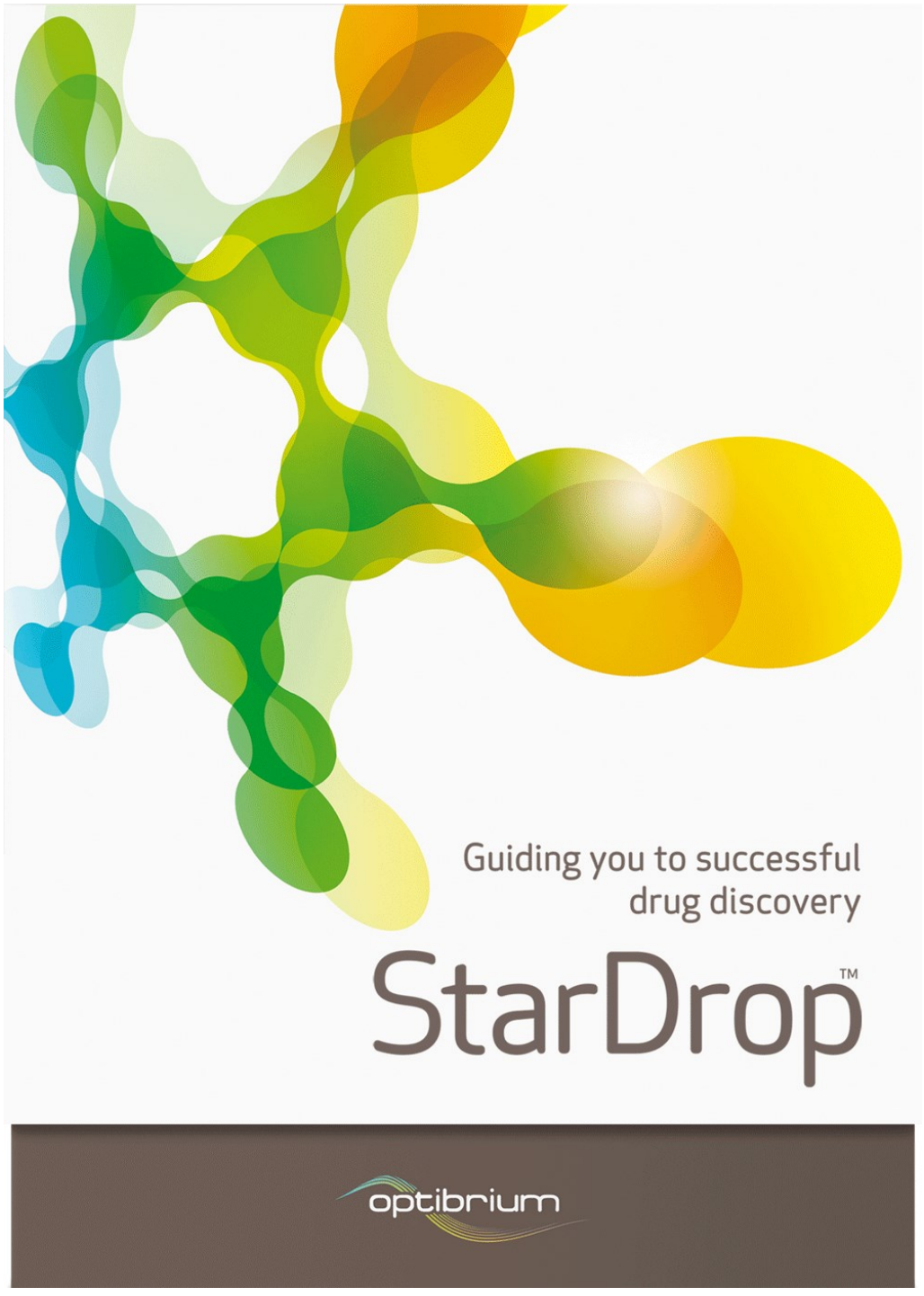
StarDrop™ Reference Guide

Version 6.3



Contents

1 Introduction

A successful, safe and efficacious drug must achieve an exquisite balance of many requirements related to its biological and physicochemical properties. In addition to having the required pharmacological activity against its intended target, it must reach the appropriate site in the body at a high enough concentration and for a sufficient period of time to be therapeutically beneficial. In order to achieve this, the molecule must overcome a variety of physiological barriers. For an oral drug to be effective against a target located within the central nervous system (CNS), the swallowed drug must disperse as it passes from the mouth, through the oesophagus and stomach, and be in solution for absorption from the intestine. It will need to be stable to gastric enzymes and the acid environment of the stomach and not be degraded by bacteria in the gut or enzymes in the gut wall. Even if it has suitable physicochemical properties to cross the gut wall into the bloodstream, this may be prevented or attenuated by active transport proteins in the gut epithelium. Once in the bloodstream, a sufficient amount of drug will need to remain free in circulation, avoiding breakdown by metabolic enzymes and excretion by the liver and kidneys, to permit an effective proportion of the dose to cross the epithelium of capillaries in the brain and reach the desired receptors in the CNS. The drug's total or partial failure at any of these steps could result in reduced or negligible efficacy.

These complex, often conflicting requirements, have led to the generation of an increasing volume of data on a wide range of properties from the early stages of drug discovery, with increasingly detailed studies performed as compounds progress through lead optimisation to candidate selection. The complexity of these data, combined with uncertainties due to experimental variability or predictive error, make it difficult to decide with confidence which lines of enquiry to pursue and which compounds to prioritise. StarDrop helps to guide these key decisions on the design and selection of high quality compounds with an increased chance of success downstream. It intuitively evaluates all of the available predicted or experimental data, rigorously taking into account the underlying uncertainties, to identify those compounds most likely to meet the property profile requirement specified for a project. This information can be explored interactively using data visualisation tools, including a view of your project's chemical space, and StarDrop's unique Card View, to identify trends and select compounds to quickly focus on the highest quality chemistries whilst mitigating risk by selecting a diverse set of compounds.

StarDrop can also help to guide the redesign of compounds to improve their properties. Easy-to-use, yet powerful, tools for R-group, molecular matched pair and activity cliff analyses, coupled with Card View help to quickly find important structure-activity relationships (SAR) in your data and identify new optimisation strategies. These new ideas can be explored in StarDrop's interactive designer, with instant feedback on their predicted properties and guided by StarDrop's Glowing Molecule that highlights regions of molecules that have a strong influence on a predicted property. The search for high quality compounds can be further stimulated by StarDrop's Nova module that can generate large numbers of relevant new compound ideas, prioritised against a project's required property profile.

1.1 StarDrop Overview

StarDrop is a platform designed to support effective decision-making in drug discovery chemistry through the prediction, analysis and visualisation of compound properties, enabling efficient optimisation and selection of compounds for an optimal balance of properties appropriate for the therapeutic goals of a project.

1.1.1 Core Capabilities

The core capabilities of StarDrop help drug discovery teams to quickly identify high quality compounds based on all available data, whether predicted or experimental. It guides key decisions on the selection and design of compounds, despite the complexity of the goals and inherent uncertainty in the underlying data, to enable confident decisions on the direction of future investigation.

Based on the available data, it should be possible to assess the probability that a compound will succeed, based on knowledge of previous failed or successful molecules. An overall likelihood of success across all properties can then be derived and it becomes possible to prioritise compounds such that resources are focused on those having the highest potential to be successful. This prioritisation is

particularly applicable at the early 'Discovery' stage of the R&D process, where typically very large numbers of molecules are assessed and many fail due to an inappropriate balance of properties.

To achieve this, StarDrop implements a proprietary probabilistic scoring algorithm that can assess the relative likelihood of success of compounds for a given target. Using the probabilistic scoring method, your project team can define the ideal profile of properties for a successful drug candidate. The scoring method assesses all of the available data, taking into account uncertainties in experimental or predicted data, to estimate the likelihood of success of each compound against your specified success criteria, enabling compounds to be rigorously and objectively prioritised. More details on this scoring algorithm and its application are given in Chapter 2 'Probabilistic Scoring'.

StarDrop provides a comprehensive, interactive environment for visualisation of your compound data to help you to quickly identify trends in your data. In particular, StarDrop's chemical space visualisation displays the distributions of properties or scored across the chemical diversity you are exploring in your project. This can also guide selection of chemotypes or individual compounds, balancing quality of compound with diversity to investigate and spread risk across a range of chemistries. This is discussed in more detail in Chapter 3 'Chemical Space and Compound Selection.'

Interpretation of your compound data can be assisted by several, powerful tools that analyse your data to help you to spot important trends or SAR within your data. These include a flexible tool for R-group decomposition and algorithms for clustering, molecular matched pair analysis and activity cliff detection. Details of these algorithms are provided in Chapter 5.

The results of algorithms such as those described above can often be difficult to interpret. However, StarDrop's unique Card View provides an easy way to understand and interact with their output, making the results clear and interpretable. Card View also provides you with the flexibility to view and manipulate your compound data in the way that you think, capturing the relationships between compounds and patterns in the data.

Of course, exploring the data for compounds that you have already considered is important. But, commonly, early compounds in hit-to-lead and lead optimisation will require further optimisation, so it is necessary to explore strategies to redesign compounds and improve their properties. StarDrop's interactive designer enables you to consider ideas for new compounds with instant feedback on the predicted changes in properties. StarDrop's Glowing Molecule goes further, to use the information captured by predictive models about the relationship between the structure of compounds and their properties and highlight key regions on a molecule that have the strongest influence on a predicted property. This helps to guide the redesign of compounds by targeting the regions and modifications that are most likely to improve a property. More details on interpretation of StarDrop's Glowing Molecule can be found in Chapter 4.

All of these features are integrated in a user-friendly, intuitive environment that includes a comprehensive range of tools to help you to manage your project data, including merging, searching and filtering data sets and the ability to perform arbitrary mathematical calculations and transformations. Furthermore, the results of your analysis and design can be easily presented and shared with colleagues by simply copying visualisations from StarDrop into your presentations and reports.

The core of StarDrop can be enhanced by a range of optional plug-in modules, summarised below, and may be further extended through a range of application program interfaces (APIs) to integrate in-house predictive models, algorithms and databases. Details of StarDrop's APIs can be found in the separate Scripting and Customisation Guide.

1.1.2 Plug-in Modules

ADME QSAR

Prior to synthesis, the only sources of data for virtual compounds are predictive *in silico* models. To aid in the design of compounds prior to synthesis, StarDrop ADME QSAR module provides a suite of high quality predictive ADME models that can be used to explore a wide range of chemistry options prior to selecting a synthetic strategy. The available models and underlying methods are described in Chapter 6.

All the StarDrop models give an indication of confidence in the prediction based upon a predicted molecule's proximity to the chemical space of the model. Additionally, all the StarDrop models provide a 'Glowing Molecule' visualisation of each result, indicating the parts of the molecule having the greatest influence on the prediction (Chapter 4).

In addition to the StarDrop QSAR models a number of molecule properties are also available:

- Molecular Weight
- Rotatable bonds
- Flexibility – ratio of rotatable bonds to total bonds
- Hydrogen bond donors
- Hydrogen bond acceptors
- Topological polar surface area (based upon oxygen and nitrogen)

The topological polar surface area is based upon the description by Ertl (Ertl, Rhodes, & Selzer, 2000).

P450

StarDrop's Cytochrome P450 metabolism models predict the regioselectivity of metabolism by the key drug metabolising isoforms of P450. Based on quantum mechanical simulations, these are more computationally intensive than QSAR models but provide detailed results identifying both the sites of metabolism by P450 enzyme and the site lability, which indicates the vulnerability of each site. This additional information provides important data to guide the redesign of compounds and improve their metabolic stability. The theory underlying these predictions is described in Chapter 7.

Auto-Modeller™

The Auto-Modeller provides an environment in which to build robust, 'local' QSAR models, tailored to specific chemistry or data, which can be used alongside or in place of the StarDrop models. The StarDrop Auto-Modeller has been designed to enable non-computational scientists to easily apply rigorous modelling techniques, while experienced modellers can control the modelling process in detail. The methods underlying the Auto-Modeller are described in detail in Chapter 8.

MPO Explorer™

The MPO Explorer module helps you develop multi-parameter optimisation strategies, enabling you to find multi-parameter scoring profiles for your project objectives, based on historic data, to optimally select successful compounds. In addition, MPO Explorer enables you to carry out sensitivity analyses on your scoring profiles, enabling you to test the robustness of your decisions to the selection criteria you have chosen. The methods underlying this approach are described in Chapter 9.

Nova™

Nova helps to quickly explore a broad range of chemistry to guide optimisation strategy and stimulate the search for high quality compounds. It achieves this through three, complementary approaches:

'Idea generation' generates new compound structures by applying established medicinal chemistry 'transformation rules' to an initial compound. You can control the generation of new compound ideas and the ideas can be automatically prioritised according to a predicted property, probabilistic score or chemical diversity. Nova can aid the rigorous exploration of chemistry around early hits, to identify those hits most likely to yield high-quality lead series; help to find strategies to overcome problems with compound properties in lead optimisation; and identify patent protection strategies or patent busting opportunities by expanding the chemistry around existing development candidates or drugs to search for compounds with improved properties. Chapter 10 describes the methods used to generate new compound ideas and the validation of the underlying transformations.

'Matched series analysis' provides an alternative approach for suggesting new compound ideas. By comparing matched series found in your data with a database of other matched series (a knowledge base), relevant predictions for new substituents that are likely to improve target activity or another property of interest can be made. The suggestions are based on the premise that a matched series with similar activity order in your data and the knowledge base implies that those groups occupy a similar binding environment created by their target proteins. Given a similar binding environment, groups that have been shown to be better binders within the knowledge base, have a strong likelihood of being

better binders to the target of the input data set. Chapter 10 describes two approaches using matched series analysis that are available within the Nova module.

These are complemented by a flexible virtual library enumeration tool that enables you to define the specific chemistry to be explored by drawing a template with substitution points and listing the modifications or substituents at each position. More details on how to define a virtual library in Nova can be found in Chapter 12 of the StarDrop User Guide.

BIOSTER™

The BIOSTER database (Digital Chemistry, n.d.) is a compilation of 23,917 transformations, corresponding to practical structural modification and replacement techniques, manually curated from the scientific literature. Available as an additional plug-in through Nova, this combination enables you to quickly and easily search the comprehensive BIOSTER database to identify transformations that are relevant to your compounds. Within Nova these can be automatically applied to generate novel structures with a high likelihood of biological activity and synthetic accessibility, prioritised against the property profile you require for your project. References to the primary literature from which each transformation was derived are provided, facilitating further chemical and biological validation of the new ideas. Details on the BIOSTER database and the generation of the corresponding transformations can be found in Chapter 11.

torch3D™

torch3D is a molecular design and SAR interpretation tool, developed by Cresset (Cresset, n.d.), which uses molecular alignment to a reference molecule in a predefined conformation as a way to make meaningful comparisons across chemical series. When used on a congeneric series the tool can help in library design and give a rationale for the prioritisation of compounds for synthesis. Using torch3D on a diverse set of active molecules can help define the requirements of the protein of interest, aiding the synthetic chemist in the design of new actives. Furthermore, torch3D can be applied as a virtual screening tool to identify structurally novel active compounds that are likely to exhibit similar 3D SAR to a known active. See Chapter 12 for more details on the methodology underlying torch3D.

Derek Nexus™

Toxicity of drug candidates is a major cause of expensive, late-stage failure in pre-clinical and clinical development. The Derek Nexus module for StarDrop provides Lhasa Limited's (Lhasa, n.d.) world-leading technology for knowledge-based prediction of key toxicities. Using data from published and donated (unpublished) sources, Derek Nexus identifies structure-toxicity relationships that alert you to the potential for your compounds to cause toxicity. The Derek Nexus module provides predictions of the likelihood of a compound causing toxicity in over 40 endpoints, including mutagenicity, hepatotoxicity and cardiotoxicity. Derek Nexus is seamlessly integrated with StarDrop's interactive designer and Glowing Molecule visualisation, to guide the redesign of compounds and reduce the potential for toxicity. The endpoints and results predicted by Derek Nexus are described in Chapter 13.

1.2 Reference Guide Summary

This Reference Guide describes the concepts underlying StarDrop and gives a number of case studies illustrating applications at different stages of drug discovery. The accompanying User Guide provides information on getting started using StarDrop and a description of the user interface.

Chapter 2, 'Probabilistic Scoring', contains an introduction to the scoring algorithm and a guide to defining scoring schemes to reflect a project's criteria for success.

Chapter 3, 'Chemical Space and Compound Selection', describes the underlying concepts and algorithms for the visualisation of chemical space and selection of compounds to balance quality and diversity.

Chapter 4, 'Glowing Molecule', explains how to further interpret the StarDrop models using the visual feedback provided indicating which parts of the molecule are influencing the prediction.

Chapter 5, 'Cheminformatics Algorithms' describes the algorithms provided by StarDrop to help with the analysis of compound data and SAR.

Chapter 6, ‘ADME QSAR Models’, provides an overview of the modelling techniques used to build the ADME QSAR models.

Chapter 7, ‘P450 Metabolism Models’, provides an overview of the science underlying the P450 metabolism models and explains how to interpret the results.

Chapter 8, ‘Auto-Modeller’, describes the techniques used in the automated model generation process.

Chapter 9, 'MPO Explorer', describes the methods that can be used to generate scoring profiles from your own data sets and analyse the robustness of your decisions to the selection criteria you have chosen.

Chapter 10, ‘Nova’, describes three approaches for generating new compounds: ‘Idea Generation’ generates and prioritises relevant new compound structures using medicinal chemistry ‘transformation rules’ and their validation, ‘Matched Series Analysis’ describes two methods for suggesting chemical substitutions that are likely to improve a compound property.

Chapter 11, ‘BIOSTER’, describes the BIOSTER database of compound transformations from Digital Chemistry.

Chapter 12, ‘torch3D’, describes the molecular design and SAR analysis technology developed by Cresset.

Chapter 13, ‘Derek Nexus’, describes the knowledge-based toxicity prediction methods developed by Lhasa Limited.

Chapter 14, 'Example Applications', provides some illustrative example applications in drug discovery projects.

Chapter 15, ‘Appendices’, provides additional, detailed information to support the information in previous chapters.

- 'ADME Models Reference' providing overviews of each model, including guidance on interpretation of their results
- ‘P450 metabolism validation results’ provides the detailed validation results of the P450 metabolism models
- ‘Descriptors’ provides the descriptors used by the Auto-Modeller and in the generation of the ADME QSAR models
- ‘Results of Lead to Drug Transformations for Nova Validation’ provides the detailed results of the validation of the transformations employed by the Nova module
- ‘File Formats’ defines the file formats for import of SMARTS and SMIRKS to define additional descriptors, filters and transformations.

2 Probabilistic Scoring

The probabilistic scoring algorithm enables prioritisation of compounds with an appropriate balance of properties to meet a project's objective. Compound scores are estimates of their likelihood of success, i.e. the likelihood that the compound will meet the project's criteria for the properties considered. As this assessment is based on the combination of multiple predictions (or measurements), each with an associated statistical uncertainty, it is essential to include the impact of this uncertainty when calculating the score. An estimate of the uncertainty in the score is also calculated.

The probabilistic scoring algorithm can be applied to any compound data, whether predicted or experimentally measured. Example applications of probabilistic scoring are given in Chapter 14.

2.1 Defining Scoring Criteria

The scoring criteria mathematically define the property profile that is required of a successful compound. Once defined, this enables all compounds and chemical series to be objectively scored against this profile within a rigorous framework, providing a strong basis for decision making.

At a qualitative level, a scoring profile can be defined simply as a set of property criteria (maximum or minimum acceptable property values or preferred category) and an importance associated with each criterion, as shown in Figure 2.1. Example scoring profile. The properties included in the profile are shown in the right-hand column. The middle column shows the criterion for each property and the sliders to the right indicate the importance of each criterion on a scale from 0 to 1. The importance value is a number between 0 and 1 that indicates how important it is that the criterion is achieved; an importance of 1 means that it is critical and that a compound that failed to meet the criterion would be rejected outright, whilst a low importance means that failure to meet the criterion would not be a major issue.

A scoring profile can be defined in a more quantitative manner that also enables more subtle criteria than a simple cut-off to be defined. The method for defining the scoring criterion for a property in this way will depend on whether the results for that property are values on a continuous scale, or classifications.

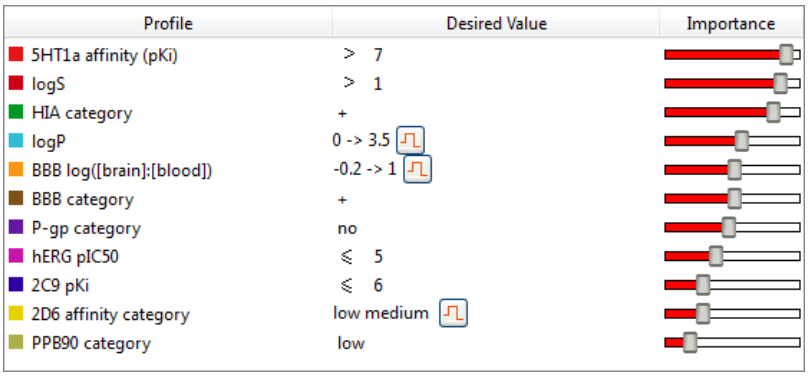


Figure 2.1. Example scoring profile. The properties included in the profile are shown in the right-hand column. The middle column shows the criterion for each property and the sliders to the right indicate the importance of each criterion on a scale from 0 to 1.

2.1.1 Continuous Values

The scoring criterion for a property with values on a continuous scale is defined in terms of a *scoring function*. The simplest example of it is a *threshold function*, as illustrated in Figure 2.2. This is defined by a *threshold* value, $X_{threshold}$, representing the value separating compounds deemed to meet the target profile from those with an inadequate property value. Two scores are defined, S_{above}, representing the score for compounds with property values exceeding the threshold and S_{below} for property values below the threshold value. Both score values must be between 0 and 1 and the score for property values on the desirable side of the threshold will generally be higher than the score for property values on the undesirable side.

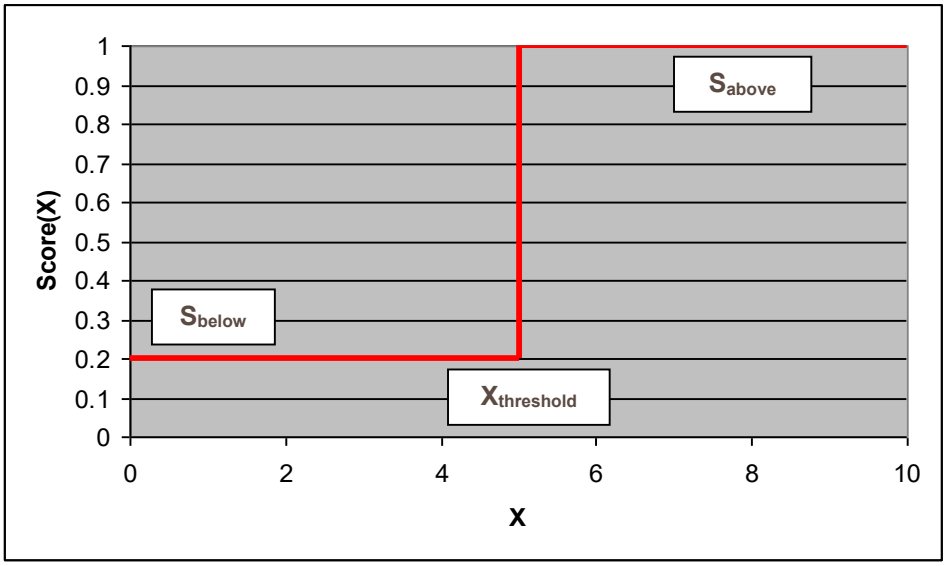


Figure 2.2 An example threshold function used to define the success criterion for a property, denoted X. The desired threshold value for the property, the score for property values exceeding this threshold and the score for the property values below this threshold are defined; $X_{threshold}$, S_{above} and S_{below} respectively.

The values of the scores above and below the threshold should reflect the likelihood of success of compounds with property values in these ranges. Commonly, the score on the desired side of the threshold is given a score of 1 because, however important the property, having the desired result will never be a problem for the compound's profile. For undesired property values, the more important properties are penalized more heavily, as a poor outcome will be associated with a higher risk and hence a greater impact on a compound's chance of success. Therefore, the importance of a property criterion is given by the difference between the ideal score of 1 and the lowest possible score for a property.

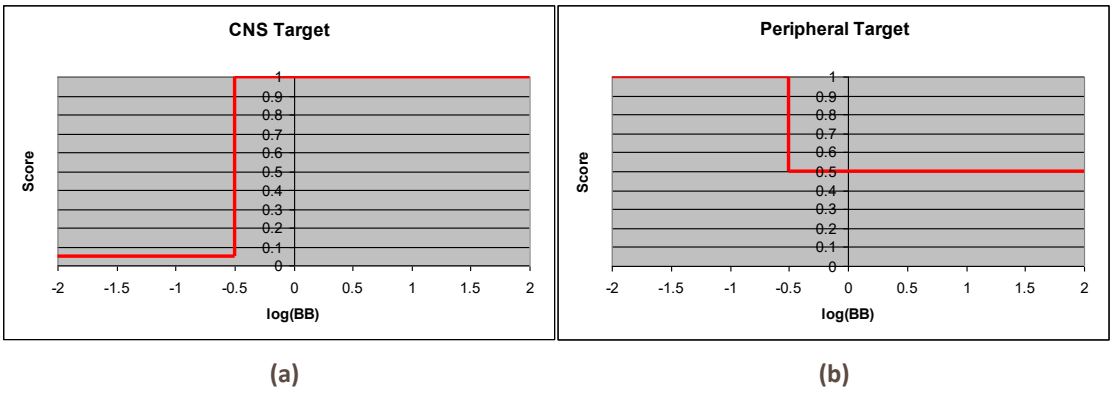


Figure 2.3 Example scoring criteria for CNS penetration, log(BB), for compounds intended for a CNS target (a) and a peripheral target (b).

As an example, consider possible scoring criteria for CNS penetration for a project with a CNS target, versus one pursuing a peripheral target (see Figure 2.3). For a drug having its effect directly within the

CNS, penetration into the brain is essential and a very low predicted brain/blood ratio would warrant a low score (approaching 0), as the chances of achieving efficacy would be very low. Conversely, a compound with a peripheral target would, ideally, not penetrate the blood-brain barrier (BBB), avoiding potential CNS side effects. However, if the compound were to penetrate the BBB, this would not significantly diminish the efficacy of the compound, although it would introduce an additional risk of side effects. Hence, the lower asymptote of the success function would be greater than zero, but still less than one. In setting the scoring function value if a desired property value is not achieved, it may help to consider the scale in percentage terms. Hence, in the example above, the project's assessment is that a compound intended for a peripheral target, which is ideal in all other properties, has a 50% chance of failing to become a successful drug (score of 0.5 if the desired value is not achieved) if a significant concentration gets into the brain.

A single threshold function is the simplest case, however, and StarDrop allows you to create scoring functions with multiple thresholds and functions combining constant and linear parts. Some examples of possible functions are given in Figure 2.4.

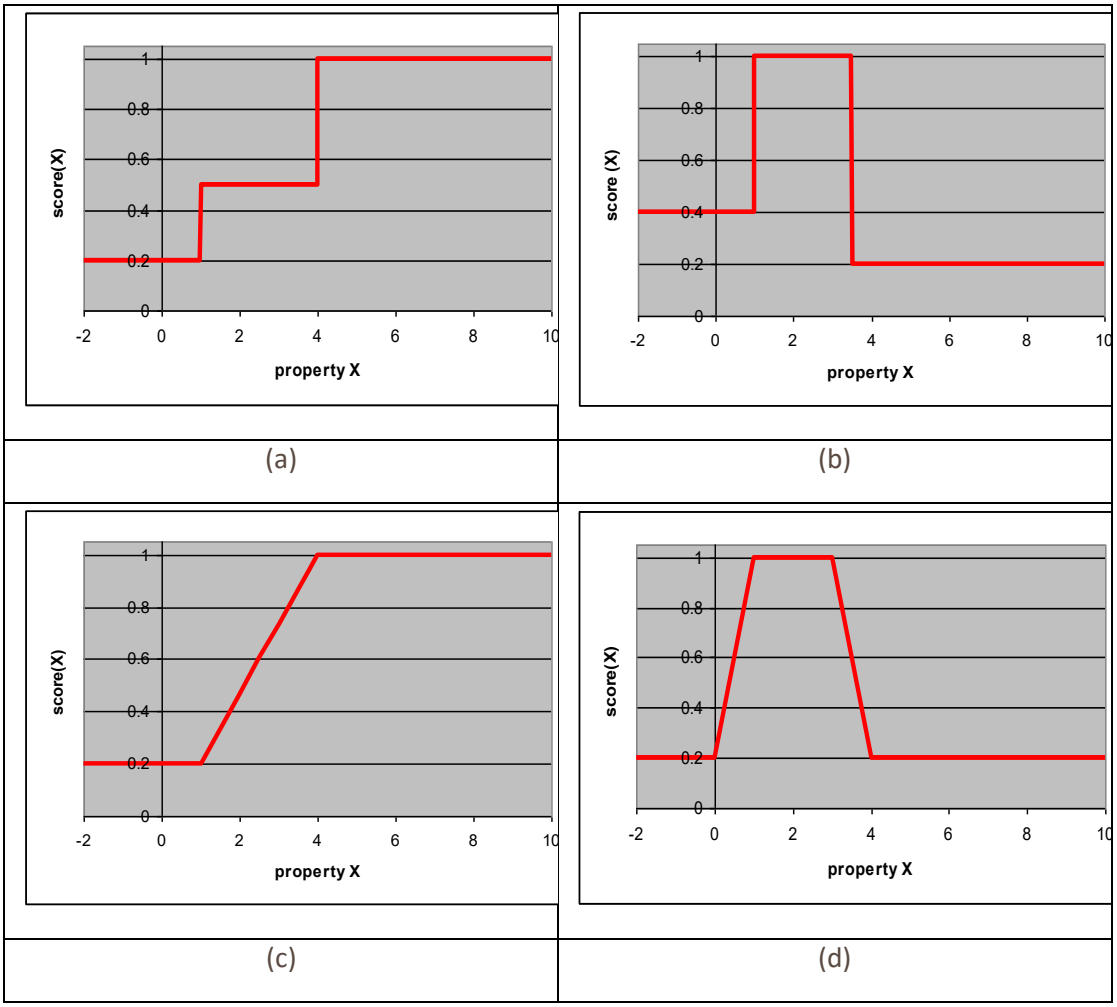


Figure 2.4 Examples of scoring functions: two-threshold function (a), 'band' function (b), piecewise linear function (c) and 'trapezoid' function (d).

2.1.2 Category Values

The success criterion for a property with values returned as a category is defined by a score value between 0 and 1 for each potential category. Thus, for human intestinal absorption (HIA), two values must be defined, S_{high} and S_{low}. Figure 2.5 illustrates possible score values for the HIA scoring criterion for a project with the goal of an orally administered compound.

The scores for desirable classes should be higher than those for undesirable classes. Commonly, the score for the most desirable class will be 1. However, in some cases this may be lowered to reflect the fact that all property values represented by a class may not be ideal. For example, in the case of the

HIA model where a classification of High represents >=30% absorption, a project where >50% is required might only give a score of 0.9 for High classifications to represent the uncertainty that this meets the criterion. In this case, it is worth noting that no compounds could achieve an overall score greater than 0.9, as this is the best possible outcome.

The scores for undesirable classes should reflect the impact of a property value in this class on the likelihood of meeting the required property profile for the project. Therefore, more important properties should receive lower scores for undesirable classes.

2.1.3 Example Scoring Profile

Consider the example of a project, the goal of which is the development of an oncology therapy for a non-CNS tumour; the current lead compound having an *in vitro* IC_{50} of 50nM against the biochemical

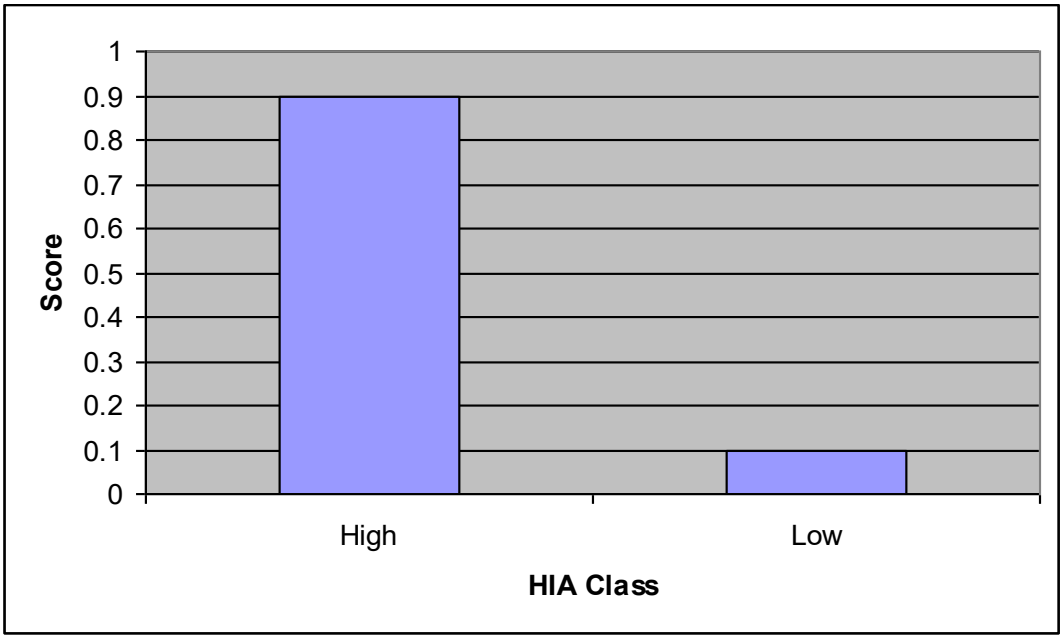


Figure 2.5 Example scoring criterion for human intestinal absorption for a project requiring oral administration. The high class represents >=30% absorption and the low class represents <30% absorption.

target. Ideally, the clinical candidate would be an orally administered drug; however, in this case IV administration is considered a viable alternative. The scoring profile used for this analysis was:

Table 1 Example Scoring Profile. In this example the scoring functions are single-threshold functions. For each property the score equals 1 if the desired value achieved.

Property	Desired Value	Scores if desired value is not achieved	Corresponding importance
Solubility (logS)	>2.0	0.2	0.8
HIA	+	0.3	0.7
P-gp transport	No	0.6	0.4
hERG Affinity (pIC_{50})	<6.3	0.7	0.3
BBB penetration category	-	0.8	0.2

CYP2D6 affinity category	Low, medium	High: 0.9 Very High: 0.8	0.2
CYP2C9 affinity (pK_i)	<6.0	0.9	0.1

The rationale for this profile was as follows:

Although the ideal candidate would be orally bioavailable, IV administration is a viable option for the target therapy. Whilst slow IV infusion of poorly soluble drugs is often accepted in chemotherapy, it is not ideal. Hence, good aqueous solubility is considered important to maximize the chance of getting oral bioavailability and minimize the volume needed for an IV dose. Consequently, solubility below 100µM was considered to have the greatest negative impact on the chance of a compound's success.

Ideally, the compound would be orally absorbed; therefore the next most important property is human intestinal absorption.

As the target is considered to be peripheral, P-gp transport would not limit efficacy due to efflux across the blood-brain barrier. Active efflux via this protein in the gut may limit oral bioavailability somewhat, particularly if absorption from the intestine is slow. However, more importantly in this therapeutic area, being a substrate for P-gp (MDR1) represents an additional risk for compounds due to the possible up-regulation of P-gp and development of drug resistance on prolonged dosing.

For a peripheral target, ideally the compound should not penetrate the blood-brain barrier, to avoid the potential for CNS side effects. Whilst such side effects are frequently seen with chemotherapeutics, it is anticipated that for an acute, fatal condition, the risk of these effects would be outweighed by the benefits of the treatment. Hence, penetration of the blood-brain barrier has not been considered to have a large negative effect on the compound's likelihood of success.

Similarly, for the target treatment, the potential for drug-drug interactions is not a primary concern. Therefore, the probability of an individual compound failing due to a high CYP2D6 or CYP2C9 is low. Nominally, a pKi greater than 6 (1 µM) would indicate a significant risk of drug-drug interactions.

Ideally a selectivity of greater that 100-fold over the hERG potassium ion channel would be sought, as inhibition of this channel is associated with the onset of Toursade-de-Pointes, a potentially lethal side-effect. As the intended therapy would be administered under medical supervision for a fatal disease, this threshold has been reduced to 10-fold. However, as patients are likely to be on a number of other drugs and their organ functions potentially disrupted, particularly the main clearance organs of liver and kidneys, hERG interaction could represent a threat to the successful compound.

2.2 Importance of Uncertainty

The probabilistic scoring algorithm automatically takes into consideration the statistical uncertainty in the data provided. The following examples illustrate the importance of this in prioritising compounds.

Figure 2.6 illustrates an example of three compounds A, B and C, for which a single property 'X' has been measured or predicted, yielding values of 5.5, 4.5 and 2.0 respectively. If one were to apply a filter to these compounds, with a threshold value of X=5, compound A would be selected for progression and compounds B and C rejected. However, if the assessment of property values had a standard error of 1, as illustrated by the probability distributions in Figure 2.6, our view of these results would be different. It would still be safe to 'reject' compound C, as there is very little probability that it exceeds the required threshold. However, we cannot confidently distinguish compounds A and B, as both have a significant chance that their true property value exceeds the required value.

In cases where the confidences in estimates of a property value vary between compounds, the importance of considering uncertainty is even more significant. For example, consider the illustration in Figure 2.7, showing two compounds, D and E with best estimates of property 'X' of 4.0 and 3.5 respectively. Both of these compounds would be rejected on the basis of a filter applied with a threshold of 5. However, in this case, the uncertainty in the estimate for compound E is significantly

higher than that for compound D. Here, the likelihood that the true property value exceeds the required threshold is higher for compound E than for compound D and, in the absence of alternatives, priority should be given to compound E.

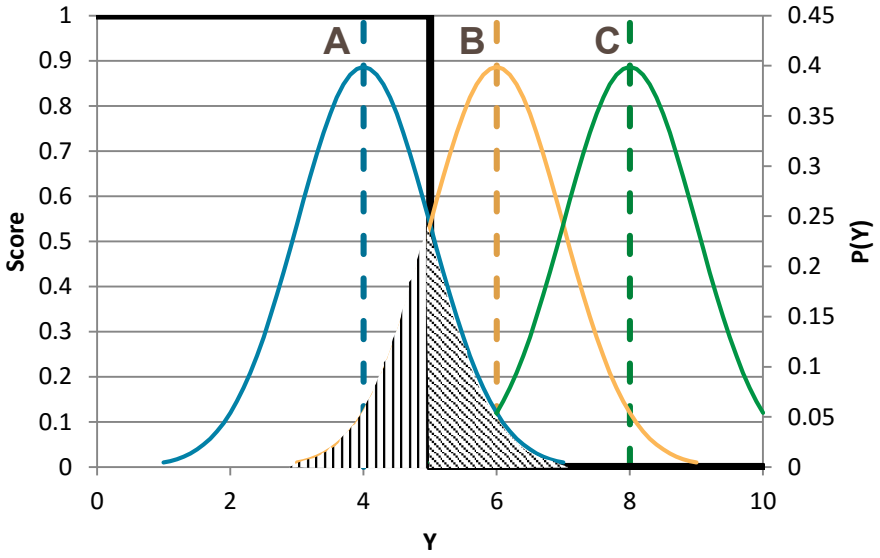


Figure 2.6 Illustration of the importance of uncertainty when selecting compounds. This shows a scoring function (bold line) corresponding to a simple filter with a criterion of <5. The dashed vertical lines indicate values of property Y for compounds labelled A, B and C. The uncertainties in these property values are illustrated by the coloured bell curves (Gaussian distributions) centred on each compound's property value. If we were to ignore the uncertainties in the property values, compound A would be accepted and B and C would be rejected. However, considering the uncertainties, we can see that, while the probability of compound C achieving the criterion is negligible, there is a significant probability (vertically hatched area), that compound B will meet the criterion and there is an equal probability (diagonally hatched area), that compound A will not meet the criterion.

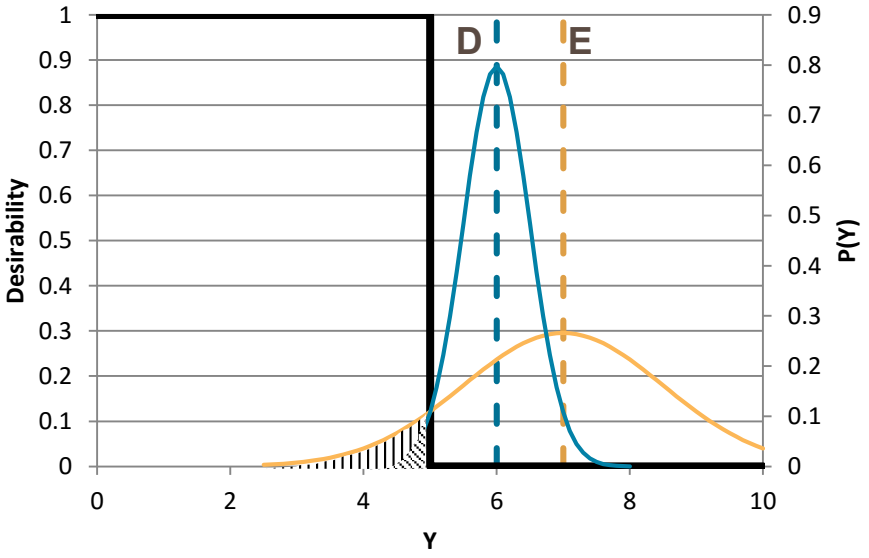


Figure 2.7 Illustration of the importance of uncertainty when selecting compounds. This shows a scoring function (bold line) corresponding to a simple filter with a criterion of <5. The dashed vertical lines indicate values of property Y for compounds labelled D and E. The uncertainties in these property values are illustrated by the coloured bell curves (Gaussian distributions) centred on each compound's property value. Here we can see that the values for both compounds D and E fail to meet the criterion. However, taking the uncertainties into account we can see that, even though the value for D is closer to the criterion than E, the probability of compound E meeting the criterion (the vertically hatched area) is actually greater than that for compound D (the diagonally hatched area).

2.3 Interpreting Scores

Two numbers are generated for a compound scored against a scoring profile:

Score: The best estimate of the likelihood of success of a compound against the scoring criteria.

Standard deviation: A measure of the uncertainty in this estimate.

In general, the 'quality' of two compounds should be compared using the scores for those compounds. However, as a rule of thumb, one cannot confidently differentiate between those compounds unless their scores differ by more than the sum of their standard deviations. One approach to visualising compounds that can be confidently differentiated is to plot a graph as shown in Figure 2.8.

In this figure the compounds are plotted in order of decreasing **score** along the x-axis and their **scores**, with error bars illustrating the **standard deviation** of the scores, on the y-axis. From this, we can see that there are groups of compounds that can be confidently separated in terms or their likelihood of success, but within these groups the available data does not distinguish between compounds.

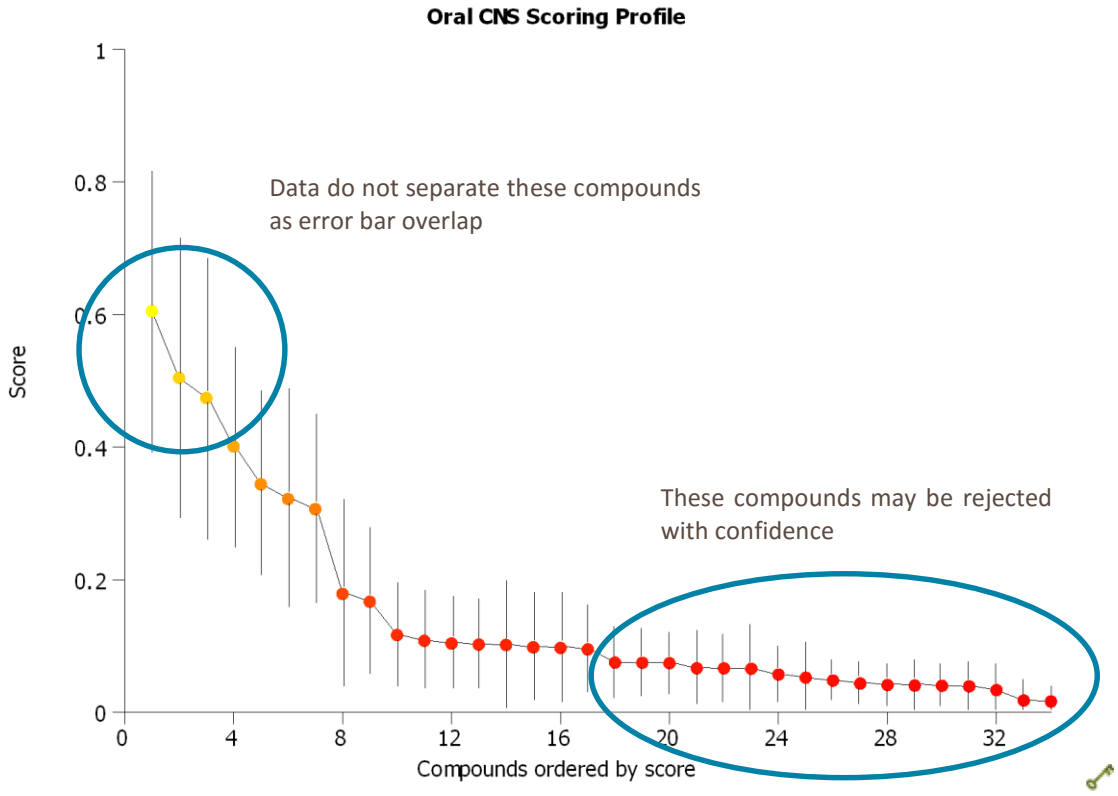


Figure 2.8 A 'snake' plot of compound scores, with error bars showing the uncertainty in each score.

