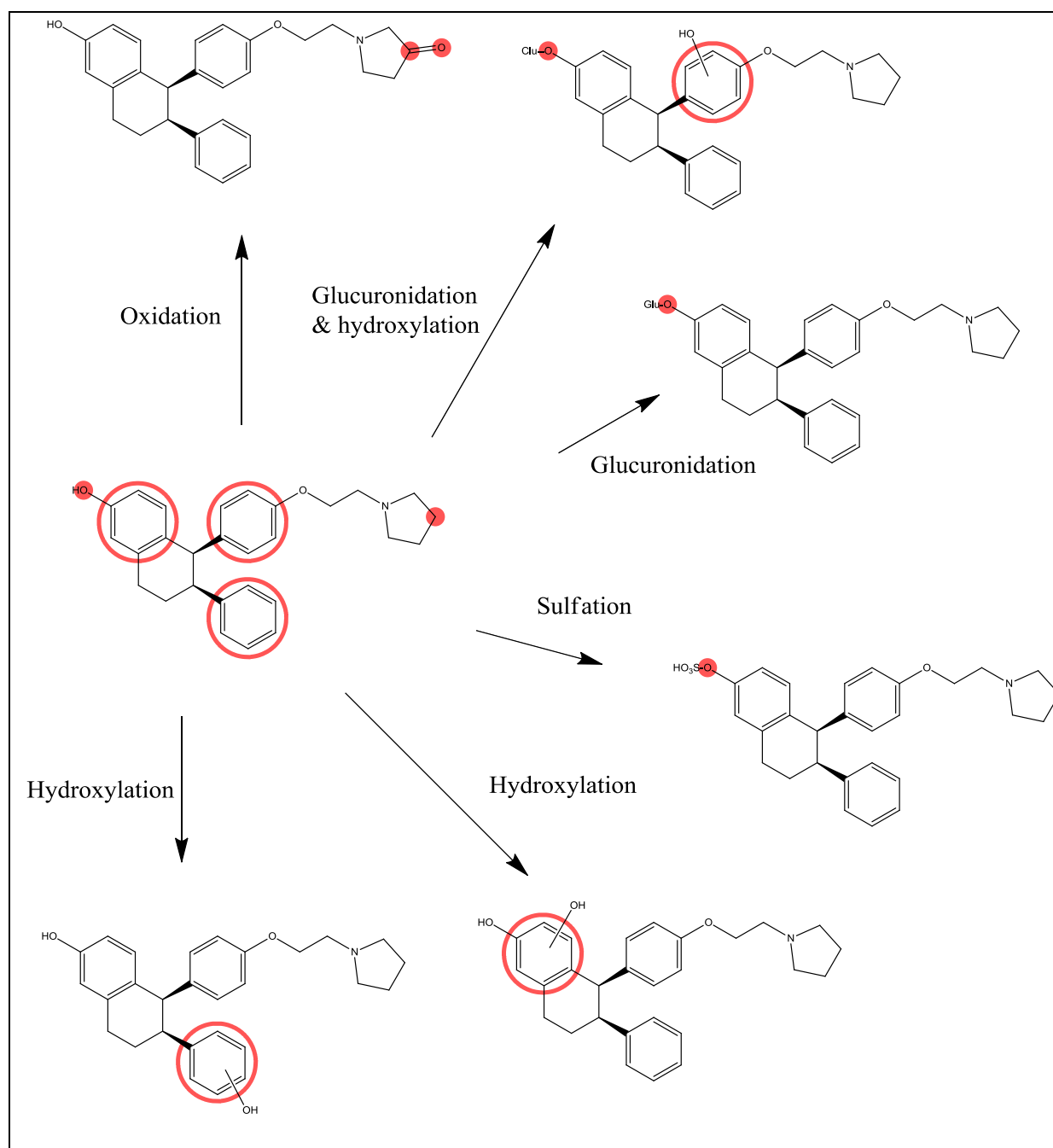
6.12.2 MetaPrint2D-React predicted transformations



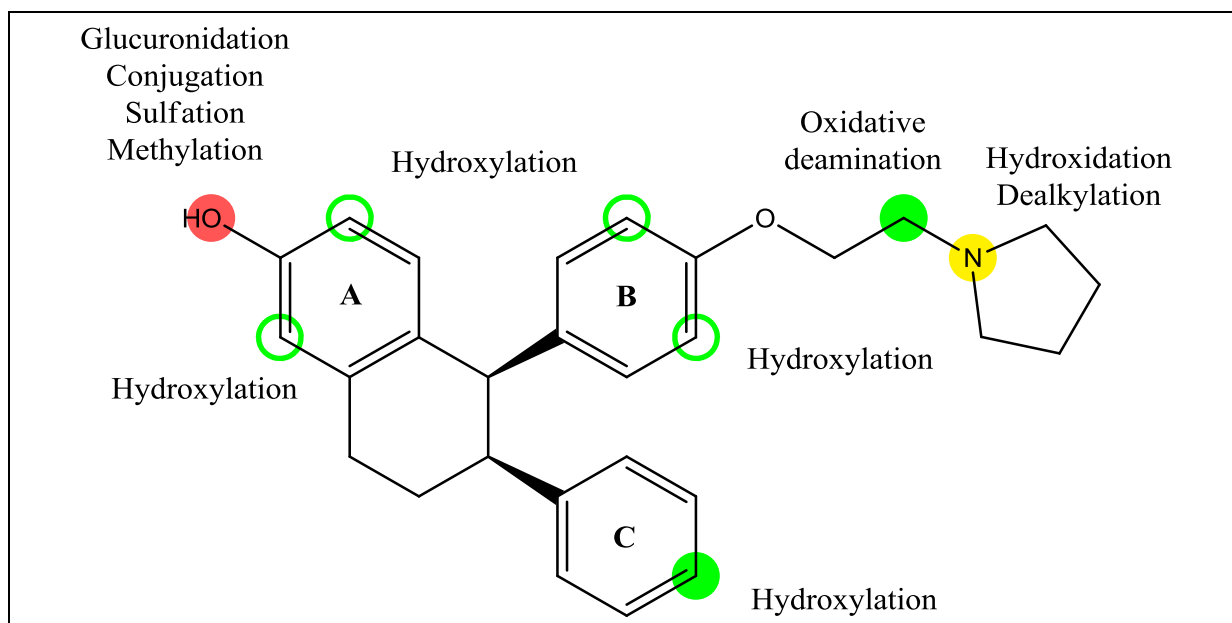
MetaPrint2D-React performed quite poorly on this compound. Of the three reported sites of hydroxylation of the quinoline ring, only two were identified and they were predicted to be generated in tandem, through epoxide formation and hydrolysis, while they are reported to occur independently of each other. The site of hydroxylation in the ring system at the opposite end of the compound was incorrectly predicted, and the deamination reaction was not predicted at all – due to the atoms occupying novel environments. The final reported hydroxylation reaction was not identified either, though oxidative deamination was predicted to occur at that site, the first step of which would likely involve the addition of a hydroxyl group.

6.13 Lasofoxifene (296)

6.13.1 Reported metabolites



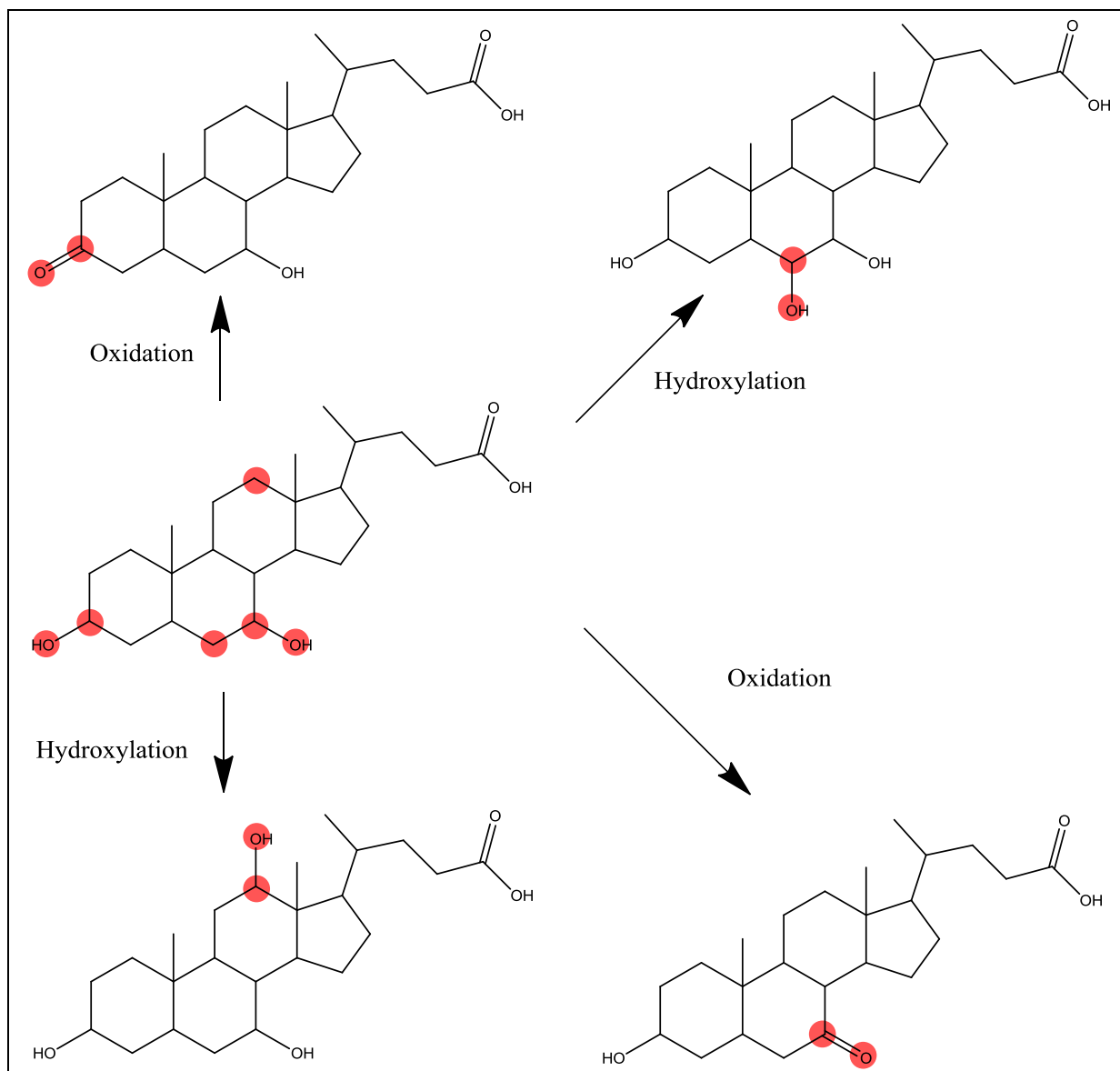
6.13.2 MetaPrint2D-React predicted transformations



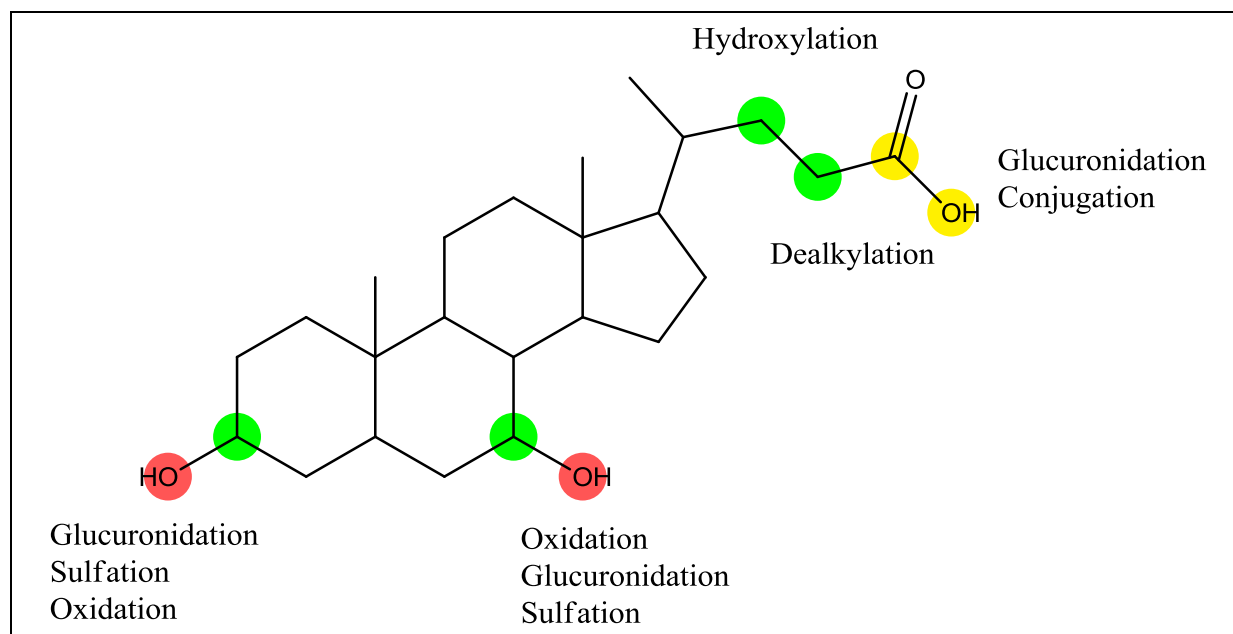
MetaPrint2D-React correctly predicted the reported glucuronidation and sulfation transformations, and suggests locations for the reported hydroxylation reactions. The model failed to predict the reported oxidation reaction, and predicted an N-dealkylation/oxidative deamination that has not been observed.