2.3.1 Score Range

If a scoring profile is defined in the manner described above, the scores should reflect the likelihood of the compounds satisfying the criteria for the properties contained in the scoring profile. Therefore, as the number of criteria in a scoring profile increases, the overall scores of success are likely to decrease because of the greater number of 'hurdles' that a successful compound must overcome.

For example, if a compound is scored against criteria for 5 properties and is 90% likely to meet each criterion, its overall likelihood of success will be 59% and the score will be (0.9^5=0.59). However, if there were criteria for 10 properties, with a 90% chance of being satisfied, the overall likelihood of success would be 35% (score 0.9^{10}=0.35).

Therefore, it can be seen that, as the number of properties in a scoring profile increases, the range of scores typically decreases.

2.3.2 Sensitivity Analysis

It is often useful to examine the effect of small changes to a criterion in a scoring profile. If such changes result in significant alterations to the set of compounds estimated to have the highest priority, this could indicate that the property values of the highest priority compounds lie close to the threshold for success for that property. This shows that this property is particularly critical to the success of these compounds and the values should be confirmed as soon as possible, typically via a high quality experimental measurement.

3 Chemical Space and Compound Selection

3.1 Introduction

Connecting the structural variations in a compound set with the properties or scores of those compounds is essential to understanding the multi-dimensional structure-activity relationships (SAR) governing an optimal balance of properties. The distribution of properties and scores across the 'chemical space' of a set of compounds can be visualised in StarDrop. This chemical space plot automatically scales to reveal an appropriate level of detail for the compound set, whether it's a diverse library or a single chemical series.

The structure of each compound in the plot can be viewed simply by pointing with the mouse at the corresponding symbol in the plot. Compounds can be selected from the plot for more detailed inspection or as candidates for progression to further analysis or experimentation.

When selecting compounds for progression, it may not be the best strategy to choose only the highest scored compounds, particularly if these all have similar structures. In many cases, it is possible to select a more diverse set of compounds, with little effect on the overall quality of the set, gaining greater knowledge of SAR across all relevant properties and spreading risk across greater chemical diversity.

Particularly in the early stages of a project, it is often useful to sample outside of the highest quality compounds in order to explore additional chemical diversity and gain additional knowledge regarding the variation in property values across a range of relevant chemistries. Data obtained in this way may be used to test the hypotheses on which predictive models are built, gain information on SAR, or sample alternative chemical series as backup, should the primary series fail for unforeseen reasons.

Coupled with the chemical space plots, StarDrop includes an algorithm to help in balancing compound quality with diversity. You may select a bias between score and diversity to reflect an appropriate selection for the requirements of your project. This automatic process can be augmented by manual selection from the chemical space plot or the datasets themselves.

This chapter discusses the concepts underlying diversity, the chemical space plots implemented in StarDrop and the methods used to automate compound selection.

3.2 Chemical Similarity and Diversity

What makes a group of compounds similar or, conversely, diverse? There is no single answer to this question. From one perspective, it may be their biological activity in one or more assays or, from another, the chemical scaffold around which they are synthesized. For this reason, numerous approaches to measuring chemical diversity have been developed and employed in this field.

In some approaches, chemical descriptors are used to define a multidimensional space in which compounds can be plotted. These chemical descriptors are similar to those used to develop QSAR models as described in Chapter 6. The coordinates of a compound on each axis of the space are defined by the values of the corresponding descriptors and the similarity of two compounds is defined by the distance between their points in the space. The diversity of a set of compounds is commonly measured by the 'volume' of the descriptor space that is occupied. Ideally, the descriptors used to define the space should correlate with the biological properties of interest, to improve the chance of finding interesting compounds in a diverse sample, or of increasing the hit rate in a focused set. For this reason, descriptor spaces 'tuned' to particular activities are sometimes used. However, generic descriptor sets are often employed that include descriptors believed to correlate with a wide range of biological endpoints. Examples of these are the BCUTS descriptors developed by Pearlman *et al.* (Pearlman, 1998) and MACCS keys (McGregor & Pallai, 1997).

Experimental measurements of biological endpoints can also be used to define a space in a similar manner to chemical descriptors. The biological activities of compounds can be used to define a space with each activity representing an axis. The underlying assumption is that compounds displaying a similar profile of biological activities are likely to have a similar activity against a novel endpoint. This approach has been employed in the BioPrint® dataset developed by Cerep (www.cerep.fr) but is limited by the fact that collections of compounds with comprehensive data on multiple properties are limited to being small in size. For this reason, biological diversity is often used in combination with another method based on chemical structure alone.

The final approach discussed here does not rely on a preconceived notion of what descriptors or properties will best define the similarity of compounds for a given purpose. This approach defines the similarity of compounds in terms of the patterns of atoms present in their chemical structures. The patterns of atoms along 'paths' through the 2D chemical structure of a compound are usually encoded in a binary 'fingerprint' and the similarity of two compounds, A and B, can then be defined in terms of the similarity between their fingerprints using a metric such as the Tanimoto coefficient:

$$T = \frac{|A \wedge B|}{|A| + |B| - |A \wedge B|}$$

where |A| is the number of bits set in the fingerprint of compound A, |B| the number of bits set in the fingerprint of compound B and |A∧B| the number of identical bits set in both A and B (logical AND). This will have a value between 1 (if A and B are identical) and 0 (if A and B have nothing in common). Examples of this approach include the Daylight fingerprint library (Daylight, n.d.) and ChemAxon's JChem package (ChemAxon, n.d.).

The advantage of a path-based fingerprint approach to similarity and diversity is that it provides a 'generic' method of comparing compounds. No assumption is made regarding the characteristics of molecules that will correlate most strongly with the biological activities of interest. Also, similarity assessed in this way usually corresponds well to a chemist's view; compounds from a chemical series will typically have high Tanimoto coefficients and *vice versa*. The main disadvantage of this approach is that there is not a natural 'geometric' interpretation of the similarity between compounds, which makes visualisation difficult. Conversely, a descriptor-based distance metric has a natural geometric interpretation and, if the most relevant descriptors are well understood for a specific activity of interest, a 'tuned' descriptor-based similarity space can provide better results.

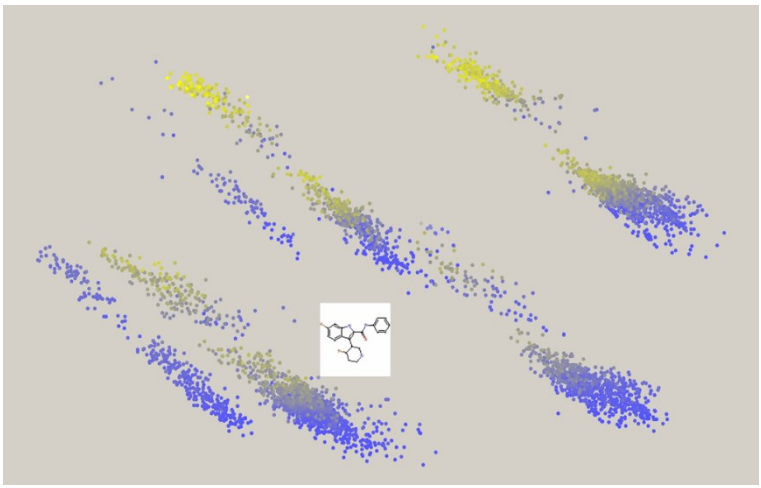


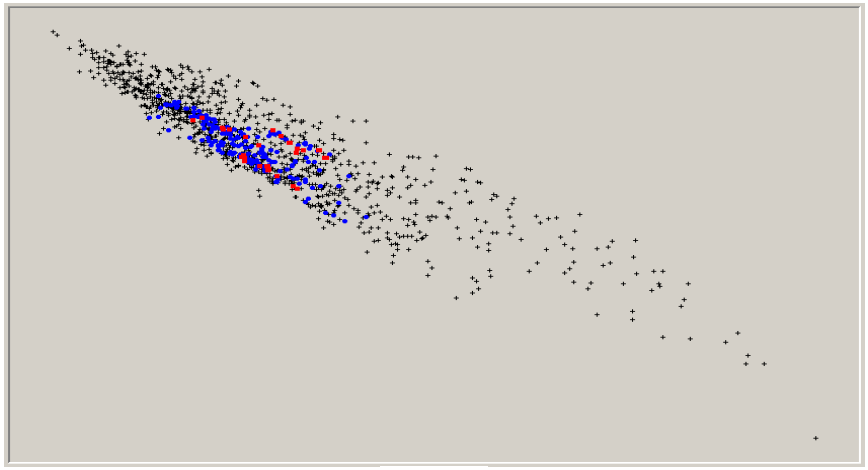
Figure 3.1 A chemical space plot illustrating the distribution of probabilistic scores across a compound set from low (blue) to high (yellow). An individual compound has been highlighted to identify its structure.

3.3 Visualising Chemical Space

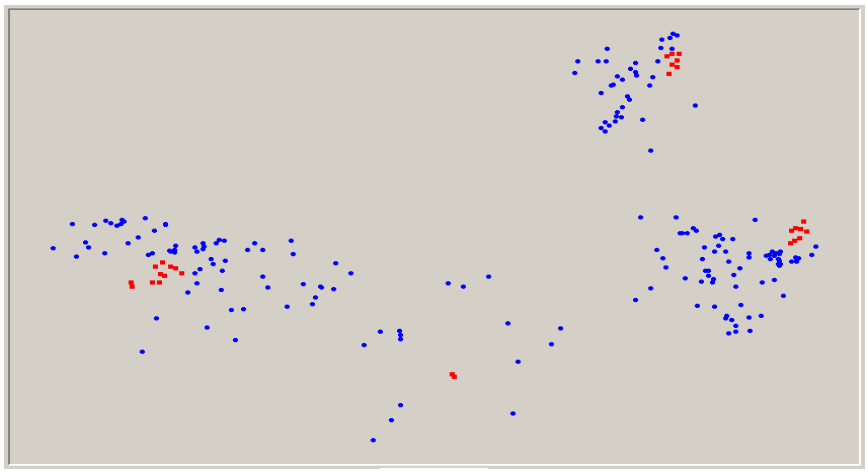
Within StarDrop, you can observe the distribution of compounds and their properties across the chemical space defined by a compound set (See Figure 3.2). In a chemical space plot, each compound is represented by a point and the similarity between two compounds by their proximity, i.e. if two points are close together they are similar in structure or properties.

A chemical space, or 'projection,' may be defined from any selected compound set. This space is automatically 'scaled' to represent an appropriate level of detail for the compound set. If the set represents a single chemotype, variations in the chemistry within that chemotype will be visualised. If the compound set represents a diverse set of compounds, a single chemotype will typically cluster together. Examples of these are shown in Figure 3.2.

Once a chemical space has been defined, any set of compounds may be visualised in that space. This may include compounds used to generate the chemical space or novel compounds.



(a)



(b)

Figure 3.2 Examples of chemical space plots.

Plot (a) shows a chemical space defined by a diverse selection of 'drug-like' compounds taken from the MDL Drug Data Repository (MDDR) plotted as black crosses. Superimposed on this are the compounds taken from a single project (blue circles) and a selection of compounds made from these (red squares).

Plot (b) shows the same project compounds plotted in the chemical space defined by the project's chemistry, along with the same selection. This demonstrates how the chemical space plot automatically represents chemical diversity on the relevant scale.

Chemical spaces may be defined within StarDrop based upon either structures or properties:

Chemical spaces defined by structure use a 'generic' definition of compound similarity constructed from path-based fingerprints and Tanimoto similarity. This gives a qualitative visualisation of chemical space, in which compounds in close proximity will be recognizably similar in chemical structure and those separated by greater distance will have less similarity. This approach is most commonly used because, when analysing data from a variety of sources or combining multiple properties in a score, an appropriate choice of descriptors that correlate with these properties is often not apparent. While a 'generic' chemical space is not guaranteed to show correlations with any specific property, trends are commonly visible in practice and can be used to understand SAR and choose appropriate compounds for further analysis.

Chemical spaces defined by properties can be based on any number of variables (greater than 1) calculated in, or imported into, StarDrop. This approach enables the visualisation of the distribution of compounds with respect to their properties. Groups of compounds with a similar spectrum of properties will cluster together in this type of space. Additionally, if specific descriptors are expected to correlate with an important property, these may be imported into StarDrop and used to define a chemical space. Thus, the diversity of selections with respect to these descriptors may be visualised to aid in library design.

StarDrop also offers two different methods for creating chemical spaces, PCA and Visual Clustering, both of which can be used to generate either 2D or 3D chemical spaces.

3.3.1 PCA (Principal Component Analysis)

This technique involves projecting the high-dimensional set of compounds to a new low-dimensional space by means of an orthogonal linear transformation. For the purpose of visualisation, we project the set of compounds to just two dimensions, which correspond to the first two *principal components*.

The transformation is constructed to ensure that the (orthogonal) principal components explain as much variance in the data as possible. We perform a full PCA analysis enabling the creation of either 2D or 3D projections.

PCA is the optimal linear dimensionality reduction algorithm with respect to mean squared error, but it is not well-suited to scenarios where the data lies on a nonlinear low-dimensional manifold embedded in a high-dimensional space. Specifically, PCA focuses on keeping the low-dimensional projections of dissimilar compounds far apart instead of keeping similar compounds close together; in these situations, we would likely achieve better results with a nonlinear dimensionality reduction algorithm.

3.3.2 Visual Clustering

Visual clustering uses an approach known as t-SNE (t-distributed Stochastic Neighbour Embedding). t-SNE is a nonlinear dimensionality reduction algorithm ideally suited to visualising high-dimensional data in two or three dimensions. The algorithm starts by converting the high- and low-dimensional similarities between *n* compounds into a set of joint probabilities. In high-dimensional space, conditional probabilities are calculated based on Gaussians centred at each high-dimensional point x_i:

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)}$$

which are then symmetrised to form joint probabilities

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2n}$$

The low-dimensional probabilities are computed based on Student's t-distribution:

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}}$$

where the y_is are the low-dimensional points.

The Kullback-Leibler divergence between these high- and low-dimensional joint probability distributions *P* and Q is given by

$$C = KL(P \,||\, Q) = \sum_{\mathrm{i}} \sum_{\mathrm{j}} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

and is then minimised over the low-dimensional points to obtain a clustering of low-dimensional points in which similar compounds are placed close together and dissimilar compounds are far apart (van der Maaten & Hinton, 2008).

As can be seen in Figure 3.3, t-SNE tends to produce superior visualisations to PCA. However, it is much more computationally expensive. As a nonparametric technique t-SNE does not provide us with a mapping that we can use to project new data sets into an existing chemical space; in our implementation, we work around this problem by "learning" a parameterised projection *post hoc* from the high-dimensional compounds and their low-dimensional counterparts. This approach is also used to enable t-SNE to automatically scale to much larger data sets by first using a random sample of the original set of compounds to compute a parameterised projection, and then applying this projection to the remaining compounds.

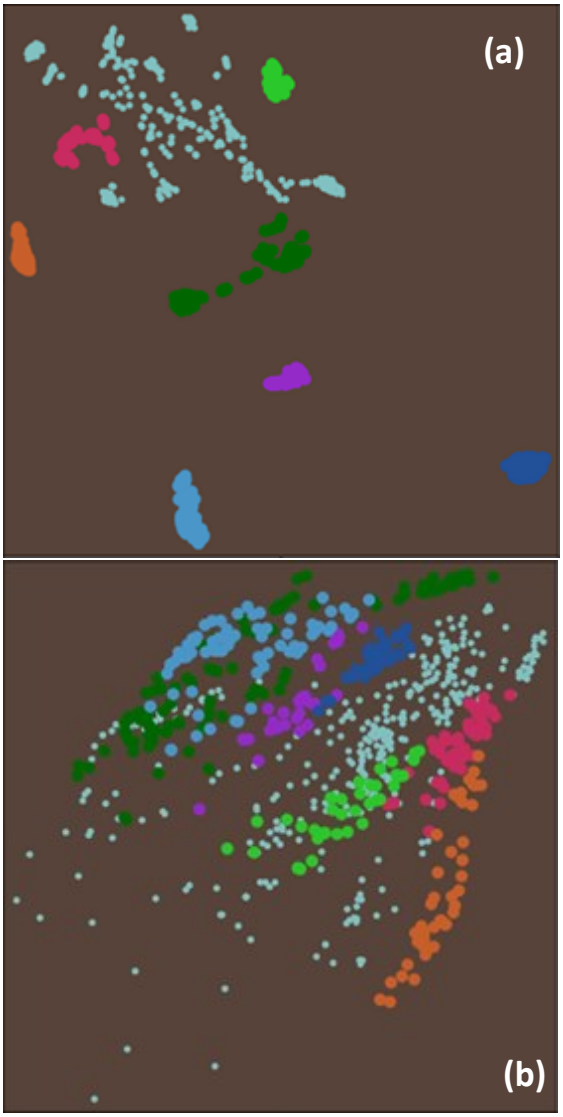


Figure 3.3 Comparison of PCA and t-SNE. Plot (a) shows a t-SNE plot of a set of dopamine actives with individual clusters highlighted. Plot (b) shows the same dataset in a PCA plot with the same clusters highlighted for comparison.

3.3.3 Interaction

Trends may be visualised by colouring the points in the chemical space plot on a user-defined scale according to any predicted or imported property or probabilistic score. Multiple compound sets may be plotted in the same space and coloured individually to visualise their relationships.

Compound structures corresponding to points in the chemical space plot may be observed and individual compounds or groups may be selected in the chemical space plot. These will be highlighted in the corresponding dataset and may then be copied to a new dataset for further investigation. Details on how this may be achieved can be found in the StarDrop User Guide.

3.4 Selecting Compounds

Essential decision points in drug discovery involve the choice of a set of compounds ('selection') for progression from a larger collection ('library'). This may be based on predicted or measured properties for the compounds in the library and other factors such as chemical diversity.

A scoring scheme, such as the probabilistic scoring method, allows the 'quality' of the compounds in the library to be assessed against the project's target profile across a broad range of properties. However, it may not be an optimal choice to choose only the 'best' compounds from the library for progression. In particular, early in a project, consideration should be given to exploring a diversity of chemistry. The purpose of this is to provide additional structure-activity data, validation of the accuracy of property predictions and, if possible, multiple chemical series to provide alternatives should an unexpected difficulty arise with the lead chemical series.

The number of possible selections increases exponentially with the size of a virtual library, e.g. there are 2.6×10^{23} ways of choosing 10 compounds from a library of 1,000.

Therefore, when considering diversity, it rapidly becomes impossible to perform an exhaustive search for the optimal selection for a given set of criteria. Instead, a 'stochastic' approach must be taken, which cannot guarantee to identify the optimal solution but will find the optimal or a near-optimal selection with high probability. Approaches such as genetic algorithms (Agrafiotis, 2001) and simulated annealing (Gillet, Khatib, Willett, Fleming, & Green, 2002) have been previously employed in this context.

StarDrop implements an automated compound selection algorithm based on a genetic algorithm to select a set of compounds based on a user-defined bias between quality and diversity. This helps to explore the balance between the score, calculated using probabilistic scoring, and diversity to select an appropriate set of compounds for your requirements (See Figure 3.4). However, you can specifically add particular compounds to a selection in order to explore certain chemistries or test specific hypotheses.

The genetic algorithm for compound selection maximises an objective function which is a weighted sum of diversity and a score or property value, i.e.

$$F = w_d D + w_s S,$$

where w_d is the weight assigned to diversity, D is a measure of the diversity of the selected set, w_s is the weight assigned to the score or property, S is the average score or property value for the selected compounds. The weights w_d and w_s are chosen by the user and must sum to 1. Different measures of diversity can be used and are summarised in Section 3.4.1.

In combination with the chemical space plots and probabilistic scoring, the compound selection algorithm provides the ability to explore the impact of different compound selection strategies on the likelihood of success. There may be a good argument for manually selecting specific compounds to test a hypothesis and this should not be ruled out. However, the objective balancing of diversity and quality encourages the majority of effort to be expended in those areas of chemistry most likely to yield progress toward a project's objectives.

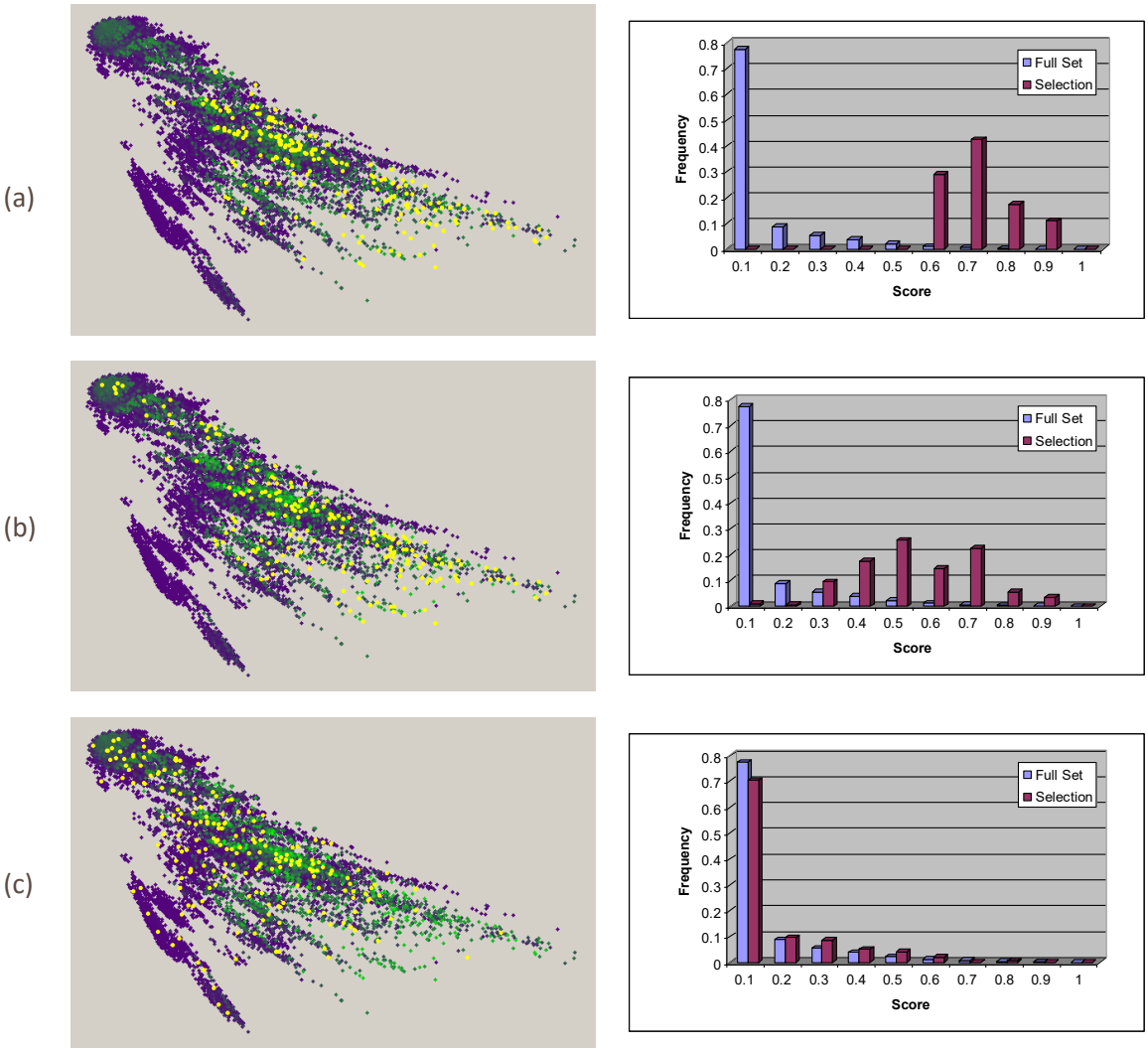


Figure 3.4 An illustration of increasing the bias from score to diversity. (a) shows the top 200 compounds selected by score, (b) the selection with a bias of score:diversity of 20:80 and (c) maximum diversity. The chemical space plots illustrate the changes in diversity and the histograms show the changes in distribution of scores for the compounds selected.

3.4.1 Diversity Metrics

The diversity of a set of compounds can be defined in terms of chemical structural, compound properties or a combination of structure and properties.

The distance between two compounds in the set, *i* and *j*, is denoted d_{ij} and depends on the chosen definition:

- Structure: $d_{ij} = 1 - T_{ij}$, where T_{ij} is the Tanimoto index between the 2D, path-based fingerprints of compounds *i* and *j*.
- Properties: $d_{ij} = \sqrt{\sum_{k=1}^{M}(p_{ik} - p_{jk})^2}$, where p_{ik} is the value of property *k* for compound *i*, and the sum over k runs over M properties selected by the user. In this calculation, the property scales are normalised such that the mean of each property scale has a mean of zero and variance of one over the full set from which the selection is being made.
- Combined structure and properties: $d_{ij} = \sqrt{\sum_{k=1}^{M}(p_{ik} - p_{jk})^2 + (1 - T_{ij})}$

The overall diversity of a selected set of *N* compounds can be calculated using a choice from three different metrics:

- max-min: $D = \min_{i,j=1,N} d_{ij}$

- max-average: $D = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} d_{ij}$
- S-optimal: $D = \frac{N(N-1)}{2 \sum_{i=1}^{N} \sum_{j=i+1}^{N} (\frac{1}{d_{ij}})}$

The default diversity metric used in the selection algorithm is max-min, based on structure.

3.5 Limitations and Tips

3.5.1 Chemical Space Visualisation

The generation of a chemical space based on compound structure alone or in combination with properties is a computationally demanding process. The computational cost increases approximately as the cube of the number of compounds in the set used to generate the space. In practice, generating a space based on more than 10,000 compounds will take an impractically long time.

If the compound set (e.g. a virtual or company library) contains significantly more compounds than this, a space can be defined in terms of a subset of compounds, selected at random or as a diverse subset of the full set. This can be achieved using the compound selection component of StarDrop (See the User Guide). Once a subset has been selected, define the chemical space relative to this set and then all compounds or selections can be plotted in this space.

3.5.2 Compound Selection

What is the appropriate balance of quality and diversity?

The answer will, of course, depend on many factors as discussed above. However, it is usually beneficial to explore the sensitivity of a selection to the degree of bias chosen before making a final decision. Often, a significant degree of added diversity can be explored for a small decrease in the overall quality of the compounds selected. In this case, it is advisable to spread the risk across diverse compounds provided synthetic resources permit. Conversely, in some cases, the selection of compounds will remain the same until a large bias toward diversity is selected. In this case, the selection of a diverse set may require an unacceptable decrease in the overall quality of the compounds.

4 Glowing Molecule™

4.1 Introduction

While the application of *in silico* models can significantly reduce the resources wasted on molecules that are unlikely to succeed, it does not itself extract the maximum utility from predictive models, which encode knowledge regarding the relationship between chemical structure and the desired properties of a successful drug. If this knowledge could be explicitly revealed, even greater efficiency could be realized by directing the design of compounds towards an appropriate balance of properties. The first questions often posed by a chemist when confronted with a prediction of a property for a molecule are "Why does this molecule have that predicted value?" and, if the property value is not appropriate, "What can I do to this molecule to improve the modelled property value?"

To answer these questions it is necessary to 'invert' the model. Rather than inputting a chemical structure to predict a property value, we must use the model to map changes in a property back to the chemical structure. However, this process is difficult to realize in practice. The models for potency or ADME/Tox properties are often Quantitative Structure Activity Relationship (QSAR) or Quantitative Structure Property Relationship (QSPR) models, which relate characteristics of a molecule's structure to an observed biological property. The descriptors (Section 6.3) used as inputs to this relationship can be difficult to relate directly to the chemical structure and often the relationship itself is complex and highly non-linear. This is particularly true of models derived using modern 'machine learning' techniques such as artificial neural networks (ANNs) or Gaussian Processes (GPs). Thus, the structural knowledge captured by these models is often hidden and uninterpretable.

The Glowing Molecule is a generally applicable process for deriving and visualising structural information regarding the causes of a prediction for any form of QSAR/QSPR model or chemical descriptors. It is necessary to base the output on a number of simplifying assumptions. Thus the output should only be used as a qualitative description of the underlying behaviour. However, despite this, the Glowing Molecule can describe for a single property, or more powerfully for a collection of properties combined into a probabilistic score (Section 2), the parts of a molecule that are influencing the property/score value either positively or negatively.

All of the ADME models that are provided in the StarDrop ADME module can generate Glowing Molecules, as can any models built using the Auto-Modeller (Section 8) that do not incorporate third party data. Additionally, any scores generated from the ADME/Auto-Modeller models will also generate Glowing Molecules.

4.1.1 Methodology

A model of a property is defined by a hypersurface in $N+1$ dimensions, where N is the number of descriptors used to build the model. A model prediction for a compound is made by finding the point on the hypersurface at the coordinates given by the descriptor values for the compound. By considering also the slope of the hypersurface at the point of prediction it is possible to infer relationships between the property and the descriptors (i.e. how much a change in a given descriptor might affect the property).

4.1.2 Linking Descriptors to the Molecule Structure

In StarDrop, all the descriptors used are either patterns of atoms or fragments or whole molecule properties that can be expressed in terms of functions of atoms or fragments. As such it is easy to understand the relationship between any one descriptor and the molecule structure (Figure 4.1). For whole molecule properties that can be expressed in terms of atoms or molecule fragments we can calculate the relationships between descriptor and molecule by considering also the relationship between the underlying fragment descriptors and the whole molecule property.

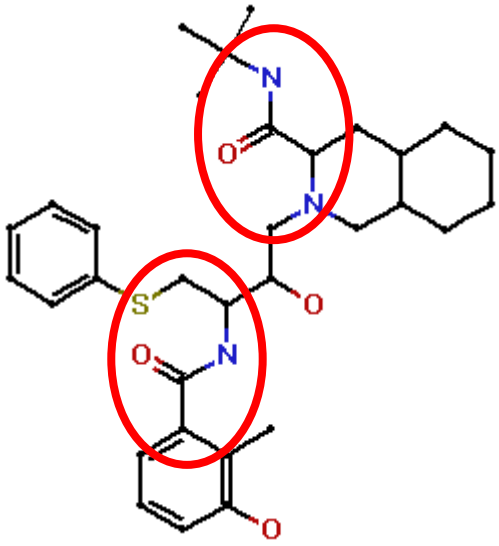


Figure 4.1 For a given molecule it is quite easy to identify the parts that contribute to an 'amide count' descriptor.

4.1.3 Relationship Between a Property and its Descriptors

Linear relationships

For a linear property, the relationship between the property and any one of the descriptors is trivial to understand. In the following example we have a linear relationship between a property P and three descriptors D_1, D_2 and D_3.

$$P = 3D_1 + 7D_2 - 4D_3$$

It is easy to see that regardless of the molecule we're testing, if descriptor D_1 or D_2 increases then the property value will also increase. Thus we can say that there is a positive relationship between the property and D_1 and D_2. Additionally we can see that the property is more sensitive to changes to D_2 than D_1. Conversely, we can also see that there is a negative relationship between property P and descriptor D_3 because any increase in D_3 will result in a decrease to the property value.

Non-linear relationships

Unlike the linear case, in non-linear cases the relationship between the property and any given descriptor will be different for two molecules that have different values for the same descriptor. The following example is a simple non-linear relationship between a property P and a descriptor D.

$$P = e^{-(5-D)^2}$$

If for molecule A, D = 4 then we know that were D to increase by a small amount (e.g. 1) we would see an increase in property P. If for molecule B, D = 6 then both molecule A and molecule B will have the same property B. However, for molecule B, an increase to descriptor D would result in a decrease to property P. Therefore, for molecule A there is a positive relationship between descriptor D and property P whereas for molecule B there is a negative relationship.

Figure 4.2 shows an example of a simple non-linear relationship between a property P and two descriptors. From this, one can also see that the relationship between the property and the descriptor depends on the value of the descriptor. Looking at a molecule for which the values of descriptors D1 and D2 are given by point X, we can see that were it possible to increase descriptor D2, this would result in an increase to property P. Conversely, a decrease in D2 would result in a decrease to P.

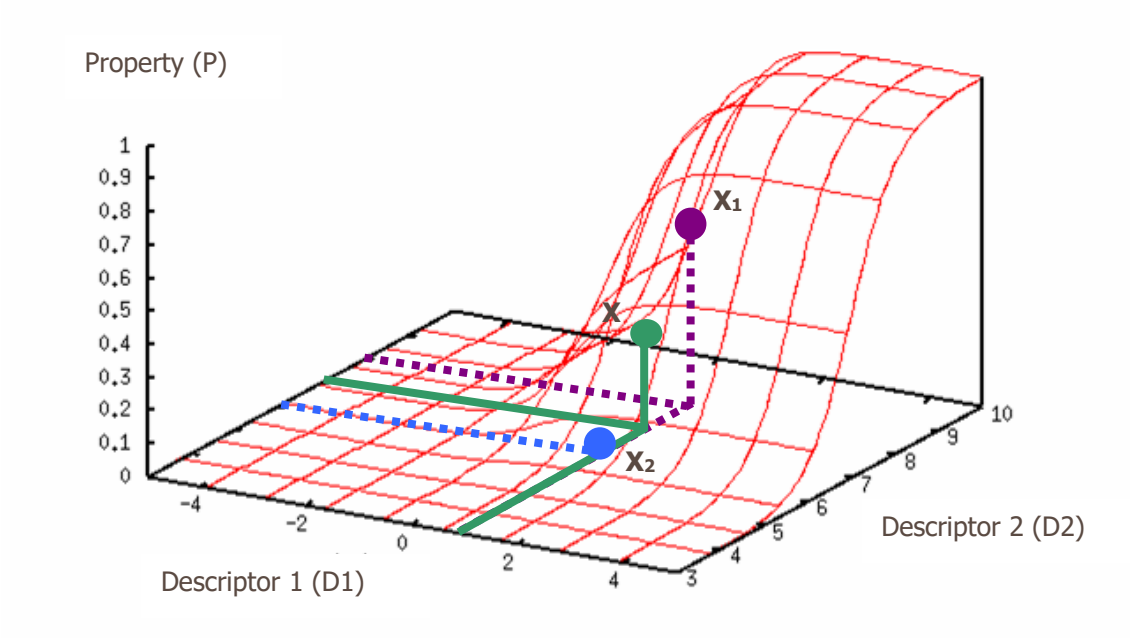


Figure 4.2 A simple hypersurface representing a model of a property (P) based upon two descriptors (D1 and D2). Point X represents a molecule for which D1=1 and D2=7 resulting in P=0.4. Point X_1 represents a molecule for which D1=1, D2=8 and P=0.7. Point X_2 represents a molecule for which D1=1, D2=6 and P=0.1.

Hence, the nature of the relationship between a descriptor and a property for a given molecule can be determined by looking at the slope of the hypersurface at the given descriptor value for that molecule. This can be done by considering each individual descriptor in turn (Figure 4.3).

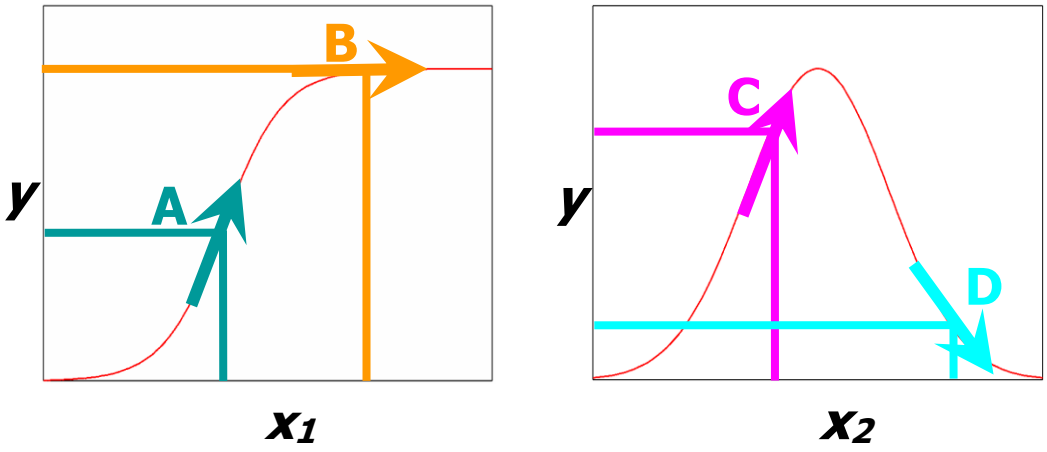


Figure 4.3 Property Y for molecule A is sensitive to descriptor X1 and is positively influenced. Property Y for molecule B is positively influenced by X1 but is not sensitive. Property Y for molecule C is sensitive to descriptor X2 and is positively influenced. Too large a change to X2 could result in an unexpected change in property Y for Molecule C. Property Y for molecule D is sensitive to X2 but negatively influenced.

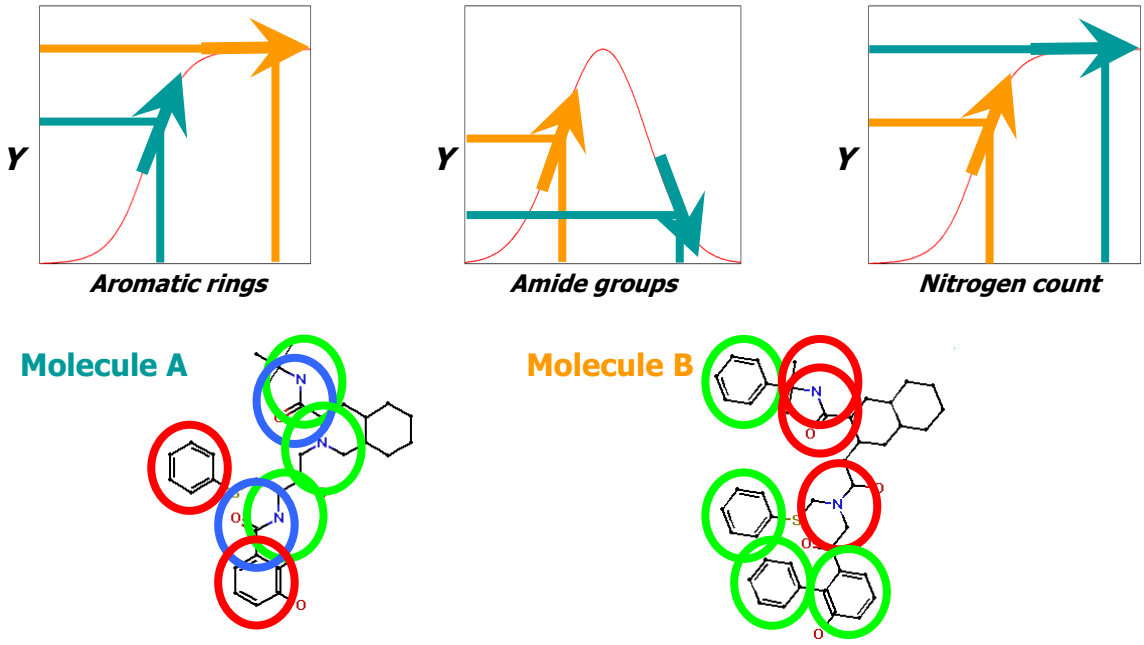


Figure 4.4 An example of a non-linear relationship between three descriptors and a property (Y). The regions having a positive influence on the property value for each molecule are coloured in red; those regions having a negative influence are coloured in blue; those regions having no significant influence are coloured in green.

4.1.4 Relationship Between the Property and the Structure

Combining this interpretation with our ability to locate the regions of the molecule each related to each descriptor, allows us to create a picture of the influence each part of the molecule is having on a property. Figure 4.4 shows an example of a non-linear relationship between three descriptors and a property showing how similar groups have a different influence on the property for different molecules.

For a non-linear model of a property based upon a larger number of descriptors we can build up a heat-map indicating the regions of the molecule having the greatest positive and negative influence on the predicted property value (Figure 4.5).

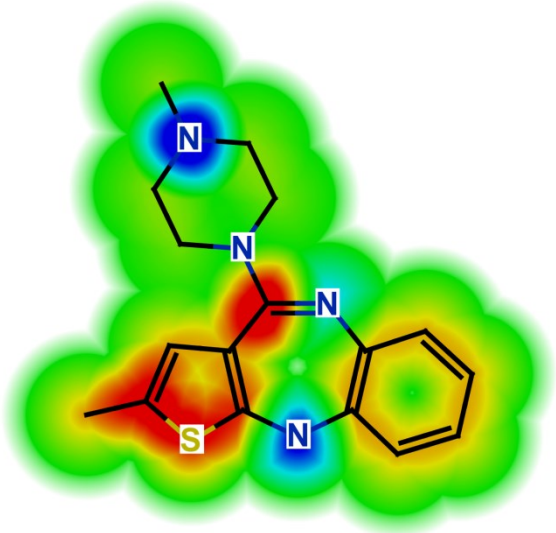


Figure 4.5 Olanzapine with a heat map indicating the influence of different parts of its structure on its predicted logP.

The heat map goes from blue (the parts having the most negative influence) through green (areas having no overall influence) to yellow and red (regions having the most positive influence). Note:

Positive and negative are used to indicate increasing and decreasing property values respectively and are not indicative of whether these changes would be considered 'good' or 'bad' in a project setting.

For category models we can apply the same logic, except that the functions are discontinuous and therefore changes in property values, as descriptors increase and decrease, are discrete jumps.

We can also take this one stage further if we consider a score that has been calculated based upon a number of properties. In this case, we must assess the regions of the molecule having greatest influence on each property and combine these based upon the relative contributions of each property to the overall score. Ultimately, a score of multiple properties is a meta-model and therefore the same approach can be used to consider the regions of the molecule having the greatest overall influence.

4.2 Interpretation

The key message to remember when interpreting a Glowing Molecule representation of property predictions is that the Glowing Molecule only indicates the structural influences on that particular prediction. Most of the StarDrop models and those that can be generated by the StarDrop Auto-Modeller are non-linear and therefore it should not be assumed that the contribution of a given region of a molecule will remain constant as other regions of the molecule are changed.

4.2.1 Basic Interpretation

A simple approach to interpreting the Glowing Molecule representation of the solubility of Olanzapine in Figure 4.5 is to consider the molecule in three sections:

In Figure 4.6 the highlighted region is blue indicating this part of the molecule is having a tending to decrease the logP. If it were necessary to try and increase the logP then this region would be an ideal target for modification.

In Figure 4.7 the highlighted region is dominated by red indicating this part of the molecule is tending to increase the logP. If it were necessary to decrease the logP then this region would be an ideal target for modification.

In Figure 4.8 the highlighted region is not dominated by any one colour as there are mixed minor influences. The aromatic ring is clearly tending to increase the logP but at the same time the presence of nitrogen in the ring system is tending to decrease the logP. Whist it might be possible to alter the influence of either of these regions through modification, the effect is likely to be less significant than making changes to the regions having larger influences (Figure 4.6 and Figure 4.7).

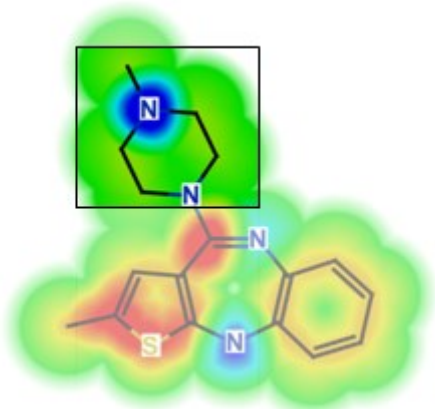


Figure 4.6 The methyl-substituted nitrogen of the piperazine ring is tending to decrease the logP of Olanzapine.

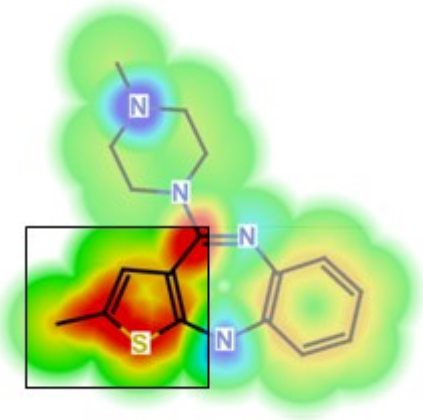


Figure 4.7 The methyl substituent on the thiophene ring is tending to increase the logP of of Olanzapine.

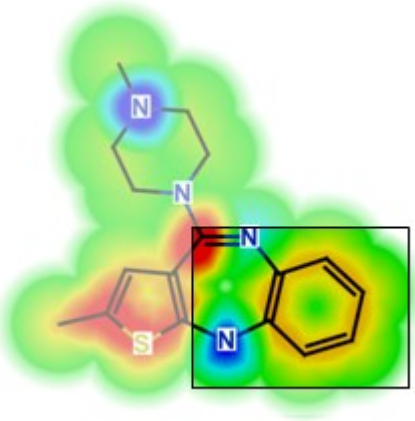


Figure 4.8 One nitrogen of the benzodiazepine fragment is also tending to decrease the logP and the benzo component of this fragment is tending to increase the logP of Olanzapine.

These basic interpretations of the Glowing Molecule representation will be appropriate for a large number of cases. However, there are a number of important considerations to bear in mind:

- The Glowing Molecule representations provide only a qualitative assessment of the influence of each part of a molecule on a property. Therefore it is inappropriate to try and infer slight differences in influence between regions of the molecule coloured similarly.
- When making modifications to a molecule, it is important to remember that some descriptors have "non-local" effects. An example of this might be a descriptor which is looking for methyl groups attached to an aromatic ring. In this case, if we have a molecule with a methyl group attached to an aromatic ring, changing the methyl group in question or even the aromatic ring to which it is attached may have an obvious effect. However, it is possible that the aromaticity of the ring is due to a larger set of fused rings across the molecule and therefore changes to a ring distant from the methyl group might also affect the descriptor if the overall aromatic state of the molecule is changed as a result.
- The larger the change made to the molecule, the harder it will be to make inferences about the property value for the resulting molecule and the greater the chance of encountering non-linear effects (see Section 4.2.3). However, as the property value becomes more extreme (i.e. particularly high or low for that property) the influence of the groups becomes less obvious.

4.2.2 Example

The most appropriate way to work with the representation is to consider that if a part of the molecule that is having a negative effect on the property can be replaced by another with a positive effect we would expect to see the property value increase. Equally, if a part of the molecule that is having a positive effect on the property can be replaced by another with a negative effect we would expect to see the property value decrease.

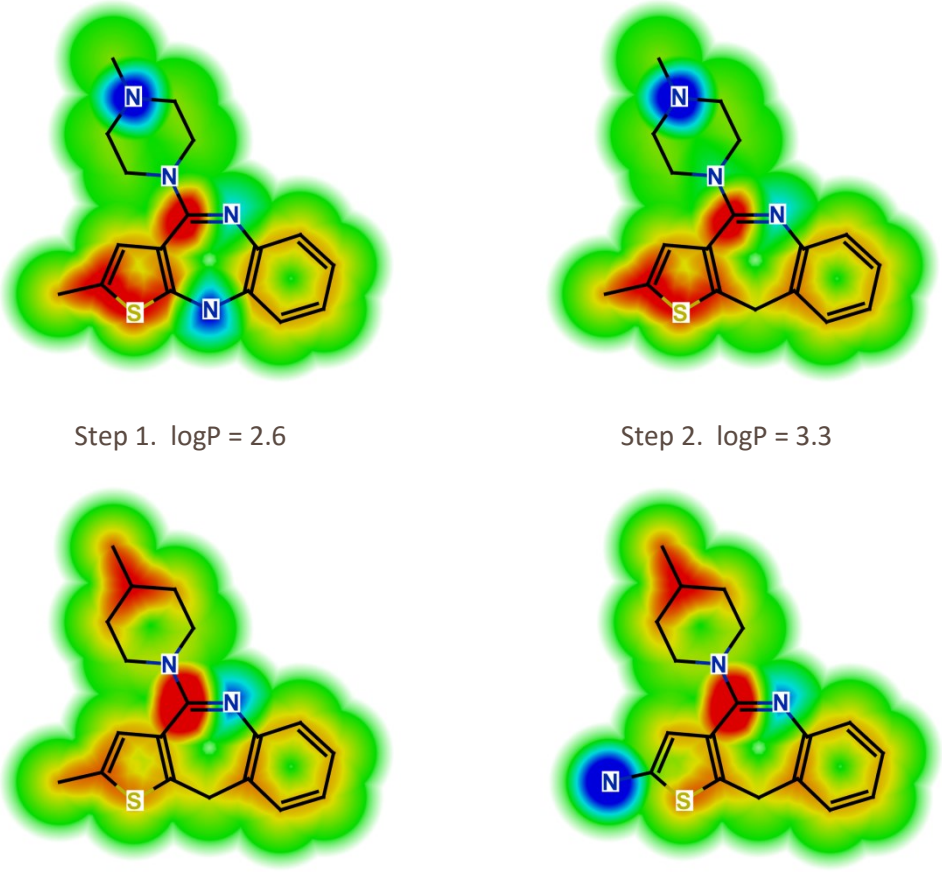


Figure 4.9 Changes to the property LogS reflect the behaviour expected when considering the indicated importance of different parts of the molecule.

Figure 4.9 is an illustration of some successive modifications made to Olanzapine. In Step 1 the logP is 2.6. At this stage the methyl-substituted nitrogen of the piperazine ring is tending to decrease the logP while the methyl substituent on the thiophene ring is tending to increase it. At the same time the one nitrogen of the benzodiazepine fragment is also tending to decrease the logP.

In Step 2 the nitrogen of the benzodiazepine unit has now been replaced by a carbon which is indicated as tending to increase the logP. As a result, the logP is now higher – but in this case only by a small amount because the region of the molecule being replaced was not having a very large influence on the logP. Meanwhile the piperazine ring continues to be having a strong influence towards decreasing the logP and the thiophene ring a tendency towards increasing it. In Step 3 the piperazine has now been replaced by a methyl-substituted pyrimidine which is now having a tendency to increase the logP. As a result, the logP increases dramatically. Finally in Step 4 the methyl substituent on the thiophene ring is replaced by an amino group which has a tendency to decrease the logP again.

4.2.3 Non-linearity

It is important to remember that, when interpreting the Glowing Molecule representations for non-linear models, the effect of large changes to a descriptor may result in a less obvious outcome.

In Figure 4.10 we can see that from a molecule at the starting point (1) a modification to the molecule resulting in a small increase in descriptor D will behave as expected (2), however a modification resulting in a larger change to descriptor D may well decrease the property value (3). Figure 4.11 shows a more complex example. From the starting point (1) we expect a modification to the molecule resulting in a small change in descriptor D to increase the property value (2).

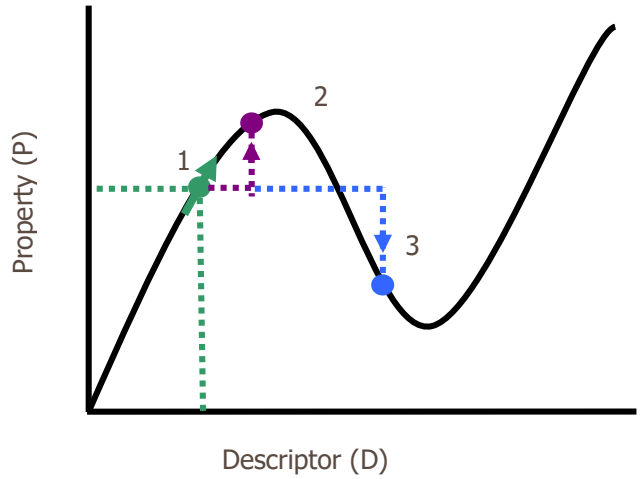


Figure 4.10 Non-linear relationship between property (P) and descriptor (D).

(1) At our starting point there is a positive relationship between P and D (illustrated by the arrow) so we expect that if we increase D, we will see an increase in P. (2) Having made a small change to D we see an increase in P as expected. (3) If we make a larger change in D it results in a decrease in P due to the non-linearity of the long-range relationship.

If our molecule is modified such that our descriptor value ends up at position (2) – a local maximum or (3) – a local minimum, we will get no feedback from the Glowing Molecule representation because a modification to the molecule resulting in a small change to the descriptor value will have little effect on the property.

Furthermore, a modification to the molecule resulting in a large change to the descriptor value from either of the points will result in a decrease in the property value from (2) and an increase from (3), *regardless* of whether the change is an increase or decrease.

Figure 4.11 demonstrates how, even when a property and descriptor have a monotonic relationship (i.e. an increase in the descriptor always corresponds to increase in the property, and *vice versa*), it is possible that within different ranges of descriptor values the reported influence of the descriptor for different molecules might be very different.

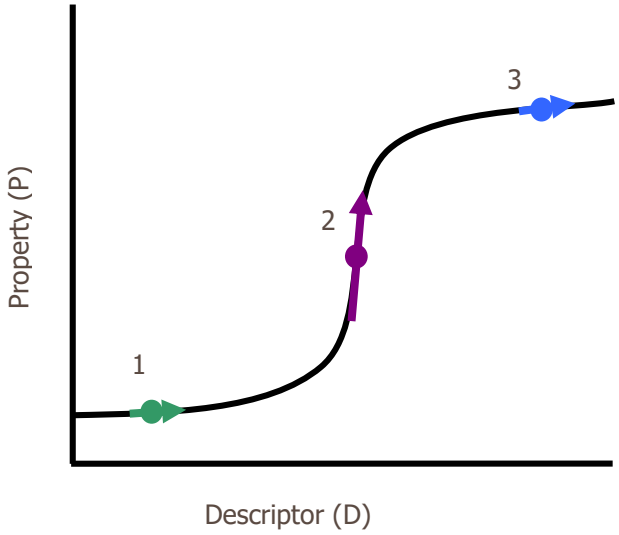


Figure 4.11 Non-linear relationship between property (P) and descriptor (D). Here there is a positive relationship between P and D – as D increases, so does P. However, it is only in the middle of this range (2) that there is a significant change in P as D increases. At points (1) and (3) there is a negligible change. A molecule with descriptor values at either point (1) or (3) would be coloured green on the basis of the influence due to this descriptor; however, at point (2) the colour would be red because a small change in D would result in a large change to P.

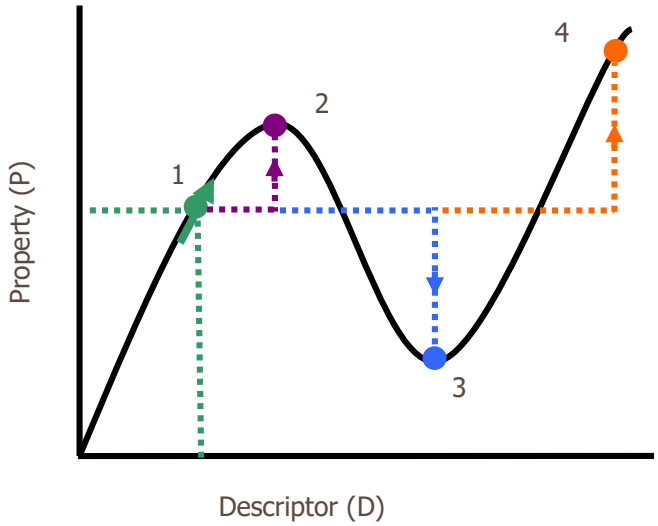


Figure 4.12 Non-linear relationship between property (P) and descriptor (D)

(1) At our starting point there is a positive relationship between P and D so we expect that if we increase D, we will see an increase in P. (2) If D increases enough we end up at a local maximum. At this point any very small change in D will result in a negligible change to P (as indicated by the absence of an arrow) and thus the Glowing Molecule representation at this point would remain green indicating little influence. A large change in D (either positive or negative) will result in a decrease in P. Only a very large positive increase in D might result in an increase in P, e.g. point (4). (3) A further increase in D ends up at a local minimum. As with the local maximum, any very small change in D will result in a negligible change to P and thus the Glowing Molecule representation at this point would also remain green, indicating little influence. From this point, any large change in D (either positive or negative) is likely to result in an increase in P.

5 Cheminformatics Algorithms

StarDrop provides a range of chemoinformatics algorithms to help you to organise your compound data sets and highlight important structure-activity relationships (SAR). These algorithms are closely coupled to StarDrop's Card View that helps to quickly interpret the output of these algorithms in an intuitive way.

The following sections provide details of these algorithms and their outputs.

5.1 Clustering

Clustering can be used to group together 'similar' compounds, for example to identify chemical series within a data set of diverse compounds, analyse SAR around hits for triaging results from high throughput screening or to identify 'regions' of chemistry that may yield good properties or scores.

StarDrop provides three different approaches to defining clusters, based on compound structure, properties or maximum common substructure. These are described in the following subsections.

5.1.1 Structure and Property-based Clustering

These clustering approaches are based on the 'dbclus' algorithm (Butina, Unsupervised Data Base Clustering Based on Daylight's fingerprint and Tanimoto Similarity: A fast and automated way to cluster small and large data set, 1999) and differ only in the measure of similarity between compounds:

- Structure: $s_{ij} = T_{ij}$, where T_{ij} is the Tanimoto index (Rogers & Tanimoto, 1960) between the 2D, path-based fingerprints of compounds *i* and *j*.
- Properties: $s_{ij} = 1 - \sqrt{\sum_{k=1}^{M}(p_{ik} - p_{jk})^2}$, where p_{ik} is the value of property *k* for compound *i*, and the sum over *k* runs over *M* properties selected by the user. In this calculation, the property scales are normalised such that the mean of each property scale has a mean of zero and variance of one over the full set of compounds.

In the dbclus method, the 'size' of the desired clusters is defined by a minimum similarity s_{min}. The dbclus method proceeds by calculating the similarity of each compound with all other compounds in the data set. The compounds are then ordered in order of decreasing number of neighbouring compounds with similarity greater than s_{min}. The first compound in this list, i.e. the compound with the largest number of neighbours, is then chosen as the centroid of the first cluster and all neighbours of this compounds with similarity greater than s_{min} are defined as members of this cluster.

The centroid and members of the first cluster are then removed from the list and highest ranked of the remaining compounds is then designated as the centroid of the second cluster and all compounds with a similarity greater than s_{min} to this compound are assigned as members of this cluster. These are then removed from the list and the process repeated until there are no compounds in the list with another compound with a similarity greater than s_{min}. These remaining compounds in the list are designated as singletons.

The dbclus algorithm has a number of advantages; in particular, the clusters that are generated are independent of the order of the compounds in the data set provided as input and it makes no assumptions regarding the 'correct' number of clusters that should arise from a given data set.

5.1.2 Maximum Common Substructure Clustering

This clustering approach uses a combination of 2D path-based fingerprints and a maximum common substructure algorithm to group together compounds containing a significant common substructure. Similar to the other cluster methods, a similarity threshold is used to control the tightness of clusters.

The search for a new cluster begins by identifying a cluster 'seed'. Molecules not yet assigned to a cluster are searched to find the two molecules that have the highest structural similarity based on 2D path-based fingerprints and Tanimoto similarity index. One of these molecules is set as the cluster 'seed'. Next, a list of cluster candidates is constructed by collecting molecules with fingerprint similarity to the seed molecule greater than 0.7 * similarity threshold. The candidate list is sorted in order of similarity to the seed.

Each candidate is considered in turn, starting with the most similar. A search is first made to find the commonality shared between the candidate and the seed using a maximum common substructure algorithm. A Jaccard similarity coefficient is calculated

$$Similarity = \frac{c}{(a_1 + a_2 - c)}$$

where *c* is the common substructure heavy atom count, a_1 is the heavy atom count in molecule 1 and a_2 is the heavy atom count in molecule 2. If this similarity is greater than or equal to the similarity threshold then the candidate is added to the cluster and the commonality that was found is recorded as the common substructure for that cluster. For subsequent candidates, a search is made to find the commonality with the common substructure for the cluster, not the seed. If the similarity threshold is met the candidate is added to the cluster and the cluster's common substructure is updated. Once all candidates have been considered these steps are repeated to find the next cluster.

If no molecules from the candidate list qualify for the current cluster then the seed molecule is designated a singleton. Candidates that are less similar to the seed molecule may display little commonality. If five consecutive candidates fail to qualify for the current cluster then the algorithm stops using a maximum common substructure algorithm and uses a faster substructure search based on the cluster's common substructure found to that point.

The common substructure algorithm identifies the largest *connected* component for each pair-wise comparison. The algorithm checks the following attributes match: atom elements, atom bond counts, atom cyclicity, bond cyclicity, bond types and atom ring chemistry. The last constraint ensures that we only match atoms if they are members of the same ring type. For example, a carbon atom in pyridine may not match a carbon atom in a phenyl ring. This prevents matching partial rings (where ring chemistries differ) and biases the algorithm so that molecules with different 'scaffolds' are more likely to be placed in separate clusters.

5.2 Molecular Matched Pair Analysis

The matched pairs method identifies molecule pairs that differ by a single, small contiguous fragment, i.e. where there is a single point of variation such as a change in R-group, linker or a ring change. Molecules with two or more points of variation are not identified as a matched pair. The analysis provides a simple way to identify and assess transformations based on existing data. The analysis can identify which transformations have been made, how common these are and what affect they have on properties (Leach, et al., 2006) (Dossetter, Griffen, & Leach, 2013). One application, for example, is to identify strategies for lead optimisation by identifying transformations that provide a consistent, significant improvement in a property of interest.

The method uses a computationally intensive maximum common substructure routine, similar to the WizePairZ method (Warner, Griffen, & St-Gallay, 2010) together with rapid screening steps that quickly identify where a matched pair cannot be present.

Screening Steps:

Molecule pairs displaying a change in heavy atom count greater than 8 atoms are dismissed as potential matched pairs.

Molecules pairs with fingerprint similarity less than a given threshold, dependent on the size of the molecules, are dismissed. For each pair of molecules the number of heavy atoms from the smaller molecule is taken. If this value is less than 15 then no similarity threshold is applied. If the heavy atom count is between 15 and 50 then a threshold equal to 0.5 + 0.007*heavy atom count is applied. If the heavy atom count is greater than 50 then a similarity threshold of 0.85 is applied. These values and this formula were arrived at empirically by running the matched pairs method without a threshold and noting the fingerprint similarity below which no matched pairs were found.

A simple fragmentation is performed to divide each molecule into rings, linkers and side chains. Strings are constructed to act as keys representing the chemistry of each fragment. These string keys provide the ability to quickly identify fragments that are different. Three difference values are calculated; a ring

difference, a linker difference and a side chain difference. For example, if molecule A has a pyridine ring and molecule B has phenyl ring, a ring difference of 2 is calculated corresponding to the removal of pyridine and the addition of phenyl. If the ring, linker or side chain difference is greater than 2 the compound pair is rejected. If the side chain difference <= 1 and linker difference == 1 and ring difference == 1 then the compound pair is not rejected. This allows a substituent to be transformed into a linker plus a ring. If the linker difference == 2 and ring difference == 1 and side chain difference == 0 then the compound pair is not rejected. This allows a linker to be replaced by two linkers and 1 ring. Finally, if the sum of the ring, linker and side chain differences is greater than 1 the compound pair is rejected.

Finding a Matched Pair:

Candidates passing the screening steps are analysed using a method based on a maximum common substructure search which identifies the commonality between two molecules. A minimum common substructure size is set to be 8 less than the number of heavy atoms in the larger molecule. Constraints are set to ensure atom elements, atom cyclicity and bond cyclicity match. Also, ring chemistry constraints are applied to match atoms only if they share the same ring chemistry. For example, a carbon atom in a pyridine ring may not match a carbon atom in a phenyl ring. The maximum number of disconnected components is set to 2.

The commonality found is inverted to identify the fragments that differ between the two molecules. The attachment points linking the fragments to the common substructure must be equivalent for both molecules. If multiple fragments are found then fragments are expanded in an attempt to merge them into one contiguous fragment. Rings are completed for fragments that form partial rings for ring sizes less than or equal to 8. Fragments are expanded, to a maximum of 3 bonds, to nearby cyclic atoms or atoms with three heavy atom neighbours. Fragments are also expanded to include any adjacent carbonyl atoms.

5.3 Activity Landscapes and Cliffs

These tools identify pairs of molecules that are structurally similar but display a significant change in some property. Where the property is activity then such pairs identify ‘activity cliffs’ (Stumpfe & Bajorath, 2012). Such cliffs can reveal structure activity relationships and suggest strategies for compound optimisation. We refer to the list of the most similar compounds in a data set as ‘nearest neighbours’.

For the case of the activity landscape tool, the most similar molecules are found by calculating similarity values for every pair of molecules using a 2D path-based fingerprints and a Tanimoto similarity index (Rogers & Tanimoto, 1960). This list of pairs is ordered by similarity. From this list the N most similar pairs are taken where N is a ‘number of nearest neighbours’ parameter chosen by the user. The ‘number of neighbours’ parameter will correspond to a particular similarity threshold.

The activity neighbourhood tool works in the same way except similarities are calculated only between the reference molecule and every other molecule.

Property differences and Structure Activity Landscape Indices (SALI) are calculated for each molecule pair in the nearest neighbour list. The SALI index is calculated as

$$SALI_{ij} = \frac{|P_i - P_j|}{(1 - sim_{(i,j)})}$$

where P_i and P_j are the properties of the i^{th} and j^{th} molecule and sim(i,j) is the similarity between the two molecules (Guha & Van Drie, 2008). This index quantifies the size of cliff between two molecules. Large cliffs identify highly similar molecules that display significant change in property. Note that in the calculation of SALI, activity values such as IC_{50} or K_i are conventionally represented in logged units such as pIC_{50} and pK_i.

6 ADME QSAR Models

This chapter describes the Quantitative Structure Activity Relationship (QSAR) models implemented in StarDrop as part of the ADME QSAR module. These models predict the following properties:

- logP (Octanol/Water)
- $logD_{7.4}$ (Octanol/Water)
    - Aqueous Solubility
    - Intrinsic Aqueous Solubility (logS)
- Solubility at pH 7.4 ($logS_{7.4}$)
- Human Intestinal Absorption (HIA) Classification
- Blood-Brain Barrier Penetration
    - Classification
    - Log([Brain]/[Blood]) (log(BB))
- Cytochrome P450 Affinities
- CYP2C9 pKi
- CYP2D6 Classification
- P-gp Transport Classification
- hERG pIC_{50}
- Plasma Protein Binding Classification

The principles on which these models are created and the interpretation of the results they generate will be discussed. More detailed descriptions of the individual models may be found in the Appendices Subsection 15.1.

6.1 Modelling Principles

A QSAR model is a mathematical relationship between the calculated characteristics of a molecule, its 'descriptors', and the biological or physicochemical property being modelled. The form of this mathematical relationship is determined by adjusting its parameters to optimise the fit to a set of molecules for which the property has been measured. This is known as the 'training set'. Thus, the three ingredients necessary to create a QSAR model are:

- A set of molecules for which the property being modelled has been accurately determined
- Descriptors which have a significant correlation with the property being modelled
- A mathematical approach to fit the form of the relationship between the descriptors and the observed property values

When a model has been generated that accurately fits the values in the training set, it is important to ensure that the relationship identified is not specific to the compounds in the training set. To be effective, a model must be sufficiently general to apply to novel molecules with a satisfactory degree of accuracy. To ensure this, the performance of a model should be confirmed against a set of molecules with known properties, which were not used in training the model. This set is known as the 'test set'. Some of the newer models in StarDrop have been built using the Auto-Modeller. In these cases the initial data set is divided into three subsets; training, validation and test, where the validation set is used to select between the models built before the test set is used to confirm the selected model's performance (see Chapter 8).

Once the generality of a model has been validated, it can be applied with confidence to molecules similar to those in the training set of the model. However, if a model is presented with a molecule with significantly different characteristics to those in the training set, it is impossible to know if the model remains valid for this novel chemistry. A prediction can be made; however, the confidence in the prediction is an unknown. The range of chemistry covered by the training set is referred to as the 'chemical space' of the model. Molecules which fall outside the chemical space of a model are explicitly automatically identified, so that the results may be treated with appropriate caution.

Each of these components and the process used to generate the StarDrop models are described in the following subsections.

6.2 Data Sets

The foundation of a high-quality model must be a set of molecules for which the property has been determined with the highest possible accuracy and consistency. The accuracy of a QSAR model cannot exceed the quality of the data with which it was built.

The source of the data used to develop the StarDrop models varies. In some cases, data from the published literature have been used, particularly for those data requiring *in vivo* measurements. Wherever possible, data sets have been generated using in-house protocols to ensure the maximum consistency between values. Regardless of the source of the data, strict quality control has been applied to the data sets used to develop the StarDrop models. In order that the model may be applied to as wide a range of chemistry as possible, the training set used to create StarDrop models covers as much chemical diversity as possible.

Prior to modelling, the data set must be divided into training and test sets (and validation set sometimes) as discussed above. Various techniques are available to design training and test sets depending on the size and chemical diversity of the overall data set. For instance, large and diverse sets can be divided by random selection. On the other hand, small and poorly diverse sets might be separated into training/test sets by means of structural similarity-based approaches to ensure that all of the chemical classes for which data are available are represented in both sets.

6.3 Descriptors

Descriptors are characteristics of molecules that can be calculated from the chemical structure. Using an appropriate set of descriptors, the most important features of the mechanism giving rise to a measured property are captured and hence a mathematical correlation with those observations can be derived.

Of highest priority when developing a model is the accuracy in predicting the observed property, and hence the correlation with the selected descriptors. However, it is also important that the results can be interpreted to provide guidance on the effects of chemical modifications on the predicted property. For this reason, wherever possible chemically interpretable descriptors are used, despite some potential penalty in accuracy.

Molecules may be described by an enormous range of characteristics. These vary in complexity and the computational cost of calculation, and can include:

- Structural descriptors

  These are typically simple descriptors that count atoms or functional groups that may be relevant to specific mechanisms. For example, hydrogen bond donors and acceptors, acidic or basic functionalities or rotatable bonds.

- Topological descriptors

  Also called molecular connectivity indices, these are 2D-descriptors computed from the molecular graph. These are a more complex form of structural descriptors that identify atoms in specific environments, capturing the effects of neighbouring atoms, for example E-state indices.

- Surface properties

  Solvents or protein binding pockets interact with the surface of a molecule. Therefore, the distributions of surface properties are commonly used to capture information on these interactions, for example, the polar surface area (PSA) or hydrophobic surface area.

  These properties should ideally be calculated from the 3-dimensional (3D) structure of the molecule, as different conformations can significantly change the accessible surface of a molecule. However, as it is computationally expensive to identify the most energetically

favourable 3D conformations of flexible molecules, surface properties are often approximated from the 2D chemical structure.

- Electronic properties

  The electronic properties of a molecule, such as the ionisation potential, dipole moment or distribution of the highest molecular orbital (HOMO) can dramatically influence the interactions of a molecule. Calculation of these properties requires computationally expensive quantum mechanical simulation, so these are rarely used in QSAR models designed to deliver high throughput.

- Whole-molecule properties

  Some properties of the whole molecule are often important in characterizing its interactions. Simple examples include the molecular weight or volume of the molecule, which may limit the size of binding pocket which can accommodate the molecule, or lipophilicity (logP) which influences the distribution of molecules between lipid and aqueous environments. The shape of the 3D conformation of a molecule will also influence binding to proteins.

When building a model to publish as part of the StarDrop platform, a wide range of descriptors are always considered, including descriptors that are calculated from the 3D molecular structure. These are evaluated using a range of statistical techniques as well as the experience of the modeller to identify those with the greatest significance for the modelled property. 3D descriptors are only used where they offer a significant improvement in the accuracy of the model, due to the additional computational cost that they incur. New descriptors are often developed to capture specific interactions that can be identified from the training set.

6.4 Fitting Methods

A number of approaches may be used to train a QSAR model. Many approaches are applied to each model developed for StarDrop and the most appropriate will be used to generate the final model. When choosing the modelling technique to employ, the highest priority is the accuracy of the resulting model. Where possible, a technique that yields an accurate model that may be easily interpreted is used, so that the results can be utilized to guide the design of compounds.

The modelling techniques used in building the models are briefly described below. More detailed information can be found from the references given and Chapter 8.

6.4.1 Linear Regression

As the name suggests, linear regression techniques are used to derive linear relationships between descriptors and the observed property values for the training set of molecules.

Multiple linear regression (MLR) (Draper & Smith, 1981) is one of the oldest methods used to find a linear relationship between the observed properties and a set of descriptors. A problem with this approach is that 4-5 samples (i.e. molecules with experimental data) are required for each descriptor used. Also, descriptors that are correlated or have skewed distributions, a common problem with chemical structures, will give poor regression models.

Rule-based Decision Trees (see below) may be combined with MLR to build ‘continuous’ models describing the QSAR for all molecules belonging to classes of molecules characterized by each rule. In other words, this approach effectively classifies a set of compounds according to structural parameters and evaluates a separate QSAR model for each subset, rather than fitting a single model to the entire set. An example of this approach is the Cubist method (Quinlan R. , Bagging, Boosting and C4.5, 1996).

Partial least squares (PLS) (Geladi, 1992) (Wold, Sjostrom, & Eriksson, 1999) has become a standard technique in this area. It overcomes most of the known problems with MLR, and allows the use of correlated descriptors. A large number of descriptors may be used, even larger than the number of molecules in the training set, and the descriptors with the most influence on the model can be

conveniently identified. PLS is part of the Auto-Modeller suite of modelling techniques (see Section 8.7 for details).

6.4.2 Non-linear Regression Techniques

Gaussian Processes is a powerful computational method for predictive quantitative structure-activity relationship (QSAR) modelling (MacKay, 2003) (Rasmussen & Williams, 2006) (Gaussian Processes website, 2007). Using a Bayesian probabilistic approach, the method is widely used in the field of machine learning but has rarely been applied in QSAR and ADME modelling (Obrezanova, Csanyi, Gola, & Segall, 2007) (Burden, 2001) (Schwaighofer, et al., 2007). This method overcomes many of the problems of existing QSAR modelling techniques. The method is suitable for modelling non-linear relationships, does not require subjective *a priori* determination of parameters such as variable importance or network architectures, works for a large number of descriptors, has built-in mechanisms to prevent over-training and does not require cross-validation. Together with each prediction the method provides an estimate of the uncertainty in prediction. The performance of Gaussian Processes compares well with, and often exceeds, that of artificial neural networks. Gaussian Processes are eminently suitable for automatic model generation and are included in the Auto-Modeller suite of modelling techniques (see Section 8.5 for details).

Radial basis functions (RBFs) have been praised for their simplicity, robustness and ease of implementation in multivariate scattered data approximation. Such techniques have been applied with success in problems ranging from training neural networks to image compression (Buhman, 2003). RBFs have not been commonly used in the QSAR field yet provide a good solution for both small and large data sets. However, they can be sensitive to noise created by excessive descriptors. In order to avoid this, a genetic algorithm (GA) can be used to run a stochastic search of the descriptor space and identify the most significant set of descriptors.

Genetic algorithms (GAs) are a class of heuristic search algorithms inspired by the mechanism of evolution (Goldberg, 1988). The set of parameters defining a model represents the 'DNA' for an initial population with appropriate genetic diversity. A combination of 'mutation' and 'cross-breeding' within this population is used iteratively to evolve the population and the 'fittest' in each generation is selected for the next iteration. The fitness of each model in the population is judged by the agreement between the predictions of the model and the data in the training set.

The RBF technique coupled with a GA is part of the Auto-Modeller suite of modelling techniques (see Section 8.6 for details).

6.4.3 Classification Methods

Where multiple biological mechanisms contribute to determining an ADME property, or a strongly non-linear relationship between the descriptors and the modelled property exists, it is often not possible to make a numerical prediction of that molecular property. Biased or noisy data sets can also make it impossible to derive a good numerical correlation with the observed data.

In these cases, classification methods are often employed which assign the property of a molecule to one of two or more classes. Decision Trees (DT) (Quinlan R. , Induction of Decision Trees, 1986) are a commonly used recursive partitioning approach to building classification models. They are suitable for categorical data and able to model non-linear relationships. They work well in the presence of many descriptors and are able to select those most relevant to a property. DT models are transparent for interpretation allowing you to understand the underlying structure–activity relationships. DT models contain a structure in which the branch taken at each intersection is determined by a rule relating to one or more descriptors. Each 'leaf' of the tree is assigned to a class. The Auto-Modeller uses this technique to build classification models (see Section 8.8 for details).

Random Forests (Breiman, 2001) is an ensemble method that makes predictions based on the output of a collection of random trees. This technique can be used to build both classification and regression models: for classification, the prediction is given by a majority vote over the committee of trees, and for regression, the prediction is set to the average output over all of the trees.

6.5 Validation

Model accuracy is assessed using well-known statistics. In the case of continuous models, these are the root-mean square error (RMSE) and the square of the correlation coefficient, R^2. The best model is the one with the highest R^2 value and the smallest RMSE value. For classification models, the model performance may be obtained from several metrics derived from a 'confusion matrix' that reflects the possible ways a compound may be classified and observed. The preferred metrics are as follows:

- Model sensitivity – the probability that a compound will be predicted to be in class X if it is actually in class X
- Model specificity – the probability that a compound will be in class X, if it is predicted to be in class X
- Model accuracy – the number of correctly classified compounds divided by the total number of compounds.

Figure 6.1 illustrates the problem of 'over-training'. The complexity of a model (i.e. the number of descriptors, or complexity of the mathematical relationship between the descriptors and measured property) can be increased until the model perfectly fits the data for the training set. However, beyond a certain point, the model typically begins to fit the idiosyncrasies of the training set molecules and loses its ability to generalize to new chemistry.

There are a variety of strategies that can be applied to avoid this situation. For example, within the training set, 'cross-validation' can be used. In this procedure, a series of models are fitted to subsets of the training set, from which small numbers of molecules have been omitted. The resulting models are used to make predictions for those compounds left out and the correlation between true and predicted values calculated. This is then compared to the correlation for the model trained on the entire training set. Similar correlations should be observed if the model is not over-trained.

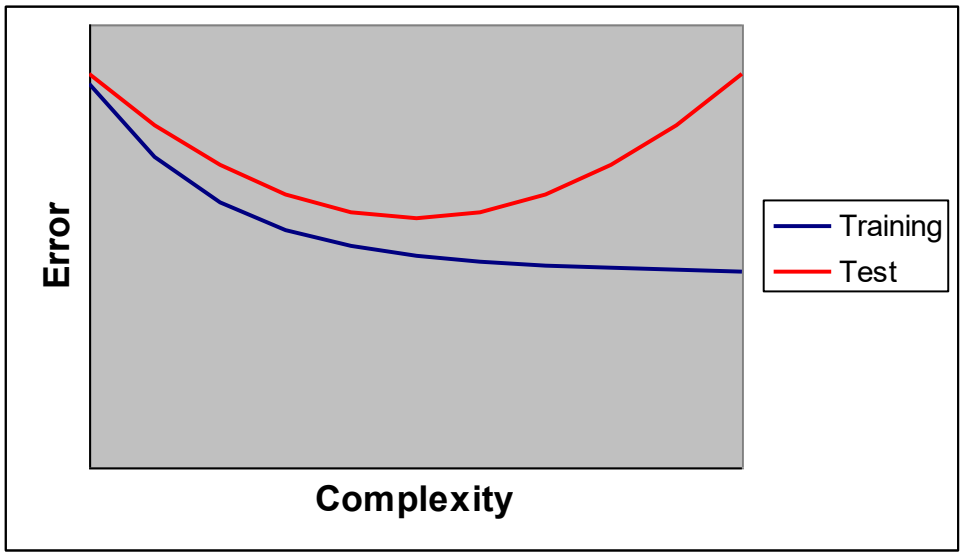


Figure 6.1 An illustration of the problem of 'over-training'. As the complexity of the model increases, the error in fit for the training set will continue to decrease until a perfect fit is achieved. However, the ability of the model to make accurate predictions for new chemistry, for example molecules in an independent test set, will eventually become worse.

Although techniques such as cross-validation are useful in the training process, in order to rigorously ensure that the model is not over-trained it is essential to validate the model's predictive power using an independent test set. This test set is removed from the data prior to modelling and is only used in the final validation step, when the predictions for this set are compared with the measured values. In order to be accepted, the correlation found for the test set must not differ significantly from that for the training set.

Once validated in this way, the results for the training and test sets are used to estimate the statistical confidence in the model.

6.6 Chemical Space

The 'chemical space' of a model, represents the range of model descriptors that are well represented by the molecules in the training set for that model. The position of a novel molecule relative to this chemical space affects the confidence with which a prediction can be made. This is illustrated in Figure 6.2.

Some modelling techniques have an inherent ability to provide an estimate of confidence in prediction. The Gaussian Processes method is an example of such a technique because it provides a standard deviation for each individual prediction.

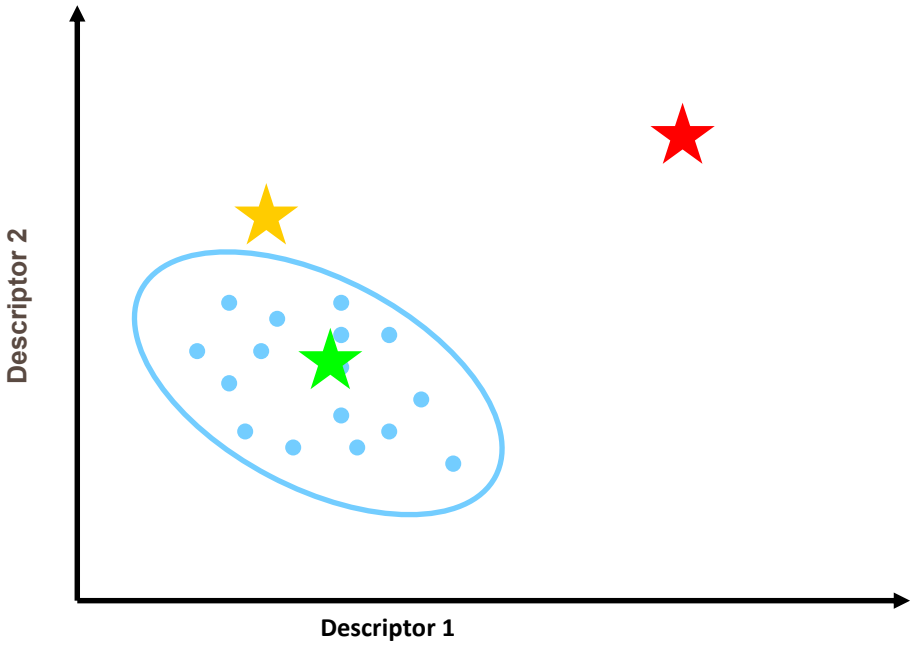


Figure 6.2 A schematic illustration of a 2-descriptor model chemical space. The training set compounds are represented by blue dots and the chemical space limit by the blue ellipse. Stars represent novel compounds for which predictions are made. The compound represented by the green star lies within the chemical space of the model; the yellow star represents a compound lying outside the chemical space but with similar characteristics to the training set compounds; and the red star represents a compound that is significantly different from the molecules in the training set.

For models created by other methods the chemical space of the model and the confidence in prediction are defined and obtained as follows.

The confidence for a compound lying within the chemical space of the model will be high, and the uncertainty in this prediction is estimated from the root-mean-square error (RMSE) of prediction. For compounds lying close to the chemical space of the model the prediction of the model will have a higher associated uncertainty as this prediction represents a small extrapolation. For compounds lying far from the chemical space, a prediction can be made, but no confidence limits can be assigned as this represents compounds with significantly different characteristics from those studied in the development of the model. These three situations are represented in the output of the confidences in predictions by the StarDrop models.

The StarDrop models employ the Hotelling's T^2 method for representing the chemical space of the model. This technique provides an estimate of the statistical confidence that a new molecule lies within the descriptor space of the model, represented by the training set compounds. By default, a confidence limit of 95% in the Hotelling's T^2 distribution is used to define the chemical space of the model. Where large numbers of data points are available in the independent test set for a model, and a statistically significant number lie outside this chemical space, an estimate of the RMSE in prediction for these

compounds is used to estimate the additional uncertainty for predictions outside the chemical space of the model.

Due to the limited size of the available data sets for some models, there are not sufficient compounds in the test set that lie outside the 95% confidence limits of the Hotelling's T^2 distribution to estimate the confidence. In these cases, the chemical space is defined by the 99% confidence limits of the Hotelling's T^2 distribution. No confidence limits are assigned to predictions made for compounds lying outside this chemical space, i.e. no claims are made regarding the accuracy of the prediction. The uncertainty in the predictions for these compounds is reported as maximal i.e. for continuous models an infinite uncertainty and for classification models an equal probability for each class.

6.7 Interpreting Model Results

The output produced by the StarDrop QSAR models depends on whether the model is a classification or continuous model. The results are displayed in the user interface (see StarDrop User Guide) and may also be exported as comma-separated variable (CSV) or MDL SD files for further analysis in third-party applications.

6.7.1 Classification Models

A classification model generates a prediction for the class into which the property of a molecule is most likely to fall. This is represented as a text string, e.g. "+", "-", "yes" or "no".

The uncertainty in the prediction is represented as a list of probabilities, one for each possible class for the model. These represent the probability that the property of the molecule falls into each class of the model. The class with the highest probability corresponds to the result displayed for the compound. Within StarDrop you can choose to see both the predicted class and the probability that the molecule lies in the predicted class. The higher the probability displayed, the greater is the confidence in the prediction.

When the results are exported to a CSV or SD file, the full list of probabilities for each class can be included.

6.7.2 Continuous Models

Continuous models predict a numerical result for a property. The predicted value is the 'most likely' value of the property, but there is a statistical uncertainty in the predictions of the model; similar to the way that any experimental result has a statistical error. The StarDrop models estimate this uncertainty from the validation and chemical space procedures outlined above. Within StarDrop you can choose to see the standard error in prediction next to the prediction. Under classical statistical assumptions, the true property value will lie within one standard error of the most likely value in 64% of cases (in 95% of cases, the value will lie within two standard errors). For continuous models, the lower the value displayed, the greater is the confidence in the prediction.

If the model statistics are displayed and no uncertainty can be calculated for a prediction because the molecule lies outside of the chemical space of the model, the standard error will be shown as "inf" for that qualitative prediction.

When the results are exported to a CSV or SD file, the standard error in a prediction will be shown in the column adjacent to the prediction.

6.8 Global versus Local Models

The StarDrop QSAR models are 'global' models of ADME properties, i.e. they are developed to cover as wide a range of chemical diversity as possible. However, as a drug discovery project progresses, the chemistry under consideration often focuses on a small number of chemical series in which the molecules are structurally similar. Global models may lack the resolution required to distinguish between molecules with subtle differences and once *in vitro* data has been generated for a chemical series, the ability of the corresponding models to discriminate within the series should be tested. If a model is found to lack discrimination, it may be possible to develop a local model to improve resolution within this limited range of chemistry. The StarDrop Auto-Modeller (Chapter 8) can be used to build local models.

7 P450 Metabolism Models

The models of Cytochrome P450 (P450) metabolism predict the likely outcome of metabolism by seven of the most important drug metabolising P450 isoforms, CYP3A4, CYP2D6, CYP2C9, CYP1A2, CYP2C19, CYP2C8 and CYP2E1. The models provide two outputs:

- **Regioselectivity**: The relative proportion of products formed by metabolism at each potential site, i.e. the most likely metabolites that will be formed **if** the molecule is a substrate for one of the isoforms modelled.
- **Site lability**: For CYP3A4, this provides an estimate of the absolute vulnerability of each potential site of metabolism, which is a measure of the efficiency of metabolism at that position.

Thus, these models can be used to aid in the redesign of a compound to overcome metabolic liabilities or, combined with other factors such as affinity or lipophilicity, to indicate a risk of high metabolic rate or 'turnover'.

These models differ from the QSAR models, discussed in the previous section, in that they are based on mechanistic models of the chemical reactions that lead to the formation of metabolites. Thus, although experimental data is used to tune the parameters of the model and validate the result, the form of the underlying model is not based on an empirical fit to a training data set and this gives greater transferability across a wide range of chemistry without loss of accuracy.

Accurate modelling of the chemical reactions requires quantum mechanical simulations, which are much more computationally expensive than the descriptor calculations employed by the QSAR models discussed earlier in Chapter 6. Consequently, the P450 metabolism models are significantly slower than the QSAR models, taking several minutes per compound. However, they provide detailed results identifying the most likely cause of metabolic instability for a compound, helping to guide chemical modifications aimed at reducing metabolic vulnerability.

7.1 Introduction to P450 Metabolism

The development of methods to predict the sites, products and rates of metabolism is an important avenue of research and finds application in the development of drugs, cosmetics, nutritional supplements and agrochemicals. It is necessary to understand the pharmacokinetics of a molecule and ensure that it has sufficient exposure at the target to exert its therapeutic effect. In this regard it would be helpful to give an absolute prediction of the rate of, or at least the lability of a compound to, metabolism, rather than just a rank ordering of sites; a factor often neglected by metabolism prediction tools. It is also important to predict the formation of toxic metabolites, which contributes to the high attrition rates experienced in the development of new chemical entities, the imposition of black-box warnings or even the withdrawal of approved pharmaceuticals. Thus, the ability to identify potential toxic metabolites early and make predictions about metabolic stability are of crucial importance in the drug discovery process.

The cytochrome P450s (CYPs) are a family of heme-containing enzymes involved in the phase-I metabolism of over 90% of drugs currently on the market (Guengerich P. , 2006). The CYP family consists of 57 isoforms (Lewis D. , 2004) with the largest contribution to xenobiotic metabolism coming from CYP3A4, the most promiscuous isoform, followed by CYP2D6 and CYP2C9. A comprehensive overview of the structure, reactivity and catalytic cycle of CYPs can be found in the review paper from Shaik et al. (Shaik, et al., 2010).
The catalytic action of P450s is predominantly that of a monoxygenase:

$$RH + O_2 + 2H+ \rightarrow ROH + H_2O$$

where RH is the substrate molecule. The most common reactions catalyzed by CYPs involve the insertion of a single oxygen into an organic molecule, such as C=C epoxidation, aromatic C oxidation and aliphatic C hydroxylation, the last example often leading to N-dealkylation or O-dealkylation if oxidation occurs on a suitable leaving group in an amine or ether moiety. The addition of oxygen into

the substrate is a precursor to excretion from the body, driving an increase in polarity and hydrophilicity and facilitating Phase II metabolism pathways such as glucuronidation.

The heme moiety at the catalytic centre of the CYPs is conserved across all isoforms, where a highly activated oxy-heme, formed by cleavage of molecular oxygen, is generated within the catalytic cycle, as shown in Figure 7.1. In addition to the main catalytic cycle there are two significant decoupling pathways, shown as (D1) and (D2) in Figure 7.1. In addition to the main catalytic cycle there are two significant decoupling pathways, labelled D1 and D2, resulting in the formation of hydrogen peroxide and water respectively and returning the active site heme to an inactivated state. The relative rate of decoupling compared to substrate metabolism will influence the observed rate of the conversion of substrate into metabolites and is an important consideration when making absolute assessments of metabolic stability.