6.14 Chenodeoxycholic acid (297)

6.14.1 Reported metabolites



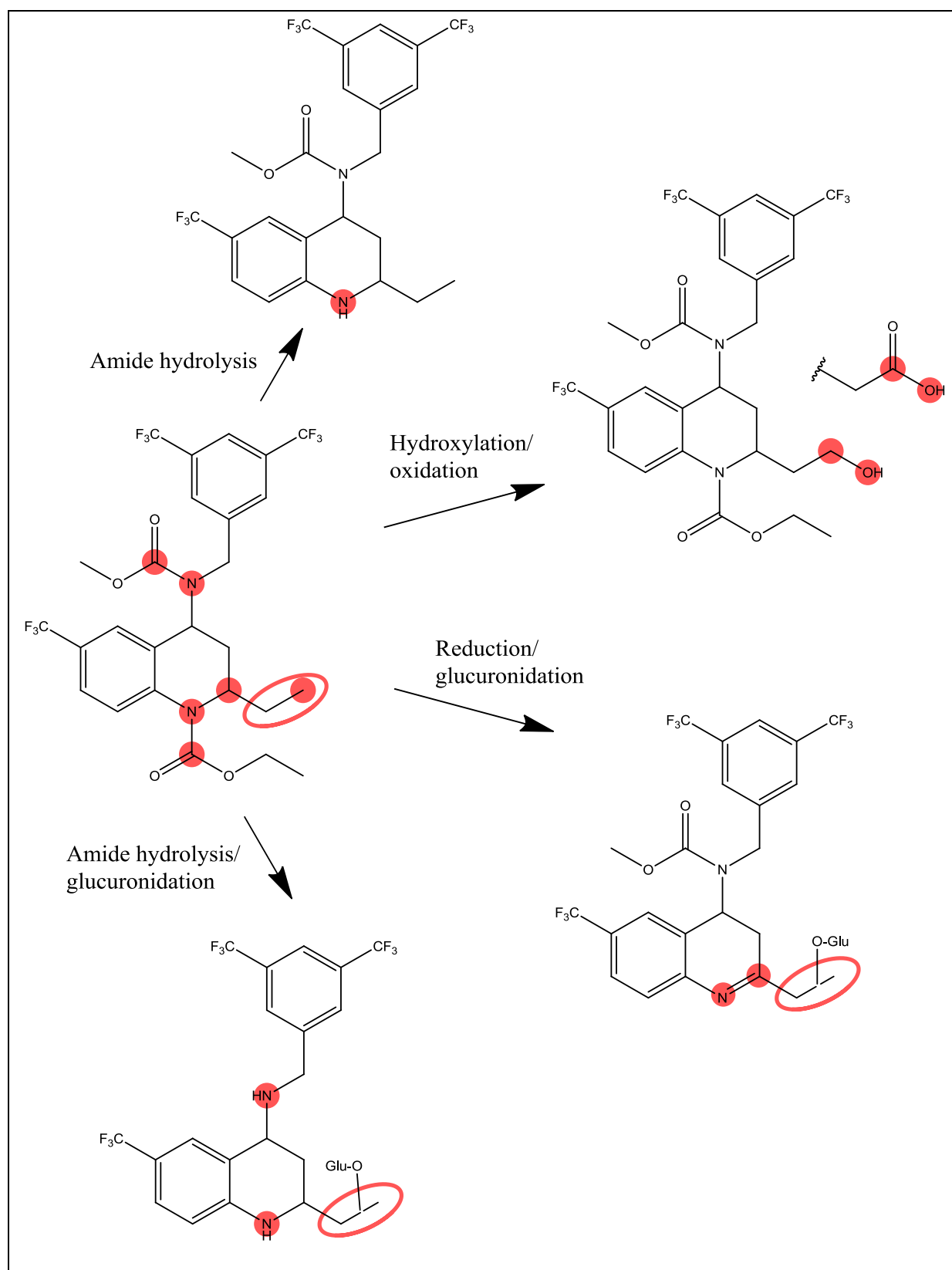
6.14.2 MetaPrint2D-React predicted transformations



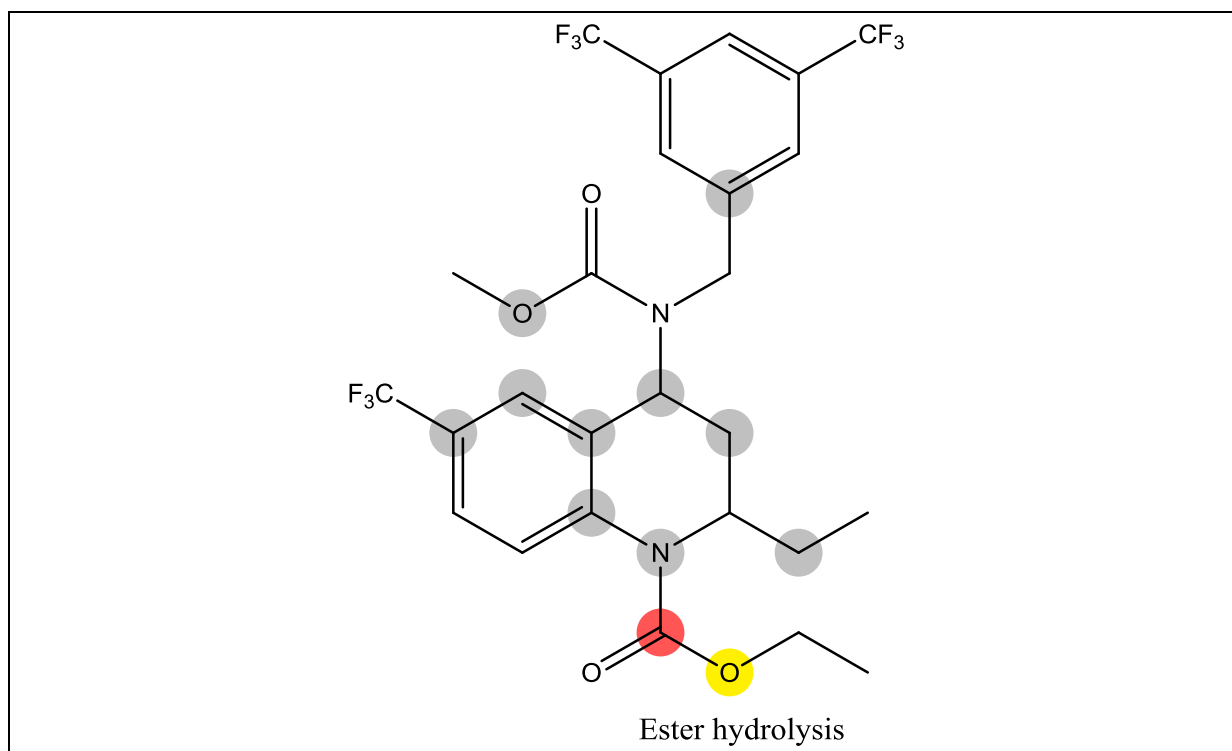
MetaPrint2D-React correctly predicted both of the hydroxyl oxidation reactions, and also suggests that glucuronidation and sulfation reactions could occur at these positions. The hydroxylation reactions were not identified; MetaPrint2D-React predicted hydroxylation at all of the vacant positions in the steroid ring system with a very low likelihood ratio, and made very little differentiation between the positions. MetaPrint2D-React also predicted that glucuronidation and other types of conjugation are likely to occur in the chain region of the molecule. The study focused on the contribution of cytochromes P450, so, like the glucuronidation and sulfation reactions mentioned earlier, these metabolites would not have been identified.

6.15 Torcetrapib (298)

6.15.1 Reported metabolites



6.15.2 MetaPrint2D-React predicted transformations



This was another compound containing a considerable number of atoms occupying environments novel to MetaPrint2D-React's model, and this has impacted heavily on the quality of the metabolite predictions, with none of the reported transformations correctly identified.

6.16 Conclusions

Overall, the quality of MetaPrint2D-React's predictions on these compounds was a little disappointing. In some cases both sites of metabolism and the types of transformations occurring at those sites were well predicted, but for many of the compounds a large number of reported transformations were missed. In large part this was due to the relatively high proportion of atoms occupying novel atom environments. Since MetaPrint2D-React is based on a data mining method, if there is no relevant or suitable data available then reliable predictions cannot be expected.

In the evaluations of MetaPrint2D and MetaPrint2D-React reported earlier, only around 3.5% of the evaluation compounds' atoms occupied novel atom environments, while for this test set the proportion was over 15%, indicating that these compounds fit less well into the model's domain of applicability. This is reflected in the proportion of transformations that were correctly predicted; 78% of the transformations found in the sample of the Symyx® Metabolite database used to evaluate MetaPrint2D-React were correctly identified, while in the case of this data only 53% of the reported transformations were predicted, although in some instances a reaction that is likely to form the first step of the reported transformation was predicted. In both cases it was checked that none of the test compounds had been used in the training of the model, in order to ensure a fair evaluation.

7. Conclusions and further work

This thesis has reported the development and evaluation of MetaPrint2D and MetaPrint2D-React. MetaPrint2D is the result of a re-development of the Substrate Product Occurrence Ratio Calculator (SPORCalc) statistical knowledge-based method for site of metabolism prediction and MetaPrint2D-React is an extension of this approach enabling the prediction of the types of reaction likely to occur at each site and the metabolites formed.

MetaPrint2D and MetaPrint2D-React have been released as freely available open source software. MetaPrint2D is accessible through a number of different user interfaces – a web site, a command line application and through the Bioclipse rich client application. The variety of interfaces to MetaPrint2D has been made possible through the abstraction of MetaPrint2D's 'calculation engine' into a library, separate from the user interface. This library provides an application programming interface (API) that other applications can use in order to integrate MetaPrint2D. MetaPrint2D-React is also available as a library, but there is currently only a single user interface available – a web site.

Extensive evaluations of MetaPrint2D and MetaPrint2D-React have been performed, in the course of which a new metric for assessing the performance of site of metabolism predictions has been proposed. This receiver operating characteristic (ROC) curve based procedure overcomes the various biases inherent to the most commonly used metrics for site of metabolism predictions – the percentage of molecules for which a true site of metabolism is found within the top one or top three predicted sites.

MetaPrint2D's predictions have been shown to be comparable in accuracy to those of other recent site of metabolism prediction tools, but with the advantage of being very fast to compute. MetaPrint2D can generate site of metabolism predictions for drug-like molecules in just tens of milliseconds, making it possible for the first time for a chemist to explore the effects of structural modifications on a compound's metabolism in a highly responsive interactive manner.

Having its basis in a data mining method, MetaPrint2D can only generate reliable prediction on compounds for which relevant data was included in the model's construction. This has

been illustrated both during the evaluation of MetaPrint2D, and during the prediction of recently reported metabolites described in the previous chapter. The accuracy of MetaPrint2D depends on how well a query compound fits the model, and it has been shown that this can be estimated from the proportion of atoms in a molecule occupying novel atom environments – environments that occur only rarely in, or are completely absent from, the training data used to develop the model. In cases where a compound does not occupy a region of chemical space characterised by MetaPrint2D an alternative method of prediction would need to be used.

A major factor affecting the performance of MetaPrint2D and MetaPrint2D-React is the quality of the data from the Symyx® Metabolite database used to train the models. As was discussed in Chapter 5, there are two problems with this data: the inconsistency with which transformations are reported, and that multiple products of a reaction are recorded in separate transformation records.

The Symyx® Metabolite database collates observed metabolic transformations as reported in the literature. Little normalization of the data appears to take place in the preparation of the database, so the manner in which a transformation is reported can vary depending on the source publication. Some metabolic schemes report only those metabolites that were characterised experimentally, but others report ‘putative metabolites’ – intermediates postulated to have been formed during the course of reactions between positively characterised compounds. The experimental methods used can determine whether intermediate metabolites are observed. This means that in some cases multiple reactions, either connected or occurring in different regions of the molecule, are reported as a single step process, while in other cases the same overall transformation is reported as a series of separate reactions.

It may be possible to improve the quality of MetaPrint2D’s models through the application of normalization procedures, pre-processing the training data. Ideally such a process should be able to identify common inconsistencies and generate a standardized version of the transformations, possibly generating missing intermediates and separating reactions occurring in independent regions of a compound.

A further limitation of both MetaPrint2D and MetaPrint2D-React is that neither makes any estimate of the likelihood of a compound undergoing metabolic transformation, predicting only the relative likelihood of transformations centred on each site in the compound if the compound is metabolised. This makes it difficult to use MetaPrint2D-React to construct trees of potential metabolites. There are a number of ways in which this could be resolved, such as through the use of simple rules, such as '*no further reactions occur after a phase II transformation*', or integration with a logP calculator together with a rule to terminate the tree once a certain hydrophilicity has been reached. Alternatively it may be possible to construct a QSAR model to predict whether a compound will undergo further metabolism by comparing the parent and intermediate compounds in metabolic schemes (from the Symyx® Metabolite database) to those at the end of metabolic scheme.

An interesting extension of MetaPrint2D-React would be the application of the statistical data mining methods used in this work to reverse metabolism prediction. In biotransformation research it is often necessary to determine the parent compound of a metabolite that has been identified. There are currently no tools designed to make these types of prediction. The only *in silico* options currently available are to predict the metabolites of all possible parent compounds and look to see whether the metabolite appears among the predictions, or to search for similar metabolites in collections of known transformations, such as the Symyx® Metabolite database. It should be possible to adapt the reaction analysis and data mining tools in MetaPrint2D-React to consider reverse transformations – from metabolite to substrate, and in this way make predictions from metabolite structures such as whether a hydroxyl group is the result of a hydroxylation, ester hydrolysis, epoxide opening or hydration reaction.

All of the work reported in this thesis was performed using data from the Symyx® Metabolite database – both for the production of models, and their evaluation. There are, however, a number of other sources of data that could be investigated. Many organisations such as pharmaceutical companies have large collections of proprietary data – the results of unpublished experiments carried out within the organisation. MetaPrint2D models could be constructed using this data alone, or by combining it with data from the Metabolite database. Incorporating an organisation's bespoke data into the model building process

should lead to the generation of models that are more relevant to the regions of chemical space on which the organisation is focussing its attention.

These tools could also be used to model other sources of transformations, in a similar manner to the work on predicting microbial catabolism using the University of Minnesota Biocatalysis/Biodegradation Database (UM-BBD) (246,247), discussed in chapter 2.