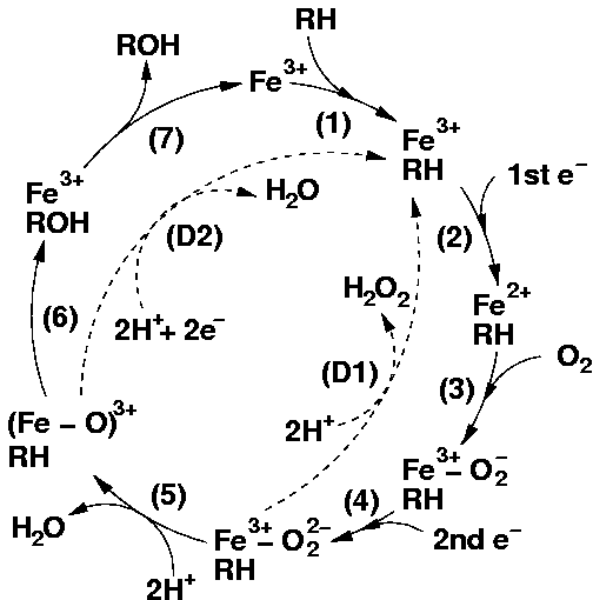


Figure 7.1 The catalytic cycle of the Cytochrome P450 enzymes. The decoupling pathways to form hydrogen peroxide and water are labelled D1 and D2 respectively.

Experimental investigation of xenobiotic metabolism can be both resource and time consuming, which has encouraged the development of computational techniques. These can be separated into two distinct categories: ligand-based and structure-based. In the first approach, structures and properties of known substrate or non-substrate ligands are modeled to develop structure-activity relationships. The second approach is focused on the structure of the metabolising CYP enzyme, its known reaction mechanisms and its interactions with substrates. For a general overview of current computational tools to predict sites of metabolism (SOM) the reader is referred to the many comprehensive review papers (Kirchmair, et al., 2012), (Kulkarni, Zhu, & Blechinger, 2005), (Tarcsay & Keseru, 2011), (Ekins, et al., 2005), (Vaz, et al., 2010).

Most metabolism prediction tools incorporate some form of reactivity and accessibility considerations. The method described herein is no exception: a ligand-based approach is used to model steric and orientation effects whilst the electronic activation energy is modeled using quantum mechanical (QM) simulations to calculate the energies of substrates and reaction intermediates for each potential SOM. This approach offers several advantages:

- QM methods are based on fundamental physical principles and therefore transfer well between chemical classes; they do not rely on specific examples being present within a training set used to fit an empirical model
- Each potential SOM is considered in the context of the whole molecular environment in which it resides, rather than identifying fragments within a substrate and treating each as a discrete uniform entity regardless of neighbouring functional groups
- QM methods can estimate the activation energy of the rate-limiting step of the oxidation reaction, allowing comparison of lability on an absolute scale.

The previously published SMARTCyp method (Rydberg, Gloriam, Zaretzki, Breneman, & Olsen, 2010) also uses a QM-based approach to predict CYP SOM. However, the approach differs from that described herein in that the SOMs are ranked based on a look-up table of small functionalities for which the activation energies have been previously calculated using ab initio density functional (DFT) methods. The use of ab initio QM methods avoids the need for detailed experimental data on which to base estimates of the activation energies. However, these calculations are computationally very expensive and the use of a look-up of pre-calculated results is required to return results in a practical time-frame. Therefore, this approach does not take into account potential long-range effects due to the

environment of the whole molecule, an important factor for a medicinal chemist developing a lead series and aiming to predict the likely impact of structural changes on metabolic stability.

The use of semi-empirical QM calculations that estimate the activation energies for each aliphatic and aromatic SOM have previously been described (Jones, Mysinger, & Korzekwa, 2002). Semi-empirical methods are significantly faster than ab initio methods and therefore can be applied to an entire substrate on a routine basis. However, they typically require detailed experimental data with which to parameterize a free energy relationship and therefore these models do not include less common, but important, pathways such as epoxidation, or N- and S-oxidation, due to the lack of sufficient experimental data.

The methods described herein build on both of these methods to achieve transferability, application to the whole substrate to explicitly consider the molecular environment of each SOM and computational efficiency, returning results in approximately 1-2 minutes per compound on a single CPU. In the following section, the theory and implementation of the models will be explained in detail. The performance of the models on independent test sets will be presented in the Results section with comparisons made to the SMARTCyp (Rydberg, Gloriam, Zaretzki, Breneman, & Olsen, 2010) method.

7.2 Theory and Implementation

7.2.1 Modelling Principles

The key factors that determine the SOM are reactivity and accessibility. The models described herein estimate the activation energy at each potential SOM in a substrate using fundamental and transferable QM methods, rather than relying on empirical pattern matching with a limited domain of applicability. The QM models are independent of isoform, reflecting the consistent reaction pathways across isoforms.

However, the binding pockets of the different CYP isoforms differ in size, shape and chemical composition and influence the orientation of the substrate relative to the reactive oxy-heme core. Electrostatic, hydrogen bonding and lipophilic interactions between substrate and CYP binding pocket have varying effects across isoforms and will cause different orientations to be favorable. In addition, the steric bulk within a substrate will influence the accessibility of sites to the reactive oxy-heme core, with those sites embedded towards the center being less accessible than those in open sites on the periphery of the substrate. These steric effects are also isoform specific, as the different sizes of the CYP binding pockets can accommodate different levels of steric hindrance. The models described herein are able to capture these orientation and steric effects with ligand-based models trained on isoform-specific data sets, enabling adjustments to be made to the QM-generated electronic activation energy that are specific to each isoform.

The relative rate of metabolism for a site can be calculated from the activation energy, E_a, for the rate-limiting step in the reaction pathway, as the rate is proportional to the negative exponential of the activation energy (the Arrhenius equation):

$$k_i \propto e^{-\frac{E_{ai}}{kT}},$$

where k_i is the relative rate of metabolism at site i, E_{ai} is the activation energy of site i, k is the Boltzmann constant and T is the temperature.

As discussed above, by directly calculating the activation energy, the lability of sites can be compared on an absolute scale between compounds, rather than just a relative ranking of sites within a compound. This is achieved by comparing the rate of product formation with that of a decoupling pathway, to give an absolute assessment of the efficiency of the product formation step in the catalytic cycle. This enables a medicinal chemist to identify likely metabolically vulnerable positions in a molecule and can be used to guide development away from compounds with potentially rapid clearance. In the remainder of this section the various aspects of the CYP metabolism prediction models will be described in detail.

7.2.2 Quantum Mechanical Models of Electronic Effects

The oxidizing species and the chemical mechanisms of oxidation are the same for all CYP isoforms. This allows the intrinsic vulnerability of the sites on a potential substrate to be calculated with reference only to the structure of that molecule.

The oxidation reactions proceed via different pathways depending on the nature of the site of metabolism on the substrate: aliphatic hydroxylation proceeds via an initial hydrogen abstraction followed by reaction of the ferryl oxygen with the alkyl radical, a process known as the rebound mechanism (Ogliaro, et al., 2000); alkene epoxidation proceeds via activation of the double bond to form an iron alkoxy radical species in a tetrahedral orientation (Shaik, de Visser, Ogliaro, Schwarz, & Schroder, 2002); aromatic C oxidation proceeds via activation of an aromatic bond (Bathelt, Ridder, Mulholland, & Harvey, 2003) followed by an intra-molecular hydrogen atom transfer known as “NIH-shift'' (de Visser & Shaik, 2003); and direct oxidation of hetero atoms such as sulphur and nitrogen proceeds via bond formation with the ferryl oxygen (Sharma, de Visser, & Shail, 2003), (Rydberg, Ryde, & Olsen, Sulfoxide, Sulfur, and Nitrogen Oxidation and Dealkylation by Cytochrome P450, 2008), although often dealkylation reactions are favorable over direct oxidation.

The reactivity model performs a QM calculation to estimate the 'electronic' activation energy, ΔHA, for every potential site of metabolism, using knowledge of the reaction pathway for that site. These calculations are performed using AM1 (Marti-Renom, et al., 1985), a quantum mechanical approach based on a semi-empirical Hamiltonian. While less accurate than a full ab initio simulation, AM1 is many times faster. Ab initio simulations have been used to identify systematic errors due to the use of AM1 and correction factors are applied within the electronic model (Rydberg, Gloriam, Zaretzki, Breneman, & Olsen, 2010), (Jones, Mysinger, & Korzekwa, 2002).

Direct calculation of activation energies is computationally very expensive due to the need to perform a transition-state search. Instead, the heat of reaction, ΔHR, is calculated from the heat of formation of the substrate and reaction intermediates and a Brönsted relationship is used to calculate an approximation to ΔHA (as a linear relationship has been shown to exist between the activation energy, ΔHA, and the heat of reaction ΔHR,). The parameters of the Brönsted relationship can be derived from detailed experimental regioselectivity data where this is available (Korzekwa, Jones, & Gillette, Theoretical studies on cytochrome P-450 mediated hydroxylation: a predictive model for hydrogen atom abstractions, 1990), (Jones, Mysinger, & Korzekwa, 2002). However, in some cases there are insufficient experimental data and, instead, high level ab initio calculations can be used to accurately calculate the activation energies which, in turn, can be used to derive the parameters of the Brönsted relationship for the faster, semi-empirical AM1 calculations.

However, due to the differences in the chemical mechanisms and methods for calculation of each of the pathways leading to oxidation, the energy scales of the activation energies will differ. Therefore, in order to compare the rates of reactions that proceed by different pathways, the activation energies must be on the same scale and therefore a normalization must be applied. To achieve this, the activation energy scale relating to the abstraction of Hydrogen from aliphatic carbon sites is used as a reference and calculations performed with other methods are transformed onto this energy scale. Figure 7.2 shows the linear relationship between the activation energies calculated using the ab initio B3LYP DFT method, as published by Rydberg et al. (Rydberg, Gloriam, Zaretzki, Breneman, & Olsen, 2010), and those estimated from ΔHR, calculated with AM1 and a Brönsted relationship, for hydrogen abstraction sites, as described below.

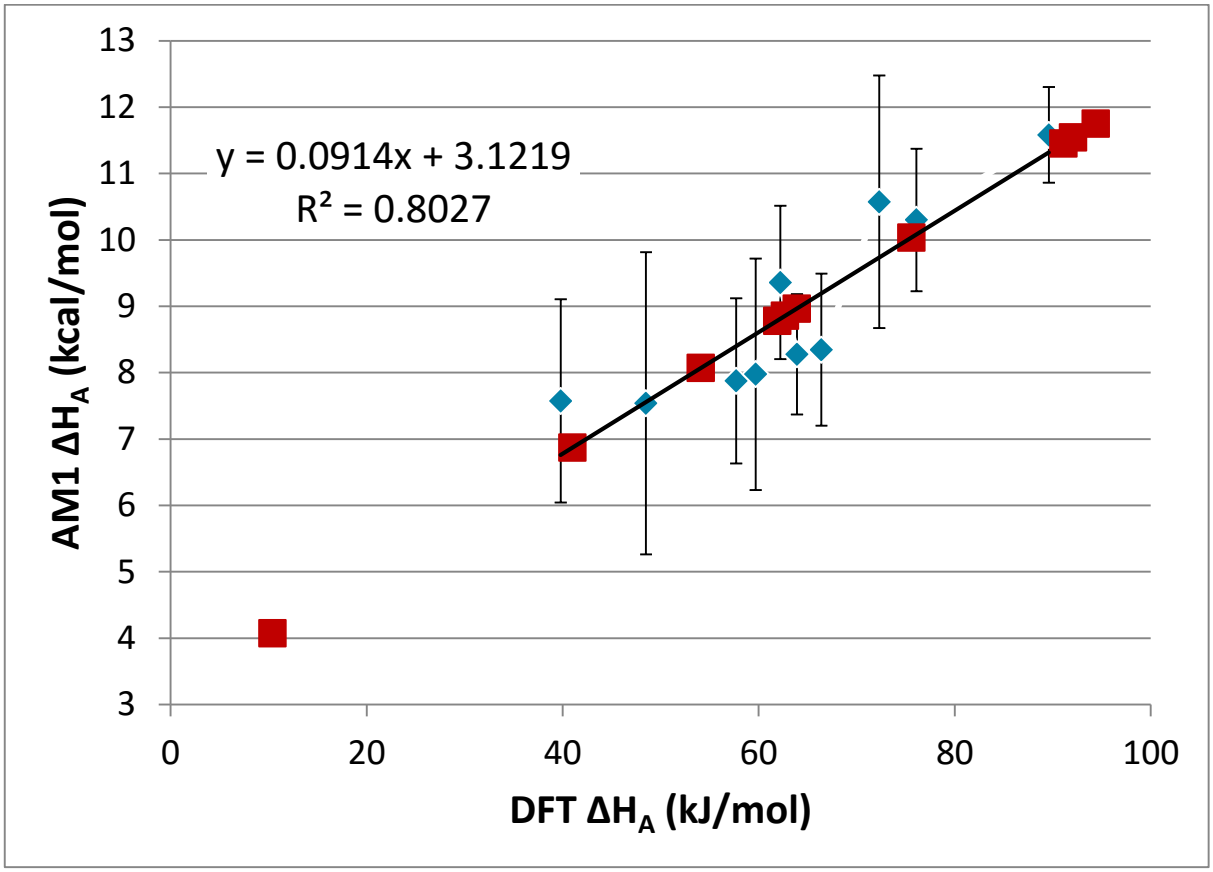


Figure 7.2 Graph showing the linear relationship between H-abstraction activation energies calculated using DFT and the models based on AM1. The points represented by blue diamonds show the average ΔHA, estimated using a Brönsted relationship based on ΔHR calculated with AM1, plotted against the single activation energy value assigned to the corresponding sites in SMARTCyp, derived from DFT transition state calculations 9. While only a single activation energy is assigned to each class of site by SMARTCyp, in practice there may be significant variations between similar sites due to different molecular environments in which they occur. To illustrate this, the error bars show one standard deviation in the ΔHA values calculated using AM1 on the full molecules. These averages were calculated over a total of 2252 sites on a wide diversity of compounds. The minimum number of sites in each class for which an average is shown was 18. The transformation of N oxidation and hydroxylation energies from DFT, on the basis of this linear relationship, is represented by the red squares.

Further details of the calculations performed for each of the reaction pathways modeled are provided in the next few subsections.

Hydrogen abstraction reactions

The rate limiting step in the formation of a metabolite by hydrogen abstraction has been identified as the abstraction of the hydrogen from the substrate by the oxy-heme and formation of an alkyl radical intermediate. In the Brönsted relationship used to estimate the activation energy, an additional linear term involving the ionisation potential has also been found to be important to capture resonance effects in the transition state (Korzekwa, Jones, & Gillette, 1990). Using detailed experimental measurements of the relative rates of product formation at different sites of the same molecule, this pathway was modelled as described in (Korzekwa, Jones, & Gillette, 1990) to estimate ΔHA for each potential site of hydrogen abstraction.

Aromatic oxidation

Aromatic oxidation progresses by formation of a tetrahedral intermediate between the substrate and oxy-heme at the site of metabolism, followed by rearrangement to form a hydroxylated product. The formation of the tetrahedral intermediate is the rate limiting step in this process and the activation energy was also found to be proportional to the heat of reaction. Using experimental measurements of the relative rates of formation of different metabolites on the same molecule, the parameters of this relationship can be determined and ΔHA calculated on the same scale as that for hydrogen abstraction, as described in detail in (Jones, Mysinger, & Korzekwa, 2002).

Double-bond Epoxidation

Similar to aromatic oxidation, the epoxidation of a carbon-carbon double bond proceeds via the formation of a tetrahedral intermediate, followed by rearrangement to form the epoxide (Kumar, Karamzadeh, Sastry, & de Visser, 2010). The formation of the tetrahedral intermediate is again the rate limiting step, with the activation energy found to be proportional to the heat of reaction.

There are insufficient experimental data with which to confidently parameterize a Brönsted relationship and, in this case, we rely on activation energies calculated with ab initio DFT calculations that were shown to agree with experimentation observations, as described in (Kumar, Karamzadeh, Sastry, & de Visser, 2010). In this case, AM1 calculations of ΔHR exhibit an excellent correlation with the ab initio activation energies, as shown in Figure 7.3. This enables the estimation of the DFT activation energy from the AM1 ΔHR, which, in turn, can be transformed to calculate ΔHA on the same energy scale as that for hydrogen abstraction, using the linear relationship shown in Figure 7.2.

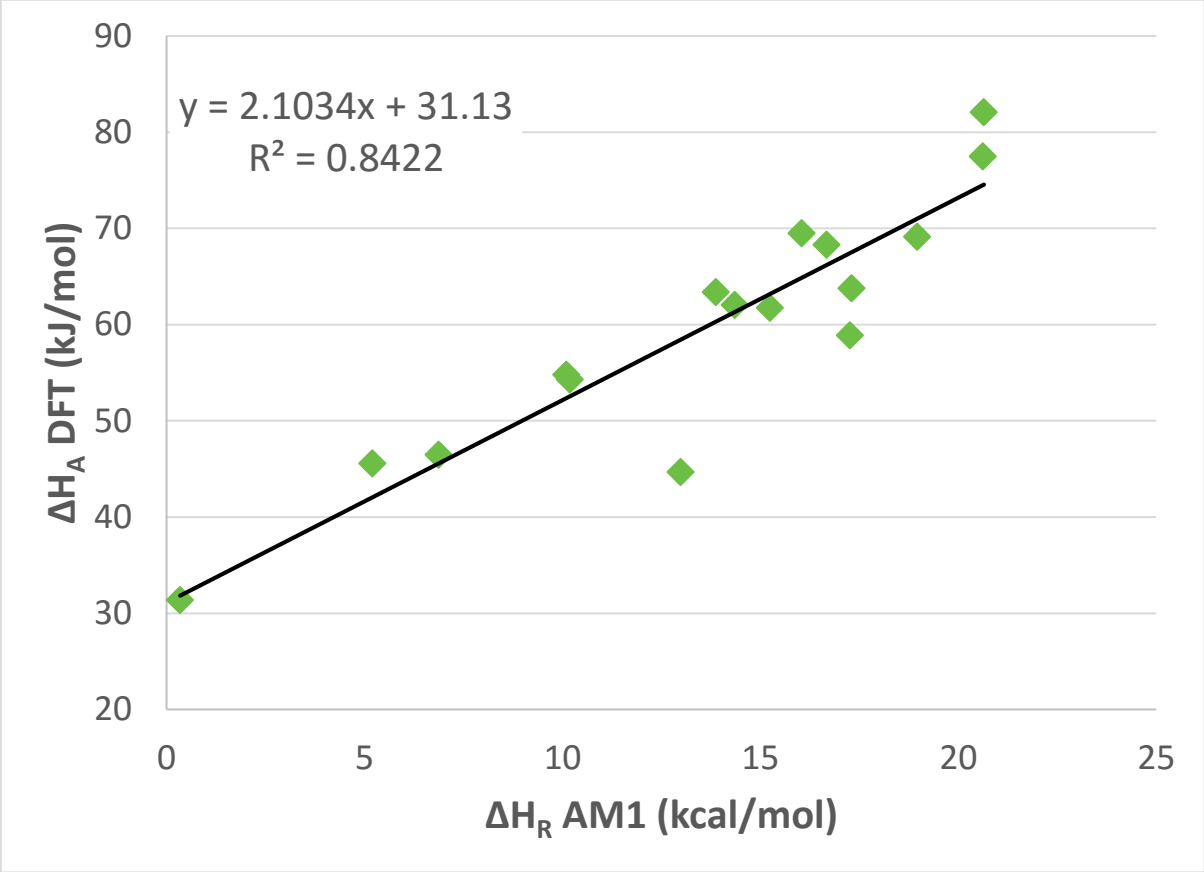


Figure 7.3 Graph showing the linear relationship between ΔHR calculated with AM1 and ΔHA calculated with DFT for potential sites of double bond epoxidation, as published by Kumar et al.19.

Epoxidation could proceed by formation of a tetrahedral intermediate with the carbon at either end of the double bond. Therefore, ΔHA is calculated for both potential sites and the lowest value is used to estimate the relative rate of epoxidation of the corresponding double bond.

Other direct oxidation pathways

For S-oxidation, N-oxidation and hydroxylation, and other pathways including desulfurization of phosphothioates, oxidation of disulfides and aldehyde oxidation/deformylation, there are limited experimental data regarding the rates of these reactions relative to other sites on the same compounds. Furthermore, ab initio DFT calculations indicate that there is less variation in these rates between similar functionalities. Therefore, for these sites, activation energies derived from ab initio DFT calculations published by Rydberg et al. (Rydberg, Gloriam, Zaretzki, Breneman, & Olsen, 2010) were transformed onto the same energy scale as the other sites described previously, using the linear relationship shown in Figure 7.2. As an illustration, the transformation of N-oxidation and -hydroxylation energies using this linear relationship is also shown in Figure 7.2.

7.2.3 Accessibility: Steric and Orientation Effects

In addition to the intrinsic vulnerability of a site in a molecule to oxidative attack, the accessibility of that site to the oxy-heme core will also influence the relative rate of metabolism. This effect is calculated as a correction to the activation energy due to the orientation of the molecule within the active site and steric hindrance by nearby atoms in the substrate.

Orientation effects are modeled by descriptors representing the topological distance to important functionalities such as acidic, basic, hydrogen bond donor/acceptor and lipophilic groups that interact with key residues in the CYP active site. The steric accessibility of a potential site of metabolism depends on the surrounding atoms in the substrate and will be influenced by nearby bulky functionalities or whether the site is part of a ring, a conjugated system or an aliphatic chain. The steric effects are modeled using descriptors representing the distance to functionalities introducing steric bulk surrounding the SOM. These functionalities are defined as SMARTS patterns (Daylight Chemical Information Systems Inc., n.d.).

Isoform-specific data sets have been carefully curated from the literature, as described in more detail in the following subsection, with the steric and orientation descriptors calculated for all sites in all molecules. Principal component regression models (Wehrens & Mevik, 2007) were trained on these data sets using knowledge about the metabolic fate of each site to set the dependent variable: 0 for a non SOM, 1 for a primary SOM, 0.5 for a secondary SOM and 0.25 for a tertiary SOM. In order to generate an adjustment to the electronic activation energy, ΔHA, it is necessary to calculate the adjustment on the same energy scale as ΔHA. This is achieved by including the ΔHA in the regression and scaling the descriptor regression coefficients relative to the ΔHA regression coefficient. These scaled coefficients can then be applied to descriptors for new molecules and the resulting correction added to the 'electronic' activation energy, ΔHA, to calculate the estimate of activation energy, Ea, adjusted for steric and orientation effects.

7.2.4 Data Curation

A detailed review of the primary literature was performed to prepare high quality datasets of isoform-specific human CYP substrates annotated with SOM. The papers were manually parsed to extract primary, secondary and tertiary SOM, along with the identity of the major and minor metabolizing CYP isoforms. The emphasis was on high quality data, retaining only human data and excluding data generated with inappropriate experimental conditions, such as un-physiological substrate concentrations. The consequence of this is that the data sets are smaller than some of those previously published (Zaretzki, et al., 2012), (Campagna-Slater, Pottel, Therrien, Cantin, & Moitessier, 2012). However, analysis of the chemical space covered by the CYP data sets against launched drug space shows that good coverage of drug-like chemical space has been achieved as illustrated in Figure 7.4 and the higher quality data is expected to result in more accurate models.

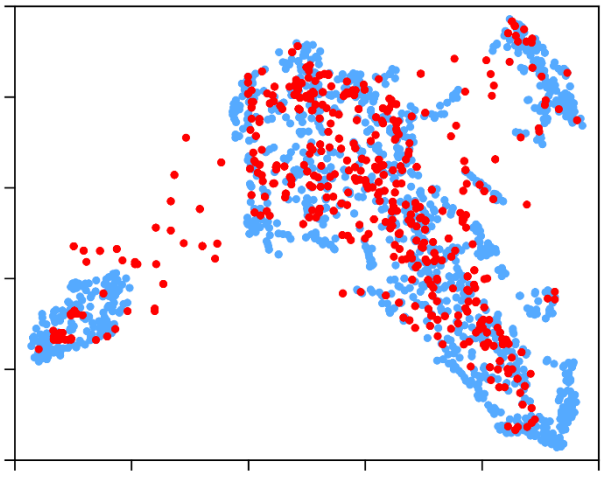


Figure 7.4 Illustration of the chemical space covered by the CYP data sets (red points) compared with approximately 1,300 launched drugs (blue points). In this chemical space plot, the proximity of two points represents the structural similarity between the corresponding compounds defined using a Tanimoto index based on a 2D path-based fingerprint. The distribution of points is generated using the t-distributed stochastic neighbour embedding algorithm (van der Maaten & Hinton, 2008).

Table 2 summarizes the number of compounds in the training set for each isoform, used to fit the contributions of the steric and orientation descriptors, and the independent test sets, used to validate each model. The data sets are available in the supplementary information for inspection, including references to the primary literature from which the SOMs were identified (see Supplementary Information below).

There is an element of judgement to be applied when classifying sites within a molecule as primary, secondary or tertiary

and identifying major/minor isoforms, with reliance placed on kinetic data from expressed supersomes and isoform-specific inhibition experiments with human liver microsomes. Variability between assays makes direct comparison of experimental data between publications challenging, but efforts have been made to make classifications across different molecules consistent.

Table 2 The numbers of compounds in the training and independent test sets of detailed regioselectivity data used to build and validate the models described herein. These data sets are provided as supplementary information.

CYP Isoform	$N_{training}$	N_{test}
CYP1A2	144	57
CYP2C8	80	27
CYP2C9	145	49
CYP2C19	136	49
CYP2D6	147	56
CYP2E1	76	30
CYP3A4	220	84

7.2.5 Calculating Regioselectivity

The regioselectivity of metabolism is the proportion of metabolism that occurs at each site. This proportion is given by the rate of metabolism at that site relative to the sum of the rates for all potential sites of metabolism. The advantage of calculating an approximation to the activation energy is that a relative rate can be generated (above), allowing the regioselectivity of metabolism at site *i* to be given by:

$$R_i = \frac{k_i}{k_{total}} \times 100,$$

where R_i is expressed as a percentage and $k_{total} = \sum_{\text{all sites}} k_i$.

7.2.6 Calculating Lability

The regioselectivity of metabolism describes the relative rate of metabolism of each potential site on a molecule. However, regioselectivity does not itself provide information on the absolute vulnerability or lability of each site to metabolism.

The lability of each site is derived by comparison of the predicted rate of the product formation step for the site with the water formation decoupling pathway in the catalytic cycle, labeled D2 in Figure 7.1. The rate of the decoupling pathway was measured using an intrinsic isotope effect methodology (Korzekwa, Trager, & Gillette, Theory for the observed isotope effects from enzymatic systems that form multiple products via branched reaction pathways: cytochrome P-450, 1989). If a site on a substrate is metabolized at a significantly higher rate than that of decoupling then metabolite formation will proceed with high efficiency. Conversely, if the rate of water formation is significantly higher than the rate of metabolism of a site then decoupling would dominate and metabolite formation at that site would not be observed. Specifically, the lability of site i is given by:

$$L_i = \frac{k_i}{k_w + k_i},$$

where kw is the rate of water formation via the decoupling pathway. The distribution of the lability of the sites on a compound is conveniently shown on a 'metabolic landscape' histogram using color as a

visual guide, based on the efficiency with which metabolism would occur at that site: red for labile (>0.80); yellow for moderately labile (between 0.35 and 0.80); green for moderately stable (between 0.05 and 0.35) and blue for stable (<0.05). An examples of this representation is shown in Figure 7.5.

The site labilities of individual sites can be combined to calculate the `composite site lability' (CSL) reflecting the overall efficiency of product formation for the molecule. This is calculated from the combined estimated rates of metabolism for all sites on the molecule:

$$CSL = k_{total}/(k_{total} + k_w),$$

An estimate of the uncertainty in the CSL estimation is also provided to allow the CSL to be used in probabilistic scoring,

$$CSL_{uncertainty} = k_w.k_{total}/(kT.(k_{total} + k_w)^2),$$

where k_{water} is the rate constant for water formation and kT = 0.616.

It should be noted that CSL is not a prediction of rate, but is one important factor influencing the rate amongst others, including reductions rates, which are often rate-limiting in the catalytic cycle, and binding affinity, which itself can be influenced by substrate properties such as size, liphophilicity and pKa. Therefore, a direct correlation between small changes to CSL and the CYP3A4 half-life or intrinsic clearance is not necessarily expected.

7.3 Interpreting Model Results

The output of the StarDrop P450 metabolism models consists of maps and tables summarizing the regioselectivity of metabolism for each isoform. A metabolic landscape and composite site lability are also included for CYP3A4 metabolism. (See StarDrop User Guide).

The following subsections discuss how these outputs may be interpreted and guide molecular redesign to overcome a metabolic liability.

7.3.1 Regioselectivity

An example regioselectivity map for the predicted metabolism of the molecule cisapride by CYP3A4 is shown in Figure 7.3. Sites are labelled with a predicted percentage of products formed due to metabolism at that position. Only sites with predicted percentages greater or equal to 1% are labelled. In the case of CYP3A4, the labels are colour-coded according to the 'site lability' of that position, in accordance with the metabolic landscape (see below).

The regioselectivity indicates the distribution of metabolites *if* the molecule is metabolised by the relevant isoform so it is assumed that in running the model you believe the molecule may be a substrate for the particular CYP.

The percentages shown are a qualitative guide to the formation of metabolites. For example, a site with greater than 50% predicted metabolism is likely to be the major metabolite formed by the corresponding isoform. Conversely, if there are a number of sites with similar predicted percentages, but no site with a significantly higher value, there is unlikely to be a single predominant metabolite. Sites predicted to contribute only a few percent to the metabolite profile of the molecule are unlikely to be observed in practice, but may be alternatives should the major predicted routes be inhibited.

7.3.2 Identifying Metabolites

In the large majority of cases, simple rules will determine the metabolites formed by oxidation at each potential site of metabolism. For example, metabolism at an aromatic carbon will generally result in the formation of a hydroxylated product at that position (e.g. sites C12 and C15 in Figure 7.3). Metabolism at a non-aromatic carbon adjacent to an oxygen or nitrogen is likely to form an O- or N-dealkylated product (e.g. site C6 in Figure 7.5). Metabolism at other aliphatic carbons will typically result in hydroxylation at that position.

These rules have been encoded as SMIRKS and are used within the P450 module to predict the metabolites formed by metabolism at each potential site. On rare occasions, rapid radical rearrangement can lead to minor metabolites formed by hydroxyl rebound at alternative positions. See for example the metabolism of debrisoquine by CYP2D6 (Lightfoot, et al., 2000)).

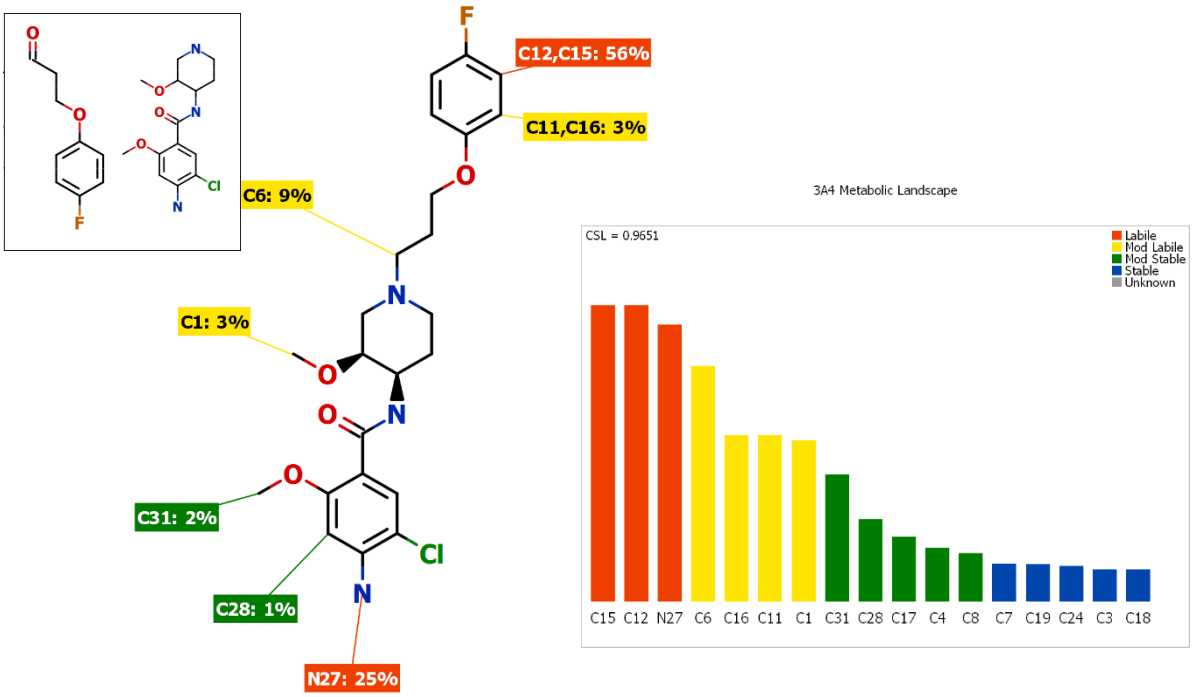


Figure 7.5 Predicted regioselectivity and metabolic landscape for P450 CYP3A4 metabolism of Cisparide. The metabolites predicted to be formed by N-dealkylation at position C6 are also shown inset.

7.3.3 Metabolic Landscape

The lability of each site on the molecule is indicated in a 'Metabolic Landscape' as vertical bars, with height indicating the degree of lability (See Figure 7.5 for an example). The category assigned to each site is indicated by the colour of the corresponding bar from red ('labile') to blue ('stable'). Two further categories, 'moderately labile' and 'moderately stable' are provided to indicate intermediate degrees

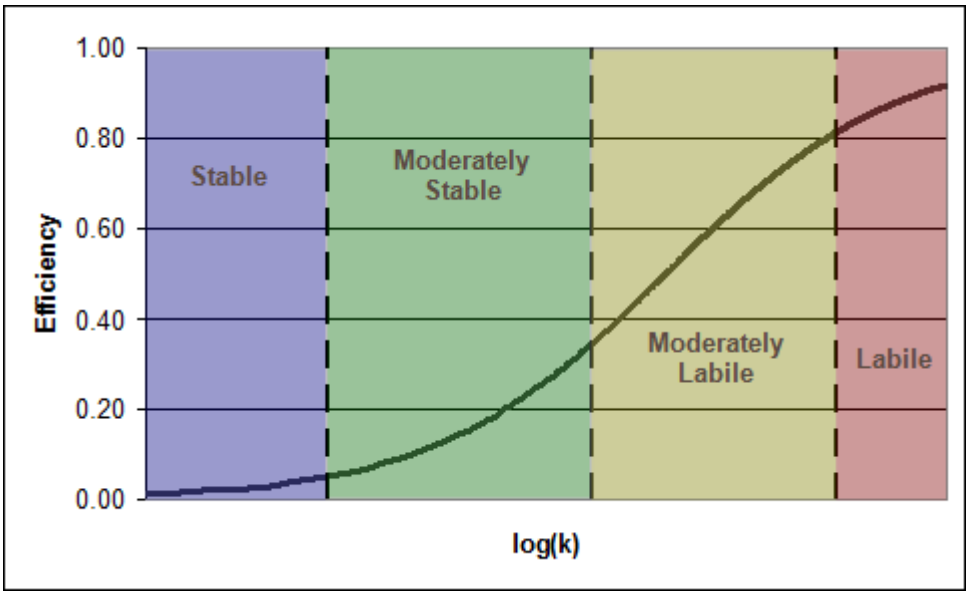


Figure 7.6 Graph illustrating the definition of metabolic site lability. The logarithm of the rate constant for the rate-limiting step of product formation at the site is plotted on the x-axis and the efficiency of metabolism, in competition with the decoupling pathway (D2), is plotted on the y-axis. The lability categories employed in the metabolic landscape are indicated by shading and labels.

of competition between metabolite formation and decoupling. The sites with the highest percentage regioselectivity will correspond to those with the highest lability on the left of the metabolic landscape.

The CSL is displayed in the top left of the metabolic landscape in the P450 display and in the 'P450' column of the dataset for molecules for which the P450 results have been calculated. The metabolic landscape can be used to guide the redesign of compounds that have a high metabolic turnover by a Cytochrome P450 enzyme (particularly CYP3A4). Modifying the chemical structure by blocking metabolism of a labile site will have the greatest impact on the rate of metabolism. However, if other sites with similar lability are present on the molecule, the impact will be reduced, as metabolism will switch to these alternative sites with little reduction in efficiency. Blocking all labile sites (and preferably moderately labile sites) may be necessary in order to have a significant impact on the rate of metabolism (See Figure 7.7 for an example). Note that 'blocking' labile sites can be achieved by altering steric and site orientation factors for a molecule as well as changing the electronic vulnerability at the specific site in question.

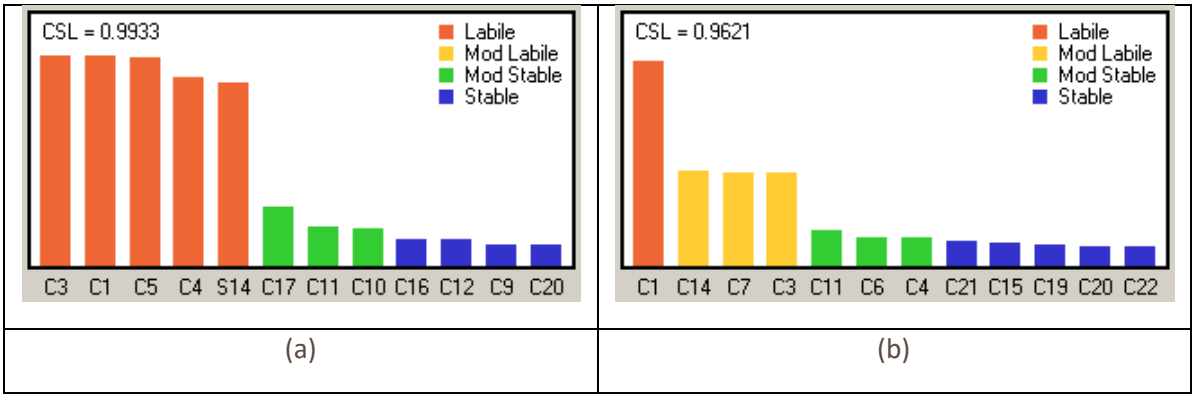


Figure 7.7 Example metabolic landscapes for two compounds (a) and (b). If these compounds had similar, high metabolic rates, the metabolic landscapes provide a guide to possible compound redesign to stabilize the molecules. In the case of compound (a), blocking one labile site, e.g. C3, is unlikely to have a significant effect on rate because metabolism will switch to alternative labile sites C1, C5, C4 or S14 with similar efficiency. However, we would expect a greater reduction in metabolic rate by blocking site C1 on compound (b), leaving only sites with significantly lower lability. Thus, compound (b) offers the better opportunity for improvement.

7.3.4 Composite Site Lability (CSL)

The CSL is an estimate of the efficiency of metabolism for the entire molecule, in competition with the decoupling pathway (D2) which leads to water formation.

In a chemical series which has high affinity for CYP3A4, one approach to decreasing the rate of metabolism would be to reduce the CSL by removing or blocking labile and moderately labile sites. Other approaches are also possible, including modifications that will reduce the affinity of the compounds for the metabolising enzyme, e.g. by reducing logP.

It is important to note that composite site lability is not, in itself, a model of rate of metabolism. Many factors contribute to determining the rate of metabolism, or 'turnover', of a molecule including;

- Affinity of the substrate for the enzyme
- The rate of the catalytic cycle, which is often limited by steps prior to product formation. This can be influenced by mechanistic inhibition, for example by Type II binding*
- Decoupling via the (D1) pathway. This results in the formation of hydrogen peroxide which limits the efficiency with which the active oxy-heme species is formed
- Decoupling via the (D2) pathway. This results in the formation of water which limits the efficiency with which product is formed

* Type II binding occurs when a substrate interacts directly with the iron in the active site haem. This dramatically reduces the rate of the first reduction, and hence the rate of metabolism of the substrate.

The CSL describes the last of these contributions to metabolic rate. Hence it is an essential component to estimate rate, but it is not itself a prediction of the rate of metabolism. In combination with other factors, such as a high binding affinity or logP, it is useful as an indicator of a higher risk of high metabolic turnover. However, it is often useful to verify the presence of such a liability through *in vitro* experiments, such as human liver microsomal (HLM) stability. If such a problem is confirmed, the total P450 metabolism models can provide an efficient guide to compound redesign to overcome this problem. Work is ongoing to develop a comprehensive model of P450 or HLM stability.

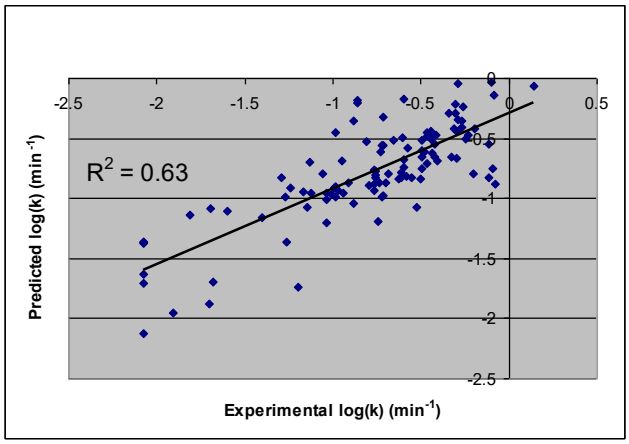


Figure 7.8 An example graph of a chemotype-specific (local) model for CYP3A4 metabolic rate. In this model F_n is the fraction of the compound in neutral form at pH 7.4, and logP is the calculated lipophilicity of the neutral form. These terms reflect binding affinity for CYP3A4. N.B. This model is specific to one chemical series and is unlikely to be applicable to different chemistry.

To date, a comprehensive, detailed model of metabolic rate for general chemistry has not proved to be tractable because not all of the contributions listed above are well understood. However, significant success has been achieved with individual chemical series, by combining CSL with other descriptors in a QSAR model of metabolic rate (See Figure 7.8 for an example).

Site lability is currently calculated only for CYP3A4 because the rate of decoupling, relative to which the lability is calculated, has only been accurately measured for CYP3A4. However, as CYP3A4 is the most promiscuous enzyme with the least contribution from orientation effects, the site lability for this enzyme can often be used as an indicator for general risk of P450 metabolism.

7.3.5 Summary of Metabolism Results

A detailed table of the predicted P450 metabolism for a compound can also be generated generated (See StarDrop User Guide) to provide further information on the contributions to the predicted metabolism of each site.

The column headings in this table are as follows:

- Metabolite – The metabolite predicted to be formed by metabolism at the site. N.B. There may be multiple metabolites for a single site, for example in dealkylation reactions.
- Site – The label of the site on the parent compound
- Parent Structure – The structure of the parent compound
- Metabolite Exact Mass – The exact mass of the metabolite to aid interpretation of metabolite ID experiments
- Hydrogen Count – The number of hydrogens bound to the site. In statistical terms, there are more chances for abstraction of a hydrogen atom from a site as the number of hydrogens increases, so this is a contributing factor to the lability of a site
- 3A4 Lability – The CYP3A4 lability category for the site, as shown on the metabolic landscape
- 3A4 ratio %, 2D6 Ratio %... - The percentage of metabolism by each isoform predicted to occur at the site (regioselectivity), if the compound is a substrate for the corresponding isoform.
- 3A4 Steric, 2D6 Steric… – The contribution of steric hindrance to the relative rate of metabolism of the site by each isoform. This is shown qualitatively as an integer indicating the magnitude with the sign indicating the direction of influence on the rate

- 3A4 Orientation, 2D6 Orientation… – The contribution of orientation effects to the relative rate of metabolism of the site by each isoform. This is shown qualitatively as an integer indicating the magnitude with the sign indicating the direction of influence on the rate

The metabolite structures and exact masses in this data set can be used to guide metabolite ID experiments. Furthermore, the properties of the predicted metabolites may be predicted, for example to identify potentially active or toxic metabolites.

The detail of the steric and orientation contributions to the rate of metabolism at each site may be used to guide chemical modification to increase the stability of molecules to metabolism by the P450 isoforms modelled. For example, a labile site could be made more stable by increasing the electronic stability of the site. However, if this is not possible, modification of the surrounding structure to increase the steric hindrance would also decrease the rate of metabolism of that site. Similarly, modifications that would alter the orientation of the molecule may place electronically labile sites away from the active oxy-heme and hence reduce the rate of metabolism.

7.4 Model Performance

The results in

Table 3 and illustrated in Figure 7.9 show the predictive performance of the models. The results show the percentage of the independent test sets where a site of metabolism (SOM) is identified in the top 2 and top 3 predictions, and also the percentage where all SOM are identified in the top 3 predictions.

Table 3 Site of metabolism (SOM) prediction performance for the independent test sets. Results show the percentage of the compounds in the independent test sets where at least one SOM is correctly identified in the top 2 and 3 predictions, and the percentage where all SOM are identified in the top 3. The average area under the curve of the ROC plots for compounds in the test set is also provided. SMARTCyp comparisons are shown where isoform specific models are available for CYP3A4, CYP2D6 and CYP2C9.

Isoform	StarDrop				SMARTCyp			
	Top 2 (%)	Top 3 (%)	All Top 3 (%)	AUC	Top 2 (%)	Top 3 (%)	All Top 3 (%)	AUC
3A4	84.5	90.5	53.6	0.87	70.2	84.5	51.2	0.89
2D6	91.1	92.9	71.4	0.91	92.9	96.4	69.6	0.95
2C9	85.7	93.9	75.5	0.91	87.8	91.8	77.6	0.95
1A2	87.7	89.5	64.9	0.87	N/A	N/A	N/A	N/A
2C8	81.5	92.6	70.4	0.86	N/A	N/A	N/A	N/A
2C19	85.7	89.8	69.4	0.89	N/A	N/A	N/A	N/A
2E1	90.0	93.3	80.0	0.84	N/A	N/A	N/A	N/A

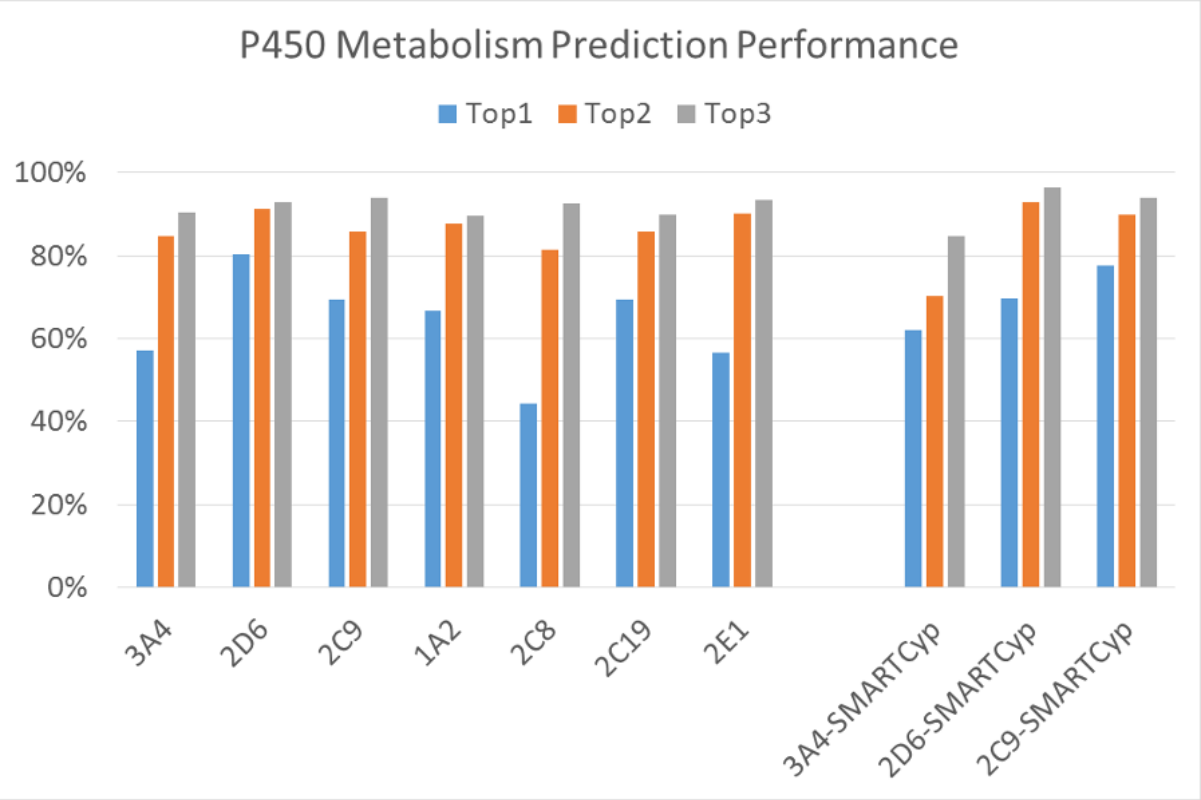
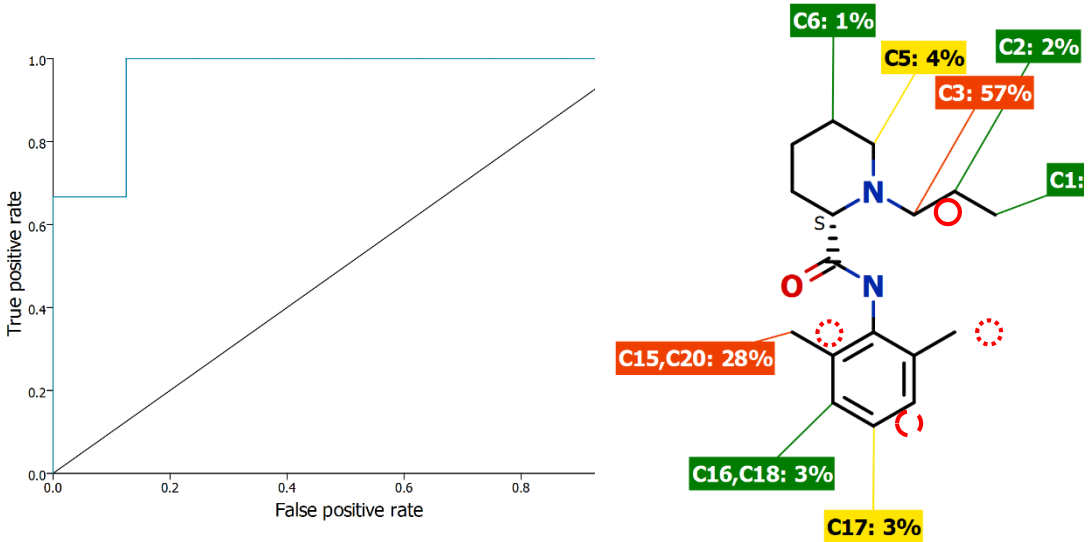
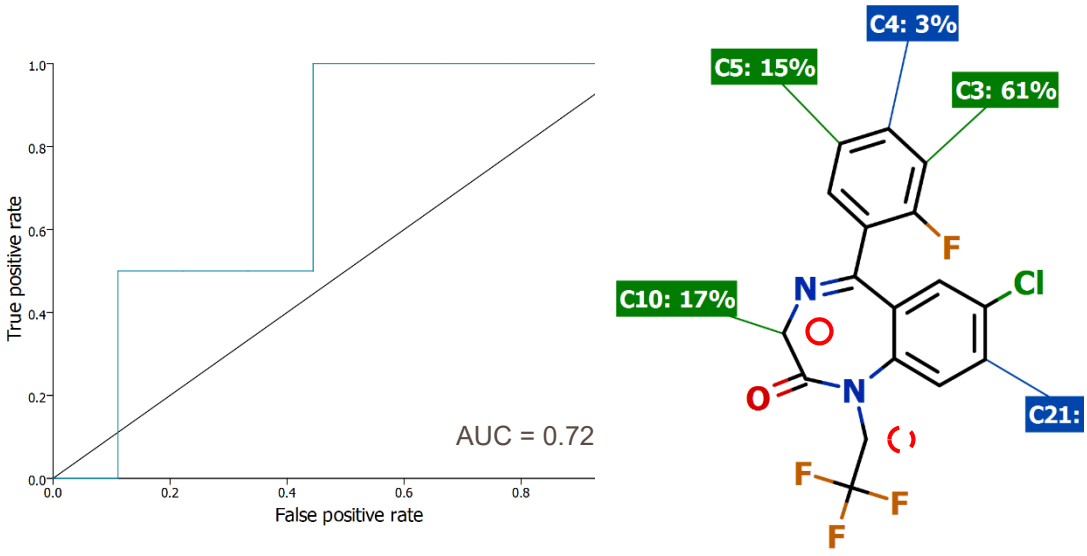


Figure 7.9 SOM prediction performance of the models described herein on independent tests. The bars labelled top-N show the percentage of an independent data set where at least one observed site of metabolism is identified in the top-N predicted sites. The performance of SMARTCyp on the same sets is shown for comparison for isoforms predicted by SMARTCyp.

In addition, receiver operating characteristic (ROC) plots have been generated for each compound, as illustrated in Figure 7.10 and the average area under the curve (AUC) for the ROC plots for the compounds in the test set for each isoform are also shown in Table 3. A greater area under the curve for a classifier indicates higher performance; the maximum possible AUC is 1 and a value of 0.5 is equivalent to the performance of random selection. The AUC for each compound and isoform in the training and validation sets is provided in the Supplementary Information.



(a)



(b)

Figure 7.10 Receiver operating characteristic (ROC) plots of the true positive rate (TPR (sensitivity)) against the false positive rate (FPR (1 - specificity)) for the prediction of SOM for two compounds. A perfect classifier would be represented by the point in the top left and a performance below the identity line (shown in black) indicates worse performance than a random classification. A greater area under the curve (AUC) for a classifier indicates higher performance; the maximum possible AUC is 1. The corresponding compounds are shown adjacent to each ROC plot, with the predicted regioselectivity indicated by the labels for each site. Primary observed sites of metabolism are highlighted on these structures by a solid red circle, secondary observed sites by a dashed red circle and tertiary observed sites by a dotted red circle. (a) shows an illustrative ROC plot for Ropivacaine which is well, but not perfectly, predicted. (b) shows an example of an ROC plot for a poorly predicted compound, in this case 2-oxo-Quazepam. The AUC is provided in the Supplementary Information for each compound in the data set.

An alternative measure of performance, Lift, was proposed by Zaretzki et al. (Zaretzki, et al., 2011). This corrects for the fact that it is easier to predict the observed SOM for compounds with a small number of potential sites than for those with a large number. The Lift measures the improvement in accuracy above that expected for random selection. Table 4 shows the Lift achieved by the models described herein.

Table 4 Lift metric for independent tests sets. Results show the improvement in performance of the models over that expected for a random model for top 2 and 3 predictions. SMARTCyp comparisons are shown where isoform specific models are available for CYP3A4, CYP2D6 and CYP2C9

Isoform	StarDrop		SMARTCyp	
	Top 2 (%)	Top 3 (%)	Top 2 (%)	Top 3 (%)
3A4	81	86	67	76
2D6	90	93	94	96
2C9	89	92	91	96
1A2	82	84	N/A	N/A
2C8	78	95	N/A	N/A
2C19	84	87	N/A	N/A
2E1	84	87	N/A	N/A

Comparative performance statistics are shown for CYP3A4, CYP2D6 and CYP2C9 from SMARTCyp, which predicts only these isoforms. Similar performance is obtained for CYP2C9 and CYP2D6 but performance on the important CYP3A4 isoform is stronger for the models presented in this paper.

It is informative to examine the contribution to overall predictive performance from the different components of the models: the electronic activation energy, ΔHA, steric hindrance affecting accessibility of each potential site of metabolism and interactions affecting the orientation of the substrate within the CYP binding pocket. The bar charts in Figure 7.11 compare the performance of different combinations of these components and it is apparent that contribution of the steric component is typically more important than the orientation component. However, the orientation component does have a notable positive influence on the performance of the 2D6 models and is able to capture the important interactions between positively charged ligand moieties and negatively charged protein residues (Glu216 and Asp301) that are known to be important for binding (Rydberg & Olsen, Ligand-Based Site of Metabolism Prediction for Cytochrome P450 2D6, 2012).

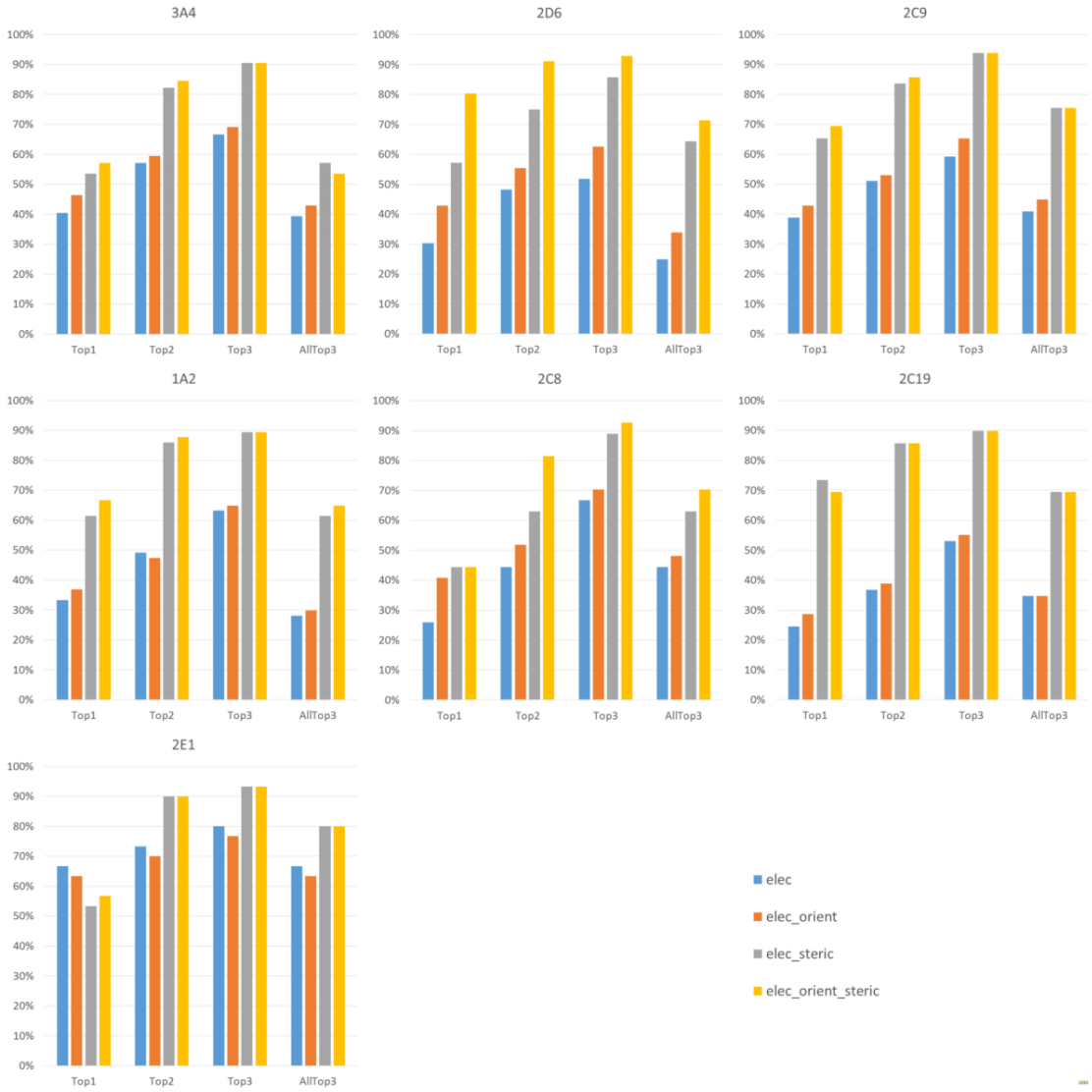


Figure 7.11 SOM prediction performance of different combinations of the three components to this method: elec (the electronic activation energy ΔH_A); steric (the effect of steric hindrance due to the structure of the substrate) and orient (the effect of interactions affecting the orientation of the substrate relative to the oxidizing oxy-heme). The performance of electronic activation energy alone is shown in blue, electronic plus orientation in red, electronic plus steric in grey and all three contributions in yellow. The bars labelled topN show the percentage of compound in an independent data set where at least one observed site of metabolism is correctly identified in the top-N predicted sites. The bars labelled AllTop3 show the percentage of compounds where all SOM are identified in the top 3 predicted sites.

8 Auto-Modeller™

8.1 Introduction

The StarDrop Auto-Modeller enables you to build predictive QSAR models of your experimental data. There are three main areas where the Auto-Modeller can be applied:

- Building 'local' models for individual chemical series or projects. Commercially available *in silico* models are normally 'global' models, designed to cover as wide a range of chemistry as possible. However, once data is available, it is often preferable to generate models tuned to proprietary chemistry.
- Generating 'corporate' models based on large compound property databases
- Iteratively improving models as more data is generated for a project

The process of model generation employed by the Auto-Modeller is largely automatic and it applies a consistent approach to model building. The main stages of this process are as follows:

1. **Input of initial data set.** You must provide a set of molecule structures together with compound names and values of a property or an activity to be modelled.
2. **Descriptor generation.** A library of molecular descriptors is provided, including whole molecule descriptors (e.g. molecular weight, logP and polar surface area) and 2D structural descriptors.
3. **Data set split into training, validation and test subsets.** A clustering algorithm based on 2D structural fingerprints is applied to split the initial dataset into three subsets. The training set is used to train models, the validation set is used to select the best model out of all those generated and the test set is used to assess the predictive power of the selected model.
4. **Application of multiple modelling techniques.** A suite of advanced and powerful methods that can be used to model either continuous or categorical data are applied to the training set data to build a series of models.
5. **Automatic selection and testing of the best model.** Based on the performance of all the models against the validation set, the best model is identified. It is then validated against the test set which has been kept completely independent of both the model building and selection processes.

The Auto-Modeller requires very little input (apart from provision of the initial data set) which allows users with less experience of computational techniques to build validated predictive models.

Alternatively, computational chemists and users experienced in model building can influence the model generation. If desired, it is possible to:

- Manually split the initial dataset or choose the splitting technique
- Input other 2D descriptors (as SMARTS) or import values for other types of descriptor, including experimental data
- Choose or refine the modelling techniques to apply

The Auto-Modeller allows you to save generated models and share them with your colleagues through StarDrop.

Models generated with the StarDrop Auto-Modeller benefit from StarDrop's unique 'Glowing Molecule' visualisation (see Chapter 4), allowing you to intuitively make the link between molecular structures and properties, highlighting possible 'problem' areas of molecules and guiding the design of molecules towards improved properties.

The Session Details and Model Details files contain information on the splitting process, descriptors used, performance statistics and model parameters.

8.2 Descriptors

The library of descriptors provided by StarDrop consists of a total of 321 SMARTS (see Section 8.2.1) based descriptors and 9 whole molecule properties such as logP, TPSA, molecular weight and the McGowan's Volume, Vx. The SMARTS based descriptors are counts of atom type (e.g. fluorine atom) and counts of functionalities (e.g. ketone). Section 15.3 in the appendices gives a full listing of descriptors provided. You can also supply your own 2D descriptors (as a SMARTS file) or use other external descriptor values such as experimental data. Models built using the StarDrop descriptors and/or imported SMARTS will be able to generate 'Glowing Molecule' visualisation. However, other external descriptor values used cannot benefit from this functionality.

8.2.1 SMARTS

SMARTS (Obrezanova, Csanyi, Gola, & Segall, 2007) (Daylight, n.d.) is a notation that allows you to specify atoms and groups of atoms using rules similar to the SMILES notation. However, SMARTS atoms and bonds are more general than in SMILES strings. SMARTS also uses symbols describing atomic properties such as the number of explicit connections (D<n>), number of total connections (X<n>) and the total bond order (V<n>) where <n> is an integer. Logical operators are also used to create an exact description of an atom (e.g. [CX4H3] is a sp3 carbon with exactly three hydrogens). SMARTS expressions can be used to defined an atomic environment by starting the string with a $ sign followed by the atom of interest. Such definitions can be considered to describe atomic properties rather than groups of atoms.

8.3 Data Set Preparation

The data set preparation comprises the following three steps:

1. The descriptors are filtered based upon the complete dataset
2. The data set is split into training, validation and test sets
3. The descriptors are filtered again based upon the training set

8.3.1 Descriptor Filtering

Calculated and imported descriptors are subjected to a feature selection step that removes descriptors with low variance and low occurrence, i.e. those descriptors that provide little information regarding the differences between the molecules in the set. Highly correlated descriptors are also removed. The default rules for descriptor exclusion are as follows:

- Descriptors with a standard deviation less than 0.0005
- Descriptors represented by less than 4% of compounds
- If the pair-wise correlation between any two descriptors exceeds 0.95, then the descriptor of the pair with the lowest correlation with the Y column is excluded

These default thresholds can all be altered if required.

8.3.2 Data Set Split

In order to be able to rigorously select the best model and then assess its predictive power, the data set is split into training, validation and test sets. The training set is used to fit models to the observed data, the validation set is used compare the models built and select the best model. Finally, the test set is used to independently assess the predictive power of the chosen model. In this way, the final test of predictive power is completely independent of the model training and selection process.

By default, 70% of compounds are assigned to the training set, 15% to validation set and 15% to test set. The percentage of compounds in each set can be altered if required.

There are three techniques available for performing the set split:

- Random
- Y-based
- Clustering

In the case of the random method the data is split randomly between the three sets in the correct proportions.

For the Y-based split the entire data set is sorted on the property values and then randomly picked from bins of similar values to go into the training, validation and test sets such that each set will have a similar spread of property values and each will be the appropriate size.

In the case of the clustering method, compounds are clustered using an unsupervised non-hierarchical clustering algorithm developed by Butina (Butina, Unsupervised Data Base Clustering Based on Daylight's fingerprint and Tanimoto Similarity: A fast and automated way to cluster small and large data set, 1999) . The cluster analysis of the chemical structures is based on 2D path-based chemical fingerprints and the Tanimoto similarity index. The algorithm identifies dense clusters where similarity within each cluster reflects the Tanimoto value (between 0 and 1) used for the clustering. If the similarity index between two compounds is greater than the Tanimoto value then these compounds belong to one cluster. The default Tanimoto value used is 0.7 but this can be altered if required.

Once the clusters are formed the centroids (the cluster centres) and singletons (compounds that are not clustered with any others) are put into the training set. Then the remaining compounds in each cluster are sorted by Y value and divided into bins. Compounds from each bin are divided randomly between the training, validation and test sets in the required proportions. If the number of centroids and singletons is greater than the number of compounds required in the training set, then the clustering information is abandoned and dataset split is instead based on Y values. Using extreme values for the Tanimoto value (i.e. close to 0 or 1) is therefore quite likely to result in a dataset split based purely on Y values.

Ideally, the division of compounds between the three sets should be based on a random sample to ensure the greatest statistical rigor. However, the data sets used to build models are typically relatively small, leading to a significant chance that a random sample will omit important chemical features from the training set. Therefore, in order to ensure that the maximum chemical space is covered by the models, the clustering technique is used by default.

The result of this approach is that the predictive performance on the validation and test set is only representative of the expected predictive power of the model within the domain of the chemical space of the training set, because the validation and test sets have been explicitly chosen to cover this space. Thus it is important to couple the models with a measure of their chemical space as described in Sections 6.6 and 8.10.

Alternatively you can choose to define your own training, validation and test sets. In this case the training, validation and test sets must be provided as separate data files in the user interface.

8.3.3 Descriptor Filtering on the Training Set

After the data set has been split, the feature selection procedure (as described in Section 8.3.1) is repeated for the training set. The descriptors excluded from the training set are also eliminated from the validation and test sets.

8.4 Modelling Techniques

The suite of modelling techniques includes:

- Partial Least Squares (PLS)
  This is a well-known robust technique for generation of continuous linear models based on multiple descriptors.

- Radial Basis Functions with a Genetic Algorithm (RBF and GA-RBF)
  An efficient numerical technique to generate non-linear models can be combined with a genetic algorithm to search for the optimal descriptor space. This technique can be used to build continuous models.

- Gaussian Processes (GP)
  This powerful 'machine learning' technique, based on the Bayesian statistical approach, is able to model non-linear relationships. This technique produces six continuous models: three of them are built on a full set of descriptors, GP Fixed (GPFixed), GP 2D search (GP2DSearch) and GP with nested sampling (GPNEST) and three models have a built-in descriptor selection tool, GP Forward Variable Selection (GPFVS), GP Rescaled Forward Variable Selection (GPRFVS) and GP Optimised (GPOPT). This technique can also be used to build classification models.

- Decision Trees (DT)
  This is a recursive partitioning approach to building classification models in cases where good 'continuous' data is not available. The technique can generate up to 20 different models by automatically varying the parameters of this method.

- Random Forests (RF)
  This is an ensemble method that makes predictions based on the output of a collection of random trees. This technique can be used to build both classification and regression models: for classification, the prediction is given by a majority vote over the committee of trees, and for regression, the prediction is set to the average output over all of the trees.

All appropriate methods are automatically applied to each modelling problem. Some of the techniques are more computationally demanding and for large training sets some techniques may be omitted. The default thresholds for the application of Gaussian Processes techniques are given in Table 3.

Table 5 Thresholds for training set size for continuous Gaussian Processes techniques.

N, training set size	GPFixed	GP2DSearch	GPFVS	GPRFVS	GPOPT	GPNEST
N≤ 300	X	X	X	X	X	X
300<N≤500	X	X	-	X	X	-
500<N≤1000	X	X	-	X	-	-
1000<N	X	-	-	-	-	-

You can override the default settings and choose which techniques to use. By default the DT, PLS, GPFixed and RBF techniques are always applied; however you can change this if desired.

8.4.1 Modelling Notation

To describe the modelling techniques it is necessary to first define some notation. We denote the property vector for the training data by Y and the matrix of descriptor values for the training set by $X = \{x_{ij}\}_{i=1\ldots N,\, j=1\ldots K}$ where x_{ij} is a value of the j-th descriptor for the i-th molecule. We will use notation $\underline{x}_i$ for the vector of descriptors for the i-th molecule. N is the number of compounds in the training set and K is the number of descriptors.

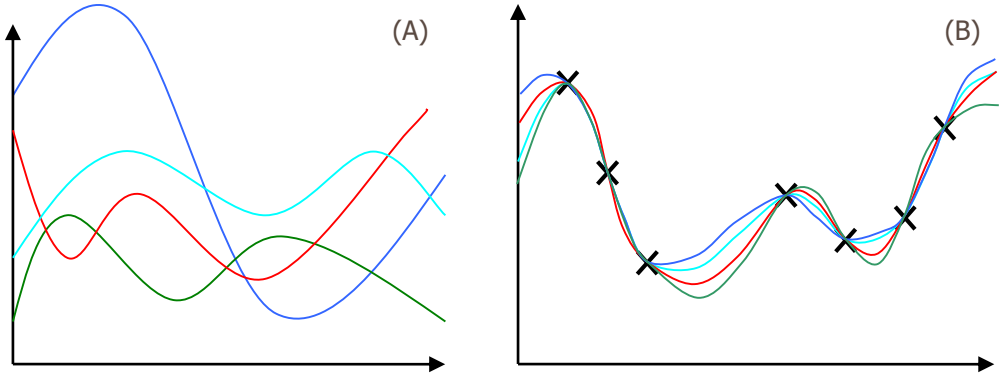


Figure 8.1 Random samples from (A) prior distribution of functions and (B) posterior distribution of functions in a one-dimensional example. Functions from the posterior distribution are conditioned to pass through the training points shown by crosses.

8.5 Gaussian Processes

Gaussian Processes is a powerful computational method for predictive quantitative structure-activity relationship (QSAR) modelling. Using a Bayesian probabilistic approach, the method is widely used in the field of machine learning but has rarely been applied in QSAR and ADME modelling. This method overcomes many of the problems of existing QSAR modelling techniques:

Most importantly, it does not require subjective *a priori* determination of parameters such as variable importance or network architectures.

It is suitable for modelling non-linear relationships.

The method has built-in mechanisms to prevent over-training and does not require cross-validation.

It works well for large numbers of descriptors.

The performance of Gaussian Processes compares well with, and often exceeds, that of artificial neural networks.

8.5.1 Theory Overview

Only a brief overview of the Gaussian Processes method is given here. Detailed descriptions of the technique can be found in the literature (MacKay, 2003) (Rasmussen & Williams, 2006) (Obrezanova, Csanyi, Gola, & Segall, 2007).

The general logic of Bayesian inference is that one assumes a prior probability distribution for the values of an unknown function and updates this probability distribution in the light of observed data to yield the posterior distribution (see Figure 8.1). The idea of Gaussian Process modelling is to place a prior directly on the space of functions.

Given any set of descriptor vectors, the prior distribution for the function values is assumed to be a multidimensional normal distribution with a zero mean and a covariance matrix Q which depends on the descriptor vectors. The elements of matrix Q are given by the covariance function $C(\underline{x}_n, \underline{x}_m)$, the role of which is to define the metric in the input space, (i.e. representing the similarity between different molecules).

$$C(\underline{x}_n, \underline{x}_m) = \theta_1 \exp\left[-\frac{1}{2}\sum_{i=1}^{K}(x_{ni} - x_{mi})^2 \Big/ r_i^2\right] + \theta_2 ,$$

where $\{\theta_1, \theta_2, r_i\ (i = 1...K)\}$ are model parameters (called hyperparameters in the Gaussian Processes framework). The $\{r_i\}$ is a set of length scale parameters, one for each descriptor. A very large value for a given r_i is equivalent to saying that differences in the corresponding descriptor do not influence the property values very much.

The central result of the method is that given a training set, the posterior predictive distribution for the property value $y' = y(\underline{x}')$ for a new molecule with descriptor vector $\underline{x}'$ is also a Gaussian distribution with the following mean and variance

$$\mu = k^T (Q + \theta_3 I)^{-1} Y, \qquad \sigma^2 = \kappa - k^T (Q + \theta_3 I)^{-1} k,$$

where the vector k with components $k_n = C(\underline{x}', \underline{x}_n)$ describes the similarity of the new molecule to the ones in the training set and $\kappa = C(\underline{x}', \underline{x}')$. Hyperparameter θ_3 describes the variance of the assumed noise in the data.

We will take the mean of this distribution as the predicted property vector for the new molecule μ. The standard deviation, σ, can be used as an indicator of where a new molecule lies within the descriptor space of the model. If this standard deviation is very large, it will indicate that the new molecule is well outside the descriptor space covered by the training data.

The hyperparameters $\Theta = \{\theta_1, \theta_2, \theta_3, r_i\ (i = 1...K)\}$ need to be learned from the training data. We want to ensure that the prediction function is smooth and matches the observed data as well as possible. Finding the *most probable* set of hyperparameters will suffice. This corresponds to finding the minimum of the log marginal likelihood (MacKay, 2003).

Determining the hyperparameters by optimising the log marginal likelihood ensures that the model is not over-trained and that the optimal trade-off between smoothness and fitting the data is achieved. This also removes the need for the cross-validation of the model.

8.5.2 Hyperparameter Tuning

We use six techniques for determining the hyperparameters, listed in order of the increasing computational time they demand. The three techniques, GPFVS, GPRFVS and GPOPT, have the ability to identify and select descriptors relevant to describing the property (Obrezanova, Csanyi, Gola, & Segall, 2007).

GP Fixed (GPFixed)

We set the hyperparameters to fixed values that depend on the standard deviation of property vector Y and the standard deviation of the columns of matrix X.

This approach is not computationally expensive and works well for the majority of sets although the chosen values might not suit some data.

- GP 2D search (GP2DSearch)
  The hyperparameters $\theta_2, r_i\ (i = 1...K)$ are fixed as in GPFixed; θ_1, θ_3 are determined by optimising the log of the marginal likelihood.

- GP Forward Variable Selection (GPFVS)
  The hyperparameters values obtained by GP2DSearch are used in this approach. A forward variable selection procedure (Everitt & Dunn, 2001) is employed to identify the most important descriptors. The final model is built on the selected subset of descriptors.

- GP Rescaled Forward Variable Selection (GPRFVS)
  This approach is similar to the GPFVS technique but the length-scale hyperparameters r_i are rescaled according to the number of descriptors used in the model. This approach leads to a slightly different and shorter list of relevant descriptors than GPFVS.

- GP Optimised (GPOPT)
  The hyperparameters $\theta_1, \theta_2, \theta_3$ are set to the values obtained by GP2DSearch. The length-scale hyperparameters r_i are optimised by a conjugate gradient method (Press, 1988). After the method has converged, i.e. the hyperparameters minimizing the marginal likelihood are found, the number of descriptors is reduced by an automatic relevance determination procedure. The final model is based upon the selected subset of descriptors.

- GP with Nested Sampling (GPNEST)
  The search for hyperparameters is performed in the full hyperparameter space $\{\theta_1, \theta_2, \theta_3, r_i\ (i = 1...K)\}$. Briefly, the idea of the method is as follows. The prior space of hyperparameters is sampled randomly. Some samples corresponding to larger values of marginal likelihood are replaced with new samples with lower values of marginal likelihood. At the end of the iterative process, we have points from the hyperparameter space which correspond to the low values of marginal likelihood; that means optimal hyperparameter values (Obrezanova, Csanyi, Gola, & Segall, 2007). This technique has a numerous advantages. It explores a wide prior space of hyperparameters and does not get 'trapped' in local minima of marginal likelihood. Although the model is built on a full set of descriptors, the length scale hyperparameters obtained can be used to identify the most important descriptors.

The three techniques, GPFVS, GPRFVS and GPOPT, all use the values for $\theta_1, \theta_2, \theta_3$ obtained by GP2DSearch. If GP2DSearch is not performed then the fixed values for $\theta_1, \theta_2, \theta_3$ from GPFixed are used.

When building classification models using Gaussian Processes, fixed values for $\theta_1, \theta_2, \theta_3$ from GPFixed are used.

8.5.3 Computational Time

The most computationally demanding step of the training process involves inverting the covariance matrix (of size N x N) which must be done each time a new set of hyperparameters is tried, resulting in complexity ($O(N^3)$). Let us denote time for one inversion of the matrix by τ. Then computational time for each Gaussian Process technique can be estimated as follows (Table 6).

Table 6 Time taken for each Gaussian Process technique

Method	Approximate Time
GPFixed	T
GP2DSearch	2500 τ
GPFVS	0.5 N(N+1) τ
GPRFVS	0.5 N(N+1) τ
GPOPT	> 200 N τ
GPNEST	~ 20000 τ

Table 7 shows typical computational time for GP methods depending on the training set size. Although computational time will depend on various factors, like the number of descriptors in the set as well as computer specifications, Table 7 can be used for guidance.

Table 7 Examples indicating the relative computational time taken for the Gaussian Processes techniques to complete. The simulations were performed on a computer with a 1 GHz CPU and 175 descriptors were used.

	Time (hours)			
Training set size	**GP2DSearch**	**GPFVS**	**GPRFVS**	**GPOPT**
200	0.03	1	1	1.5
400	0.5	6	6	15
600	3	25	25	90

8.5.4 GP Model Output

The Model Details file generated by the StarDrop Auto-Modeller for the GP techniques contains information about the values obtained for θ_1, θ_2 and θ_3 along with the descriptors used and their corresponding length scales. In some cases, such as for the GPOPT and GPNEST models, the file contains the normalized length scales which can be used to identify the most important descriptors. The descriptors which are more relevant for describing the property will have smaller normalized length scales. The model performance statistics (R^2 and RMSE) for each of the training, validation and test sets are also included. See Section 8.9.1 for more information about the model performance statistics.

8.6 Radial Basis Functions with a Genetic Algorithm

Radial basis functions (RBFs) have been praised for their simplicity, robustness and ease of implementation in multivariate scattered data approximation. Such techniques have been applied with success in problems ranging from training neural networks to image compression (Buhman, 2003). RBFs have not been commonly used in the QSAR field.

RBFs provide a good solution for both small and large data sets. However, they can be sensitive to noise created by excessive descriptors. In order to avoid this, the StarDrop Auto-Modeller applies a genetic algorithm (GA) to run a stochastic search of the descriptor space and identify the most significant set of descriptors that best represent the property being modelled. For a number of compounds N in a training set with K descriptors, if the ratio N/K is less than five, then the GA is combined with RBF to enable descriptor selection.

RBF technique can be applied with or without GA descriptor selection. While capable of decreasing noise and providing a better insight into the investigated property, the GA-RBF approach is computationally expensive. Data sets with a large number of compounds and descriptors might take weeks to achieve an optimal solution. Therefore, by default, when the ratio of compounds to descriptors (N/K) is greater than five, only the RBF technique is applied without a GA descriptor selection.

8.6.1 RBF Overview

A radial basis function is a non-linear transfer function. It operates by measuring the Euclidian distance between an input vector and the function centre. The aim of the RBF is to approximate a real valued function $y(\underline{x})$ by $\Psi(\underline{x})$ given the set of sample values $Y = (y_1,\ldots, y_N)$ at the distinct points $X = \{\underline{x}_1,\ldots,\underline{x}_N\}$. We choose $\Psi(\underline{x})$ to be of the form:

$$\Psi(\underline{x}) = \sum_{i=1}^{N} a_i \phi(\|\underline{x} - \underline{x}_i\|)$$

where a_i is a real valued weight, ϕ is a basic function and $\|\cdot\|$ denotes the Euclidian distance metric. Vectors $\underline{x}_i$ represent N points where the radial basis functions are centred. Therefore $\underline{x}_i$ represent the descriptor vectors for the N compounds in the training set.

In fitting the RBF to the data, the conventional method is to require that $\Psi(\underline{x})$ pass through all the training data points, which gives the following linear system of equations:

$$y_j = \sum_{i=1}^{N} a_i \phi(\|\underline{x}_j - \underline{x}_i\|), \qquad j = 1,\ldots,N.$$

This linear system can be solved for the weights a_i, $(i = 1 \ldots N)$. Then the fitted function $\Psi(\underline{x})$ can be used for predicting new data points.

Multiple training data points with identical values for all descriptors, even if resulting from different molecules and with different Y values, can cause a numerical instability in the training algorithm. If such a case arises, the first of the identical data points will be retained and the subsequent identical data points discarded.

In application, a variety of basis function types might typically be used, including linear, cubic, multi-quadratic and Gaussian. Based on the results of experiments with different forms of basis function, the StarDrop Auto-Modeller uses linear radial basis functions.

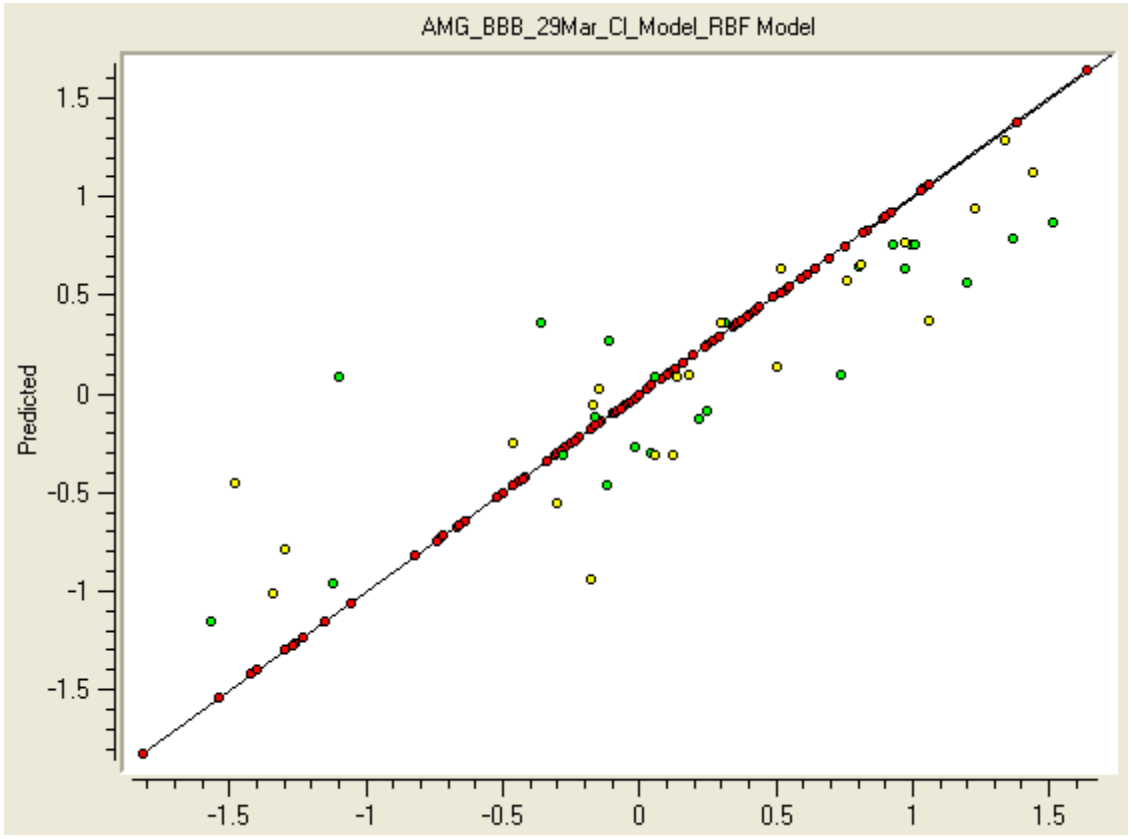


Figure 8.2 The predicted logBB values versus observed for blood-brain penetration. The model is built by the RBF technique. The molecules in the training set are shown in red, the validation set in green and the test set in yellow.

Because a fitted RBF is required to pass through all the training data points, the compounds of the training set are always perfectly predicted (see Figure 8.2).

Note: where compounds have been excluded from the training set because they have identical descriptors, such points will not be perfectly predicted when StarDrop creates a plot of predicted vs. observed values such as Figure 8.2.

8.6.2 Genetic Algorithms

The application of genetic algorithms (GAs) to predictive models is based on an analogy with Darwin's principle of 'survival of the fittest'. The GA will produce an initial generation of random RBF models each of which is then evaluated for fitness. The better models are more likely to be selected for cross-breeding and will tend to survive while the weaker models die out. By applying the basic genetic principles of selection, combination and mutation, the GA increases the diversity of models evaluated over successive generations. This search of possible solutions continues until some pre-defined stopping criteria is met, at which time the fittest model is selected and approximates the best solution to the problem. The following section describes GAs and the evolutionary process.

8.6.3 Genetic Encoding

A chromosome is represented by a binary bit string whose length corresponds to the number of input descriptors. A bit equals 0 if the corresponding descriptor is not selected, otherwise the bit equals 1. A model is built using the selected descriptors from each chromosome in order to evaluate that chromosome.

8.6.4 Fitness Function

The GA method is based on minimising the difference between calculated and observed data while penalising models in which the ratio of compounds per descriptor is less than a chosen threshold (default five). The fitness function F is based on the predictive ability of the RBF model, measured by R^2 (defined in Section 8.9.1) as well as the number of descriptors selected by the GA. The GA-RBF algorithm searches for the best RBF model in such a way that from one generation to another the algorithm tries to minimise the penalty function whilst increasing the R^2.

$$F = 10 * R^2 - penalty\ ,\ penalty = a - \frac{a}{(1 + \exp(-c * (x - t)))}$$

where x is the number of training compounds divided by the number of descriptors (N/K). The parameters a, c and t describe the sigmoid function that decreases as x increases. The parameter t defines the inflection point, i.e. where the magnitude of the gradient is maximal. The parameter c controls the gradient around the inflection point. The parameter a controls the weight of the penalty relative to the quality of the model as measured by R^2.

The default values of t, c and a are t=4, c=1 and a=5. Representations of the score and penalty values are shown in Figure 36 and Figure 35.

8.6.5 Initialisation

The GA starts by creating an initial population of 50 chromosomes by randomly populating them with 1's and 0's. For each chromosome this corresponds to selecting indiscriminately a set of descriptors from the initial pool.

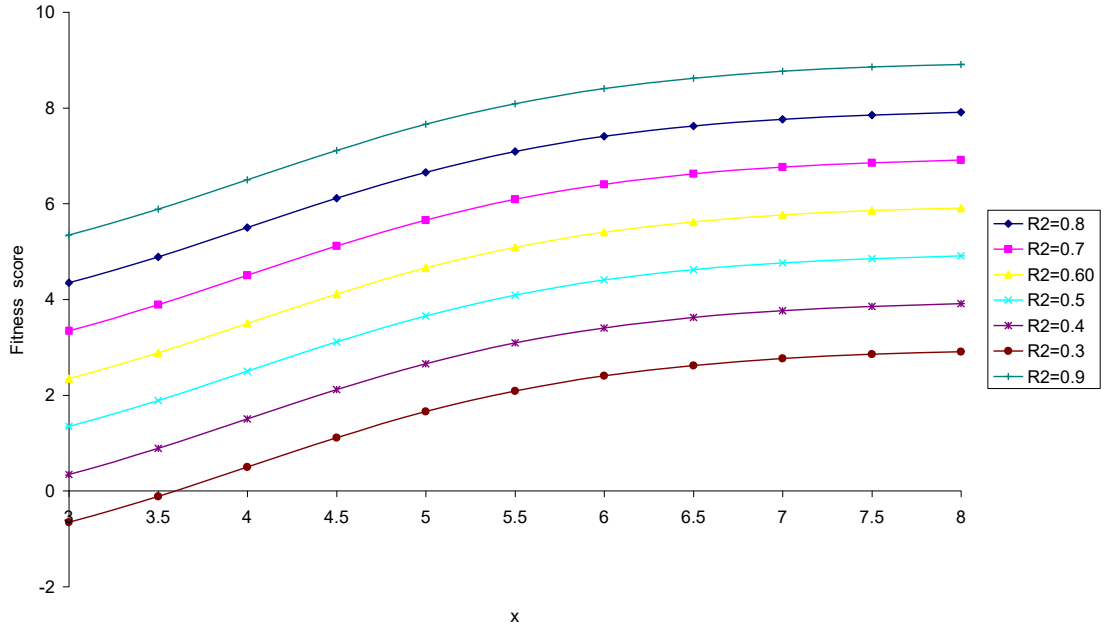


Figure 8.3 Fitness score values against x ratio for various R^2 values.