It may also be possible to adapt the reaction analysis tools of MetaPrint2D-React to other uses. For example, it could be possible to produce a tool for predicting potential side-reactions during organic syntheses through a similar type of reaction analysis and modelling to that described here, performed on reaction schemes reported in synthetic chemistry journals and theses.

As already discussed, during the course of this work it has been shown that the reliability of predictions is dependent on the amount of data on which the prediction is based. Currently users can access this information in the form of the raw values behind the calculated occurrence ratio, and SPORCalc's 'traffic-light' visualization has been extended to highlight regions of the molecule occupying novel environments – having fingerprints with little or no related data on which to base a prediction.

More information could be presented to the user. Care must be taken not to make the display of results too complicated, but it may be possible to provide some indication of the quantity of data behind each atom's occurrence ratio, and hence the confidence in predictions at that site, through varying the shade or the size of the coloured highlights of the atoms. Alternatively, finer discrimination between predicted sites of metabolism could be indicated through use of varying shades of highlighting.

In conclusion, two new tools for making predictions of xenobiotic metabolism have been developed and made freely available. Xenobiotic metabolism is of great importance to the safety and efficacy both of pharmaceutical compounds and within the wider chemical industry. It is hoped that MetaPrint2D and MetaPrint2D-React will help to make it easier to identify and address potential metabolic liabilities.

8. Bibliography

This bibliography is formatted according to the directions of *The ACS style guide: effective communication of scientific information*, 3rd ed.; Coghill, A. M., Garson, L. R., Eds.; American Chemical Society: Washington, DC, 2006.

1. Sneader, W. *Drug Discovery. A History*; John Wiley & Sons Ltd: Chichester, UK, 2005.
2. Forbes. The World's Ten Best-Selling Drugs.
http://www.forbes.com/2006/03/21/pfizer-merck-amgen-cx_mh_pk_0321topdrugs.html (accessed Sept 08, 2009).
3. Drews, J. Drug discovery: a historical perspective. *Science* **2000**, 287 (5460), 1960-1964 (DOI:[10.1126/science.287.5460.1960](https://doi.org/10.1126/science.287.5460.1960)).
4. Johnson, D. The discovery-development interface has become the new interfacial phenomenon. *Drug Discov. Today* **1999**, 4 (12), 535-536 (DOI:[10.1016/S1359-6446\(99\)01423-3](https://doi.org/10.1016/S1359-6446(99)01423-3)).
5. Lipper, R. A. E pluribus product. *Modern Drug Discovery* **1999**, 2, 55-60.
6. Dickson, M.; Gagnon, J. P. Key factors in the rising cost of new drug discovery and development. *Nature Reviews Drug Discovery* **2004**, 3, 417-429 (DOI:[10.1038/nrd1382](https://doi.org/10.1038/nrd1382)).
7. Food and Drug Administration (U.S). *From test tube to patient: Improving health through human drugs*; HHS Publication No. (FDA) 99-3168; FDA: Rockville, MD, 1999.
8. Glover, G. J. *Testimony for the Pharmaceutical Research and Manufacturers of America*; Washington, DC, March 19, 2002.
9. DiMasi, J. Trends in drug development costs, times and risks. *Drug Inf. J.* **1995**, 29, 375-384.
10. DiMasi, J. A.; Hansen, R. W.; Grabowski, H. G. The price of innovation: new estimates of drug development costs. *Journal of Health Economics* **2003**, 22 (2), 151-185 (DOI:[10.1016/S0167-6296\(02\)00126-1](https://doi.org/10.1016/S0167-6296(02)00126-1)).

11. Lester M. Crawford, Acting Commissioner of the FDA. Speech before Cleveland Clinic Foundation's 2004 Medical Innovation Summit.
<http://www.fda.gov/NewsEvents/Speeches/ucm053331.htm> (accessed Sept 17, 2009).
12. European Parliament, Council. *Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency*, 2006.
13. Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberge, L. E. Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors. *J. Med. Chem.* **1996**, 39 (16), 3049-3059 (DOI:[10.1021/jm960290n](https://doi.org/10.1021/jm960290n)).
14. Ng, R. *Drugs. From Discovery to Approval.*; John Wiley and Sons Inc.: Hoboken, NJ, 2004.
15. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **2004**, 431, 931-945 (DOI:[10.1038/nature03001](https://doi.org/10.1038/nature03001)).
16. Irwin, J. J. How good is your screening library? *Current Opinion in Chemical Biology* **2006**, 10 (4), 352-356 (DOI:[10.1016/j.cbpa.2006.06.003](https://doi.org/10.1016/j.cbpa.2006.06.003)).
17. Djulbegovic, B.; Kumar, A.; Soares, H. P.; Hozo, I.; Bepler, G.; Clarke, M.; Bennett, C. L. Treatment Success in Cancer: New Cancer Treatment Successes Identified in Phase 3 Randomized Controlled Trials Conducted by the National Cancer Institute–Sponsored Cooperative Oncology Groups, 1955 to 2006. *Arch. Intern. Med.* **2008**, 168 (6), 632-642.
18. Frantz, S. Pharma's year of trouble and strife. *Nature Reviews Drug Discovery* **2006**, 5, 7-9 (DOI: [10.1038/nrd1944](https://doi.org/10.1038/nrd1944)).
19. Couzin, J. Withdrawal of Vioxx casts a shadow over COX-2 inhibitors. *Science* **2004**, 306 (5695), 384-385 (DOI:[10.1126/science.306.5695.384](https://doi.org/10.1126/science.306.5695.384)).

20. Krumholz, H. M.; Ross, J. S.; Presler, A. H.; Egilman, D. S. What have we learnt from Vioxx? *BMJ* **2007**, No. 334, 120-123 (DOI:[10.1136/bmj.39024.487720.68](https://doi.org/10.1136/bmj.39024.487720.68)).
21. Vioxx settlement to total \$4.85bn. *BBC News*, November 9, 2007, <http://news.bbc.co.uk/1/hi/business/7087348.stm>.
22. Solmajer, T.; Zupan, J. Optimization algorithms and natural computing in drug discovery. *Drug Discov. Today: Technologies* **2004**, 1 (3), 247-252 (DOI:[10.1016/j.ddtec.2004.11.011](https://doi.org/10.1016/j.ddtec.2004.11.011)).
23. Hann, M. What we can and cannot do in computational chemistry, and why we should continue to try! *Virtual Discovery, RSC Meeting*, London, March 21, 2006.
24. Young, S. S.; Sheffield, C. F.; Farmen, M. Optimum Utilization of a Compound Collection or Chemical Library for Drug Discovery. *J. Chem. Inf. Comput. Sci.* **1997**, 5 (892-899), 37 (DOI: [10.1021/ci970224+](https://doi.org/10.1021/ci970224+)).
25. Bajorath, J. Integration of virtual and high-throughput screening. *Nature Reviews Drug Discovery* **2002**, 1 (11), 882-894 (DOI:[10.1038/nrd941](https://doi.org/10.1038/nrd941)).
26. Lengauer, T.; Lemmen, C.; Rarey, M.; Zimmermann, M. Novel technologies for virtual screening. *Drug Discov. Today* **2004**, 9 (1), 27-34 (DOI:[10.1016/S1359-6446\(04\)02939-3](https://doi.org/10.1016/S1359-6446(04)02939-3)).
27. Briem, H.; Lessel, U. F. In vitro and in silico affinity fingerprints: Finding similarities beyond structural classes. *Perspect. Drug Discovery Des.* **2004**, 20 (1), 231-244 (DOI:[10.1023/A:1008793325522](https://doi.org/10.1023/A:1008793325522)).
28. Warmuth, M. K.; Liao, J.; Ratsch, G.; Mathieson, M.; Putta, S.; Lemmen, C. Active Learning with Support Vector Machines in the Drug Discovery Process. *J. Chem. Inf. Comput. Sci.* **2003**, 43 (2), 667-673 (DOI:[10.1021/ci025620t](https://doi.org/10.1021/ci025620t)).
29. Keiser, M. J.; Setola, V.; Irwin, J. J.; Laggner, C.; Abbas, A. I.; Hufeisen, S. J.; Jensen, N. H.; Kuijter, M. B.; Matos, R. C.; Tran, T. B.; Whaley, R.; Glennon, R. A.; Hert, J.; Thomas, K. L. H.; Edwards, D. D.; Shoichet, B. K.; Roth, B. L. Predicting new molecular targets for known drugs. *Nature* **2009**, 462 (7270), 175-181 (DOI: [10.1038/nature08506](https://doi.org/10.1038/nature08506)).

30. Bergström, C. A. S.; Norinde, U.; Luthman, K.; Artursson, P. Experimental and Computational Screening Models for Prediction of Aqueous Drug Solubility. *Pharmaceutical Research* **2002**, *19* (2), 182-188 (DOI:[10.1023/A:1014224900524](https://doi.org/10.1023/A:1014224900524)).
31. Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *5* (868-873), 39 (DOI: [10.1021/ci990307l](https://doi.org/10.1021/ci990307l)).
32. Xing, L.; Glen, R. C. Novel Methods for the Prediction of logP, pKa, and logD. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (4), 796-805 (DOI:[10.1021/ci010315d](https://doi.org/10.1021/ci010315d)).
33. Boyer, S.; Zamora, I. New methods in predictive metabolism. *J. Comput. Aided Mol. Des.* **2002**, *16*, 403-413 (DOI:[10.1023/A:1020881520931](https://doi.org/10.1023/A:1020881520931)).
34. Santos-Filho, O. A.; Hopfinger, A. J.; Zheng, T. Characterization of Skin Penetration Processes of Organic Molecules Using Molecular Similarity and QSAR Analysis. *Molecular Pharmaceutics* **2004**, *1* (6), 466–476 (DOI: [10.1021/mp049924+](https://doi.org/10.1021/mp049924+)).
35. Sanderson, D. M.; Earnshaw, C. G. Computer prediction of possible toxic action from chemical structure; the DEREK system. *Hum. Exp. Toxicol.* **1991**, *10* (4), 261-273 (DOI: [10.1177/096032719101000405](https://doi.org/10.1177/096032719101000405)).
36. Derek; Lhasa Ltd.: Leeds, UK.
37. Prival, M. J. Evaluation of the TOPKAT system for predicting the carcinogenicity of chemicals. *Environ. Mol. Mutagen.* **2001**, *37* (1), 55-69 (DOI: [10.1002/1098-2280\(2001\)37:13.O.CO;2-5](https://doi.org/10.1002/1098-2280(2001)37:13.O.CO;2-5)).
38. Williams, D. P.; Naisbitt, D. J. Toxicophores: Groups and metabolic routes associated with increased safety risk. *Current Opinion in Drug Discovery & Development* **2002**, *5* (1), 104-115.
39. Kennedy, T. Managing the drug discovery/development interface. *Drug Discov. Today* **1997**, *2* (10), 436-444 (DOI:[10.1016/S1359-6446\(97\)01099-4](https://doi.org/10.1016/S1359-6446(97)01099-4)).

40. Gillet, V. J.; Newell, W.; Mata, P.; Myatt, G.; Sike, S.; Zsoldos, Z.; Johnson, A. P. SPROUT: Recent developments in the de novo design of molecules. *J. Chem. Inf. Comput. Sci.* **1994**, *34* (1), 207–217 (DOI: [10.1021/ci00017a027](https://doi.org/10.1021/ci00017a027)).
41. Johnson, A. M., Maggiora, G. M., Eds. *Concepts and Applications of Molecular Similarity*; John Wiley & Sons: New York, 1990.
42. Bender, A.; Jenkins, J. L.; Li, Q.; Adams, S. E.; Cannon, E. O.; Glen, R. C. Molecular Similarity: Advances in Methods, Applications and Validations in Virtual Screening and QSAR. *Annu. Rep. Comput. Chem.* **2006**, *2*, 141-168 (DOI: [10.1016/S1574-1400\(06\)02009-3](https://doi.org/10.1016/S1574-1400(06)02009-3)).
43. Todeschini, R.; Consonni, V. *Methods and Principles in Medicinal Chemistry: Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000; Vol. 11.
44. Hong, H.; Xie, Q.; Ge, W.; Qian, F.; Fang, H.; Shi, L.; Su, Z.; Perkins, R.; Tong, W. Mold², Molecular Descriptors from 2D Structures for Chemoinformatics and Toxicoinformatics. *J. Chem. Inf. Model.* **2008**, *48* (7), 1337-1344 (DOI: [10.1021/ci800038f](https://doi.org/10.1021/ci800038f)).
45. MOE: *Molecular Operating Environment*; Chemical Computing Group: Montreal, Canada.
46. SYBYL *Molecular Modeling Software*; Tripos Associates Inc.: St Louis, MO, USA.
47. Senese, C. L.; Duca, J.; Pan, D.; Hopfinger, A. J.; Tseng, Y. J. 4D-Fingerprints, Universal QSAR and QSPR Descriptors. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (5), 1526–1539 (DOI: [10.1021/ci049898s](https://doi.org/10.1021/ci049898s)).
48. Vedani, A.; Dobler, M. 5D-QSAR: The Key for Simulating Induced Fit? *J. Med. Chem.* **2002**, *45* (11), 2139-2149 (DOI: [10.1021/jm011005p](https://doi.org/10.1021/jm011005p)).
49. Lill, M. A. Multi-dimensional QSAR in drug discovery. *Drug Discov. Today* **2007**, *12* (23-24), 1013-1017 (DOI: [10.1016/j.drudis.2007.08.004](https://doi.org/10.1016/j.drudis.2007.08.004)).

50. Fujita, T. Recent Success Stories Leading to Commercializable Bioactive Compounds with the Aid of Traditional QSAR Procedures. *Quant. Struct.-Act. Relat.* **1997**, 16 (2), 107-112 (DOI:[10.1002/qsar.19970160202](https://doi.org/10.1002/qsar.19970160202)).
51. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews* **2001**, 46 (1-3), 3-26 (DOI:[10.1016/S0169-409X\(00\)00129-0](https://doi.org/10.1016/S0169-409X(00)00129-0)).
52. Wang, R.; Fu, Y.; Lai, L. A New Atom-Additive Method for Calculating Partition Coefficients. *J. Chem. Inf. Comput. Sci.* **1997**, 37 (3), 615-621 (DOI:[10.1021/ci960169p](https://doi.org/10.1021/ci960169p)).
53. *MACCS Keys*; Symyx Software: San Ramon, CA.
54. Bawden, D. Computerized chemical structure-handling techniques in structure-activity studies and molecular property prediction. *J. Chem. Inf. Comput. Sci.* **1983**, 23 (1), 14-22 (DOI:[10.1021/ci00037a003](https://doi.org/10.1021/ci00037a003)).
55. Daylight Chemical Information Systems, Inc. *Daylight Theory Manual*; Aliso Viejo, CA, USA, 2006.
56. *UNITY Reference Manual*; Tripos Inc.: St. Louis, MO, USA.
57. Bremser, W. Hose - a novel substructure code. *Analytica Chimica Acta* **1978**, 103 (4), 355-365 (DOI: [10.1016/S0003-2670\(01\)83100-7](https://doi.org/10.1016/S0003-2670(01)83100-7)).
58. Faulon, J.-L. The Signature Molecular Descriptor. 1. Using Extended Valence Sequences in QSAR and QSPR Studies. *J. Chem. Inf. Comput. Sci.* **2003**, 43 (3), 707-720 (DOI: [10.1021/ci020345w](https://doi.org/10.1021/ci020345w)).
59. Faulon, J.-L.; Churchwell, C. J. The Signature Molecular Descriptor. 2. Enumerating Molecules from Their Extended Valence Sequences. *J. Chem. Inf. Comput. Sci.* **2003**, 43 (3), 721-734 (DOI: [10.1021/ci020346o](https://doi.org/10.1021/ci020346o)).
60. Xing, L.; Glen, R. C.; Clark, R. D. Predicting pKa by Molecular Tree Structured Fingerprints and PLS. *J. Chem. Inf. Comput. Sci.* **2003**, 43 (3), 870-879 (DOI:[10.1021/ci020386s](https://doi.org/10.1021/ci020386s)).

61. Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naïve Bayesian Classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (1), 170-178 (DOI:[10.1021/ci034207y](https://doi.org/10.1021/ci034207y)).
62. Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (5), 1708-1718 (DOI:[10.1021/ci0498719](https://doi.org/10.1021/ci0498719)).
63. *Pipeline Pilot*; Scitegic Inc.: San Diego, CA.
64. Rogers, D.; Brown, R. D.; Hahn, M. Using Extended-Connectivity Fingerprints with Laplacian-Modified Bayesian Analysis in High-Throughput Screening Follow-Up. *Org. Biomol. Chem.* **2005**, *10* (7), 682-686 (DOI:[10.1177/1087057105281365](https://doi.org/10.1177/1087057105281365)).
65. Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures - A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5* (2), 107-113 (DOI: [10.1021/c160017a018](https://doi.org/10.1021/c160017a018)).
66. Holliday, J. D.; Jelfs, S. P.; Willet, P.; Gedeck, P. Calculation of Intersubstituent Similarity Using R-Group Descriptors. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (2), 406-411 (DOI: [10.1021/ci025589v](https://doi.org/10.1021/ci025589v)).
67. Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25* (2), 64-73 (DOI:[10.1021/ci00046a002](https://doi.org/10.1021/ci00046a002)).
68. Judson, P. N. Structural similarity searching using descriptors developed for structure-activity relationship studies. *J. Chem. Inf. Comput. Sci.* **1992**, *32* (6), 657-663 (DOI:[10.1021/ci00010a012](https://doi.org/10.1021/ci00010a012)).
69. Melville, J. L.; Hirst, J. D. TMACC: Interpretable Correlation Descriptors for Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model.* **2007**, *47* (2), 626-634 (DOI: [10.1021/ci6004178](https://doi.org/10.1021/ci6004178)).

70. Young, S. S.; Gombara, V. K.; Emptagea, M. R.; Carielloa, N. F.; Lambert, C. Mixture deconvolution and analysis of Ames mutagenicity data. *Chemometr. Intell. Lab. Syst.* **2002**, 60 (1-2), 5-11 (DOI: [10.1016/S0169-7439\(01\)00181-2](https://doi.org/10.1016/S0169-7439(01)00181-2)).
71. Nigsch, F.; Mitchell, J. B. O. How To Winnow Actives from Inactives: Introducing Molecular Orthogonal Sparse Bigrams (MOSBs) and Multiclass Winnow. *J. Chem. Inf. Model.* **2008**, 48 (2), 306–318 (DOI: [10.1021/ci700350n](https://doi.org/10.1021/ci700350n)).
72. Willet, P.; Winterman, V.; Bawden, D. Implementation of Nearest-Neighbor Searching in an Online Chemical Structure Search System. *J. Chem. Inf. Comp. Sci.* **1986**, 26 (1), 36-41 (DOI: [10.1021/ci00049a008](https://doi.org/10.1021/ci00049a008)).
73. Bender, A.; Mussa, H. Y.; Glen, R. C. Screening for Dihydrofolate Reductase Inhibitors Using MOLPRINT 2D, a Fast Fragment-Based Method Employing the Naïve Bayesian Classifier: Limitations of the Descriptor and the Importance of Balanced Chemistry in Training and Test Sets. *J. Biomol. Scr.* **2005**, 10 (7), 658-666 (DOI: [10.1177/1087057105281048](https://doi.org/10.1177/1087057105281048)).
74. Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.* **2004**, 2, 3256-3266 (DOI: [10.1039/B409865J](https://doi.org/10.1039/B409865J)).
75. Mason, J. S.; Good, A. C.; Martin, E. J. 3-D Pharmacophores in Drug Discovery. *Curr. Pharm. Des.* **2001**, 7 (7), 567-597 (DOI: [10.2174/1381612013397843](https://doi.org/10.2174/1381612013397843)).
76. Karnachi, P.; Kulkarni, A. Mining, Application of Pharmacophore Fingerprints to Structure-based Design and Data. In *Pharmacophores and Pharmacophore Searches*; Mannhold, R., Kubinyi, H., Folkers, G., Eds.; Wiley-VCH: Weinheim, Germany, 2006; Vol. 32, pp 193-204.
77. Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angewandte Chemie International Edition* **1999**, 38 (19), 2894-2896 (DOI: [10.1002/\(SICI\)1521-3773\(19991004\)38:193.0.CO;2-F](https://doi.org/10.1002/(SICI)1521-3773(19991004)38:193.0.CO;2-F)).

78. Morris, R. J.; Najmanovich, R. J.; Kahraman, A.; Thornton, J. M. Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. *Bioinformatics* **2005**, *21* (10), 2347-2355 (DOI:[10.1093/bioinformatics/bti337](https://doi.org/10.1093/bioinformatics/bti337)).
79. Ballester, P. J.; Richards, W. G. Ultrafast shape recognition for similarity search in molecular databases. *Proc. R. Soc. A* **2007**, *463* (2081), 1307-1321 (DOI: [10.1098/rspa.2007.1823](https://doi.org/10.1098/rspa.2007.1823)).
80. Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28* (7), 849-857 (DOI: [10.1021/jm00145a002](https://doi.org/10.1021/jm00145a002)).
81. *GRID*; Molecular Discovery, Ltd: London.
82. Cramer III, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110* (18), 5959-5967 (DOI:[10.1021/ja00226a005](https://doi.org/10.1021/ja00226a005)).
83. Carbo, R.; Calabuig, B. Quantum similarity measures, molecular cloud description, and structure-properties relationships. *J. Chem. Inf. Comput. Sci.* **1992**, *32* (6), 600-606 (DOI:[10.1021/ci00010a005](https://doi.org/10.1021/ci00010a005)).
84. Cheeseright, T.; Mackey, M.; Rose, S.; Vinter, A. Molecular Field Extrema as Descriptors of Biological Activity: Definition and Validation. *J. Chem. Inf. Model.* **2006**, *46* (2), 665-676 (DOI:[10.1021/ci050357s](https://doi.org/10.1021/ci050357s)).
85. *CORINA*; Molecular Networks GmbH: Erlangen, Germany.
86. *CONCORD*; Tripos Inc.: St. Louis, MO.
87. Bultinck, P.; Kuppens, T.; Gironés, X.; Carbó-Dorca, R. Quantum similarity superposition algorithm (QSSA): a consistent scheme for molecular alignment and molecular similarity based on quantum chemistry. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (4), 1143-1150 (DOI: [10.1021/ci0340153](https://doi.org/10.1021/ci0340153)).

88. Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. GRIND-INdependent Descriptors (GRIND): A Novel Class of Alignment-Independent Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **2000**, *43* (17), 3233–3243 (DOI: [10.1021/jm000941m](https://doi.org/10.1021/jm000941m)).
89. Bender, A.; Mussa, H. Y.; Gill, G. S.; Glen, R. C. Molecular Surface Point Environments for Virtual Screening and the Elucidation of Binding Patterns (MOLPRINT 3D). *J. Med. Chem.* **2004**, *47* (26), 6569–6583 (DOI: [10.1021/jm049611i](https://doi.org/10.1021/jm049611i)).
90. Beger, R. D.; Freeman, J. P.; Lay Jr., J. O.; Wilkes, J. G.; Miller, D. W. Producing ¹³C NMR, Infrared Absorption, and Electron Ionization Mass Spectrometric Data Models of the Monodechlorination of Chlorobenzenes, Chlorophenols, and Chloroanilines. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (6), 1449–1455 (DOI: [10.1021/ci000331v](https://doi.org/10.1021/ci000331v)).
91. Beger, R. D. Computational modeling of biologically active molecules using NMR spectra. *Drug Discov. Today* **2006**, *11* (9-10), 429–435 (DOI: [10.1016/j.drudis.2006.03.014](https://doi.org/10.1016/j.drudis.2006.03.014)).
92. Bursi, R.; Dao, T.; van Wijk, T.; de Gooyer, M.; Kellenbach, E.; Verwer, P. Comparative Spectra Analysis (CoSA): Spectra as Three-Dimensional Molecular Descriptors for the Prediction of Biological Activities. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (5), 861–867 (DOI: [10.1021/ci990038z](https://doi.org/10.1021/ci990038z)).
93. Willighagen, E. L.; Denissen, H. M. G. W.; Wehrens, R.; Buydens, L. M. C. On the Use of ¹H and ¹³C 1D NMR Spectra as QSPR Descriptors. *J. Chem. Inf. Model.* **2006**, *46* (2), 487–494 (DOI: [10.1021/ci050282s](https://doi.org/10.1021/ci050282s)).
94. Kauvar, L. M.; Higgins, D. L.; Villar, H. O.; Sportsman, J. R.; Engqvist-Goldstein, Å.; Bukar, R.; Bauer, K. E.; Dilley, H.; Rocke, D. M. Predicting ligand binding to proteins by affinity fingerprinting. *Chemistry & Biology* **1995**, *2* (2), 107 – 118 (DOI: [10.1016/1074-5521\(95\)90283-X](https://doi.org/10.1016/1074-5521(95)90283-X)).
95. Briem, H.; Kuntz, I. D. Molecular Similarity Based on DOCK-Generated Fingerprints. *J. Med. Chem.* **1996**, *39* (17), 3401–3408 (DOI: [10.1021/jm950800y](https://doi.org/10.1021/jm950800y)).

96. Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, 38 (6), 983-996 (DOI:[10.1021/ci9800211](https://doi.org/10.1021/ci9800211)).
97. Rogers, D. J.; Tanimoto, T. T. A Computer Program for Classifying Plants. *Science* **1960**, 132 (3434), 1115-1118 (DOI: [10.1126/science.132.3434.1115](https://doi.org/10.1126/science.132.3434.1115)).
98. Jaccard, P. La distribution de la flore dans la zone alpine. *Revue générale des Sciences pures et appliquées* **1907**, 18, 961-967.
99. Jaccard, P. The Distribution of the Flora in the Alpine Zone. *New Phytologist* **1912**, 11 (2), 37-50 (DOI:[10.1111/j.1469-8137.1912.tb05611.x](https://doi.org/10.1111/j.1469-8137.1912.tb05611.x)).
100. Holliday, J. D.; Ranade, S. S.; Willett, P. A Fast Algorithm For Selecting Sets Of Dissimilar Molecules From Large Chemical Databases. *Quantitative Structure-Activity Relationships* **1995**, 14 (6), 501-506 (DOI:[10.1002/qsar.19950140602](https://doi.org/10.1002/qsar.19950140602)).
101. Brown, R. D.; Martin, Y. C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand-Receptor Binding. *J. Chem. Inf. Comput. Sci.* **1997**, 37 (1), 1-9 (DOI:[10.1021/ci960373c](https://doi.org/10.1021/ci960373c)).
102. Brown, R. D.; Martin, Y. C. Use of Structure-Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, 36 (3), 572-584 (DOI:[10.1021/ci9501047](https://doi.org/10.1021/ci9501047)).
103. Hurst, T.; Heritage, T. HQSAR - A Highly Predictive QSAR Technique Based on Molecular Holograms. *213th ACS Natl. Meeting April 13-17*, San Francisco, CA, 1997.
104. Fechner, U.; Paetz, J.; Schneider, G. Comparison of Three Holographic Fingerprint Descriptors and their Binary Counterparts. *QSAR & Combinatorial Science* **2005**, 24 (8), 961-967 (DOI:[10.1002/qsar.200530118](https://doi.org/10.1002/qsar.200530118)).
105. Flower, D. R. On the Properties of Bit String-Based Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci.* **1998**, 38 (3), 379-386 (DOI:[10.1021/ci970437z](https://doi.org/10.1021/ci970437z)).
106. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The WEKA Data Mining Software: An Update. *SIGKDD Explorations* **2009**, 11 (1), 10-18 (DOI: [10.1145/1656274.1656278](https://doi.org/10.1145/1656274.1656278)).