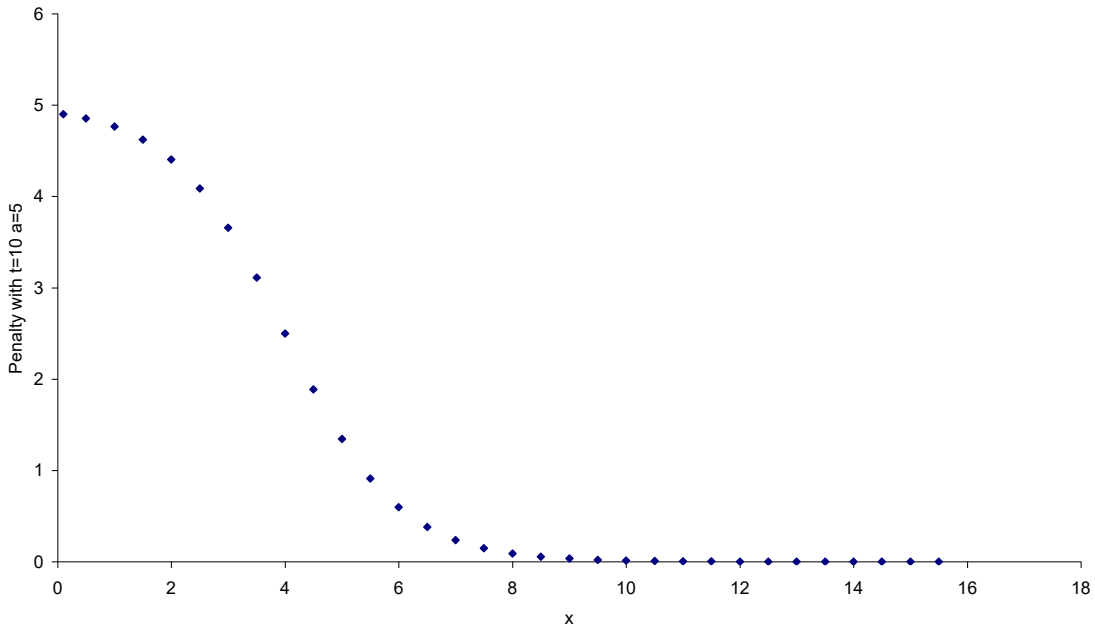


Figure 8.4 Penalty value against x ratio.

8.6.6 Selection

A selection mechanism in a GA is a process that favours the selection of better chromosomes in the population for the mating pool. These selected chromosomes will then be used to generate new offspring for the next generation. The selection rate is the degree to which the fittest chromosomes are favoured: the higher the rate, the more often the better chromosomes are favoured. This selection process drives the GA to improve the population fitness over successive generations with higher selection rates resulting in higher convergence rates, see Figure 8.5. The selection rate value is between 0 and 1, with the default set to 0.9.

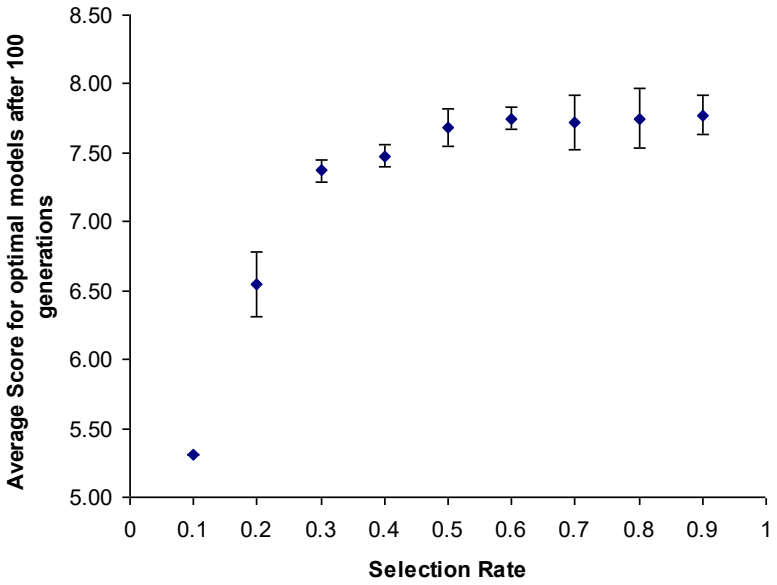


Figure 8.5 of the average score for 3 optimal models after 100 generations. Penalty function criteria: t = 4, c = 1 and a = 5. Pool size = 8. There were 199 compounds in the training set and 164 descriptors. The average score values were obtained by repeating 3 GA-RBF models for each selection rate.

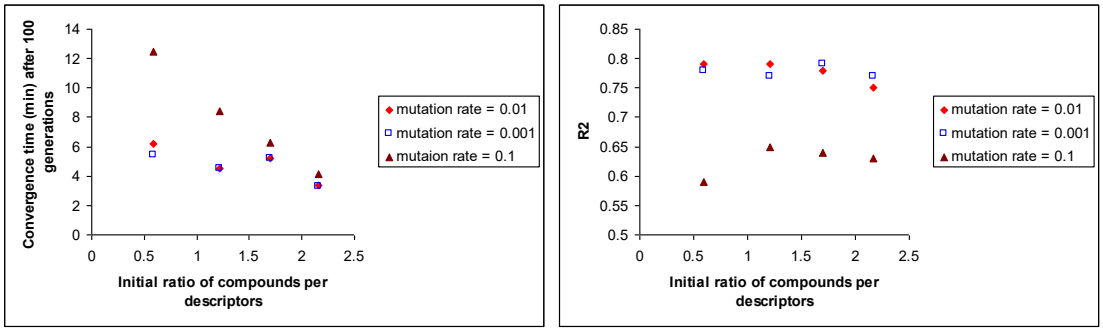


Figure 8.6 Results of GA-RBF models applied to a set of 199 compounds and varying number of descriptors from 338 to 92. GA operators are: pool size = 8, selection rate = 0.9, combination rate = 0.9 and fitness score criteria: t=4, c=1 and a=5. Note the influence of the ratio of data set size to descriptors on the convergence time for a mutation rate of 0.1.

Tournament selection provides the method for selecting a chromosome from the current generation and is applied every time a chromosome must be selected for combination or cloning. Either a random selection from the current generation is made or M random chromosomes are pooled with the fittest chromosome from the pool being selected. M is the mating pool size which by default is eight. The 'Selection rate' determines how often a chromosome is selected from the mating pool rather than being chosen at random.

8.6.7 Combining Chromosomes

The process of combining two chromosomes, parent1 and parent2, to produce two child chromosomes, child1 and child2, (also known as crossover) starts by randomly choosing two points along the length of a chromosome. The parents are then split at these points into front, middle and back sections. Child1 is made up of the front and back of parent1 and the middle of parent2. Child2 is made up of the front and back of parent2 and the middle of parent1. Thus each child contains some data from each parent.

8.6.8 Mutation

Mutation provides genetic diversity and enables the genetic algorithm to search a broader space. When mutated, a random bit in the chromosome is reversed meaning that one descriptor is either added or removed at random.

8.6.9 Evolution

The following process is used to create a new generation of chromosomes (the same size as the previous generation):

The fittest chromosome in the current generation is automatically inserted into the next generation. This is never mutated. Successive chromosomes are then selected using the method described above. The 'Combination rate' (between 0 and 1) determines whether a selected chromosome is combined (as above) with another selected chromosome to produce two children. The greater the combination rate the more often two chromosomes are combined. If combined, then both the children are added to the next generation. Otherwise, if not combined then the chromosome itself is added to the next generation. The 'Mutation rate' determines how often any of the not combined or child chromosomes are mutated before being added to the next generation. The mutation rate can take values between 0 and 1. A very small mutation rate may lead to premature convergence in a local optimum. A mutation rate that is too high may lead to loss of good solutions. Figure shows the effect of mutation rate on the speed of convergence and on the R^2 values. A mutation rate of 0.1 leads to lower R^2 values than the lower mutation rates. The default mutation rate is 0.01.

8.6.10 Termination

Successive generations are evolved until a termination condition has been reached. Terminating conditions are either:

- A fixed number of generations has been reached, the default value is 200, or
- A fixed number of successive generations are evolved with no improvement in the fitness value of the best chromosome; the default value is 2.

The size of the training set significantly impacts the speed of convergence. The size of a data set, i.e. the number of compounds, is an important factor to take into account when selecting GA-RBF, as the more compounds there are the longer it will take to reach optimal solutions (see Figure 8.7). The number of descriptors also affects the speed of convergence. The more descriptors there are the larger is the descriptor space to explore (see Figure 8.7).

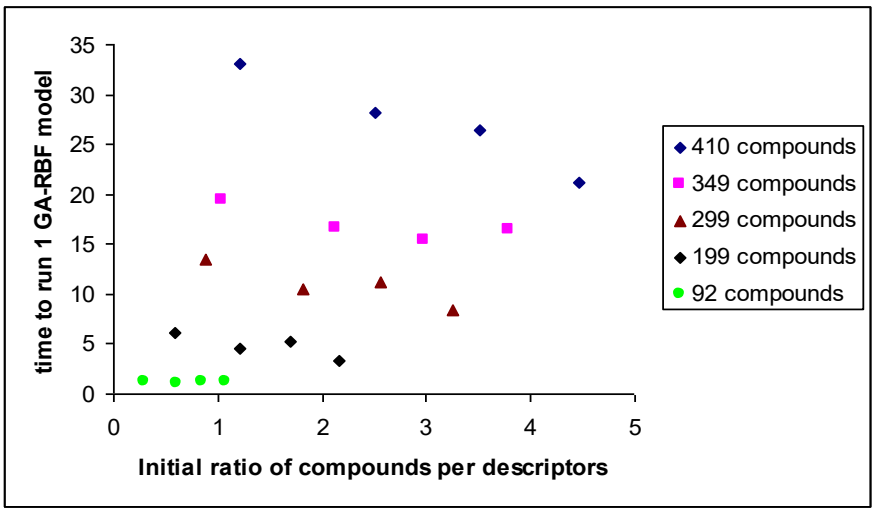


Figure 8.7 Ratio of compounds per descriptors versus time (min). Five data sets of varying size (from 92 compounds to 410 compounds) and descriptor count. GA-RBF models were generated for these data sets using the default GA settings.

You can change the StarDrop Auto-Modeller defaults to decrease the number of compounds in the training set by lessening the percentage of compounds to form the training set. The default is currently set to 70%. Additionally, the number of descriptors can be reduced by making the descriptor selection options stricter (Section 8.3.1). For example, a much lower descriptor pair-wise correlation coefficient could be used instead of the default 0.95. Both these approaches will decrease the number of descriptors used. Should the ratio of compounds per descriptor become five or higher, the GA-RBF is automatically replaced by simple RBF and no stochastic search of the descriptor space is made.

However, you should be aware that this could lead to loss of important descriptors and decreasing the size of the training set would have the effect of decreasing the chemical space of the final model.

8.6.11 Genetic Algorithm-RBF (GA-RBF)

The goal of combining the RBF with a GA is to find an optimal or near optimal set of descriptors to use in the RBF model. However, GAs sometimes find local minimums of the descriptor space rather than the global minimum. To overcome this deficiency a k-fold validation method is applied. This provides for a better search of the descriptors and facilitates the identification of a statistically sound set of descriptors that best represent the property being modelled.

To do this, the training data set is divided into five mutually exclusive subsets, each of equal size. One subset is left out as a validation set while 50 GA-RBF models are developed using a set consisting of the remaining four subsets. This procedure is repeated with each subset being used as a validation set for the others, resulting in a total of 250 models being obtained. A statistical test, based upon the binomial distribution, is then applied to identify the descriptors that occur with greater than random frequency and hence are significant. The significance level α (alpha) for the statistical test is set by default to 0.005 but can be altered by a user. The lower this value the more often a descriptor must have occurred to make the final selection. Finally, an RBF model is developed using the entire training data using just the most significant descriptors.

This approach is computationally expensive. It can take from two hours on a small training set of less than 100 compounds to more than a week on training set greater than 400 compounds.

8.6.12 GA-RBF and RBF Model Output

The Model Details file generated by the StarDrop Auto-Modeller for the GA-RBF technique contains information about the settings used for the GA along with a list of the descriptors selected. The model performance statistics (R^2 and RMSE) for each of the training, validation and test sets are also included. See Section 8.9.1 for more information about the model performance statistics.

8.7 Partial Least Squares

The Partial Least Squares (PLS) method is a well-known and widely used tool for QSAR modelling (Wold, Sjostrom, & Eriksson, 1999). PLS is able to describe linear relationship but can cope with some non-linearity as well. It is suitable for modelling sets with many, noisy, correlated descriptors. Models built by PLS give an insight into the relative importance of descriptors and computationally it is a very fast method.

Because PLS is such a well-known technique we will not document the details of the method (see reference (Wold, Sjostrom, & Eriksson, 1999) for details). The central idea of this approach is that the PLS algorithm forms new 'latent' variables that are most relevant for describing property Y, and then Y is expressed as a linear combination of these latent variables. In order to avoid over-training, the number of PLS-components (latent variables) to use in the model is determined by a cross-validation technique in which the training set is split into seven subsets. Each subset is excluded in turn and a model is built on the remaining compounds (the other six subsets) and tested on the excluded subset. The number of PLS components is identified as that which maximises the Q-squared, which is the R-squared for prediction of the compounds in each excluded subset (this is equivalent to minimising the RMSEP). Note that the separate validation and test sets defined by the Auto-Modeller are not used in this process; they are maintained as external prediction sets.

8.7.1 PLS Model Output

The Model Details file generated by the StarDrop Auto-Modeller for the PLS technique contains coefficients and scaled coefficients for each descriptor. The predicted property is a linear combination of descriptor values multiplied by their associated coefficients. The scaled coefficients are adjusted for mean-centred data and are scaled to unit-variance. These can be used to understand the relative importance of the descriptors. The bigger the absolute value of a scaled coefficient the more important the corresponding descriptor is.

The Model Details file also gives the number of PLS components used in the model and the cross-validation Q^2 statistic which indicates the predictive ability of the model. The Q^2 statistic is equivalent to the coefficient of determination R^2 but it is obtained by cross-validation.

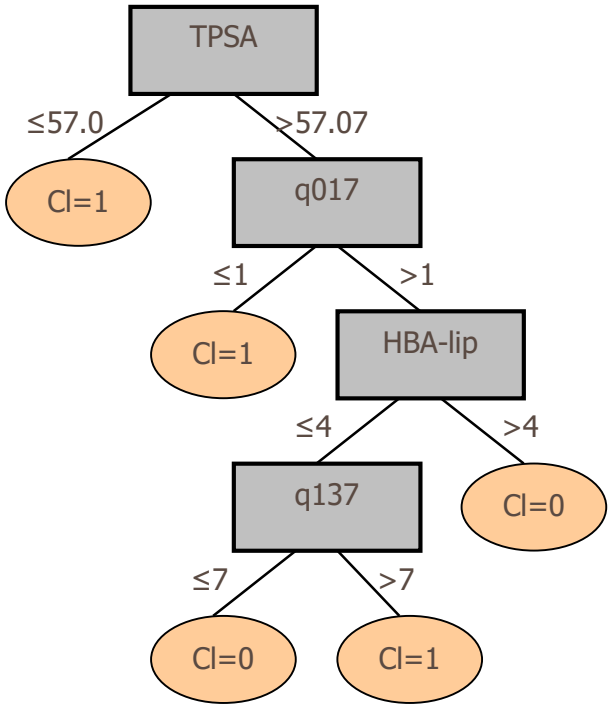
8.8 Decision Trees

Decision Trees (DT) provides a recursive partitioning approach to building classification models. It is suitable for categorical data (i.e. the observed property or activity takes values YES/NO, Low/High, 0/1, Class 0-4, etc.) and is able to model non-linear relationships. It works well in the presence of many descriptors and able to select those most relevant to a property. DT creates models that are transparent for interpretation allowing you to understand the underlying structure–activity relationships.

It should be noted that in building a satisfactory classification model it is desirable to have a uniform spread of compounds among classes. Also, in general, it is more difficult to build satisfactory multi-class models than two-class models although this is still possible with the StarDrop Auto-Modeller.

8.8.1 Building a Decision Tree

The StarDrop Decision Trees (DT) technique is based on the C4.5 algorithm introduced by Quinlan (Quinlan J. R., 1993). The idea of the method is to attempt to divide the set of compounds into single-class subsets. The tree is built by recursively partitioning training data based on the value of a particular descriptor. To make each partition it is necessary to choose the best descriptor and the best threshold. This is achieved by maximising the information gain or the information gain ratio (see Quinlan's book (Quinlan J. R., 1993) for definitions). These two criteria produce different trees. An example of a decision tree is given in Figure 8..



TPSA <= 57.07
→ **class=1**
TPSA > 57.07
→
q017 <= 1
→ **class=1**
q017 > 1
→
HBA-lip <= 4
→
q137 <= 7,
→ **class=0**
q137 > 7,
→ **class=1**
HBA-lip > 4
→ **class=0**

Figure 8.8 Decision tree for blood-brain barrier penetration. The classification boundary is set at logBB of -0.5. Class 0 = CNS- ; Class 1=CNS+ . The graphical representation of the tree is shown on the left side; the text description is on the right side.

8.8.2 Training Techniques

Up to 20 different DT models can be produced by the StarDrop Auto-Modeller. This is achieved by applying three different techniques for training:

8.8.3 Building Decision Trees Using Stopping Conditions

Stopping conditions prescribe when to stop partitioning the data. Within the StarDrop Auto-Modeller there are restrictions on the minimum number of compounds on a leaf (e.g. the data will not be partitioned further if a leaf contains less than 3 cases) and on the relative frequency of the majority class on a leaf (e.g. it will not split further if 90% of the compounds belong to the same class). Different thresholds for these two stopping conditions are employed to build 12 different DT models (shown as models 1 – 12 in the output of the Auto-Modeller).

8.8.4 Pruning Decision Trees

Rather than using stopping conditions, full-length decision trees are built and then pruned in order to decrease the risk of over-training, i.e. some branches are replaced by leaves using a pessimistic pruning technique (Quinlan J. R., 1993). Four additional DT models are obtained by this approach (shown as models 13, 14, 17, 18 in the output of the Auto-Modeller).

8.8.5 Converting Decision Trees to Rule Sets

The full decision trees are built and converted into a set of rules (Quinlan J. R., 1993) (e.g. HBD-lip<=0 and HBA-lip>6 results in category: 1). The rules are then simplified by statistical tests of independence based upon the contingency tables to remove unnecessary conditions in each rule. Empty, conflicting and inaccurate rules are then removed from the ruleset before the rules are sorted in order of decreasing accuracy, as defined by the Laplace Ratio (Quinlan J. R., 1993). A default rule is created based upon the majority class of the compounds in the training set that do not trigger any rule when run through the model. If all the compounds in the training set trigger a rule, the overall majority class of the training set is used as the default.

When a ruleset is used to classify a compound, several rules may apply to the compound, sometimes predicting different classes. There are two ways to resolve such conflicts: Apply the rule with highest accuracy, or use a voting system to obtain the prediction. In the first case the rules are applied in order and a compound is progressed down the list of rules only if it has not been classified by earlier rules (models 15, 19 in the output of the Auto-Modeller). In the latter case each applicable rule votes for its predicted class with a weight equal to its accuracy, after which all the votes are added up. The class with highest total vote is chosen as the predicted class. It should be noted that the default rule is not used in voting and is used only if no other rule applies (models 16, 20 in the output of the Auto-Modeller).

8.8.6 Decision Tree Model Output

The Model Details file generated by the StarDrop Auto-Modeller for the DT technique contains information about the model classes and the descriptors used. A depiction of the tree or ruleset shows how molecules are classified on the basis of the descriptor values. For each leaf or rule the information on number of compounds on this leaf and number of misclassified compounds is given for the training, validation and test sets. The model performance statistics (Kappa-statistic, Overall accuracy, Specificity, Sensitivity) are also included along with the confusion matrices for the training, validation and test sets. See Section 8.9.2 for more information about the model performance statistics. In addition to the above you can generate information on which compound belongs to which leaf/rule via the option 'Create Split Data Sets'.

8.8.7 Model Analysis and Selection

Performance measures for continuous models

To estimate the model accuracy for continuous models we use R^2, the coefficient of determination:

$$R^2 = 1 - \frac{\sum_i (y_i^{pred} - y_i^{obs})^2}{\sum_i (y_i^{obs} - \overline{y^{obs}})^2},$$

and the Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i^{pred} - y_i^{obs})^2}\ ,$$

where N is a number of compounds in a set. The RMSE is expressed in the same units as the observed property values.

The coefficient of determination ranges from 0 to 1 and the closer it is to 1 the better the model. R^2 describes the proportion of the variation in the observed property values that is explained by the fitted regression, e.g. if we have $R^2 = 0.85$ this means that 85 % of the variation in Y is explained by the model. Note that this definition of R^2 is different from the Pearson correlation coefficient,

$$R^2_{Pearson} = \frac{\left(\sum_i (y_i^{pred} - \bar{y}^{pred})(y_i^{obs} - \bar{y}^{obs})\right)^2}{\sum_i (y_i^{pred} - \bar{y}^{pred})^2 \sum_i (y_i^{obs} - \bar{y}^{obs})^2}\ ,$$

which is often used in other modelling systems. The Pearson correlation coefficient, quantifies how well the predicted versus observed values fit to a straight line, but not the ideal line of unity.

You can view the graph of predicted property values versus observed values for compounds from the training, validation and test sets. An example of such a graph is given in Figure 8..

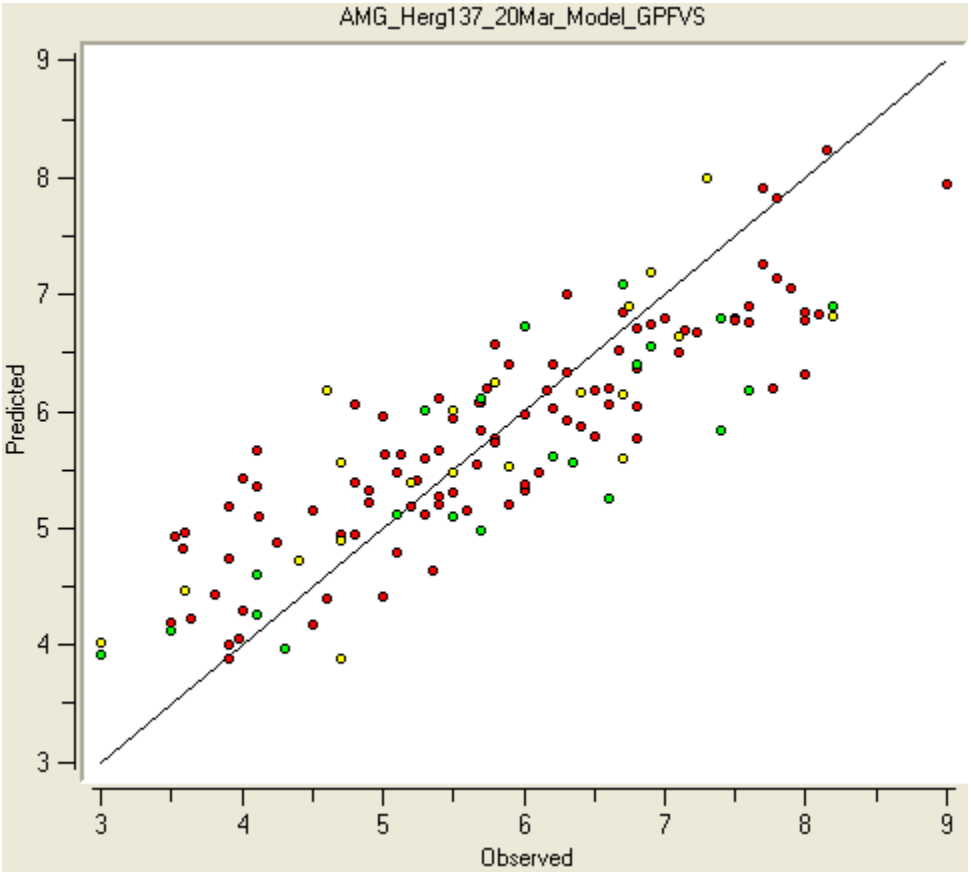


Figure 8.9 The predicted versus observed pIC50 values for hERG inhibition. The compounds from the training set are shown in red, the validation set in green and the test set in yellow.

Performance measures for classification models

An overview of the model performance for a categorical model is given by the 'confusion matrix' showing how the misclassified cases are distributed among classes. An example of the confusion matrices for both the training and validation sets for a two-class model is given in Table 5. Obviously,

for a good model the diagonal elements of the confusion matrix will be much greater than the non-diagonal elements.

To measure the performance of classification models we use the kappa-statistic (κ) as well as the overall accuracy. Here, the overall accuracy is a ratio of the number of correctly predicted cases versus the total number of cases in a set. The kappa-statistic summarizes all the information from the confusion matrix in one number. It assesses the model's improvement in prediction over chance and measures the agreement between observed and predicted classifications with adjustment for chance:

κ = (Observed Agreement- Chance Agreement)/ (Total – Chance Agreement).

Table 8 Example confusion matrices for the training and validation sets for a two-class model. The top left cell of the confusion matrix contains the number of compounds that were observed and predicted in Class 1; the bottom left cell contains the number of compounds that were observed in Class 0 and predicted in Class 1; etc. The overall accuracy and kappa-statistic are also given for each set.

Training set				Validation set			
		Predicted				Predicted	
		Class 1	Class 0			Class 1	Class 0
Observed	Class 1	69	0	Observed	Class 1	17	1
	Class 0	11	61		Class 0	2	10
Overall accuracy = 92.2 % Kappa-statistic = 0.844 Specificity for class 1 =0.86 (86%) Specificity for class 0 =1 (100%) Sensitivity for class 1 =1 (100%) Sensitivity for class 0 = 0.85 (85%)				Overall accuracy = 90 % Kappa-statistic = 0.789 Specificity for class 1 =0.89 (89%) Specificity for class 0 =0.91 (91%) Sensitivity for class 1 =0.94 (94%) Sensitivity for class 0 = 0.83 (83%)			

To define κ for a two-class model, let the confusion matrix be

	Predicted in Class 1	**Predicted in Class 0**
Observed in Class 1	a	b
Observed in Class 0	c	d

In this case the kappa-statistic is expressed in the following way in terms of the elements of the confusion matrix:

$$\kappa = \frac{(a+d)-\eta}{N-\eta},$$

where $\eta = \frac{1}{N}\left[(a+c)(a+b)+(d+b)(d+c)\right]$ is the chance agreement. The extension of these formulae to the case of multiple class models is straightforward.

Interpretation of the kappa-statistic varies in the literature. We recommend using the following ranges:

- ı< 0.5 poor or fair agreement,

- 0.55≤ κ <0.6 moderate agreement,
- 0.6≤ κ <0.8 good agreement,
- 0.8≤ κ <1 very good agreement.

Measures of specificity and sensitivity give an assessment of model accuracy within classes. Specificity for a class is defined as a number of correctly predicted compounds in this class divided by a total number of compounds predicted in this class. Sensitivity for a class is defined as a number of observed compounds in this class which were predicted correctly divided by a total number of compounds observed in this class.

For example, for the above case of two-class model, we have

- Specificity for class 1 = a/(a+c)
- Specificity for class 0 = d/(b+d)
- Sensitivity for class 1 = a/(a+b)
- Sensitivity for class 0 = d/(c+d).

Selection of the best model
Selection of the best model is made by its performance on the validation set. In the case of continuous models the best model is the one with smallest RMSE on the validation set. For the categorical models the best model is the one with highest kappa-statistic on the validation set. In cases where there are multiple models with equal kappa-statistic values for the validation set the model with greatest kappa-statistic for the training set is chosen.

Tips on model analysis
The Auto-Modeller will automatically select the best model as described in herein. Sometimes due to limitations of the initial data set or an 'unlucky' data set split into training, validation and test sets, the best model might still not be satisfactory. You should take a decision on whether to accept the best model by analysing the performance measures, graphs and confusion matrices for the best model as well as the other available models. The following points are worth mentioning.

A model built on a selected subset of descriptors might be preferable to one built on a full descriptor set, even if it has a slightly smaller R^2.

It is more difficult to build multi-class classification models than two-class models. In these cases you should expect to achieve a lower overall accuracy than in the case of two-class models.

8.9 Random Forests

Random Forests (RF) is an ensemble method that seeks to reduce the variance of a large collection of noisy-but-unbiased trees by combining their outputs. In other methods (e.g. bagging), the potential reduction in variance is limited by the fact that using a large number of trees will typically lead to pairs of trees exhibiting a substantial degree of correlation. Random Forests aims to reduce this correlation – and hence improve the variance reduction – by only using a small, randomly-chosen selection of *m* input variables at each splitting step in the tree-growing process. Typically, reducing *m* will have the effect of decreasing the variance at the cost of slightly increasing the overall bias of the ensemble. In our implementation, we use a heuristic to automatically calculate an appropriate value of *m* based on the data.

For classification, Random Forests builds a model whose prediction at an input point is given by the majority class vote over the committee of trees; for regression, however, the model built will make a prediction by *averaging* the output over all trees.

Random Forests requires very little tuning, with just one user-configurable parameter in our implementation; it is also scalable to large data sets, robust in the presence of many descriptors, and produces simple estimates for descriptor importances as well as the uncertainty in predictions at new input points.

See (Breiman, 2001) for more details on the workings of Random Forests.

8.9.1 Building a Random Forest

In our implementation of Random Forests, we use different underlying "weak learners" depending on whether we are performing classification or regression. In the case of classification, the weak learner we use is a standard classification tree with minimum cross-entropy used as the splitting criterion, whereas for regression we use a regression tree with minimum variance as the splitting criterion. We grow the trees fully without pruning in both cases.

For a training set of size *N* and a random forest comprising *B* trees, Random Forests starts by drawing *B* bootstrapped samples of size *N* from the training data. For each of these bootstrapped samples, we grow a new classification or regression tree in the standard manner – except that whenever we have to choose a split point for a node, we only consider a random selection of *m* variables from the *p* possible descriptors. Upon completion of this procedure, we return the ensemble of *B* random trees.

8.9.2 Descriptor Importances

We calculate descriptor importances by determining each tree's prediction accuracy on out-of-bag samples. Specifically, for each tree in the ensemble, we first compute its prediction accuracy over all data points not in its bootstrapped sample; we then randomly permute the values for a single descriptor *j* and again compute the prediction accuracy of the tree over the out-of-bag samples. Because this randomisation effectively voids the effect of descriptor *j*, we determine this descriptor's importance by averaging the percentage decrease in prediction accuracy over all trees after performing the randomisation.

8.10 Confidence in Prediction

With each predicted value of a property, StarDrop reports a confidence in that prediction. For continuous models this is the standard error of prediction between predicted and observed values; for classification models it is the probability of belonging to the predicted class. Gaussian Processes models provide individual confidences in prediction for each compound. For the rest of modelling techniques the confidence in the prediction for each compound is calculated using the chemical space of the model as described in Section 6.6. For models generated by the Auto-Modeller the confidences are obtained in the following way:

Continuous models

Gaussian Processes techniques provide standard deviation in prediction, σ together with each prediction (see Section 8.5.1). This provides an estimate of confidence in prediction for a 'new' compound with an unknown observed value. For a modelling set compound the variance in prediction can be obtained by adding hyperparameter θ_3 to the variance σ^2, the former accounting for observational error (Obrezanova, Csanyi, Gola, & Segall, 2007).

For the other modelling techniques, the confidence for a compound within the chemical space of the model is calculated based on the combined RMSE for all the compounds in the validation and test sets that are within the chemical space of the model. The confidence for compounds that lie close to, but not in, the chemical space of the model is calculated based on the combined RMSE for all the compounds in the validation and test sets that also lie close to, but not in, the chemical space of the model. (See Section 6.6 for details on chemical space.)

Categorical models

The confidence is calculated as the Laplace Ratio (Quinlan J. R., 1993) and is based on the joint training, validation and test sets of the model. The chemical space of the model is taken into account in the same way as for continuous models. It is necessary to include the training set compounds when calculating the confidence in categorical models to ensure there are an appropriate number of compounds predicted at each node in the decision tree (or for each rule in the ruleset) to calculate a confidence for that node (or rule).

9 MPO Explorer™

StarDrop provides a rigorous approach to multi-parameter optimisation called Probabilistic Scoring (see Chapter 2). This allows a project team to define a scoring profile, specifying the property values that would be satisfied by an ideal compound for their project's objectives, and the importance of each criterion, which defines the acceptable property trade-offs if a perfect compound cannot be found.

However, as the range and volume of experimental and calculated data generated in early drug discovery increases, it is not immediately obvious how to choose an appropriate scoring profile based on more complex multi-dimensional data. This is particularly difficult when trying to satisfy a variety of different drug discovery objectives, such as lowering the risk of adverse events or identifying compounds for alternative routes of administration. A subjective approach may not necessarily yield the *optimal* property profile and one cannot use this approach to construct profiles for objectives where expert knowledge might be lacking or where the range of available experimental and calculated properties is very large.

Identifying an appropriate scoring profile is also challenging because, while it is relatively straightforward to identify trends for individual properties, a successful compound may require multiple criteria to be satisfied simultaneously. Furthermore, if some properties in the profile are highly correlated and therefore redundant, we would like to remove the redundant properties from the profile in order to focus only on the relevant subset of profile properties. This is particularly important when data are obtained experimentally, because we do not want to spend valuable time and resources generating data that add little value to our ability to select successful compounds.

MPO Explorer provides a Profile Builder that can be applied to compounds for which the outcome of the chosen objective is known, to find easily interpretable multi-parameter property criteria that best distinguish successful from unsuccessful compounds. Furthermore, the importance of each criterion to selecting successful compounds is also determined, enabling resources to be focussed on generating the most critical data. The resulting property criteria and their importances are represented by a scoring profile within StarDrop, which can then be applied prospectively to new data to select compounds with a high probability of meeting the objective.

Whether a scoring profile is developed in MPO Explorer's Profile Builder or created in another way, it is also important to know if the specific property criteria defined in the profile may be artificially distorting decisions about which compounds to pursue. If a small change in a criterion or its importance would lead to a different decision, this can highlight new avenues for exploration and avoid missed opportunities. The Sensitivity Analysis tool in MPO Explorer enables the robustness of decisions, based on a scoring profile, to be easily tested and helps to explore the impact of sensitive parameters on project strategy.

The following sections describe the methods underlying both the Profile Builder and Sensitivity Analysis tool and the interpretation of their outputs. Illustrative applications of MPO Explorer can be found in Sections 14.10 and 14.11.

9.1 Profile Builder

The Profile Builder in MPO Explorer generates scoring profiles via a process similar to that used by the StarDrop Auto-Modeller and requires little input from the user besides the data set from which the profile is to be built.

The procedure for building a profile is as follows:

1. **Data set:** The user provides a data set of property values, one of which must be the objective for which the rules will be derived. All the other property values in the data set provide the parameters on which the rules may be based. The objective and properties can be either numerical or categorical.
   Note: It does not matter if there are missing data for some of the properties, but any rows with missing data for the objective value will not be used.

2. **Desired objective value:** The user should specify whether to search for 'high' or 'low' values of the objective; accordingly, MPO Explorer will then search for property criteria that maximise (or minimise) the objective. Where the objective is categorical the user will be asked to confirm the relative order of the possible categories to ensure that MPO Explorer understands how these relate to the desired 'high' or 'low' objective.
3. **Data set split:** The data set is split into training, validation, and test sets; these sets can either be provided by the user or generated automatically. The training set is used to train MPO Explorer, which will generate a number of possible rules; the validation set is used to choose the 'best' of these; and the test set is used to assess the predictive power of the final rule (the test set is wholly independent of the rule building process).
4. **Property selection:** The user can specify which of the available properties should be used to train MPO Explorer. Optionally, the user can choose to use only the most predictive subset of these selected properties to train MPO Explorer; this is particularly useful in cases where the number of provided properties is very high.
5. **Profile coverage:** The user can specify how 'big' the eventual rule should be by stipulating a minimum value for the *coverage* of the rule. The coverage is defined as the percentage of compounds in the data set that satisfy the rule; typically, there will be a trade-off between the coverage of the rule and the desirability of compounds covered by the rule.
6. **Rule visualisation and interaction:** After generating a single rule for a profile, MPO Explorer displays the rule in a 'grid view' that allows the user to visualise the relationship between the objective and the rule's multi-dimensional property criteria. At this stage, the rule can easily be modified based on the user's domain knowledge, and the effect the user's changes have on the rule's performance will be reflected in the displayed statistics.
7. **Generating additional rules:** Once the user has decided to accept a rule, MPO Explorer can then be instructed to search for another rule by discarding the compounds selected by the current rule and searching for alternative rules elsewhere in property space. In this manner, MPO Explorer can generate a scoring profile comprised of multiple independent rules.

Upon completion of the profile building process, the user can save, modify, and run the new scoring profile just as any other profile within StarDrop. Scoring profiles generated by the Profile Builder (as well as manually created profiles) can also be analysed and modified *post hoc* via the interactive visualisations used in profile building process. At this stage, the user can also modify the objective to see how the profile might perform against different drug discovery objectives. For more details on using and interacting with MPO Explorer please see the StarDrop User Guide.

The Profile Builder applies a method called the Patient Rule Induction Method (PRIM) (Friedman & Fisher, 1999) to a set of compounds, each with data for multiple properties and a value representing the objective we would like to achieve. The objective value may be a category, e.g. good/bad, or a numerical value that we would like to maximise or minimise. PRIM searches for regions of property space over which the mean objective value is significantly higher (or lower) than the mean over the full property space, i.e. compounds that satisfy the property criteria identified by PRIM will have a higher chance of success for the objective than the average for the compounds in the full set. This method is described in detail in the following sections.

9.1.1 Terminology

Given a set of n (categorical or numerical) properties and associated objective values, we define a *box* (or *rule*) over property space *S* to be the Cartesian product of individual *property criteria* $C_1, \ldots, C_m$ ($m \leq n$):

$$B = C_1 \times \ldots \times C_m$$

where a property criterion C_i for property i is defined to be the closed interval

$$C_i = [c_{i_1}, c_{i_2}]$$

if property *i* is numerical, and the restricted set of categories

$$C_i = \{category_{i_1}, \ldots, category_{i_k}\}$$

if property *i* is categorical.

A *box covering* over *S* is defined to be a union of individual boxes $B_1, \ldots, B_p$ as shown in Figure 9.1.Together with an importance value for each property criterion, a box covering can be used as a standard scoring profile within StarDrop.

The *mean* over a box *B* ($\overline{y^B}$) is defined to be the average objective value of points within the box, and the *support* (or *coverage*) of the box is the percentage of points in the overall data set that lie within the box.

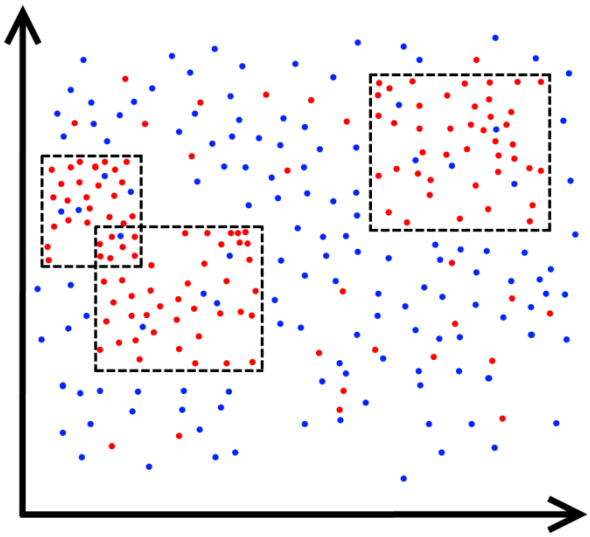


Figure 9.1 A box covering in 2D property space. Red compounds are desirable and blue compounds are undesirable; note that the box covering contains significantly more good than bad compounds compared to the full data set.

9.1.2 Theory Overview

PRIM seeks to find a box covering in property space over which the mean objective value is significantly higher than the mean over the full property space. To construct a single box in the covering, PRIM applies a top-down 'peeling' process followed by a bottom-up 'pasting' procedure, resulting in a *peeling sequence* of boxes from which the optimal box can be selected. We will describe the steps to construct a single box B_1 in more detail in the following sections.

9.1.3 Top-down Peeling

Intuitively, we can think of the box construction strategy as a top-down process called 'peeling', setting B_1 to be equal to the entire property space of the training set *S* and remembering that each face corresponds to an upper or lower bound on an individual compound property value (if the property is numerical). At each step, PRIM compresses the box along a single face, as shown in Figure 9.2; the face chosen for compression is the one that will result in the largest mean $\overline{y_i^{B_1}}$ in the newly compressed box B_1. PRIM then repeats this process until it reaches one of a set of predefined stopping criteria (e.g. if the support of the box B_1 becomes too small).

Specifically, a single peeling step involves considering each property *j* in turn ($1 \leq j \leq n$):

- If the *j*th property is numerical, the peeling step for box B_1 involves considering removing either all the compounds whose property values are below the γ-quantile of property *j* or those compounds with property values above the (1 - γ)-quantile of property *j*, depending on which removal will result in the higher mean for the remaining compounds in the compressed box. Here, γ is the 'peeling fraction' specifying the proportion of compounds to remove in each step.

- If the *j*th property is categorical, the peeling step for box B_1 involves considering removing all the compounds whose property values are equal to one of the *j*th property's possible category

values; PRIM removes the category that will result in the highest mean for the remaining compounds in the compressed box.

- After considering each property *j*, the final choice of the box face to compress is based on which of the above candidates for removal results in the highest mean for the remaining compounds in the compressed box when removed.

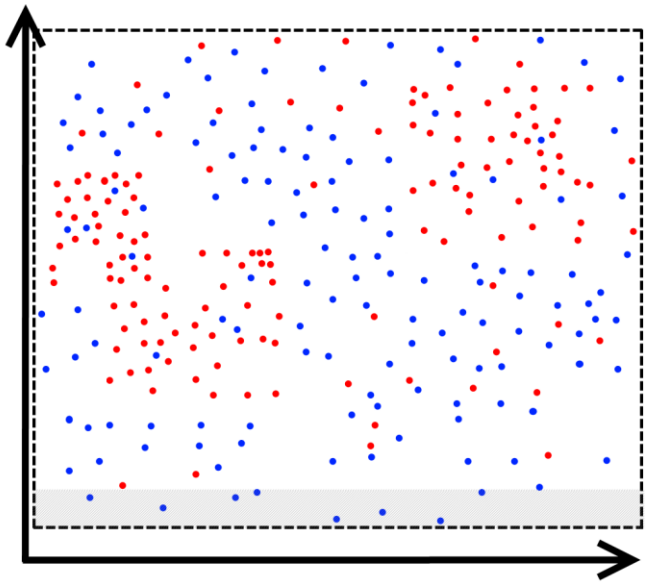


Figure 9.2 Illustration of the first step in top-down peeling. The box covers the whole training set; PRIM chooses to peel the shaded region of the box as it contains only undesirable (blue) compounds.

9.1.4 Bottom-up Pasting

Because top-down peeling greedily chooses the next face for compression, it is possible that PRIM might be able to increase the mean of box B_1 mean still further via a bottom-up 'pasting' strategy, as illustrated in Figure 9.3; this is essentially the inverse of the top-down peeling process. PRIM iteratively expands B_1 along whichever face results in the largest increase in the mean $\overline{y_t^{B_1}}$, stopping when the next expansion will result in a decrease in the box mean.

Specifically, a single pasting step involves considering each property *j* in turn ($1 \leq j \leq n$):

- If the *j*th property is numerical, the pasting step proceeds by considering extending either the lower or upper boundary of B_1 on the *j*th property, thus adding βN_{B_1} of the previously peeled compounds to B_1 where β is the 'pasting fraction' and N_{B_1} is the number of compounds in B_1.

- If the *j*th property is categorical, pasting proceeds by considering adding the compounds whose property values are equal to one of the categories for property *j* not represented in the current box B_1; PRIM adds the category that will result in the highest mean for the new set of compounds in the expanded box once added.

- After considering each property *j*, the final choice of the box face to expand is based on which of the above candidates for addition results in the highest mean for the new set of compounds in the expanded box when added.

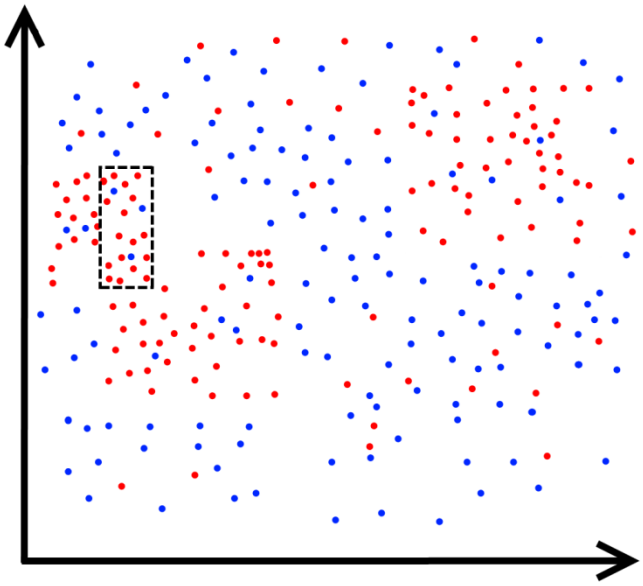


Figure 9.3 Illustration of bottom-up pasting. The box can be extended along the left edge to include several desirable (red) compounds to increase its mean further.

9.1.5 Constructing a Box Covering

As a result of the top-down peeling process followed by bottom-up pasting, there now exists a 'peeling sequence' of boxes induced from the training set. The PRIM algorithm then uses the validation set to compute the mean objective value over each box in the peeling sequence, and the optimal box B_1 is taken to be that with the highest validation set mean.