107. Mierswa, I.; Wurst, M.; Klinkenberg, R.; Scholz, M.; Euler, T. YALE: Rapid Prototyping for Complex Data Mining Tasks. *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, Philadelphia, PA, USA, 2006; pp 935--940.
108. R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna Austria, 2009;; <http://www.R-project.org>.
109. Hansch, C.; Fujita, T. ρ - σ - π Analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.* **1964**, *86* (8), 1616-1626 (DOI: [10.1021/ja01062a035](https://doi.org/10.1021/ja01062a035)).
110. Hansch, C.; Dunn III, W. J. Linear relationships between lipophilic character and biological activity of drugs. *J. Pharm. Sci.* **1972**, *61*, 1-19 (DOI: [10.1002/jps.2600610102](https://doi.org/10.1002/jps.2600610102)).
111. Free, S. M.; Wilson, J. W. A Mathematical Contribution to Structure-Activity Studies. *J. Med. Chem.* **1964**, *7* (4), 395-399 (DOI: [10.1021/jm00334a001](https://doi.org/10.1021/jm00334a001)).
112. Olah, M.; Bologa, C.; Oprea, T. I. An automated PLS search for biologically relevant QSAR descriptors. *J. Comput. Aided Mol. Des.* **2004**, *18* (7), 437-449 (DOI: [10.1007/s10822-004-4060-8](https://doi.org/10.1007/s10822-004-4060-8)).
113. Hansch, C.; Leo, A.; Mekapati, S. B.; Kurup, A. QSAR and ADME. *Bioorganic & Medicinal Chemistry* **2004**, *12* (12), 3391-3400 (DOI: [10.1016/j.bmc.2003.11.037](https://doi.org/10.1016/j.bmc.2003.11.037)).
114. Blum, A. L.; Langley, P. Selection of relevant features and examples in machine learning. *Artif. Intell.* **1997**, *97* (1-2), 245-271 (DOI: [10.1016/S0004-3702\(97\)00063-5](https://doi.org/10.1016/S0004-3702(97)00063-5)).
115. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157-1182.
116. Bender, A.; Glen, R. C. A Discussion of Measures of Enrichment in Virtual Screening: Comparing the Information Content of Descriptors with Increasing Levels of Sophistication. *J. Chem. Inf. Model.* **2005**, *45* (5), 1369-1375 (DOI: [10.1021/ci0500177](https://doi.org/10.1021/ci0500177)).

117. Shannon, C. E. A mathematical theory of communication. *Bell System Technical Journal* **1948**, 27, 379-423.
118. Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews Drug Discovery* **2004**, 3 (11), 935-949 (DOI:[10.1038/nrd1549](https://doi.org/10.1038/nrd1549)).
119. Mohan, V.; Gibbs, A. C.; Cummings, M. D.; Jaeger, E. P.; DesJarlais, R. L. Docking: Successes and Challenges. *Curr. Pharm. Des.* **2005**, 11, 323-333.
120. Tame, J. R. H. Scoring functions: A view from the bench. *J. Comput. Aided Mol. Des.* **2004**, 13 (2), 99-108 (DOI:[10.1023/A:1008068903544](https://doi.org/10.1023/A:1008068903544)).
121. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, 267 (3), 727-748 (DOI:[10.1006/jmbi.1996.0897](https://doi.org/10.1006/jmbi.1996.0897)).
122. Lengauer, T. The FLEX Approach: An Alternative for Receptor-Ligand Docking and Computing Crystal Conformations. In *Computer-Assisted Lead Finding and Optimization*; van de Waterbeemd, H., Testa, B., Folkers, G., Eds.; Verlag Helvetica Chimica Acta: Zürich, 2007; pp 397-420.
123. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method using an Incremental Construction Algorithm. *J. Mol. Biol.* **1996**, 261 (3), 470-489 (DOI:[10.1006/jmbi.1996.0477](https://doi.org/10.1006/jmbi.1996.0477)).
124. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, 47 (7), 1739-1749 (DOI:[10.1021/jm0306430](https://doi.org/10.1021/jm0306430)).
125. Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **2004**, 47 (7), 1750-1759 (DOI:[10.1021/jm030644s](https://doi.org/10.1021/jm030644s)).

126. Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins. *J. Med. Chem.* **1999**, *42* (25), 5100-5109 (DOI:[10.1021/jm990352k](https://doi.org/10.1021/jm990352k)).
127. Meagher, K. L.; Carlson, H. A. Incorporating Protein Flexibility in Structure-Based Drug Discovery: Using HIV-1 Protease as a Test Case. *J. Am. Chem. Soc.* **2004**, *126* (41), 13276-13281 (DOI:[10.1021/ja0469378](https://doi.org/10.1021/ja0469378)).
128. Caffrey, M. Membrane protein crystallization. *Journal of Structural Biology* **2003**, *142* (1), 108-132 (DOI:[10.1016/S1047-8477\(03\)00043-1](https://doi.org/10.1016/S1047-8477(03)00043-1)).
129. Rasmussen, S. G. F.; Choi, H.-J.; Rosenbaum, D. M.; Kobilka, T. S.; Thian, F. S.; Edwards, P. C.; Burghammer, M.; Ratnala, V. R. P.; Sanishvili, R.; Fischetti, R. F.; Schertler, G. F. X.; Weis, W. I.; Kobi. Crystal structure of the human β 2 adrenergic G-protein-coupled receptor. *Nature* **2007**, *450* (7168), 383-387 (DOI: [10.1038/nature06325](https://doi.org/10.1038/nature06325)).
130. Cherezov, V.; Rosenbaum, D. M.; Hanson, M. A.; Rasmussen, S. G. F.; Thian, F. S.; Kobilka, T. S.; Choi, H.-J.; Kuhn, P.; Weis, W. I.; Kobilka, B. K.; C., S. R. High-Resolution Crystal Structure of an Engineered Human β 2-Adrenergic G Protein–Coupled Receptor. *Science* **2007**, *318* (5854), 1258-1265 (DOI: [10.1126/science.1150577](https://doi.org/10.1126/science.1150577)).
131. Rosenbaum, D. M.; Cherezov, V.; Hanson, M. A.; Rasmussen, S. G. F.; Thian, F. S.; Kobilka, T. S.; Choi, H.-J.; Yao, X.-J.; Weis, W. I.; Stevens, R. C.; Kobilka, B. K. GPCR Engineering Yields High-Resolution Structural Insights into β 2-Adrenergic Receptor Function. *Science* **2007**, *318* (5854), 1266-1273 (DOI: [10.1126/science.1150609](https://doi.org/10.1126/science.1150609)).
132. Hooft, R. W. W.; Vriend, G.; Sander, C.; Abola, E. E. Errors in protein structures. *Nature* **1996**, *381*, 272 (DOI:[10.1038/381272a0](https://doi.org/10.1038/381272a0)).
133. Davies, A. Organising our Ignorance: The Secret of Successful Modelling. *Virtual Discovery, RSC Meeting*, London, March 21, 2006.
134. Johnson, S. R. The Trouble with QSAR (or How I Learned To Stop Worrying and Embrace Fallacy). *J. Chem. Inf. Model.* **2008**, *48* (1), 25-26 (DOI:[10.1021/ci700332k](https://doi.org/10.1021/ci700332k)).

135. Doweyko; M., A. QSAR: dead or alive. *J. Comput. Aided Mol. Des.* **2008**, *22* (2), 81-89 (DOI:[10.1007/s10822-007-9162-7](https://doi.org/10.1007/s10822-007-9162-7)).
136. Maggiora, G. M. On Outliers and Activity Cliffs - Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46* (4), 1535 (DOI:[10.1021/ci060117s](https://doi.org/10.1021/ci060117s)).
137. Medina-Franco, J. L.; Marti'nez-Mayorga, K.; Bender, A.; Marín, R. M.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A. Characterization of Activity Landscapes Using 2D and 3D Similarity Methods: Consensus Activity Cliffs. *J. Chem. Inf. Model.* **2009**, *49* (2), 477–491 (DOI: [10.1021/ci800379q](https://doi.org/10.1021/ci800379q)).
138. Maldonado, A. G.; Doucet, J. P.; Petitjean, M.; Fan, B.-T. Molecular similarity and diversity in chemoinformatics: From theory to applications. *Mol. Div.* **2006**, *10* (1), 39-79 (DOI:[10.1007/s11030-006-8697-1](https://doi.org/10.1007/s11030-006-8697-1)).
139. Hobbs, D. W.; Guo, T. Library Design Concepts and Implementation Strategies. In *Combinatorial Library Design and Evaluation. Principles, Software Tools, and Applications in Drug Discovery*; Ghose, A. K., Viswanadhan, V. N., Eds.; Marcel Dekker: New York, 2001; pp 1-50.
140. Roderick, S. L.; Fournie-Zaluski, M. C.; Roques, B. P.; Matthews, B. W. Thiorphan and retro-thiorphan display equivalent interactions when bound to crystalline thermolysin. *Biochemistry* **1989**, *28* (4), 1493–1497 (DOI: [10.1021/bi00430a011](https://doi.org/10.1021/bi00430a011)).
141. Kubini, H. A General View on Similarity and QSAR Studies. In *Computer-Assisted Lead Finding and Optimization: Current Tools for Medicinal Chemistry*; Waterman, H. v. d., Testa, B., Folkers, G., Eds.; VHCA & Wiley-VHC: Basel, 1997; pp 7-28.
142. Gillet, V. J.; Willett, P.; Bradshaw, J. Identification of Biological Activity Profiles Using Substructural Analysis and Genetic Algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (2), 165-179 (DOI: [10.1021/ci970431+](https://doi.org/10.1021/ci970431+)).
143. Mannhold, R.; Poda, G. I.; Ostermann, C.; Tetko, I. V. Calculation of molecular lipophilicity: State-of-the-art and comparison of log P methods on more than 96,000 compounds. *J. Pharm. Sci.* **2009**, *98* (3), 861-893 (DOI:[10.1002/jps.21494](https://doi.org/10.1002/jps.21494)).

144. Manchester, J.; Czermiński, R. SAMFA: Simplifying Molecular Description for 3D-QSAR. *J. Chem. Inf. Model.* **2008**, *48* (6), 1167-1173 (DOI: [10.1021/ci800009u](https://doi.org/10.1021/ci800009u)).
145. Bergstrom, C. A. S.; Wassvik, C. M.; Norinder, U.; Luthman, K.; Artursson, P. Global and Local Computational Models for Aqueous Solubility Prediction of Drug-Like Molecules. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (4), 1477-1488 (DOI: [10.1021/ci049909h](https://doi.org/10.1021/ci049909h)).
146. Nettles, J. H.; Jenkins, J. L.; Bender, A.; Deng, Z.; Davies, J. W.; Glick, M. Bridging Chemical and Biological Space: "Target Fishing" Using 2D and 3D Molecular Descriptors. *J. Med. Chem.* **2006**, *49* (23), 6802-6810 (DOI: [10.1021/jm060902w](https://doi.org/10.1021/jm060902w)).
147. Matter, H.; Pötter, T. Comparing 3D Pharmacophore Triplets and 2D Fingerprints for Selecting Diverse Compound Subsets. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (6), 1211-1225 (DOI: [10.1021/ci980185h](https://doi.org/10.1021/ci980185h)).
148. Hillebrecht, A.; Klebe, G. Use of 3D QSAR Models for Database Screening: A Feasibility Study. *J. Chem. Inf. Model.* **2008**, *48* (2), 384-396 (DOI: [10.1021/ci7002945](https://doi.org/10.1021/ci7002945)).
149. Brooks, W. H.; Daniel, K. G.; Sung, S.-S.; Guida, W. C. Computational Validation of the Importance of Absolute Stereochemistry in Virtual Screening. *J. Chem. Inf. Model.* **2008**, *48* (3), 639-645 (DOI: [10.1021/ci700358r](https://doi.org/10.1021/ci700358r)).
150. Feher, M.; Williams, C. I. Effect of Input Differences on the Results of Docking Calculations. *J. Chem. Inf. Model.* **2009**, *49* (7), 1704-1714 (DOI: [10.1021/ci9000629](https://doi.org/10.1021/ci9000629)).
151. Chung, J. Y.; Hah, J.-M.; Cho, A. E. Correlation between Performance of QM/MM Docking and Simple Classification of Binding Sites. *J. Chem. Inf. Model.* **2009**, *49* (10), 2382-2387 (DOI: [10.1021/ci900231p](https://doi.org/10.1021/ci900231p)).
152. Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br. J. Pharmacol.* **2008**, *153* (S1), S7-S26 (DOI: [10.1038/sj.bjp.0707515](https://doi.org/10.1038/sj.bjp.0707515)).
153. Guha, R.; Dutta, D.; Jurs, P. C.; Chen, T. Local Lazy Regression: Making Use of the Neighborhood to Improve QSAR Predictions. *J. Chem. Inf. Model.* **2006**, *46* (4), 1836-1847 (DOI: [10.1021/ci060064e](https://doi.org/10.1021/ci060064e)).