In this implementation, the user can then either accept this box as the only one in the eventual profile, or start the peeling-and-pasting procedure again with the entire property space minus the training compounds from box B_1 (i.e. $S - B_1$) to get a second box B_2. This process can then be repeated until we have a box B_{p+1} that the user decides to reject, typically because the mean or support are not sufficiently high. The final result will be a 'covering' of boxes $B_1, \ldots, B_p$ that collectively describes the region of the property space where the mean objective value is high.

9.1.6 Property Criterion Importance Values

In the following discussion, we assume that the objective y_i is binary, so that for compounds meeting the drug discovery objective, y_i is set to 1, and for those that do not meet the objective, y_i is set to 0. In StarDrop, PRIM handles numerical and multi-category objectives by requesting that the user specify the 'desired values' of the objective, which are then used to determine which compounds should be assigned a y_i value of 1 and which should be given the value of 0.

PRIM calculates the importance of each property criterion by determining the ratio between the probability that a training compound meeting the criterion achieves the objective and the probability that a compound *not* meeting the criterion achieves the objective.

Let $h_{jk}(x_{ij})$ be the indicator function for whether the training compound property value x_{ij} lies within box B_k, so that we can define an overall classification function $g_k(\boldsymbol{x_i}) = \prod_j h_{jk}(x_{ij})$, where $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{im})$.

Now consider generalising $h_{jk}(x_{ij})$ to $\bar{h}_{jk}(x_{ij})$ so that instead of being a zero-one indicator, $\bar{h}_{jk}(x_{ij})$ is the α_j-one indicator defined by

$$\bar{h}_{jk}(x_{ij}) = \begin{cases} 1 & \text{if } x_{ij} \text{ satisfies property criterion } j \\ \alpha_j & \text{otherwise} \end{cases}$$

The constant α_j is defined to be the false-negative rate of criterion *j*, i.e. the probability that a training compound whose *j*th property value does not satisfy its associated criterion does in fact satisfy the objective.

This generalisation of $h_{jk}(x_{ij})$ to $\bar{h}_{jk}(x_{ij})$ leads to an associated generalisation of g_k to $\overline{g_k}(\boldsymbol{x_i}) = \prod_j \bar{h}_{jk}(x_{ij})$. The function $\overline{g_k}$ defines a likelihood over the sets of values $X = \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_N}\}$ and $Y = \{y_1, \ldots, y_N\}$:

$$L(X, Y) = \prod_{i=1}^{N} [y_i + (1 - y_i)\alpha_j]$$

where *N* is the number of training compounds. Note that the function *L* is convex in α_j and so this optimisation is tractable. We will now explain how it allows us to estimate property criteria importance values as a constrained maximum likelihood optimisation performed over the full set of training compounds.

For each compound *I* with property values $\boldsymbol{x}_i$, we can define a vector of $h_{jk}(x_{ij})$-indicators specifying whether the compound property value x_{ij} lies within the box B_k. If the optimisation problem is non-degenerate, then the probability that a particular compound *i,* with property values $\boldsymbol{x}_i$, satisfies the objective of interest – i.e. $P(y_i = 1 \mid \boldsymbol{x_i})$ – is a monotonically decreasing function of $h_{jk}(x_{ij})$. Furthermore, if we add the restriction that $P(y_i = 1 \mid \boldsymbol{x_i}) = 1$ when all relevant conditions are fulfilled, we can see that our generalised classifier $\overline{g_k}$ is actually the constrained maximum likelihood estimate $\overline{g_k}(\boldsymbol{x}_i) \approx P(y_i = 1 \mid \boldsymbol{x_i})$.

The desired property criterion importance λ_{jk} is equal to $1 - \alpha_j$, and represents the probability that applying a certain property criterion would lead to a compound being mistakenly rejected – this is analogous to the notion of statistical power. Importantly, λ_{jk} as defined above is correlation corrected, so that given two highly correlated variables, the one with higher explanatory power will have high importance and the other will have low importance (as it has low residual explanatory power).

9.1.7 Optimal Property Selection

In situations where the number of properties in the data set is prohibitive, PRIM will probably generate a profile much faster if it is only trained on the most predictive property subset. Furthermore, the profile generated will probably have superior predictive performance to one built from training PRIM on every property in the data set.

To search for the most predictive property subset of the full set of *n* properties, the Profile Builder uses forward selection combined with a cross-validation approach. Starting with an empty set of properties, the Profile Builder iteratively adds one property at a time to the set used to train PRIM; at each step, the property chosen for inclusion is the one that provides the most improvement in the performance of the profiles generated by PRIM. This process is stopped when it is no longer possible to significantly improve the performance of the generated profiles by adding additional properties.

To assess the performance of the profiles generated by PRIM from a single set of properties, the Profile Builder splits the training set into *k* different folds, using $k - 1$ folds to train PRIM and the remaining fold for validation. For each of the *k* cross-validation runs, the Profile Builder computes the mean of the generated profile over the validation set; it then uses the average of these *k* means as its metric for determining the predictive power of the current property set.

9.1.8 Generating 'Soft' Box Boundaries

By default, the Profile Builder will generate continuous box boundaries as hard cut-offs, so that each compound in the data set will either be strictly inside or outside the box. However, the Profile Builder also has the ability to compute 'soft' box boundaries that reflect cases where a hard cut-off is not appropriate or when sparse data does not enable the box thresholds to be determined with confidence.

Soft box boundaries are computed by finding confidence intervals for each box boundary. To compute a single confidence interval for a box boundary, the Profile Builder uses *k*-fold cross-validation on the combined training and validation set, dividing it into *k* different folds with *k* – 1 folds used to train PRIM and the remaining fold for validation. For each of the *k* cross-validation runs, the Profile Builder computes the numerical value of the box boundary threshold, giving a set of *k* different thresholds for the box boundary. The soft threshold for the box boundary is then given by the 95% confidence interval for the mean value of the generated box boundary thresholds, assuming that the box boundary thresholds are normally distributed.

9.1.9 Interpretation of Profile Builder Results

Several statistics are produced by the Profile Builder with which to assess the quality of the rules derived by PRIM. These may be summarised as follows:

- **Mean improvement** – The improvement in the mean objective value for compounds in the box defined by the rule criteria, compared with the mean over the entire data set, i.e. how much better is it to choose a compound within the box than choosing one randomly from the entire data set. This is expressed as a percentage.
- **Coverage** – The percentage of the total number of compounds in the data set that lie within the box defined by the rule criteria.
- **Odds ratio** – The ratio between the odds of randomly selecting a 'good' compound from within the box compared with the odds of randomly selecting a 'good' compound from outside the box.
- **P-value** – The statistical significance of these results expressed as a probability of achieving the result by chance, i.e. a low value is better.

Additional statistics, including confusion matrixes for each set, can also be seen in a report.

In the case of multi-class or continuous objectives, where the goal is to maximise or minimise the mean objective within the box, the odds ratios and confusion matrices are calculated based on a binary classification. The classification boundary for this can be defined using the parameters "True/False Threshold" (for continuous objectives) or "Minimum/Maximum desirable category" (for classification objectives) when setting the profile parameters in the Profile Builder.

When comparing PRIM with traditional classification methods, it should be noted that the mean improvement is not directly comparable with statistics such as accuracy, specificity and sensitivity; it is a better measure of the *improvement* provided by the rule criteria. This is particularly important when considering biased sets; for example, if we consider a binary classification objective in which a data set has 90% 'good' compounds and 10% 'bad', the null model that classifies all compounds as 'good' would have a specificity of 90% for selecting 'good' compounds. But this would not, in practice, be any better than a random selection. In this scenario, a rule with a mean improvement of 10% would imply a specificity for selecting 'good' compounds of 99%, i.e. 99% of compounds that obey all the rule criteria will be 'good'.

9.2 Sensitivity Analysis

The Sensitivity Analysis tool within MPO Explorer enables the user to determine if a small change in any property criteria or their importance in a scoring profile will have a significant impact on the compounds selected from a given data set. To achieve this, the tool considers changes in the priorities of the highest scoring compounds to identify when those compounds that would be selected will change significantly. A 'sensitivity score' between 0 and 1 is reported for each parameter in the scoring profile, where a parameter is defined to be either a property's desired value or its importance. A high sensitivity score means that a small change in the parameter will have a significant effect on the choice of top-scoring compounds from the data set. The sensitivity score also takes into account the effect of uncertainty in the compound property values in the data set, to identify when a change in compound selection is statistically significant.

For a continuous property in a profile, two sensitivity scores will be reported: one is the 'value sensitivity' score, which quantifies the sensitivity of the selection of compounds to a change in the desired property value, and the other is the 'importance sensitivity' score, which quantifies the sensitivity of the selection of compounds to changes in the importance of the property. A categorical property has only an importance sensitivity score because the value ranges cannot be changed.

To calculate the value sensitivity score for an individual continuous property criterion, we first define a number of perturbations to the desired value represented by 'shifts' to the original scoring function for the property. In a simple example, if the original desired values for logP are in the range (0, 3.5), a shift of 0.5 units to the right would change the desired value range to (0.5, 4). An example for a more complex scoring function is shown in Figure 9.4.

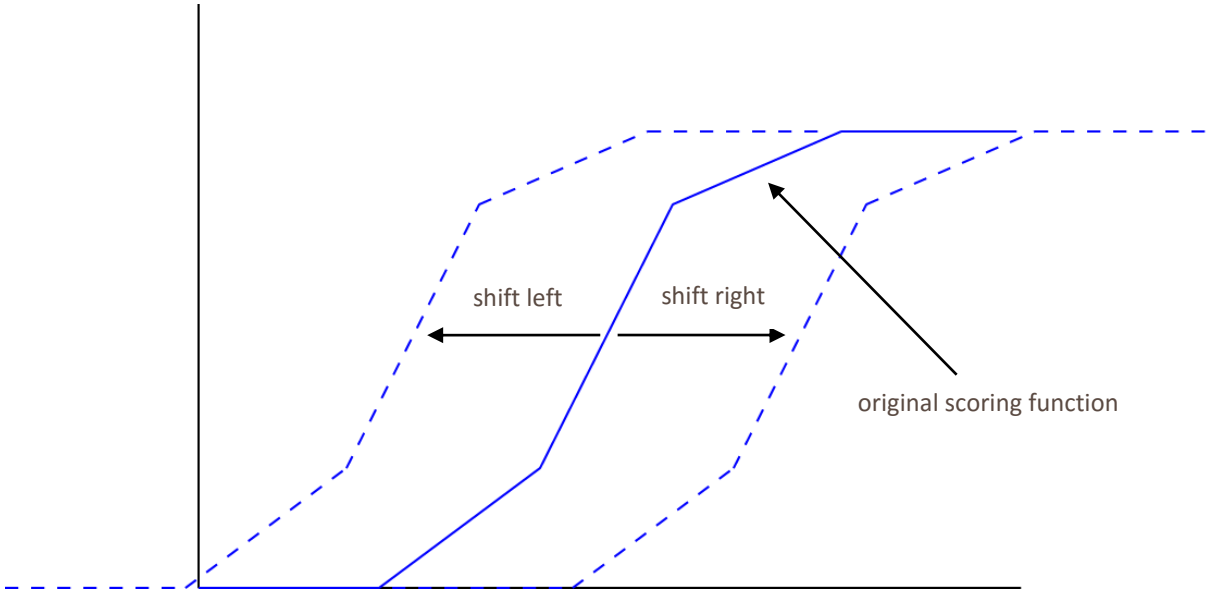


Figure 9.4 Original scoring function together with two shifted versions of the original function.

To calculate the sensitivity score due to a shift in a scoring function, we start by computing the list of scores, L_{old}, of every compound in the data set using the original scoring profile with the unmodified scoring function for the property. Next, we compute the list of scores L_{new} of every compound in the data set using the scoring profile with the new, shifted scoring function for the property.

We then compute Spearman's rank correlation coefficient between L_{old} and L_{new}. To do this, we determine the rank of each compound in L_{old} and L_{new}; let $\{x_1, \dots, x_n\}$ be the list of ranks for each compound in L_{old} and $\{y_1, \dots, y_n\}$ be the list of ranks for each compound in L_{new}.

Spearman's rank correlation coefficient between L_{old} and L_{new} is then defined as Pearson's correlation coefficient between the lists of ranks, i.e.

$$\boldsymbol{\rho} = \frac{\sum_{\mathbf{i}}(\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{y_i} - \bar{\mathbf{y}})}{\sqrt{\sum_{\mathbf{i}}(\mathbf{x_i} - \bar{\mathbf{x}})^{\,2} \sum_{\mathbf{i}}(\mathbf{y_i} - \bar{\mathbf{y}})^{\,2}}}$$

However, note that we only consider the top-scoring compounds (i.e. those compounds with the highest *original* scores) in the computation of the correlation coefficient. By default the top 10% of the compounds in the data set are considered, but this can be modified by the user, if required. The sensitivity score for the single shift to the scoring function is then defined to be $1 - \rho$.

To account for the effect of uncertainty in the score values, we adjust the standard computation of Spearman's correlation coefficient by assigning the maximum possible correlation contribution to compounds with old and new scores that are not statistically significantly different when their uncertainties are taken into account. Specifically, if compound *i* has the original score s_i and new score t_i, we calculate $P(|s_i - t_i| > 0)$, assuming the individual scores are normally distributed. If this

probability is below a specified significance threshold (by default this is 0.75, but this may be changed by the user), we consider the score change to be insignificant. In this case, if the compound has original rank x_i, we change its new rank from y_i to $x_i - (\bar{x} - \bar{y})$, i.e. the compound is given the same rank translated by the difference between the means of the original and new sets of ranks.

We then consider multiple incremental shifts within a 'window' around the original scoring function. The size of the window is defined by a fraction of the total range of the property in the data set (by default this is 50%, but this can be changed if required). To determine an overall value sensitivity score for the property, we first calculate a sensitivity score for each individual scoring function corresponding to a shift within the window. The value sensitivity score for the property is then defined to be the maximum sensitivity score over shifts within this window. Figure 46 shows a plot of the value sensitivities for multiple shifts in the desired value of 5HT1a affinity (pK_i) and illustrates the window of shifts considered in the calculation of the overall value sensitivity for this property. In this case, the sensitivity of the compound selection to shifts in the desired value for this property is very high. The overall value sensitivity score for the property is 1, because there is at least one individual shift within the window that results in the maximum sensitivity score of 1.

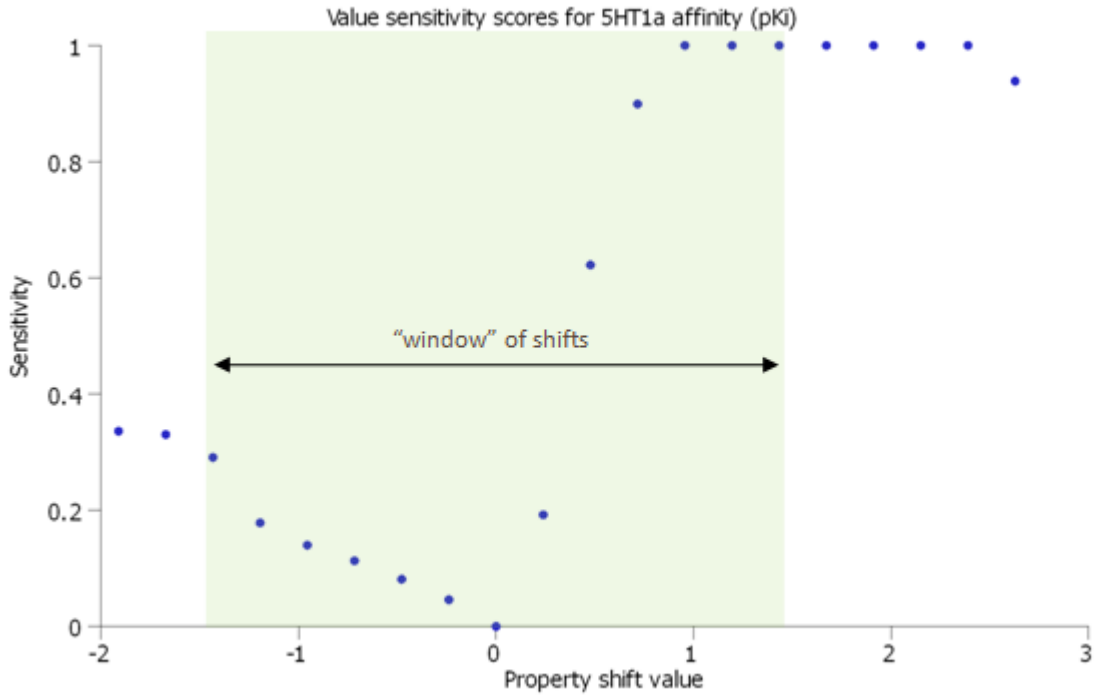


Figure 9.5 Graph showing how the sensitivity score varies with shift to the property's original scoring function. The shift of the scoring function is plotted on the x-axis and the resulting sensitivity score is plotted on the y-axis. Note that a shift of 0 corresponds to the original scoring profile and hence, by definition, the sensitivity will be 0. The window of shift values used to compute the property's overall sensitivity score is also shown.

The sensitivity score for a property criterion's importance is computed in a similar manner. However, instead of a window of incremental shifts to the property's original scoring function, we define a window of shifts to the property's original importance. For example, if the property's original importance value is 0.6, by default we define a window of shifts varying from -0.25 to 0.25 around the original importance value, giving a number of importance values ranging from 0.35 to 0.85 (by default the window is 50% of the total importance range from 0 to 1, but this can be changed by the user, if required). The importance sensitivity score for the property is then defined to be the maximum sensitivity score over all importance values within this window. The method for calculation of the sensitivity score of a single importance within the window is the same as that used for a single scoring function shift, as described above. Figure 9.6 shows a plot of the importance sensitivities for multiple shifts in importance value of P-gp category and illustrates the window of importance values considered in the calculation of the overall importance sensitivity for this property. In this case, the sensitivity of

the compound selection to changes in the importance of this property is low. The overall importance sensitivity score for the property is 0.035, because this is the maximum sensitivity value for an importance value within the window.

The Sensitivity Analysis tool will display a table showing the value and importance sensitivities for each property in a scoring profile, as shown in Figure 9.7. This is sorted, such that the property with the highest sensitivity (either for value or importance) will be at the top, helping to easily identify the property criteria that should be most carefully considered.

By selecting an overall sensitivity value from the table, the Sensitivity Analysis tool will also display a graph (see Figure 9.6), showing how the sensitivity score varies as the shift to the property's original scoring function is changed. For a value sensitivity, each individual point in this graph represents a single scoring function corresponding to a shift in the desired value, with the x-coordinate giving the shift value and the y-coordinate giving the sensitivity score for this single scoring function. For the importance of the property criterion, an analogous graph is produced showing the sensitivities corresponding to shifts in the importance (see Figure 9.6). These graphs can be used to identify the range of desired values or importance values over which the selection of compounds is not sensitive. If you are confident that the 'correct' criterion or importance lies within this range, there is no need to consider changes to this parameter.

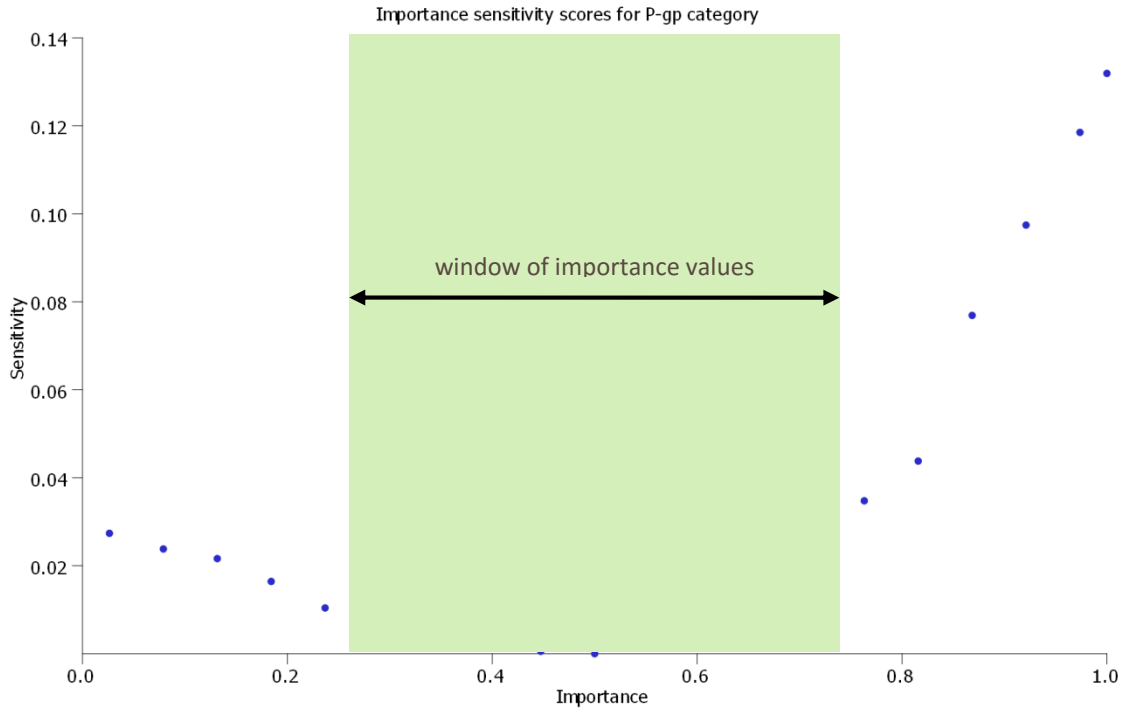


Figure 9.6 Graph showing how the sensitivity score varies with shift to the property criterion's importance. The importance of the property criterion is plotted on the x-axis and the resulting sensitivity score is plotted on the y-axis. Note that the importance of this property in the original profile was 0.5 and hence, by definition, the sensitivity will be 0. The window of shift values used to compute the property's overall sensitivity score is also shown.

	Value Sensitivity	Importance Sensitivity
5HT1a affinity (pKi)	1.000	0.008
logP	0.310	0.096
BBB log([brain]:[blood])	0.249	0.015
hERG pIC50	0.096	0.207
2D6 affinity category	N/A	0.107
BBB category	N/A	0.055
logS	0.040	0.002
P-gp category	N/A	0.035
PPB90 category	N/A	0.034
2C9 pKi	0.008	0.002
HIA category	N/A	0.0

Figure 9.7 Example output of value and importance sensitivities for a scoring profile, generated by the Sensitivity Analysis tool.

If the selection of compounds is sensitive to small changes in the desired value or importance of a property, it is important to consider the impact on the compounds that would be selected. Clicking on a point in the graph of sensitivity versus shift will bring up a second graph (Figure 9.8) showing the new scores of every compound in the data set (i.e. the scores computed using the new, shifted scoring function or importance value) plotted against the original scores (i.e. the scores computed using the original scoring function). These points are coloured according to how much their corresponding compounds' ranks have changed: very red compounds have greatly decreased their rank, very yellow compounds have greatly increased their rank, and grey compounds have not significantly changed their rank. Clicking on any point in this graph will highlight the corresponding compound in the data set itself. This can be used to easily identify compounds that would significantly change in priority, given a small change in the scoring profile. This can highlight potential new research directions or missed opportunities that would be worth more detailed investigation.

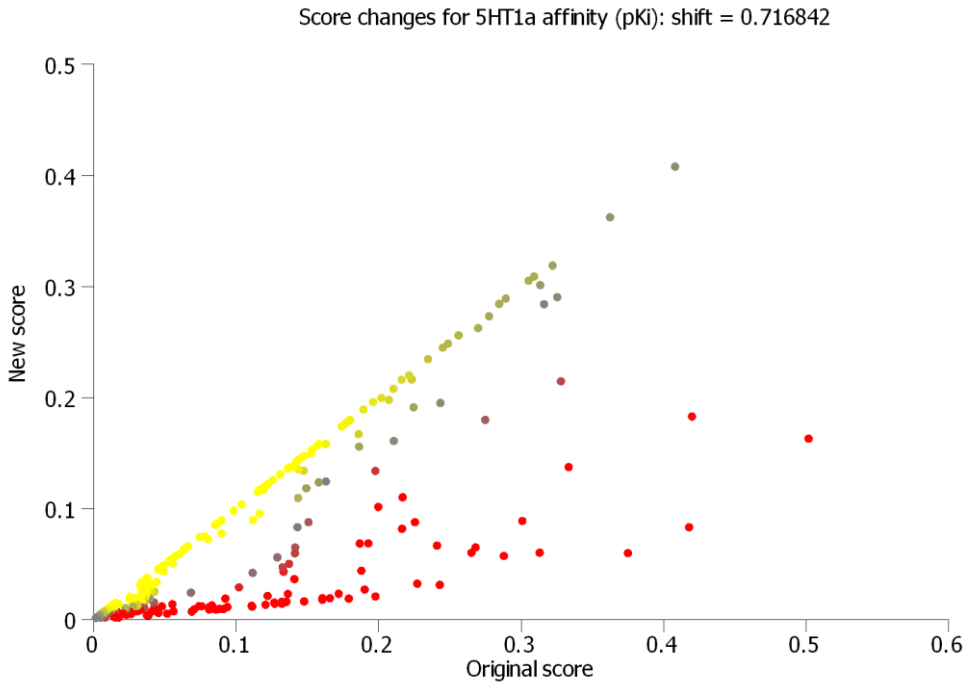


Figure 9.8 Graph showing the new scores of every compound in the data set plotted against the original scores. The new scores are calculated with the scoring profile with the shifted parameter value. The colours of the points indicate the change in priorities of the compounds; yellow points represent compounds that will have significantly higher priority with the new profile and those in red correspond to compounds that will have significantly lower priority with the new profile.

Ideally, we would like to find that the compounds selected using a scoring profile are not sensitive to any small changes to the parameters defining the scoring profile. In this case, we can be confident that

the specific choice of parameters does not artificially distort the compounds that will be selected. However, if the prioritisation of compounds is sensitive to one or more parameters of the scoring profile then these parameters should be considered in more detail, because the decision we would make regarding the selection of compounds will depend significantly on the specific values we have chosen. If we are confident that the values in the profile are 'correct' we can proceed with the selection of compounds as usual. However, if we are uncertain of the most appropriate values for these parameters, we should consider alternative compounds that would be selected by alternative profiles in which the sensitive parameters have been changed within reasonable ranges. Investigation of these variations may identify alternative compounds that would be valuable to consider and downstream testing of these may help to determine the most appropriate scoring profile for selecting further high quality compounds.

10 Nova™ Idea Generation

10.1 Introduction

StarDrop's models, probabilistic scoring, chemical space and Glowing Molecule methods provide the capability to quickly assess a large number of ideas for potential leads or candidate drugs and prioritise those with the highest chance of downstream success. These methods can assess a very large number of potential ideas and hence provide the opportunity to explore a wide range of possibilities to find compounds with a good balance of properties. Therefore, the limitation to this exploration can be the time and experience necessary to generate a wide diversity of compound ideas and enter these structures into StarDrop for analysis.

One approach that Nova provides is 'Idea generation' where new compound structures are generated by applying established medicinal chemistry 'transformation rules' to an initial compound. Many generations of ideas can be explored and the resulting suggestions can be automatically prioritised according to a predicted property, probabilistic score or chemical diversity. Section 10.2 describes the methods used by this approach to generate new compound ideas and the validation of the underlying transformations.

An alternative approach, also available within Nova, uses 'Matched series analysis' to suggest new compound ideas. By comparing matched series found in your data with a database of other matched series (a knowledge base), relevant predictions for new substituents that are likely to improve target activity or another property of interest can be made. The suggestions are based on the premise that a matched series with similar activity order in your data and the knowledge base implies that those groups occupy a similar binding environment created by their target proteins. Given a similar binding environment, groups that have been shown to be better binders within the knowledge base, have a strong likelihood of being better binders to the target of the input data set. Section 10.3 describes two approaches using matched series analysis that are available within the Nova module.

These are complemented by a flexible virtual library enumeration tool that enables you to define the specific chemistry to be explored by drawing a template with substitution points and listing the modifications or substituents at each position. More details on how to define a virtual library in Nova can be found in Chapter 12 of the StarDrop User Guide.

10.2 Medicinal Chemistry Transformations

Methods for automatically applying medicinal chemistry 'transformation rules' to generate new compound structures have been previously described (Stewart, Shiroda, & James, 2006) (Ekins, Honeycutt, & Metz, Evolving molecules using multi-objective optimization: applying to ADME/Tox., 2010). These typically accept an initial 'parent' structure as input and generate 'child' structures by applying transformations based on collective medicinal chemistry experience. Examples of transformation rules range from simple substitutions or bioisostere replacements to more dramatic modifications of the molecular framework such as ring opening or closing. A computer can store and apply many more rules than a single chemist and can 'learn' from historical examples of transformations between molecules (Raymond, Watson, & Mahoui, 2009). Applying a set of transformations iteratively to produce multiple 'generations' of compound ideas can result in a large number of molecules – too many to be examined visually by a chemist to select the most interesting for further consideration.

The Nova module provides an algorithm to generate compound ideas by applying transformations to an initial molecule, integrated with predictive models and probabilistic scoring to quickly prioritise those ideas most likely to satisfy the required property profile for detailed consideration. The goal is a tool to support experts and stimulate the process of innovation – achieving a creative combination of a computer's ability to cover a wide breadth of possibilities with the experience and detailed knowledge of a chemist.

The following list describes the main principles of Nova:

- It must generate a wide diversity of chemistry, as the objective is to explore many ideas in the search for an optimal solution.

- The compound structures generated must be relevant. In particular, the number of 'nonsensical', e.g. chemically unstable or infeasible, compounds must be kept to a minimum. Also, the chemist must be able to control the generation process, for example by specifying a region that must not be modified or restricting the transformations that will be applied.
- The transformations that are applied should include a broadly representative set of those applied successfully in the past to optimise successful drugs.
- The method used to prioritise the resulting compound ideas should reliably identify high quality compounds within those given the highest rank in the generated set.

In the following sections, we will describe the methods used to create and apply the set of transformations provided with Nova. Furthermore, we will describe the validation of this method to ensure that the transformations cover a broad range of 'drug like' chemistry and that the resulting structures are relevant and not unstable or infeasible. Section 14.6 provides an example of the application of Nova in combination with StarDrop's predictive models, probabilistic scoring and chemical space algorithm. This example illustrates how known drug and similar compounds can be efficiently identified, starting from a lead compound.

10.2.1 Transformations

Two hundred and six transformations were generated manually, as SMIRKS codes, by study of medicinal chemistry literature (Burger, 1970) (Bonnet & Robins, 1993) (Binder, et al., 1987) (Roehrig, et al., 2005) (Patani & LaVoie, 1996) (Black, Duncan, & Shanks, 1965) (Walsh, Franzyshen, & Yanni, 1989) (Fournié-Zaluski, et al., 1994) (Larsen & Lish, 1964) (Rocheblave, et al., 2002) (Yoshino, Kohno, Morita, & Tsukamoto, 1989) (Uno, et al., 1990) (Americ, et al., 1994) (Americ, et al., 1994) (Hynes Jr., et al., 2008) (Sun, et al., 1995) (Parks, et al., 2005) (Cox, et al., 2005) and observation of the optimisation steps between known drugs and the lead molecules from which they were derived. SMIRKS is a reaction transform language designed by Daylight Chemical Information Systems which uses SMILES and SMARTS notations to specify a generic reaction or transformation (Weininger, 1998).

The transformations were divided into seven broad groups: Functional Group Addition, Linker Modification, Remove Atom, Ring Addition, Ring Modification, Ring Removal, Terminal Group Exchange. The distribution of transformations between the groups is shown in Table 6 and examples of each are shown in Table 10.

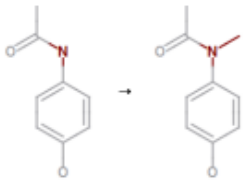
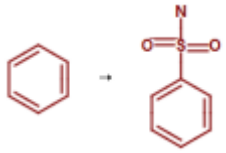
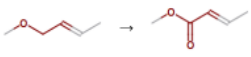
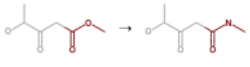
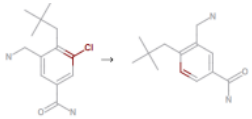
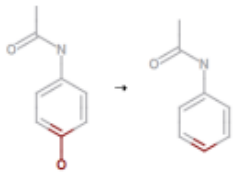
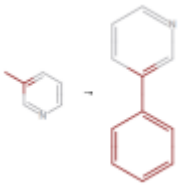
Table 9 Distribution of transformation between groups.

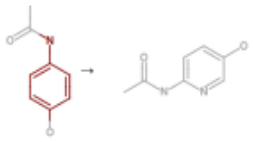
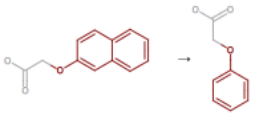
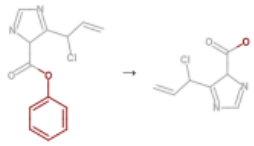
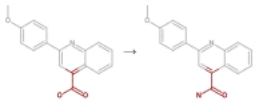
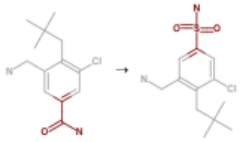
Group	Number of transformations
Functional Group Addition	20
Linker Modification	54
Remove Atom	5
Ring Addition	13
Ring Modification	26
Ring Removal	4
Terminal Group Exchange	84
Total	**206**

The transformations do not necessarily correspond to specific chemical reactions or synthetic routes; rather they are intended to describe changes to molecules that a medicinal chemist might consider in the course of an optimisation project. A single transformation might require multiple synthetic steps or

the synthesis of new building blocks. However, the transformations are typically not major rearrangements – they are relatively feasible moves in chemical space.

Table 10 Example Transformation Rules.

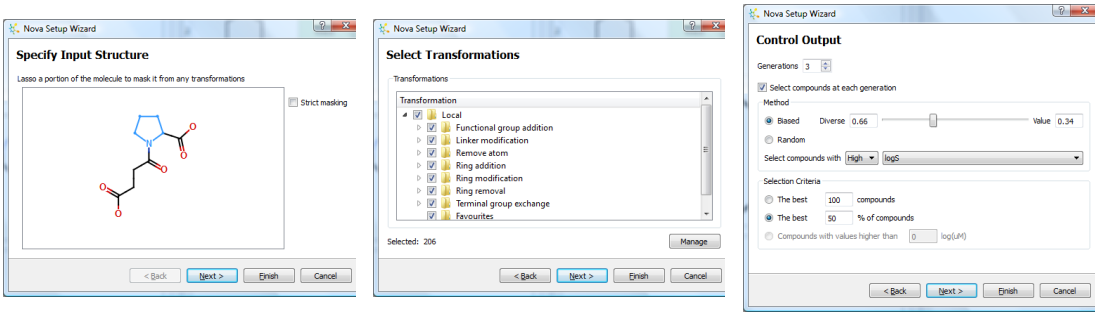
Group	Transformation Name	Illustration	SMIRKS
Functional Group Addition	Methyl addition to amine		[N:1][H]>>[N:1]C
	Sulfonamide addition to benzene		[c:1]1[c:2][c:3][c:4][c:5][c:6]1[H]>>[c:1]1[c:2][c:3][c:4][c:5][c:6]1S(N)(=O)=O
Linker Modification	Secondary carbon to carbonyl		[*;!#1:1][CH2][*;!#1:2]>>[*;!#1:1]C(=O)[*;!#1:2]
	Ester to amide linker		[#6:1]O[C;!R:3](=O)[#6:2]>>[#6:1]N[C;!R:3](=O)[#6:2]
Remove Atom	Remove halogen		[C,c:1][F,Cl,Br,I]>>[C,c:1]
	Remove hydroxyl		[C,c:1][OH]>>[C,c:1]
Ring Addition	Methyl to phenyl		[*;!#1:1][CH3]>>[*!#1:1]c1ccccc1
	Six membered aromatic to indole		[c:1]([H])1[c:2]([H])[a:3][a:4][a:5][a:6]1>>[C:1]12[a:6]=[a:5][a:4]=[a:3][C:2]=1[nH]C=C2

Ring Modification	Phenyl to 3-pyridine		[*;!#1:1][c:2]1[c:3][c:4][c:5][cH][c:6]1>>[*;!#1:1][c:2]1[c:3][c:4][c:5][n][c:6]1
	NC-switch		[*:1]1:[c]([*:2]):[c:10]([*:3]):[n]([*:4]):[*:5]1>>[*:1]1:[n]([*:2]):[c:10]([*:3]):[c]([*:4
Ring Removal	Napthalene to benzene		[*;!#1:7][c:1]1[cH]c2c([cH][c:6]1)[c:5][c:4][c:3][c:2]2>>[*;!#1:7][c:1]1[c:2][c:3][c:4][c:5][c:6]1
	Remove phenyl		[*;!#1:1]c1[cH][cH][cH][cH][cH]1>>[*;!#1:1]
Terminal Group Exchange	Carboxyl to amide		[*;!#1:1][C:2](=O)[OH]>>[*;!#1:1][C:2](=O)N
	Amide to sulphonamide		C(=O)([NH2])[*;!#1:1]>>S(=O)(=O)([NH2])[*;!#1:1]

10.2.2 Generation of Compound Structures

Nova applies the transformations, encoded as SMIRKS, to a parent compound structure encoded as a SMILES string. While doing so, Nova allows a fragment of the parent to be specified as a SMARTS pattern, such that this fragment will not be modified during the generation process and any transformations that would modify this region will be ignored.

You can specify the parent structure in the Nova wizard. The typical workflow is illustrated in Figure 10.1: You can specify a region of the compound that must not be modified; the transformations to be applied can be selected; the number of generations of transformations to be applied can be specified; and finally, because this process generates a number of compounds that grows exponentially with the number of generations, you can control this growth by specifying criteria based upon a property, a score, chemical diversity or randomness to select a subset of the compounds in each generation. The process for selecting compounds at the end of each generation has the same flexibility for finding a balance between properties and chemical diversity as the standard StarDrop compound selection algorithm (Section 3.5.2). In addition, you have the opportunity to apply filters at the end of the process to remove compounds that contain specific functional groups. Some default filters are available but you can also define your own.



(a) (b) (c)

Figure 10.1 Illustration of workflow to initiate the generation of new compound structures. (a) Specify the input structure. A region of the molecule can be chosen to be 'frozen' (shown in light blue), in which case no modifications will be made to this region. (b) The transformations to apply can be selected, either individually or as groups. The groups can be managed to create groups tailored to specific objectives or to add new transformations. (c) The number of generations can be specified and a criterion for selection can be defined to limit the growth of the number of compounds generated. The selection can be defined as a minimum threshold for a property or score or a maximum number or percentage of each generation that will be used as the basis for subsequent generations.

10.2.3 Visualisation

Due to the large number of compounds and volume of associated data that this process can generate, it is important to provide visual tools to guide the exploration of the rich data set generated. In addition to typical scatter plots and histograms it can be valuable to explore the parent-child relationships between generated molecules to identify transformations that have a large impact on predicted properties. An example of such a visualisation is shown in Figure 10.2.

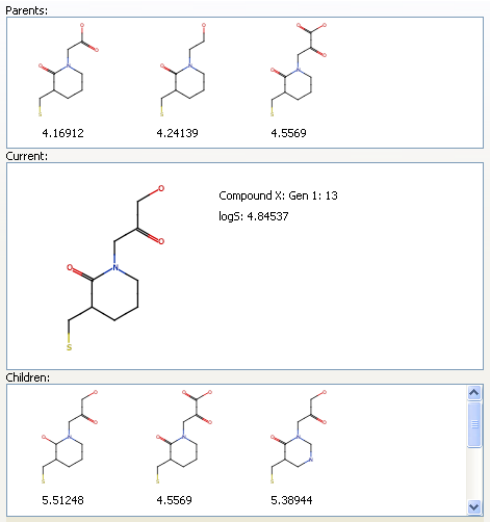


Figure 10.2 A view of the relationships between compounds in a dataset generated by the algorithm. The currently selected compound is shown in the middle, the parent compounds from which it was generated by different transformations are shown above and child compounds are shown below. The network of related compounds can be navigated by selecting compounds above or below the current compound. The value of a property, in this case logS, is shown with each compound allowing transformations that give rise to large changes in the property to be easily identified.

StarDrop's chemical space visualisation, as discussed in Chapter 3, can also be useful to visualise the diversity of the compounds generated and trends in properties and scores across this diversity.

10.2.4 Transform Set Validation

Coverage

In order to ensure that the set of transformations covers a wide range of 'drug-like' chemistry, enabling the exploration of a diverse range of potential modifications, each transformation should apply to a

wide range of molecules; a transformation that uniquely applies to a single molecule is not of interest. Furthermore, when the full set of transformations is applied to a 'typical' drug-like parent molecule, a large number of child molecules should be generated.

To test these requirements, the 206 transformations were applied to a set of 3,211 drug molecules (the "drug" set) derived as follows: Version 2.5 of the DrugBank Small Molecule database (Wishart, et al., 2008) was obtained on August 23, 2010. This initial set containing 4,854 molecules was reduced by removing molecules containing atoms other than C, H, N, O, P, S, Cl, or F, molecules with molecular weight less than 200 Da and 140 molecules which contained poorly specified SMILES (127 aromaticity errors and 13 valence errors), resulting in 3,214 compounds. Finally, 3 additional molecules (insulin, inulin and DB05413) were removed, as these are very large, not representative of the compounds to which we expect this method to be applied and likely to skew the validation statistics due to their size. 40 compounds were slightly edited to remove small cofactors or counter-ions or to select only one isomer where multiple isomers were specified.

The 206 transformations were applied to the drug set resulting in 584,124 child compounds; thus, on average, 182 child compounds were generated from each parent. Furthermore, on average, each transformation applied at least once to 31% of the molecules in the drug set.

These statistics indicate that the set of transformations have broad applicability to drug-like compounds and will generate a wide range of child compounds.

Quality

As discussed above, the transformation rules should be sufficiently general. However, there is a trade-off in that a more general transform is more likely to apply in an occasionally inappropriate chemical context. This can generate undesirable or infeasible compound structures. The desirability of compound structures is, to some extent, subjective. Therefore, the quality of the compound structures generated was assessed by asking two independent medicinal chemists to examine a set of 1,500 compounds generated using the 206 transformations.

The quality assessment set was generated as follows: 400 compounds were randomly selected from the drug set described above. All of the 206 transformations were applied to the 400 selected molecules to generate a set of child compounds. From the full set of child compounds, 1,500 were selected at random for assessment by the medicinal chemists.

The medicinal chemists were asked to assess each child compound to determine whether it was undesirable. They were not asked to determine if they could identify a synthetic route to the product – an ideal compound that was synthetically challenging may be worth the effort of devising a difficult synthetic route or may spark further ideas that are more accessible.

From the same set of 1,500 child compounds, one chemist flagged 7% of the structures as undesirable while the other flagged 4.1%. This demonstrates that desirability is, to some extent, subjective. However, an average acceptance rate of 94% was considered to be more than sufficient. It would be possible to filter out some of the undesirable structures before they are output. However, it was decided to retain this small proportion of poor compound structures as though they may be a minor distraction, they may stimulate ideas for similar compounds that are chemically feasible.

Hit-like to Drug-like Transformation Series

The transformations in the set should be representative of those used in practice to optimise leads into drug molecules. To assess this, a data set containing 60 marketed drugs and the initial leads from which they were derived, published by Perola (Perola, 2010), was used (we will refer to these lead/drug pairs as the "Perola" set).

For each lead/drug pair in the Perola set, the lead was used as the initial parent and the 206 transformations were applied iteratively to explore the 'universe' of compounds that are accessible from the lead. The goal of this was to identify the closest compound structure in this universe to the corresponding drug. This is challenging, as many of the derivations of drugs in the Perola set from their corresponding leads include the exchange or incorporation of large or relatively uncommon fragments. A result of the coverage requirements described above is that most of the transforms involve smaller fragments. Therefore, many iterative applications of the transformations may be required, creating

many generations of child compounds, to move from a lead to a compound similar to the corresponding drug and, even then, it may not be possible to find an exact match to the drug.

As the number of compounds generated increases exponentially with the number of generations, it is impractical to exhaustively enumerate all offspring compound structures. For example, if 182 compounds are generated on average from a single parent, the third generation will contain more than 6 million compounds. Therefore, a 'beam' search was implemented, whereby the 100 compounds with the greatest similarity to the target drug were retained after each iteration and a total of five iterations were applied. The closest match to the corresponding drug was identified from the resulting child compounds. The disadvantage of this approach is that it does not guarantee to find the closest match that could be achieved, as it may be necessary to initially move away from the drug in order to ultimately generate the most similar compound. Furthermore, it may be possible to find a closer child compound if more than five iterations were applied.

Similarity was measured using the Tanimoto index calculated between topological path-based fingerprints, with a maximum path length of 7 and a fingerprint size of 2048 bits. This was performed using the RDKit toolkit (RDKit: Cheminformatics and Machine Learning Software, n.d.).

Out of the 60 Perola lead/drug pairs, 7 exact matches were achieved within the compounds generated from the initial lead. On average, the similarity of the drug with closest match in the child compounds generated from the corresponding lead was 0.85 compared with an average similarity between the drugs and leads of 0.64. The structures of the initial leads, corresponding drugs and closest identified child compounds are provided in Appendix Sub-section 15.4. This demonstrates that the transformations are representative of those used to move from lead-like to drug-like compounds.

10.2.5 Application

There is a wide range of potential applications of this technology. These include: aiding the rigorous exploration of chemistry around early hits, to identify those hits most likely to yield high quality lead series; helping to find strategies to overcome problems with compound properties in lead optimisation; and identifying patent busting opportunities by expanding the chemistry around existing development candidates or drugs to search for compounds with improved properties. An example of the application of this method, coupled with predictive models and probabilistic scoring is provided in Section 14.6.