154. Zhang, S.; Golbraikh, A.; Oloff, S.; Kohn, H.; Tropsha, A. A Novel Automated Lazy Learning QSAR (ALL-QSAR) Approach: Method Development, Applications, and Virtual Screening of Chemical Databases Using Validated ALL-QSAR Models. *J. Chem. Inf. Model.* **2006**, *46* (5), 1984-1995 (DOI:[10.1021/ci060132x](https://doi.org/10.1021/ci060132x)).
155. Tetko, I. V. Neural Network Studies. 4. Introduction to Associative Neural Networks. *J. Chem. Inf. Comput. Sci.* **717-728**, *42* (3), 2002 (DOI:[10.1021/ci010379o](https://doi.org/10.1021/ci010379o)).
156. Tetko, I. V.; Tanchuk, V. Y. Application of Associative Neural Networks for Prediction of Lipophilicity in ALOGPS 2.1 Program. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (5), 1136-1145 (DOI:[10.1021/ci025515j](https://doi.org/10.1021/ci025515j)).
157. Rodgers, S. L.; Davis, A. M.; Tomkinson, N. P.; van de Waterbeemd, H. QSAR Modeling Using Automatically Updating Correction Libraries: Application to a Human Plasma Protein Binding Model. *J. Chem. Inf. Model.* **2007**, *47* (6), 2401-2407 (DOI:[10.1021/ci700197x](https://doi.org/10.1021/ci700197x)).
158. Rodgers, S. L.; Davis, A. M.; van de Waterbeemd, H. Time-Series QSAR Analysis of Human Plasma Protein Binding Data. *QSAR & Combinatorial Science* **2007**, *24* (4), 511-521 (DOI:[10.1002/qsar.200630114](https://doi.org/10.1002/qsar.200630114)).
159. Surowiecki, J. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*; Little, Brown: London, 2004.
160. Buskirk, E. V. How the Netflix Prize Was Won, 2009. wired.com. <http://www.wired.com/epicenter/2009/09/how-the-netflix-prize-was-won/> (accessed Nov 15, 2009).
161. Willett, P. Enhancing the Effectiveness of Ligand-Based Virtual Screening Using Data Fusion. *QSAR & Combinatorial Science* **2006**, *25* (12), 1143-1152 (DOI: [10.1002/qsar.200610084](https://doi.org/10.1002/qsar.200610084)).
162. Beroza, P.; Villar, H. O.; Wick, M. M.; Martin, G. R. Chemoproteomics as a basis for post-genomic drug discovery. *Drug Discov. Today* **2002**, *7* (15), 807-814 (DOI: [10.1016/S1359-6446\(02\)02371-1](https://doi.org/10.1016/S1359-6446(02)02371-1)).

163. Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discov. Today* **2007**, 7 (11), 903-911 (DOI: [10.1016/S1359-6446\(02\)02411-X](https://doi.org/10.1016/S1359-6446(02)02411-X)).
164. Feher, M. Consensus scoring for protein–ligand interactions. *Drug Discov. Today* **2006**, 11 (9-10), 421-428 (DOI: [10.1016/j.drudis.2006.03.009](https://doi.org/10.1016/j.drudis.2006.03.009)).
165. Teramoto, R.; Fukunishi, H. Supervised Consensus Scoring for Docking and Virtual Screening. *J. Chem. Inf. Model.* **2007**, 48 (2), 526–534 (DOI: [10.1021/ci6004993](https://doi.org/10.1021/ci6004993)).
166. Carlsson, L.; Helgee, E. A.; Boyer, S. Interpretation of Nonlinear QSAR Models Applied to Ames Mutagenicity Data. *J. Chem. Inf. Model.* **2009**, ASAP, Web Publication Date: October 14, 2009 (DOI: [10.1021/ci9002206](https://doi.org/10.1021/ci9002206)).
167. Jacobsson, M.; Liden, P.; Stjernschantz, E.; Bostrom, H.; Norinder, U. Improving Structure-Based Virtual Screening by Multivariate Analysis of Scoring Data. *J. Med. Chem.* **2003**, 46 (26), 5781-5789 (DOI: [10.1021/jm030896t](https://doi.org/10.1021/jm030896t)).
168. Fontaine, F.; Pastor, M.; Zamora, I.; Sanz, F. Anchor-GRIND: filling the gap between standard 3D QSAR and the GRid-INdependent descriptors. *J. Med. Chem.* **2005**, 48 (7), 2687-2694 (DOI: [10.1021/jm049113+](https://doi.org/10.1021/jm049113+)).
169. *Molecular Drug Data Report (MDDR)*; Symyx Software: San Ramon, CA.
170. Olah, M.; Mracec, M.; Ostopovici, L.; Rad, R.; Bora, A.; Hadaruga, N.; Olah, I.; Banda, M.; Simon, Z.; Mracec, M.; Oprea, T. I. WOMBAT: World of Molecular Bioactivity. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, Germany, 2005; pp 221-239.
171. ChEMBL Database. <http://www.ebi.ac.uk/chembl/db/index.php> (accessed Nov 17, 2009).
172. Llinàs, A.; Glen, R. C.; Goodman, J. M. Solubility Challenge: Can You Predict Solubilities of 32 Molecules Using a Database of 100 Reliable Measurements? *J. Chem. Inf. Model.* **2008**, 48 (7), 1289-1303 (DOI: [10.1021/ci800058v](https://doi.org/10.1021/ci800058v)).

173. Austin, C. P.; Brady, L. S.; Insel, T. R.; Collins, F. S. NIH Molecular Libraries Initiative. *Science* **2004**, 306 (5669), 1138-1139 (DOI: [10.1126/science.1105511](https://doi.org/10.1126/science.1105511)).
174. PubChem BioAssay Home. <http://www.ncbi.nlm.nih.gov/pcassay> (accessed Nov 17, 2009).
175. Zhang, Q.; Hughes-Oliver, J. M.; Ng, R. T. A Model-Based Ensembling Approach for Developing QSARs. *J. Chem. Inf. Model.* **2009**, 49 (8), 1857–1865 (DOI: [10.1021/ci900080f](https://doi.org/10.1021/ci900080f)).
176. Jorissen, R. N.; Gilson, M. K. Virtual Screening of Molecular Databases Using a Support Vector Machine. *J. Chem. Inf. Model.* **2005**, 45 (3), 549-561 (DOI: [10.1021/ci049641u](https://doi.org/10.1021/ci049641u)).
177. Krejsa, C. M.; Horvath, D.; Rogalski, S. L.; Penzotti, J. E.; Mao, B.; Barbosa, F.; Migeon, J. C. Predicting ADME properties and side effects: the BioPrint approach. *Curr. Opin. Drug Discov. Devel.* **2003**, 6 (4), 470-480.
178. Xu, X. New concepts and approaches for drug discovery based on traditional Chinese medicine. *Drug Discov. Today: Technologies* **2006**, 3 (3), 247-253 (DOI: [10.1016/j.ddtec.2006.09.008](https://doi.org/10.1016/j.ddtec.2006.09.008)).
179. Ehrman, T. M.; Barlow, D. J.; Hylands, P. J. Virtual Screening of Chinese Herbs with Random Forest. *J. Chem. Inf. Model.* **2007**, 47 (2), 264–278 (DOI: [10.1021/ci600289v](https://doi.org/10.1021/ci600289v)).
180. McCahey, F. *Connecting Traditional Chinese and Western Medicines through Molecular Informatics*, 2007.
181. Saeh, J. C.; Lyne, P. D.; Takasaki, B. K.; Cosgrove, D. A. Lead Hopping Using SVM and 3D Pharmacophore Fingerprints. *J. Chem. Inf. Model.* **2005**, 45 (4), 1122-1133 (DOI: [10.1021/ci049732r](https://doi.org/10.1021/ci049732r)).
182. Cartmell, J.; Enoch, S.; Krstajic, D.; Leahy, D. E. Automated QSPR through Competitive Workflow. *J. Comput. Aided Mol. Des.* **2005**, 19 (11), 821-833 (DOI: [10.1007/s10822-005-9029-8](https://doi.org/10.1007/s10822-005-9029-8)).

183. Kong, D.-X.; Ren, W.; Lü, W.; Zhang, H.-Y. Do Biologically Relevant Compounds Have More Chance To Be Drugs? *J. Chem. Inf. Model.* **2009**, *49* (10), 2376-2381 (DOI: [10.1021/ci900229c](https://doi.org/10.1021/ci900229c)).
184. Gibson, G. G.; Skett, P. *Introduction to drug metabolism*, 3rd ed.; Nelson Thornes: London, 2001.
185. Ionescu, C.; Caira, M. R. *Drug Metabolism: Current Concepts*; Springer: Dordrecht, NL, 2005.
186. Smith, D. A.; Waterbeemd, H. v. d. Pharmacokinetics and metabolism in early drug discovery. *Curr. Opin. Chem. Biol.* **1999**, *3* (4), 373-378 (DOI: [10.1016/S1367-5931\(99\)80056-8](https://doi.org/10.1016/S1367-5931(99)80056-8)).
187. Bugrim, A.; Nikolskaya, T.; Nikolsky, Y. Early prediction of drug metabolism and toxicity: systems biology approach and modeling. *Drug Discov. Today* **2004**, *9* (3), 127-135 (DOI: [10.1016/S1359-6446\(03\)02971-4](https://doi.org/10.1016/S1359-6446(03)02971-4)).
188. Coleman, M. D. *Human Drug Metabolism: An Introduction*; John Wiley & Sons Ltd: Chichester, UK, 2005; p 2.
189. Rowley, M.; Hallett, D. J.; Goodacre, S.; Moyes, C.; Crawforth, J.; Sparey, T. J.; Patel, S.; Marwood, R.; Patel, S.; Thomas, S.; Hitzel, L.; O'Connor, D.; Szeto, N.; Castro, J. L.; Hutson, P. H.; Ma. 3-(4-Fluoropiperidin-3-yl)-2-phenylindoles as High Affinity, Selective, and Orally Bioavailable h5-HT2A Receptor Antagonists. *J. Med. Chem.* **2001**, *44* (10), 1603-1614 (DOI: [10.1021/jm0004998](https://doi.org/10.1021/jm0004998)).
190. Josephy, P. D.; Mannervik, B. *Molecular Toxicology*, 2nd ed.; Oxford University Press: New York, 2006; p 251.
191. Paakkari, I. Cardiotoxicity of new antihistamines and cisapride. *Toxicology Letters* **2002**, *127* (1-3), 279-284 (DOI: [10.1016/S0378-4274\(01\)00510-0](https://doi.org/10.1016/S0378-4274(01)00510-0)).
192. Lee, W. M. Drug-induced hepatotoxicity. *N. Engl. J. Med.* **2003**, *349*, 474-485.

193. Park, K.; Williams, D. P.; Naisbitt, D. J.; Kitteringham, N. R.; Pirmohamed, M.
Investigation of toxic metabolites during drug development. *Toxicology and Applied Pharmacology* **2005**, *207* (2 suppl. 1), 425-434 (DOI: [10.1016/j.taap.2005.02.029](https://doi.org/10.1016/j.taap.2005.02.029)).
194. TOPKAT; Accelrys Inc.: Birmingham, MA.
195. Ekins, S.; Andreyev, S.; Ryabov, A.; Kirillov, E.; Rakhmatulin, E. A.; Sorokina, S.; Bugrim, A.; Nikolskaya, T. A combined approach to drug metabolism and toxicity assessment. *Drug Metab. Dispos.* **2006**, *34*, 495-503 (DOI: [10.1124/dmd.10](https://doi.org/10.1124/dmd.10)).
196. Muster, W.; Breidenbach, A.; Fischer, H.; Kirchner, S.; Müller, L.; Pähler, A.
Computational toxicology in drug development. *Drug Discov. Today* **2008**, *13* (7-8), 303-310 (DOI: [10.1016/j.drudis.2007.12.007](https://doi.org/10.1016/j.drudis.2007.12.007)).
197. Gunatilleka, A. D.; Poole, C. F. Models for estimating the non-specific toxicity of organic compounds in short-term bioassays. *Analyst* **2000**, *125* (1), 127-132 (DOI: [10.1039/a907235g](https://doi.org/10.1039/a907235g)).
198. Expert Scientific Group on Phase One Clinical Trials. *Final Report*; The Stationary Office: London, 2006,
http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_063117.
199. Stella, V. J. Prodrug Approaches to Enhancing the Oral Delivery of Poorly Permeable Drugs. In *Prodrugs: Challenges and Rewards Part 1*; Springer,, DOI: 10.1007/978-0-387-49785-3.
200. Stella, V. J. Prodrug Strategies for Improving Drug-Like Properties. In *Optimizing the "Drug-Like" Properties of Leads in Drug Discovery*; Springer,, DOI: [10.1007/978-0-387-44961](https://doi.org/10.1007/978-0-387-44961).
201. Rautio, J.; Kumpulainen, H.; Heimbach, T.; Oliyai, R.; Oh, D.; Järvinen, T.; Savolainen, J.
Prodrugs: design and clinical applications. *Nature Reviews Drug Discovery* **2008**, *7*, 255-270 (DOI: [10.1038/nrd2468](https://doi.org/10.1038/nrd2468)).

202. Mendel, D. B.; Tai, C. Y.; Escarpe, P. A.; Li, W.; Sidwell, R. W.; Huffman, J. H.; Sweet, C.; Jakeman, K. J.; Merson, J.; Lacy, S. A.; Lew, W.; Williams, M. A.; Zhang, L.; Chen, M. S.; Bischofberger, N. Oral Administration of a Prodrug of the Influenza Virus Neuraminidase Inhibitor GS 4071 Protects Mice and Ferrets against Influenza Infection. *Antimicrob. Agents Chemother.* **1998**, *42*, 640-646.
203. Stella, V. J., Borchardt, R. T., Hageman, M. J., Oliyai, R., Maag, H., Tilley, J., Eds. *Prodrugs: Challenges and Rewards*; Springer: New York, 2007.
204. Chan, W. K.; Nguyen, L. T.; Miller, V. P.; Harris, R. Z. Mechanism-based inactivation of human cytochrome P450 3A4 by grapefruit juice and red wine. *Life Sciences* **1998**, *62* (10), PL135-PL142 (DOI:[10.1016/S0024-3205\(98\)00013-7](https://doi.org/10.1016/S0024-3205(98)00013-7)).
205. Baillie, T. A. Metabolism and Toxicity of Drugs. Two Decades of Progress in Industrial Drug Metabolism. *Chem. Res. Toxicol.* **2008**, *21*, 129-137 (DOI: [10.1021/tx7002273](https://doi.org/10.1021/tx7002273)).
206. Anari, M. R.; Sanchez, R. I.; Bakhtiar, R.; Franklin, R. B.; Baillie, T. A. Integration of Knowledge-Based Metabolic Predictions with Liquid Chromatography Data-Dependent Tandem Mass Spectrometry for Drug Metabolism Studies: Application to Studies on the Biotransformation of Indinavir. *Anal. Chem.* **2004**, *76* (3), 823-832 (DOI: [10.1021/ac034980s](https://doi.org/10.1021/ac034980s)).
207. Daley-Yates, P. T.; Price, A. C.; Sisson, J. R.; Pereira, A.; Dallow, N. Beclomethasone dipropionate: absolute bioavailability, pharmacokinetics and metabolism following intravenous, oral, intranasal and inhaled administration in man. *British Journal of Clinical Pharmacology* **2001**, *51* (5), 400-409 (DOI: [10.1046/j.0306-5251.2001.01374.x](https://doi.org/10.1046/j.0306-5251.2001.01374.x)).
208. Williams, J. A.; Hyland, R.; Jones, B. C.; Smith, D. A.; Hurst, S.; Goosen, T. C.; Peterkin, V.; Koup, J. R.; Ball, S. E. Drug-drug interactions for UDP-glucuronosyltransferase substrates: a pharmacokinetic explanation for typically observed low exposure (AUC_i/AUC) ratios. *Drug Metab. Dispos.* **2004**, *32* (11), 1201-1208 (DOI: [10.1124/dmd.104.000794](https://doi.org/10.1124/dmd.104.000794)).

209. Guengerich, F. P. Cytochrome P450 and Chemical Toxicology. *Chem. Res. Toxicol.* **2008**, 21 (1), 70–83 (DOI: [10.1021/tx700079z](https://doi.org/10.1021/tx700079z)).
210. Glue, P.; Clement, R. P. Cytochrome P450 Enzymes and Drug Metabolism—Basic Concepts and Methods of Assessment. *Cellular and Molecular Neurobiology* **1999**, 19 (3), 309–323 (DOI: [10.1023/A:1006993631057](https://doi.org/10.1023/A:1006993631057)).
211. Schoch, G. A.; Yano, J. K.; Sansen, S.; Dansette, P. M.; Stout, C. D.; Johnson, E. F. Determinants of Cytochrome P450 2C8 Substrate Binding: Structures of Complexes with Montelukast, Troglitazone, Felodipine, and 9-Cis-Retinoic Acid. *J. Biol. Chem.* **2008**, 283, 17227–17237 (DOI: [10.1074/jbc.M802180200](https://doi.org/10.1074/jbc.M802180200)).
212. Guengerich, F. P.; Macdonald, T. L. Chemical mechanisms of catalysis by cytochromes P-450: a unified view. *Acc. Chem. Res.* **1984**, 17 (1), 9–16 (DOI: [10.1021/ar00097a002](https://doi.org/10.1021/ar00097a002)).
213. Guengerich, F. P. Oxidative, Reductive, and Hydrolytic Metabolism of Drugs. In *Drug Metabolism in Drug Design and Development: Basic Concepts and Practice*; Zhang, D., Zhu, M., Humphreys, W. G., Eds.; John Wiley & Sons, Inc.: Hoboken, NJ, 2008; pp 15–36.
214. Cosme, J.; Johnson, E. F. Engineering Microsomal Cytochrome P450 2C5 to Be a Soluble, Monomeric Enzyme. *J. Biol. Chem.* **2000**, 275 (4), 2545–2553 (DOI: [10.1074/jbc.275.4.2545](https://doi.org/10.1074/jbc.275.4.2545)).
215. Williams, P. A.; Cosme, J.; Sridhar, V.; Johnson, E. F.; McRee, D. E. Mammalian Microsomal Cytochrome P450 Monooxygenase: Structural Adaptations for Membrane Binding and Functional Diversity. *Mol. Cell* **2000**, 5 (1), 121–131 (DOI: [10.1016/S1097-2765\(00\)80408-6](https://doi.org/10.1016/S1097-2765(00)80408-6)).
216. Rowland, P.; Blaney, F. E.; Smyth, M. G.; Jones, J. J.; Leydon, V. R.; Oxbrow, A. K.; Lewis, C. J.; Tennant, M. G.; Modi, S.; Eggleston, D. S.; Chenery, R. J.; Bridges, A. M. Crystal Structure of Human Cytochrome P450 2D6. *J. Biol. Chem.* **2006**, 281 (11), 7614–7622 (DOI: [10.1074/jbc.M511232200](https://doi.org/10.1074/jbc.M511232200)).

217. de Groot, M. J.; Vermeulen, N. P. E.; Kramer, J. D.; van Acker, F. A. A.; Donne-Op den Kelder, G. M. A Three-Dimensional Protein Model for Human Cytochrome P450 2D6 Based on the Crystal Structures of P450 101, P450 102, and P450 108. *Chem. Res. Toxicol.* **1996**, *9* (7), 1079-1091 (DOI: [10.1021/tx960003i](https://doi.org/10.1021/tx960003i)).
218. Afzelius, L.; Arnby, C. H.; Broo, A.; Carlsson, L.; Isaksson, C.; Jurva, U.; Kjellander, B.; Kolmodin, K.; Nilsson, K.; Raubacher, F.; Weidolf, L. State-of-the-art Tools for Computational Site of Metabolism Predictions: Comparative Analysis, Mechanistical Insights, and Future Applications. *Drug Met Rev* **2007**, *39* (1), 61-86 (DOI: [10.1080/03602530600969374](https://doi.org/10.1080/03602530600969374)).
219. Korzekwa, K. R.; Jones, J. P.; Gillette, J. R. Theoretical studies on cytochrome P-450 mediated hydroxylation: a predictive model for hydrogen atom abstractions. *J. Am. Chem. Soc.* **1990**, *112* (19), 7042–7046 (DOI: [10.1021/ja00175a040](https://doi.org/10.1021/ja00175a040)).
220. Jones, J. P.; Mysinger, M.; Korzekwa, K. R. Computational Models for Cytochrome P450: A Predictive Electronic Model for Aromatic Oxidation and Hydrogen Atom Abstraction. *Drug Metab. Dispos.* **2002**, *30* (1), 7-12 (DOI: [10.1124/dmd.30.1.7](https://doi.org/10.1124/dmd.30.1.7)).
221. Olsen, L.; Rydberg, P.; Rod, T. H.; Ryde, U. Prediction of Activation Energies for Hydrogen Abstraction by Cytochrome P450. *J. Med. Chem.* **2006**, *49* (22), 6489–6499 (DOI: [10.1021/jm060551l](https://doi.org/10.1021/jm060551l)).
222. Rydberg, P.; Ryde, U.; Olsen, L. Prediction of Activation Energies for Aromatic Oxidation by Cytochrome P450. *J. Phys. Chem. A* **2008**, *112* (50), 13058-13065 (DOI: [10.1021/jp803854v](https://doi.org/10.1021/jp803854v)).
223. Rydberg, P.; Ryde, U.; Olsen, L. Sulfoxide, Sulfur, and Nitrogen Oxidation and Dealkylation by Cytochrome P450. *J. Chem. Theory Comput.* **2008**, *4* (8), 1369-1377 (DOI: [10.1021/ct800101v](https://doi.org/10.1021/ct800101v)).
224. Rydberg, P.; Olsen, L. The Accuracy of Geometries for Iron Porphyrin Complexes from Density Functional Theory. *J. Phys. Chem. A* **2009**, *113* (43), 11949-11953 (DOI: [10.1021/jp9035716](https://doi.org/10.1021/jp9035716)).

225. Rydberg, P.; Vasanthanathan, P.; Oostenbrink, C.; Olsen, L. Fast Prediction of Cytochrome P450 Mediated Drug. *ChemMedChem* **2009**, *4* (12), 2070-2079 (DOI: [10.1002/cmdc.200900363](https://doi.org/10.1002/cmdc.200900363)).
226. Ekins, S.; De Groot, M. J.; Jones, J. P. Pharmacophore and threedimensional QSAR methods for modeling cytochrome P450 active sites. *Drug Metab. Dispos.* **2001**, 936-944.
227. Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aided Mol. Des.* **2001**, *15* (5), 411-428 (DOI: [10.1023/A:1011115820450](https://doi.org/10.1023/A:1011115820450)).
228. Vasanthanathan, P.; Hritz, J.; Taboureau, O.; Olsen, L.; Jørgensen, F. S.; Vermeulen, N. P. E.; Oostenbrink, C. Virtual Screening and Prediction of Site of Metabolism for Cytochrome P450 1A2 Ligands. *J. Chem. Inf. Model.* **2009**, *49* (1), 43-52 (DOI: [10.1021/ci800371f](https://doi.org/10.1021/ci800371f)).
229. de Graaf, C.; Oostenbrink, C.; Keizers, P. H. J.; van der Wijst, T.; Jongejan, A.; Vermeulen, N. P. E. Catalytic Site Prediction and Virtual Screening of Cytochrome P450 2D6 Substrates by Consideration of Water and Rescoring in Automated Docking. *J. Med. Chem.* **2006**, *49* (8), 2417–2430 (DOI: [10.1021/jm0508538](https://doi.org/10.1021/jm0508538)).
230. Singh, S. B.; Shen, L. Q.; Walker, M. J.; Sheridan, R. P. A Model for Predicting Likely Sites of CYP3A4-mediated Metabolism on Drug-like Molecules. *J. Med. Chem.* **2003**, *46* (8), 1330–1336 (DOI: [10.1021/jm020400s](https://doi.org/10.1021/jm020400s)).
231. Sheridan, R. P.; Korzekwa, K. R.; Torres, R. A.; Walker, M. J. Empirical Regioselectivity Models for Human Cytochromes P450 3A4, 2D6, and 2C9. *J. Med. Chem.* **2007**, *50*, 3173-3184 (DOI: [10.1021/jm0613471](https://doi.org/10.1021/jm0613471)).
232. Lewis, D. F. V.; Dickins, M. Factors influencing rates and clearance in P450-mediated reactions: QSARs for substrates of the xenobiotic-metabolizing hepatic microsomal P450s. *Toxicology* **2002**, *170* (1-2), 45-53 (DOI: [10.1016/S0300-483X\(01\)00524-8](https://doi.org/10.1016/S0300-483X(01)00524-8)).

233. Zamora, I.; Afzelius, L.; Cruciani, G. Predicting Drug Metabolism: A Site of Metabolism Prediction Tool Applied to the Cytochrome P450 2C9. *J. Med. Chem.* **2003**, *46* (12), 2313–2324 (DOI: [10.1021/jm021104i](https://doi.org/10.1021/jm021104i)).
234. Zhou, D.; Afzelius, L.; Grimm, S. W.; Andersson, T. B.; Zauhar, R. J.; Zamora, I. Comparison of methods for the prediction of the metabolic sites for CYP3A4-mediated metabolic reactions. *Drug Metab. Dispos.* **2006**, *34* (6), 976–983 (DOI: [10.1124/dmd.105.008631](https://doi.org/10.1124/dmd.105.008631)).
235. Boyer, S.; Arnby, C. H.; Carlsson, L.; Smith, J.; Stein, V.; Glen, R. C. Reaction Site Mapping of Xenobiotic Biotransformations. *J. Chem. Inf. Model.* **2007**, *47* (2), 583–590 (DOI: [10.1021/ci600376g](https://doi.org/10.1021/ci600376g)).
236. Klopman, G.; Dimayuga, M.; Talafous, J. META. 1. A Program for the Evaluation of Metabolic Transformation of Chemicals. *J. Chem. Inf. Comput. Sci.* **1994**, *34* (6), 1320–1325 (DOI: [10.1021/ci00022a014](https://doi.org/10.1021/ci00022a014)).
237. Talafous, J.; Sayre, L. M.; Mieyal, J. J.; Klopman, G. META. 2. A Dictionary Model of Mammalian Xenobiotic Metabolism. *J. Chem. Inf. Comput. Sci.* **1994**, *34* (6), 1326–1333 (DOI: [10.1021/ci00022a015](https://doi.org/10.1021/ci00022a015)).
238. Klopman, G.; Tu, M.; Talafous, a. J. META. 3. A Genetic Algorithm for Metabolic Transform Priorities Optimization. *J. Chem. Inf. Comput. Sci.* **1997**, *37* (2), 329–334 (DOI: [10.1021/ci9601123](https://doi.org/10.1021/ci9601123)).
239. *Meteor*; Lhasa Ltd.: Leeds, UK.
240. Button, W. G.; Judson, P. N.; Long, A.; Vessey, J. D. Using Absolute and Relative Reasoning in the Prediction of the Potential Metabolism of Xenobiotics. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (5), 1371–1377 (DOI: [10.1021/ci0202739](https://doi.org/10.1021/ci0202739)).
241. Payne, M. P.; Long, A.; Marchant, C. A. Improving structure-metabolism relationships for the Meteor Knowledge base system.
<http://www.lhasalimited.org/documents/KL11.pdf> (accessed Oct 14, 2009).

242. Ridder, L.; Wagener, M. SyGMA: Combining Expert Knowledge and Empirical Scoring in the Prediction of Metabolites. *ChemMedChem* **2008**, 3 (5), 821-832 (DOI: [10.1002/cmdc.200700312](https://doi.org/10.1002/cmdc.200700312)).
243. Embrechts, M. J.; Ekins, S. Classification of Metabolites with Kernel-Partial Least Squares (K-PLS). *Drug Metab. Dispos.* **2007**, 35 (3), 325-327 (DOI: [10.1124/dmd.106.013185](https://doi.org/10.1124/dmd.106.013185)).
244. *MetaDrug database*; GeneGo Inc.: St. Joseph, MI.
245. Hou, B. K.; Wackett, L. P.; Ellis, L. B. M. Microbial Pathway Prediction: A Functional Group Approach. *J. Chem. Inf. Comput. Sci.* **2003**, 43 (3), 1051-1057 (DOI: [10.1021/ci034018f](https://doi.org/10.1021/ci034018f)).
246. Ellis, L. B. M.; Roe, D.; Wackett, L. P. The University of Minnesota Biocatalysis/Biodegradation Database: the first decade. *Nucleic Acids Res.* **2006**, 34 (Database issue), D517-D521 (DOI: [10.1093/nar/gkj076](https://doi.org/10.1093/nar/gkj076)).
247. University of Minnesota Biocatalysis/Biodegradation Database. <http://umbbd.msi.umn.edu/> (accessed December 8, 2009).
248. Fenner, K.; Gao, J.; Kramer, S.; Ellis, L.; Wackett, L. Data-driven extraction of relative reasoning rules to limit combinatorial explosion in biodegradation pathway prediction. *Bioinformatics* **2008**, 24 (18), 2079-2085 (DOI: [10.1093/bioinformatics/btn378](https://doi.org/10.1093/bioinformatics/btn378)).
249. OpenEye Scientific Software, Inc., Santa Fe, NM, USA. OEChem. www.eyesopen.com.
250. Symyx Technologies, Inc., Sunnyvale, CA, USA. Metabolite Database. <http://www.symyx.com/products/databases/bioactivity/metabolite/index.jsp> (accessed Aug 07, 2009).
251. Tripos Mol2 File Format. http://www.tripos.com/tripos_resources/fileroot/pdfs/mol2_format2.pdf (accessed May 2009).

252. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, 28 (1), 31-36 (DOI: [10.1021/ci00057a005](https://doi.org/10.1021/ci00057a005)).
253. Open Babel. <http://openbabel.sourceforge.net> (accessed May 2009).
254. Symyx Technologies, Inc., Sunnyvale, CA, USA. MDLNUMBER. In *Metabolite Browser Help*.
255. IUPAC. The IUPAC International Chemical Identifier. <http://www.iupac.org/inchi> (accessed Aug 07, 2009).
256. Stein, S. E.; Heller, S. R.; Tchekhovskoi, D. An Open Standard for Chemical Structure Representation: The IUPAC Chemical Identifier. *International Chemical Information Conference*, Nimes, 2003; pp 131-143.
257. Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, 32 (3), 244-255 (DOI: [10.1021/ci00007a012](https://doi.org/10.1021/ci00007a012)).
258. Symyx Technologies, Inc, San Ramon, CA. CTFile Formats. http://www.symyx.com/solutions/white_papers/ctfile_formats.jsp (accessed Nov 12, 2009).
259. Symyx Technologies, Inc, San Ramon, CA. Atom-atom maps. In *Metabolite Browser Help*.
260. Negishia, M.; Pedersenc, L. G.; Petrotchenkoa, E.; Shevtsova, S.; Gorokhovc, A.; Kakutad, Y.; Pedersen, L. C. Structure and Function of Sulfotransferases. *Arch Biochem Biophys* **2001**, 390 (2), 149-157 (DOI: [10.1006/abbi.2001.2368](https://doi.org/10.1006/abbi.2001.2368)).
261. McGregor, J. J.; Willett, P. Use of a maximum common subgraph algorithm in the automatic identification of ostensible bond changes occurring in chemical reactions. *J. Chem. Inf. Comput. Sci.* **1981**, 21 (3), 137-140 (DOI: [10.1021/ci00031a005](https://doi.org/10.1021/ci00031a005)).

262. Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem* **1996**, 39 (15), 2887–2893 (DOI: [10.1021/jm9602928](https://doi.org/10.1021/jm9602928)).
263. Huang, X.; Lai, J.; Jennings, S. F. Maximum common subgraph: some upper bound and lower bound results. *BMC Bioinformatics* **2006**, 7 (Suppl 4), S6 (DOI: [10.1186/1471-2105-7-S4-S6](https://doi.org/10.1186/1471-2105-7-S4-S6)).
264. Krissinel, E. B.; Henrick, K. Common subgraph isomorphism detection by backtracking search. *Software: Practice and Experience* **2004**, 34 (6), 591-607 (DOI: [10.1002/spe.588](https://doi.org/10.1002/spe.588)).
265. Raymond, J. W.; Willett, P. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput. Aided Mol. Des.* **2002**, 16 (7), 521-533 (DOI: [10.1023/A:1021271615909](https://doi.org/10.1023/A:1021271615909)).
266. Ullmann, J. R. An Algorithm for Subgraph Isomorphism. *Journal of the ACM* **1976**, 23 (1), 31-42 (DOI: [10.1145/321921.321925](https://doi.org/10.1145/321921.321925)).
267. Barrow, H. G.; Burstall, R. M. Subgraph isomorphism, matching relational structures and maximal cliques. *Inf. Proc. Lett.* **1976**, 4 (4), 83-84 (DOI: [10.1016/0020-0190\(76\)90049-1](https://doi.org/10.1016/0020-0190(76)90049-1)).
268. Bron, C.; Kerbosch, J. Algorithm 457: finding all cliques of an undirected graph. *Commun. ACM* **1973**, 16 (9), 575-577 (DOI: [10.1145/362342.362367](https://doi.org/10.1145/362342.362367)).
269. HashMap. <http://java.sun.com/j2se/1.5.0/docs/api/java/util/HashMap.html> (accessed Oct 17, 2009).
270. Gosling, J.; Joy, B.; Steele, G. In *The Java Language Specification*, First Edition ed.; Addison Wesley Publishing Company, 1996; §20.1.4.
271. JME Molecular Editor, courtesy of Peter Ertl, Novartis. <http://www.molinspiration.com/jme/> (accessed Sept 17, 2009).
272. Amazon Elastic Compute Cloud (EC2). <http://aws.amazon.com/ec2/> (accessed Sept 17, 2009).

273. Tomasic, M. G. CambridgeSoft Corporate Overview. *CambridgeSoft Conference and User Meeting*, London, March 9, 2009.
274. Spjuth, O.; Helmus, T.; Willighagen, E.; Kuhn, S.; Eklund, M.; Wagener, J.; Murray-Rust, P.; Steinbeck, C.; Wikberg, J. Bioclipse: an open source workbench for chemo- and bioinformatics. *BMC Bioinformatics* **2007**, *8* (1), 59 (DOI: [10.1186/1471-2105-8-59](https://doi.org/10.1186/1471-2105-8-59)).
275. Eclipse Project. <http://www.eclipse.org/> (accessed May 18, 2009).
276. Cruciani, G.; Carosati, E.; Boeck, B. D.; Ethirajulu, K.; Mackie, C.; Howe, T.; Vianello, R. MetaSite: Understanding Metabolism in Human Cytochromes from the Perspective of the Chemist. *J. Med. Chem.* **2005**, *48* (22), 6970-6979 (DOI: [10.1021/jm050529c](https://doi.org/10.1021/jm050529c)).
277. Swets, J. A. Measuring the accuracy of diagnostic systems. *Science* **1988**, *240* (4857), 1285-1293 (DOI: [10.1126/science.3287615](https://doi.org/10.1126/science.3287615)).
278. Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters* **2006**, *27* (8), 861-874 (DOI: [10.1016/j.patrec.2005.10.010](https://doi.org/10.1016/j.patrec.2005.10.010)).
279. Hanley, J. A.; McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29-36.
280. Nikolic, D.; van Breemen, R. B. New metabolic pathways for flavanones catalyzed by rat liver microsomes. *Drug Metab. Dispos.* **2004**, *32*, 387-397 (DOI: [10.1124/dmd.32.4.387](https://doi.org/10.1124/dmd.32.4.387)).
281. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* **1945**, *1* (6), 80-83.
282. Mann, H. B.; Whitney, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann. Math. Statist.* **1947**, *18* (1), 50-60.
283. Xu, Q.-S.; Liang, Y.-Z. Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems* **2001**, *56* (1), 1-11 (DOI: [10.1016/S0169-7439\(00\)00122-2](https://doi.org/10.1016/S0169-7439(00)00122-2)).
284. Zhu, M.; Whigan, D. B.; Chang, S. Y.; Dockens, R. C. Disposition and Metabolism of [14C]Brasofensine in Rats, Monkeys, and Humans. *Drug Metab. Dispos.* **2008**, *36* (1), 24-25 (DOI: [10.1124/dmd.107.016139](https://doi.org/10.1124/dmd.107.016139)).

285. Sargentini-Maier, M. L.; Espié, P.; Coquette, A.; Stockis, A. Pharmacokinetics and Metabolism of ¹⁴C-Brivaracetam, a Novel SV2A Ligand, in Healthy Subjects. *Drug Metab. Dispos.* **2008**, *36* (1), 35-45 (DOI: [10.1124/dmd.107.017129](https://doi.org/10.1124/dmd.107.017129)).
286. Musick, T. J.; Gohdes, M.; Duffy, A.; Erickson, D. A.; Krieter, P. A. Pharmacokinetics, Disposition, and Metabolism of Bicifadine in the Mouse, Rat, and Monkey. *Drug Metab. Dispos.* **2008**, *36* (2), 241-251 (DOI: [10.1124/dmd.107.017863](https://doi.org/10.1124/dmd.107.017863)).
287. Helsby, N. A.; Goldthorpe, M. A.; Tang, M. H. Y.; Atwell, G. J.; Smith, E. M.; Wilson, W. R.; Tingle, M. D. Influence of Mustard Group Structure on Pathways of in Vitro Metabolism of Anticancer N-(2-Hydroxyethyl)-3,5-dinitrobenzamide 2-Mustard Prodrugs. *Drug Metab. Dispos.* **2008**, *36* (2), 353-360 (DOI: [10.1124/dmd.107.018739](https://doi.org/10.1124/dmd.107.018739)).
288. Blech, S.; Ebner, T.; Ludwig-Schwellinger, E.; Stangier, J.; Roth, W. The Metabolism and Disposition of the Oral Direct Thrombin Inhibitor, Dabigatran, in Humans. *Drug Metab. Dispos.* **2008**, *36* (2), 386-399 (DOI: [10.1124/dmd.107.019083](https://doi.org/10.1124/dmd.107.019083)).
289. Yan, R.; Ko, N. L.; Li, S.-L.; Tam, Y. K.; Lin, G. Pharmacokinetics and Metabolism of Ligustilide, a Major Bioactive Component in Rhizoma Chuanxiong, in the Rat. *Drug Metab. Dispos.* **2008**, *36* (2), 400-408 (DOI: [10.1124/dmd.107.017707](https://doi.org/10.1124/dmd.107.017707)).
290. Deo, A. K.; Bandiera, S. M. Biotransformation of Lithocholic Acid by Rat Hepatic Microsomes: Metabolite Analysis by Liquid Chromatography/Mass Spectrometry. *Drug Metab. Dispos.* **2008**, *36* (2), 442-451 (DOI: [10.1124/dmd.107.017533](https://doi.org/10.1124/dmd.107.017533)).
291. Kotsuma, M.; Tokui, T.; Ishizuka-Ozeki, T.; Honda, T.; Iwabuchi, H.; Murai, T.; Ikeda, T.; Saji, H. CYP2D6-Mediated Metabolism of a Novel Acyl Coenzyme A:Cholesterol Acyltransferase Inhibitor, Pactimibe, and Its Unique Plasma Metabolite, R-125528. *Drug Metab. Dispos.* **2008**, *36* (3), 529-534 (DOI: [10.1124/dmd.107.018853](https://doi.org/10.1124/dmd.107.018853)).
292. McClue, S. J.; Stuart, I. Metabolism of the Trisubstituted Purine Cyclin-Dependent Kinase Inhibitor Seliciclib (R-Roscovitrine) in Vitro and in Vivo. *Drug Metab. Dispos.* **2008**, *36* (3), 561-570 (DOI: [10.1124/dmd.107.019232](https://doi.org/10.1124/dmd.107.019232)).

293. Kuuranne, T.; Leinonen, A.; Schänzer, W.; Kamber, M.; Kostianen, R.; Thevis, M. Aryl-Propionamide-Derived Selective Androgen Receptor Modulators: Liquid Chromatography-Tandem Mass Spectrometry Characterization of the in Vitro Synthesized Metabolites for Doping Control Purposes. *Drug Metab. Dispos.* **2008**, *36* (3), 571-581 (DOI: [10.1124/dmd.107.017954](https://doi.org/10.1124/dmd.107.017954)).
294. Xu, L.; Adams, B.; Jeliaskova-Mecheva, V. V.; Trimble, L.; Kwei, G.; Harsch, A. Identification of Novel Metabolites of Colchicine in Rat Bile Facilitated by Enhanced Online Radiometric Detection. *Drug Metab. Dispos.* **2008**, *36* (4), 731-739 (DOI: [10.1124/dmd.107.019463](https://doi.org/10.1124/dmd.107.019463)).
295. Minato, K.; Suzuki, R.; Asagarasu, A.; Matsui, T.; Sato, M. Biotransformation of 3-Amino-5,6,7,8-tetrahydro-2-(4-[4-(quinolin-2-yl)piperazin-1-yl]butyl)quinazolin-4(3H)-one (TZB-30878), a Novel 5-Hydroxytryptamine (5-HT)_{1A} Agonist/5-HT₃ Antagonist, in Human Hepatic Cytochrome P450 Enzymes. *Drug Metab. Dispos.* **2008**, *36* (5), 831-840 (DOI: [10.1124/dmd.107.018168](https://doi.org/10.1124/dmd.107.018168)).
296. Prakash, C.; Johnson, K. A.; Gardner, M. J. Disposition of Lasofoxifene, a Next-Generation Selective Estrogen Receptor Modulator, in Healthy Male Subjects. *Drug Metab. Dispos.* **2008**, *36* (7), 1218-1226 (DOI: [10.1124/dmd.108.020404](https://doi.org/10.1124/dmd.108.020404)).
297. Deo, A. K.; Bandiera, S. M. Identification of Human Hepatic Cytochrome P450 Enzymes Involved in the Biotransformation of Cholic and Chenodeoxycholic Acid. *Drug Metab. Dispos.* **2008**, *36* (10), 1983-1991 (DOI: [10.1124/dmd.108.022194](https://doi.org/10.1124/dmd.108.022194)).
298. Prakash, C.; Chen, W.; Rossulek, M.; Johnson, K.; Zhang, C.; O'Connell, T.; Potchoiba, M.; Dalvie, D. Metabolism, Pharmacokinetics, and Excretion of a Cholesteryl Ester Transfer Protein Inhibitor, Torcetrapib, in Rats, Monkeys, and Mice: Characterization of Unusual and Novel Metabolites by High-Resolution Liquid Chromatography-Tandem Mass Spectrometry and ¹H Nuclear Magnetic Resonance. *Drug Metab. Dispos.* **2008**, *36* (10), 2064-2079 (DOI: [10.1124/dmd.108.022277](https://doi.org/10.1124/dmd.108.022277)).