Finally, while we have focussed on the creation and validation of an initial set of transformations this set can be extended with new transformations based on the experience of medicinal chemists or designed around specific chemistry available within an organisation. Furthermore, transformations can be organised into groups, perhaps tailored to specific objectives such as improving metabolic stability or reducing plasma protein binding. Thus, this approach can be used as a tool to capture and share knowledge between medicinal chemists or even as an educational resource for less experienced scientists.

10.3 Matched Series Analysis

Matched molecular pair analysis (MMPA) (Warner, Griffen, & St-Gallay, 2010) (Dossetter, Griffen, & Leach, 2013) is a well-established method to investigate structure-activity relationships in experimental data, as discussed in Section 5.2. A matched molecular pair is two compounds that differ only in one small substitution and MMPA seeks to find substitutions that give rise to a consistent and significant change in a property across a data set of interest. Such a substitution, or transformation, corresponds to interesting SAR that may provide insights into strategies for further optimisation.

However, the application of MMPA is limited as a method when predicting the impact of a substitution on target activity because the impact of a transformation is often strongly context dependent, for example limited to a specific scaffold or target. Hajduk and Sauer (Hajduk & Sauer, 2006) found that potency changes associated with the majority of matched pair transformations, across 84,000 compounds with potency data against 30 protein targets, were approximately normally distributed with an average of zero. An example of this is shown in Figure 10.3(a) for the replacement of an ethyl group with butyl across all corresponding pairs in the ChEMBL database (ChEMBL, n.d.). This means that if we ask the question, "Would it be a good idea to replace an ethyl group with butyl in our series in order to improve the potency of our compounds", these matched pair data do not provide an answer. Greater

success may be achieved with MMPA for physicochemical properties or where the context is better defined, such as restricted to a single target or for closely related scaffolds where the substituents are likely to bind in a similar location within a binding site.

However, if we have more information, in the form of additional members of a matched *series*, this can yield a statistically significant prediction of the impact of a new substitution, as illustrated in Figure 10.3(b). Here we can see that the distribution of changes in potency for the replacement of an ethyl group with butyl is positively biased when we know that the propyl derivative is more active than ethyl. This indicates that, in this scenario, there is a high likelihood that substitution of butyl will increase potency.

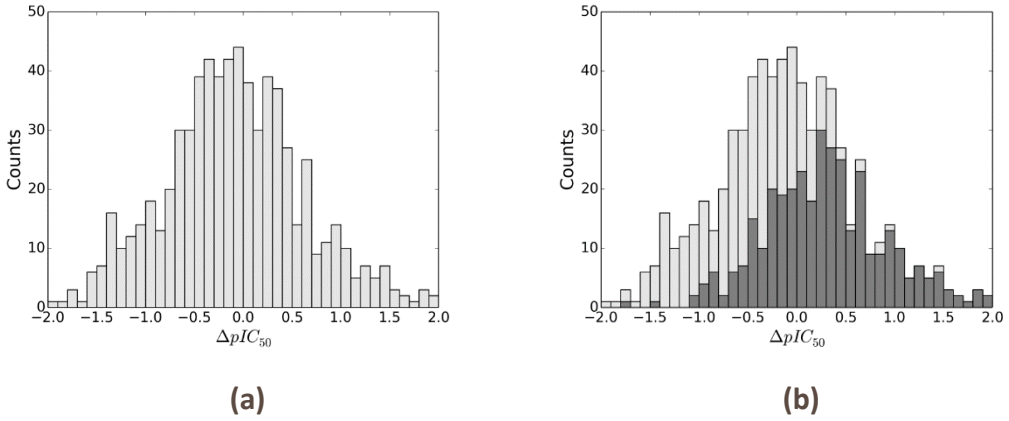


Figure 10.3 Distributions of the change in pIC_{50} for replacement of an ethyl group with butyl in the ChEMBL database (version 19). (a) shows the distribution for all examples of this transformation. In (b) the cases where the propyl derivative is more active than the ethyl are highlighted.

Matched molecular series, first proposed by Wawer and Bajorath (Wawer & Bajorath, 2011), are a generalisation of matched molecular pairs. Where a matched pair consists of two compounds that differ only by a single small substitution, a matched series of length *N* contains *N* molecules that are identical except for different substituents at the same position. An illustrative example of a matched series of length 3 is shown in Figure 10.4, corresponding to Cl, F and NH_2 substituents at the para position of the phenyl. In common with the paper by O'Boyle *et al.* (O'Boyle, Bostrom, Sayle, & Gill, 2014) we will use the notation [Cl, F, NH_2] to denote such a series, where there is no implied ordering of the compounds with corresponding substitutions. However, when activity values have been measured for the compounds in a matched series, these define an ordering; in the example shown in Figure 10.4 this would be denoted [Cl>F>NH_2].

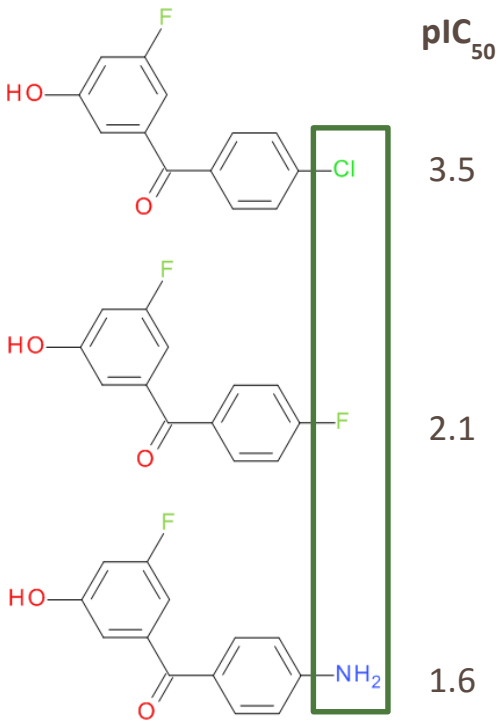


Figure 10.4. Example of a matched series of length 3, denoted [Cl>F>NH_2] according to the order of the measured activities of the corresponding compounds. The three compounds are identical with the exception of the substituent at the para position of the phenyl.

Comparing matched series found in your data with a database of other matched series (a knowledge base), enables more relevant predictions for new substituents that are likely to improve target activity or another property of interest. The suggestions are based on the premise that a matched series with similar activity order in your data and the knowledge base implies that those groups occupy a similar binding environment created by their target proteins. Given a similar binding environment, groups that have been shown to be better binders within the knowledge base, have a strong likelihood of being better binders to the target of the input data set. In this Section we will describe two methods that are implemented in Nova, the Matsy™ algorithm (O'Boyle, Bostrom, Sayle, & Gill, 2014) and SAR transfer (Wasserman & Bajorath, 2011) (Gupta-Ostermann, Wawer, Wassermann, & Bajorath, 2012) (Zhang, Wasserman, Vogt, & Bajorath, 2012).

10.3.1 Matsy™

The Matsy algorithm (O'Boyle, Bostrom, Sayle, & Gill, 2014) works by comparing short matched series (typically *N≥3*) from the input data (a query series) with matched series in a knowledge base. If the corresponding series in the knowledge base has additional members with greater activity than the substituents in the query series (an extended series), these indicate suggestions for new substituents that may increase the activity over the compounds in the query series.

The relevance of a suggestion is indicated by three statistics:

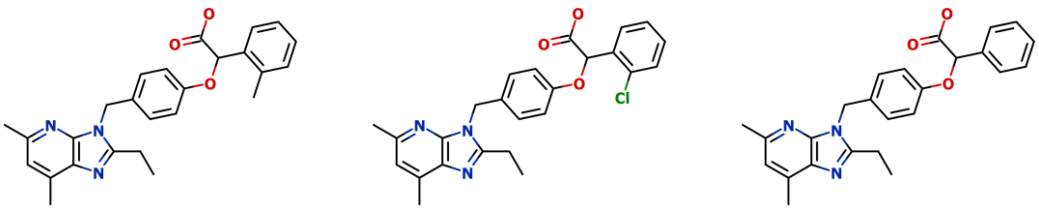
- Number of occurrences: The number times with which the extended series occurs in the knowledge base, with the same order as the query series. For example, for a query series [A > B > C] that identifies a suggestion R, the number of occurrences is the number of series [R, A, B, C] such that [A > B > C], i.e. sum of the numbers of matched series [R > A > B > C], [A > R > B > C], [A > B > R > C] and [A > B > C > R], in the knowledge base.
- % that improve: The percentage of observations in the knowledge base for which the suggested substituent increases the activity over all of the other members of the series. In the

example above, this would be given by the number of matched series [R > A > B > C] in the knowledge base divided by the number of occurrences and expressed as a percentage.

- Enrichment: This is a measure of how preferred the ordering of the extended, ordered series is relative to what one would expect if the order of the series was random. In the example above, this would correspond to the number of times the matched series [R > A > B > C] is found in the knowledge base divided by the total number of times the series [R, A, B, C] is found in any order, relative to what we would expect if the series were ordered randomly. There are *N!* ways in which a series of length *N* may be ordered, so the expected fraction is *1/N!*, in this case 1/24.

By default, the minimum number of occurrences that will be considered is 20 to ensure a statistically relevant sample, although this may be changed. The substituent with the highest percentage that improve is considered the most likely to improve the activity. O'Boyle *et al.* also noted that higher enrichments were typically found for longer series.

Figure 10.5 shows the results generated by a query series [2-methyl-phenyl > 2-chloro-phenyl > phenyl] corresponding to the input compounds shown. For each suggestion, the supporting evidence can be retrieved from the knowledge base, including the scaffold for each series, the individual activity measurements and the corresponding target, as illustrated in Figure 10.6. This enables the relevance of a suggestion to be quickly confirmed by comparison of the chemistry and target classes for which the extended series were found.



	Structure	RGroup	Scaffold	% that impro	Enrichment	Observations
1				45	1.749	40
2				31.25	1.048	32
3				30.95	1.068	42
4				29.17	1.012	24

Figure 10.5. Example suggestions for substituted phynyl replacements resulting from the query series shown at the top of the figure [2-methyl-phenyl > 2-chloro-phenyl > pheny]. The top 4 suggestions are shown

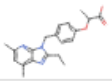
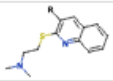
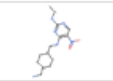
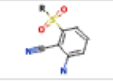
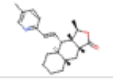
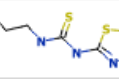
Scaffold	Target				
			8.222	8	7.538
	Serotonin 1 (5-HT1) receptor	6.2	5.9	5.8	5.7
	Protein kinase C theta	7.77	7.47	7.34	6.3
	Human immunodeficiency virus type 1 reverse transcriptase	5.85	5.35	5.23	5.16
	Proteinase-activated receptor 1	7.96	7.85	7.59	7.57
	Human immunodeficiency virus type 1 reverse transcriptase	7.4	7.1	6.22	6.05

Figure 10.6 The supporting series identified in the knowledge base for the top-ranked suggestion generated by query series shown in Figure 10.5. The top row corresponds to the query series, showing the scaffold and the input activity data; a blank is shown for the suggested compound. The remaining rows show the data for the series in the knowledge base in which the suggested substitution increases activity. In each case the scaffold and target for which the activity was measured is also shown. The top 5 of a total of 18 rows is shown.

10.3.2 SAR Transfer

In contrast to Matsy, SAR transfer uses longer query series to identify corresponding matched series in the knowledge base. By default, the minimum length of query series that will be considered is 8, although this may be modified.

For longer query series, the number of corresponding series in the knowledge base is likely to be small, therefore a different approach to identifying relevant series must be taken. In SAR transfer the correlation between the observed activities in the query series and the corresponding matched series in the knowledge base is used. If this correlation is high, it indicates that the SAR from the knowledge base is likely to be transferable to the query series. In this case, more active members of the series in the knowledge base are likely to improve the activity over those substituents in the query.

The correlation between the query series and those in the knowledge base is calculated using a Spearman's rank correlation coefficient. By default, the minimum correlation that will be reported is 0.7, but this may be modified. It is notable that as the minimum number of derivatives in the series is reduced any deviations from the order in the query series will have a larger influence on the correlation. For short series, if the order of only one or two derivatives do not match the query series then the correlation will fall below the minimum acceptable value, so these two parameters need to be considered at the same time.

An example of the output of SAR transfer for a single suggestion is shown in Figure 10.7. As for Matsy, the supporting evidence can be retrieved from the knowledge base, including the scaffold for each series, the individual activity measurements and the corresponding target. This enables the relevance of a suggestion to be quickly confirmed by comparison of the chemistry and target classes for which the extended series were found.

Scaffold	Target	Correlation											
				8.854	8.699	8.62	8.495	8.444	8.387	8.181	8.113	7.824	7.745	7.658
	P2X purinocepto...	0.9108	8.3	8.1	7.9	7.7	7.2	6.9	7.3	7.3	6.9	6.8	6.8	6.4

Figure 10.7 Example output for a single suggestion generated by SAR transfer. The top row corresponds to the query series, showing the scaffold and the input activity data; a blank is shown for the suggested compound. The next row shows the data for the series in the knowledge base in which the suggested substitution increases

activity. For this series, the scaffold and target for which the activity was measured and the correlation with the query series are shown.

10.3.3 Matched Series Knowledge Base

The knowledge base provided with StarDrop is generated using pIC_{50} values from the ChEMBL database (ChEMBL, n.d.). With each release of StarDrop the knowledge base will be updated with the latest version of ChEMBL.

The ChEMBL database covers a diverse range of chemistries and targets, derived primarily from the medicinal chemistry literature. The transferability of matched series, as discussed above, means that suggestions can be derived from a diverse range of targets. However, due to the nature of this data used in this knowledge base, it is most useful for generating suggestions to improve or reduce target binding.

The knowledge based provided with StarDrop can be extended or replaced with additional knowledge bases using data from in-house or other sources of data. These can also be generated for properties other than binding affinity to generate suggestions for optimisation of other properties, for example metabolic stability or physicochemical properties. The generation of knowledge bases lies beyond the scope of StarDrop and this capability is provided by our partners at NextMove Limited.

10.3.4 Conclusions

Matched series analysis provides two approaches, Matsy and SAR transfer, which provide empirical, data-drive suggestions for substitutions to improve target activity. Of course, target activity is not the only criterion by which a suggestion should be prioritised; physicochemical, ADME and safety properties should also be taken into consideration. For this reason, the metrics generated by matched series analysis can be included in StarDrop's Probabilistic Scoring method for multi-parameter optimisation, as discussed in Chapter 2. The resulting scores can be used to prioritise compounds that are likely to improve target activity and also retain or improve the other properties required in a successful compound for a project's objectives, as predicted by *in silico* models. All of these parameters can also be used in StarDrop's other visualisation and data analysis features to help to quickly identify the most relevant ideas to progress.

11 BIOSTER™

The BIOSTER database (Ujváry & Hayward, 2012) is a compilation of 29,012 pairs of compounds, published in the scientific literature, representing structural modifications made in the course of chemistry projects. Thus, BIOSTER is a chemically and synthetically validated database encompassing a wide range of chemical transformations including: bioisosteric replacements of functional groups, linker replacements, homologisation, introduction of conformational constraints and reversible derivatisations (e.g. pro-drugs).

11.1 The BIOSTER Database

The first version of BIOSTER was completed in 1992 and it has been updated regularly ever since. Sources for BIOSTER include review papers, authoritative textbooks such as Burger's Medicinal Chemistry, monographs such as Annual reports in Medicinal Chemistry and publications such as Journal of Medicinal Chemistry and Chemical and Pharmaceutical Bulletin. The contents of BIOSTER extend beyond medicinal chemistry and include publications from fields such as pest control (e.g. Chemie fer Pflanzenshtuz- und Schädlingsbekämpfungsmittel) and agrochemicals (Journal of Agrochemical and Food Chemistry). Tens of thousands of papers from approximately 100 scholarly journals and periodicals have been analysed for novel analogous structures.

Each BIOSTER record is in the form of a pseudo reaction, relating a pair of compounds, with a manually designated 'reaction centre', indicating the replacement. An example is shown in Figure 11.1. Associated with each transformation are a number of fields containing relevant information.

A unique ID code is assigned to each transformation, in which the first three letters specify the chemical type of the starting lead fragment. This three-letter code can be used for a quick search for replacements for a common coded functionality. A summary of the three-letter ID prefixes are shown in Table 11.

Citations from the literature are provided for each pair. The first contains the original reference in which the pair of compounds shown was described and subsequent lines may be provided, indicating more recent studies related to the application of the specified transformation in chronological order. In some cases lines, with a "see also" notation provide references to related modifications or the 'reverse' transformation.

Additional, searchable keywords are also provided, describing: the biological activity or therapeutic category mentioned in the original publication, the mode of action (if known) and additional information (where available) on changes in the properties or biological activity of the new analogue (e.g. conformationally constrained, nonmutagenic, peptide beta turn mimic, prodrug, water soluble…).

The specific and general names of the fragment replacement of the pair of compounds are also given. For simplicity and to enable easier searching, numberings, indicators of degree of saturation and, for polyheterocyclic systems, fusion descriptors are typically omitted.

11.2 Creating Bioisosteric Transformations

It is challenging to identify those bioisosteric replacements that may be applicable to a chemistry of interest and assess the potential results of making a similar modification. Therefore, in StarDrop, the BIOSTER module may be used in combination with Nova to automatically find and apply bioisosteric replacements and generate novel compound structures that are likely to preserve the required biological activities. As described in Chapter 10, the properties of the resulting compounds can be predicted and the most promising ideas prioritised for further consideration to identify those that are most likely to have a good balance of the properties required in a high quality drug.

Wagener and Lommerse (Wagener & Lommerse, 2006) attempted to extract bioisosteric substructures automatically from the BIOSTER database by fragmenting the molecules and removing identical fragments from both sides, but the different substituents on each side meant that they were only able to successfully do this for about 14% of the records in the database.

Table 11 List of the three-letter ID codes for the transformations in the BIOSTER database

Three-letter ID Code	Example fragments
ACE	acetal, furanose, glycoside, hemiacetal, ketal
AC]	Amino acid, anhydride, carboxylic acid
ACY	acylfluoride, acylguanidine, acylhydrazone, acylurea
ALD	aldehyde, aminoaldehyde, hydroxyaldehyde
AMI, AMN	amidine, amine, carboxamide, hydrazide, lactam, oxalamide, squaramide
ANI	anilide, aniline, diphenylamine
ARG	arginine, guanidine
AZ]	azide, aziridine
AZO	azobenzene, semicarbazone, triazene, triazole
BEN	benzene, benzhydryl, benzophenone, benzyl, phenyl
BOR	boran, borate, borole, boronic acid
BRO	Bromo
CAR	benzyloxycarbonyl, carbamate, carbazate, carbonate
CHA, CHN	alkane, alkene, aminoalkyl, aralkyl, chain, polyene, styryl
CYA	cyanamide, cyano, cyanoguanidine, isocyanate, nitrile
DOP	catechol, dihydroxyphenyl, diphenol
EPO	epoxide, oxirane
EST	ester, lactone, orthoester
ETH	alkoxy, benzyloxy, ether, morpholine, oxetane, phenoxy
HAL	bromo, chloro, fluoro, halogen, iodine, trifluoromethyl
HYD	alcohol, hydrazone, hydrogen, hydroxamic acid, hydroxy
IMI	benzimidazole, imidazole, imide, imine
IND	indan, indazole, indole, indolizine, isatin, oxindole, tryptophan
ISO	isoxazole, isothiourea
KET	alkenone, butyrophenone, ketone, ketoacid, ketoamide
MET	metal, methine, methoxy, methyl, methylene
NIT	nitrate, nitric oxide, nitrile, nitro, nitromethylene, nitroso
OXA	benzoxazine, benzoxazole, oxadiazole, oxazoline, sydnone
OXY	N-oxide, oxime ether, peroxide, trioxide
PEP, PET	acylproline, amide, amino acid, macrocyclic peptide, peptide
PHE	biphenyl, naphthyl, phenol, phenoxy, phenyl, phenylalanine, tyrosine
PHO	phosphate, phosphole, phosphonate, phosphoramidate, phosphoryl, pyrophosphate, triphosphate
PLA	Platinum
PUR	guanine, nucleoside, purine, xanthine
PYR	pyranone, pyrazole, pyridazine, pyridine, pyrimidine, pyrrole, pyrrolidine
QUI	anthracenedione, quinoline, quinone, quinone methide, quinoxaline
RIG, RIN, RIQ, RIV	ring replacements (steroids, heterocycles, macrocycles, etc.)
SEL	diselenide
SUL	disulphide, sulfamate, sulphonamide, sulfone, sulfonylurea, sofloxide, sultam
TER	isoprene, terpene
TET	Tertrahydrofuran, tetrazole
THI	isothiourea, sulphide, thiazole, thioether, thiol, thiocarbamate, thiophene, thiophenol, thiourea
URE	acylurea, urea

In order to make BIOSTER available within StarDrop's Nova framework, a transformation was generated for each record, representing the atoms in the bioisosteric substructures in Daylight's SMIRKS notation (Daylight, n.d.) using the process outlined in Figure 52 implemented in using a customised version of Digital Chemistry's MOLSMART program (Digital Chemistry, n.d.). Of the 29,012 records in the latest version of the BIOSTER database, it is currently possible to generate SMIRKS for 23,917 (82.4%).

The substituent groups around these substructures are not necessarily identical in both molecules.

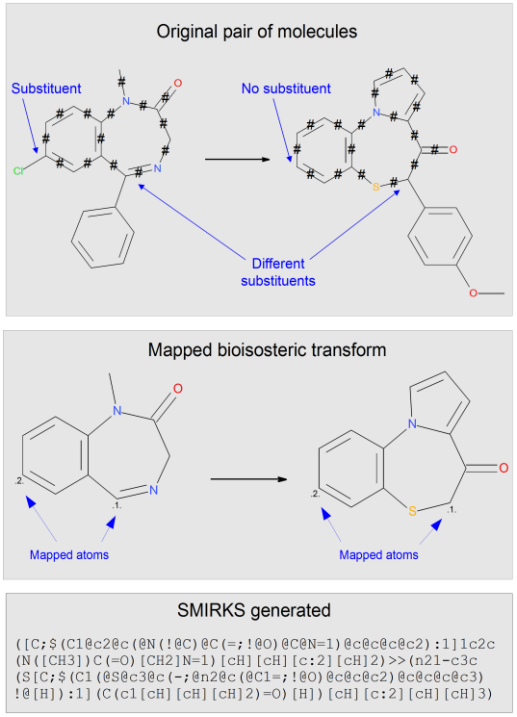


Figure 11.1 Illustration of the process of generating a transformation from a pair of bioisosteric compounds. At the top, the initial pair of compounds is shown, with the bioisosteric replacement highlighted as hashed bonds. The points of substitution on these compounds are used to define the mapped atoms in the transformation, resulting in the SMIRKS string shown at the bottom.

Therefore, equivalent substitution positions on the 'reactant' and 'product' sides were determined heuristically on the basis of:

- Chemical similarity of the substituent groups or substituted element types
- Spatial orientation of the substituent groups in the original BIOSTER diagrams
- Avoidance of valency violations

In bioisosteric replacement, the same atoms do not generally appear on both sides of the transformation. Therefore, in the SMIRKS, only the substitutable atoms are 'mapped'.

To minimise the number of inappropriate or 'promiscuous' transforms generated:

- Substitution is permitted only where there is a substituent on at least one side of the 'reaction'
- Atoms are designated aliphatic or aromatic based on the original molecules
- Bonds are designated as 'ring' or 'chain' based on the original molecules

11.3 Predictive Application of Bioisosteric Transformations

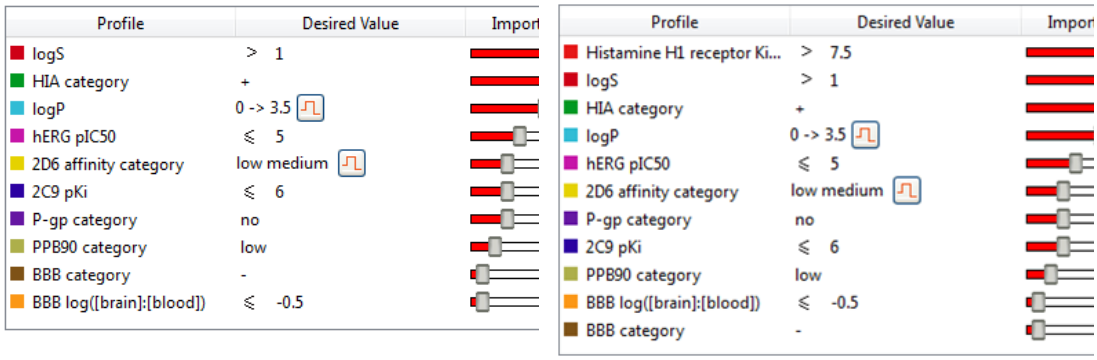
In the following two retrospective examples, the bioisosteric transformations were applied using the Nova module and prioritised using *in silico* models in the ADME QSAR module and probabilistic scoring (See Chapter 2).

11.3.1 Lead Optimisation: Dipeptidyl Peptidase IV Inhibitor

The BIOSTER transformations were applied to the lead compound from the project that resulted in the discovery of the anti-diabetic Dipeptidyl Peptidase IV (DPP IV) inhibitor Alogliptin (Feng, et al., 2007). This resulted in the generation of 230 compounds that were prioritised against the scoring profile shown in Figure 11.2(a) and some illustrative results are shown in Figure 11.3. It is notable that the product shown in the centre of Figure 11.3 is a close analogue of Alogliptin (also shown in Figure 11.3 for comparison).

11.3.2 Fast Follower: Histamine H1 Receptor Antagonist

Application of the BIOSTER transformations to the antihistamine drug Azatadine yielded a total of 89 compounds that were prioritized against the profile shown in Figure 11.2(b), including pK_i against the Histamine H1 receptor, predicted using a QSAR model. Some illustrative results are shown in Figure 11.4 and it is notable that the product on the right above represents the core replacement that led to the candidate compound Hivenyl (Janssens, et al., 2005) (also shown in Figure 11.4 for comparison).



(a) (b)

Figure 11.2 Scoring profiles used in the example BIOSTER application: (a) defines appropriate ADMET properties for an orally dosed compound for a central nervous system (CNS) target; (b) defines an appropriate balance of properties for a potent inhibitor intended for a peripheral target, in this case the Histamine H1 receptor.

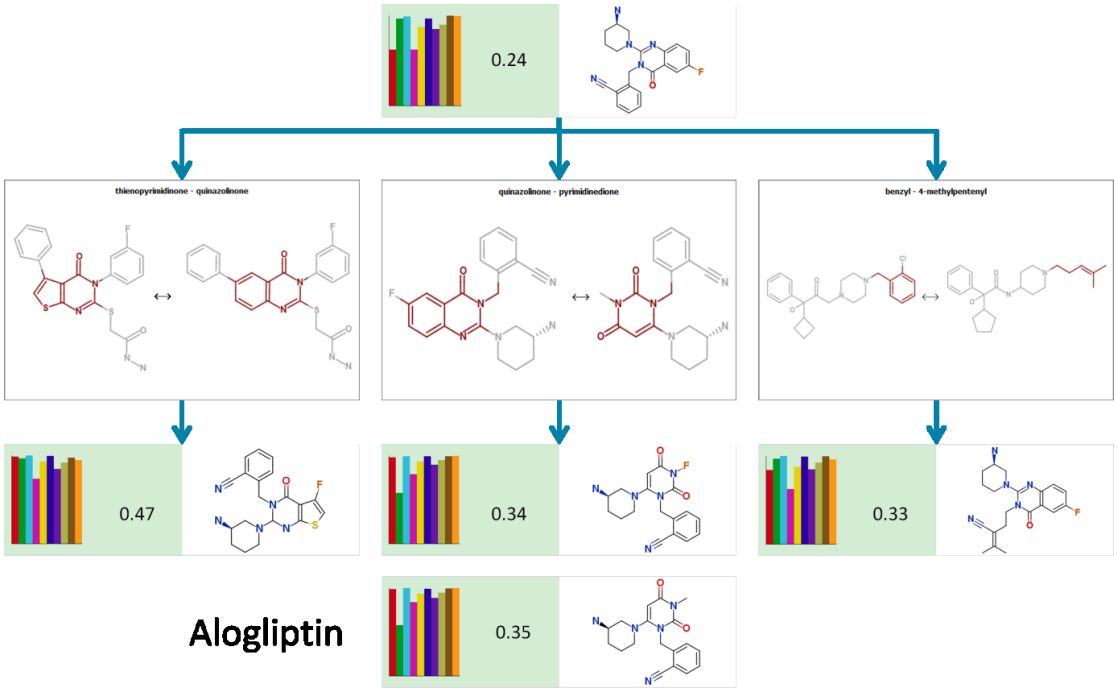


Figure 11.3 Illustrative examples of the application of the BIOSTER transformations to a lead compound from which the DPP IV inhibitor Alogliptin was discovered. The scores for each compound were generated using the scoring profile shown in Figure 3(a); the colours in the histograms correspond to the key shown in this figure and show the impact of each property on the overall score. The structure of Alogliptin is shown for comparison.

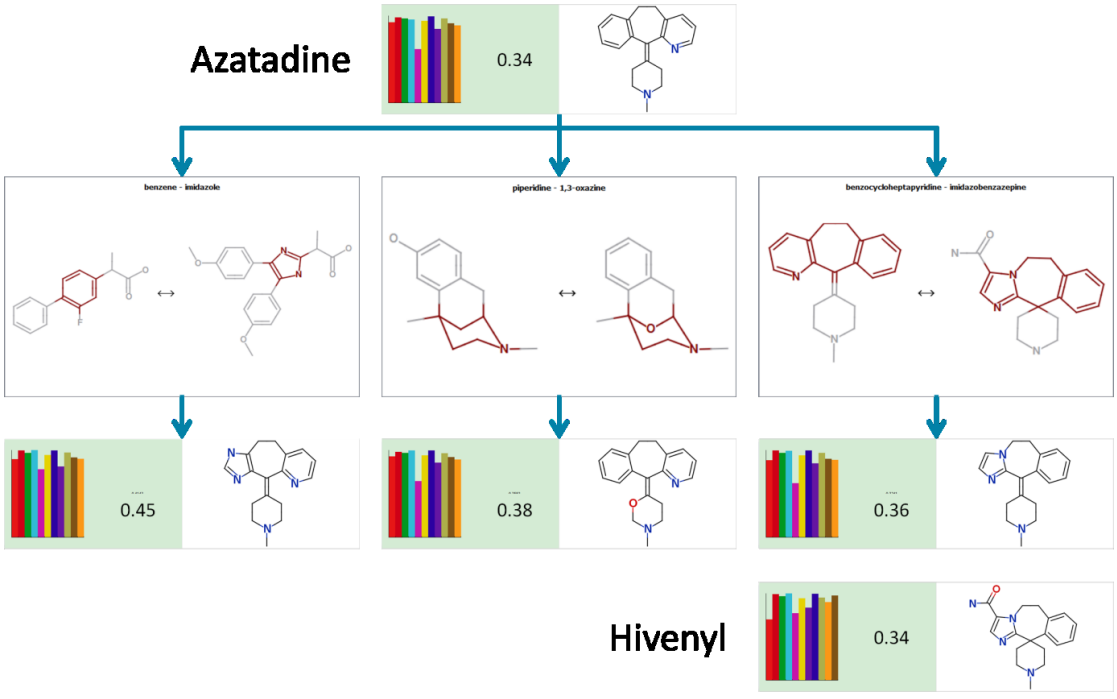


Figure 11.4 Illustrative examples of the application of the BIOSTER transformations to the drug Azatadine (top). The scores for each compound were generated using the scoring profile shown in Figure 3(b); the colours in the histograms correspond to the key shown in this figure and show the impact of each property on the overall score. The structure of Hivenyl is shown for comparison.

11.4 Conclusions

Bioisosteric transformations are an excellent source of new ideas for compound design, providing access to increased chemical diversity whilst maintaining a high likelihood of biological activity. Automatically applying bioisosteric transformations from a large database of precedented replacements enables efficient exploration of new chemical space in the search for new optimisation strategies. This may result in a large number of new ideas, which can be prioritised to highlight those most likely to succeed against a project's objectives. Furthermore, links to the primary literature, from which the transformations were derived, make it easy to follow-up the most interesting ideas to find synthetic routes and investigate the underlying biological data.

This approach can be applied throughout the drug discovery process, including expansion around initial hits, exploring scaffold hopping opportunities in lead optimisation and patent protection.

12 torch3D™

12.1 Introduction

torch3D is a molecular design and SAR interpretation tool, developed by Cresset (Cresset, n.d.), which uses molecular alignment to a reference molecule in a predefined conformation as a way to make meaningful comparisons across chemical series. When used on a congeneric series the tool can help in library design and give a rationale for the prioritisation of compounds for synthesis. Using torch3D on a diverse set of active molecules can help define the requirements of the protein of interest, aiding the synthetic chemist in the design of new actives.

torch3D aligns molecules based on their molecular fields, not on their structure. The interaction between a ligand and a protein involves electrostatic fields and surface properties (e.g. hydrogen bonding, hydrophobic surfaces and so on). Two molecules which both bind to a common active site tend to make similar interactions with the protein and hence have highly similar field properties. Accordingly, aligning and scoring molecules based on the similarity in these properties is a powerful tool for the medicinal chemist as it concentrates on the aspects of the molecules that are important for biological activity. The alignments provided give ideas on how molecules with different structures could interact with the same protein, and the scores for those alignments provide insights into SAR and ideas for further synthesis.

torch3D is a valuable tool to align members of a congeneric series of compounds prior to QSAR analysis. It can resolve queries related to selecting the best orientation of rotatable groups for example, so as to give a consistent alignment based on the best matching of molecular fields, without the need to define (possibly arbitrary) series-specific alignment rules.

torch3D can also be used to align structurally diverse compounds. This can be useful when comparing the SAR of two known active series and looking for comparable substitution sites. torch3D also serves as a useful tool for compound design. For example, you can use it to design analogues of a known active compound and see how the modifications affect the field pattern, giving insight into how activity can be interpreted in terms of field pattern. A further application is for library design. Small virtual libraries can be compared to a known active molecule to help prioritise scaffold and reagent selection.

torch3D takes a single or small set of molecules in a predefined conformation to use as a “reference”. It then aligns a series of database molecules to the reference based on molecular fields. The process that is followed for each database molecule is shown in Figure 12.1.

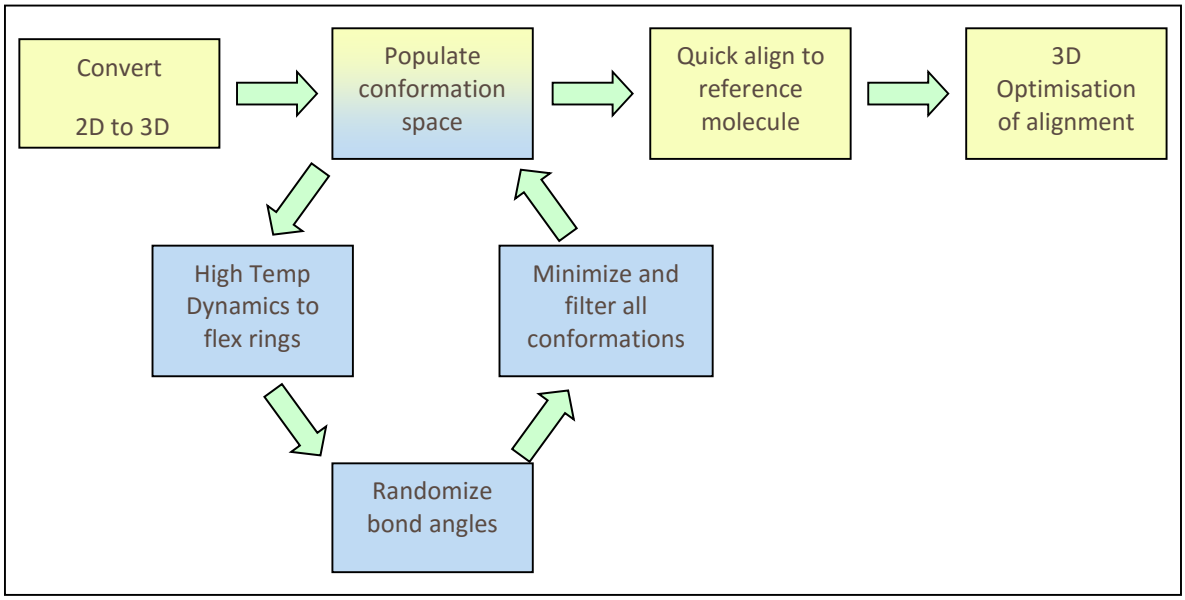


Figure 12.1 The process followed for each database molecule in torch3D

12.2 What are Field Points?

For computational efficiency, Cresset's field technology condenses the molecular fields down to a set of points around the molecule, termed "field points". Field points are the local extrema of the electrostatic, van der Waals and hydrophobic potentials of the molecule. They can be thought of as extended pharmacophores, with the advantages that their position is directly calculated from the molecule's physical properties, and they have size/strength information associated with them (so that e.g. not all H-bond donors are treated the same: some make stronger bonds than others). The generation of field points is described in detail (Cheeseright T, 2006). The four field types are used in unison to describe all the potential interactions that a ligand in a specified conformation can make to a protein.

12.3 Interpretation of Field Point Patterns

A representative field point pattern is shown in Figure 12.2. Larger field points represent stronger points of potential interaction. Throughout Cresset's software the field points are coloured as follows:

- Blue: Negative field points (like to interact with positives/H-bond donors on a protein)
- Red: Positive field points (like to interact with negatives/H-bond acceptors on a protein)
- Yellow: van der Waals surface field points (describing possible surface/vdW interactions)
- Gold/Orange: Hydrophobic field points (describe regions with high polarisability/hydrophobicity)

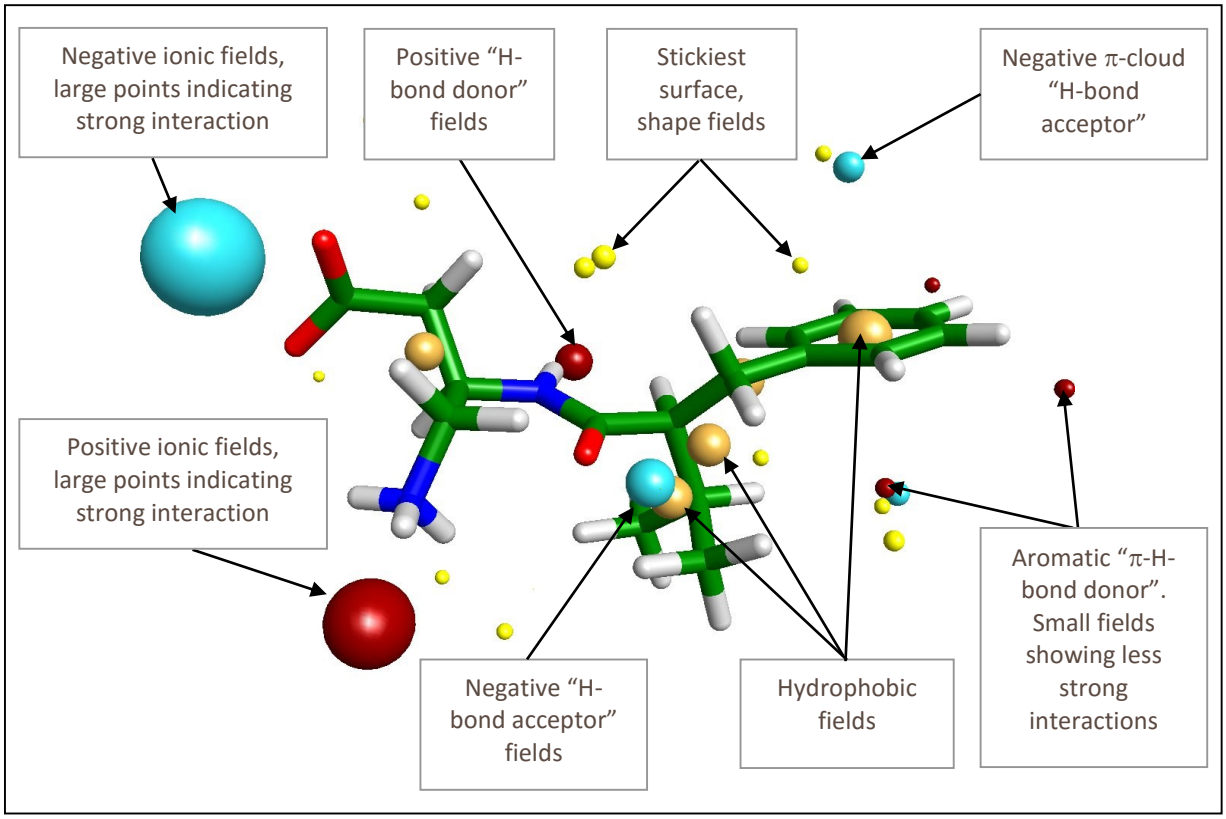


Figure 12.2 Interpretation of a field point pattern. The size of the point indicates the potential strength of the interaction

ionic groups give rise to the strongest electrostatic fields. Hydrogen bonding groups also give strong electrostatic fields. Aromatic groups encode both electrostatic and hydrophobic fields. Aliphatic groups such as the *iso*-propyl group give rise to hydrophobic and surface points but are essentially electrostatically neutral.

12.4 Reference Molecules

Suitable reference molecules are highly active molecules, preferably in the bioactive (protein bound) conformation for the protein of interest. This bioactive conformation could come from a protein-ligand x-ray crystal structure or from a dock of the ligand into the protein. In the absence of protein data the information could come from Cresset's FieldTemplater program or from a pharmacophore model. Lastly a reasonable guess of the conformation can work well in cases where the structural diversity of the ligands is low. If you choose to use a 2D molecule then torch3D will convert this into a 3D conformation before proceeding.

The reference molecule should be in a **defined 3D conformation**. The file can be in sdf format (which is the same as MDL mol format), mol2 format or Cresset's own XED format. If you use a 2D molecule then it will be converted to 3D before proceeding.

12.5 Conformer Generation

As part of the process of aligning your molecules with the reference molecule, torch3D generates a series of conformations (the exact number depending on your choice of running fast or slow calculations). The 3D conformation ensembles are generated by stochastic sampling of torsions followed by energy minimization using the XED force field; only conformations within a 6.0 kcal/mol threshold from the global minimum are retained. To implicitly model solvent, attractive van der Waals and electrostatic interactions are turned off during the geometry optimization stage, in order to avoid a prevalence of folded conformations due to hydrophobic collapse and electrostatic attraction between moieties with opposite charges, if present. Ring conformations are normally taken from a ring library; for ring systems which are not represented in the ring library, ring conformations are sampled through high temperature molecular dynamics followed by optimization at 298 K. Finally, only conformations which have a heavy atom RMSD higher than 0.5 Å are considered different and retained.

12.6 torch3D Scores

The score is an important factor in deciding the validity and potential activity of particular alignments and molecules. However, it is not the only factor to be considered before embarking on the synthesis of a compound designed using torch3D. The top-scoring result is the one that is the most similar to the target molecule in terms of fields and shape. That doesn't necessarily mean that it is the most likely to be active, and certainly doesn't mean that it's the one you want to make first.

The absolute value of the scores isn't that informative in isolation, largely because the scores provided are the similarity of the result molecule to the target molecule. If you are replacing only a small part of a large molecule, then the large number of atoms in common between the target and the results will mean that the similarity values may all fall in a range of 0.8-0.99. In other words, the scores are useful for ranking the results (higher-scoring result molecules are more similar to the target than lower-scoring ones), but don't pay too much attention to the absolute numbers and don't compare the numbers between different target molecules.

Sometimes it may seem that the field points of 2 molecules in a particular alignment don't match up. This is probably because the scoring algorithm uses field points as sampling points of the true field around a molecule. To score two molecules the field of B is sampled at the locations of the field points of A and *vice versa*. Thus the field points for the result and the target molecules may not be exactly coincident, but if the true fields show similar properties at the field point locations, the field similarity and hence the score will be high. Viewing the field surfaces for the target and result molecules can be instructive.

13 Derek Nexus™

Derek Nexus (Sanderson & Earnshaw, 1991) (Ridings, et al., 1996) (Greene, Judson, Langowski, & Marchant, 1999) is a knowledge-based toxicity prediction tool developed by Lhasa Limited (Lhasa, n.d.). Data from both published and donated (unpublished) sources are used to develop the knowledge base on which Derek Nexus predictions are based. This ensures that the accuracy of predictions is reflective of the current knowledge of structure-toxicity relationships, offering expert decision support to scientists in a variety of industries including the pharmaceutical, cosmetic and chemical industries. Using structure-activity relationships created by Lhasa Limited scientists, Derek Nexus provides early indications of the potential toxicities of your compounds in over 40 endpoints, including mutagenicity, hepatotoxicity and cardiotoxicity.

Each prediction takes the form of a structural alert, identifying the structural feature giving rise to the predicted risk of toxicity. Furthermore, Lhasa Limited's reasoning-based system provides an estimate of the level of likelihood associated with the alert, based on precedence from experimental data (Judson, Stalford, & Vessey, 2013). The potential results for a Derek Nexus prediction are as follows:

- 'No report' indicates that the query compound does not contain any structural alerts associated with that endpoint and there are no reasons based on the physical properties of the compound to predict either activity or inactivity
- 'Equivocal' is defined as meaning that there is an equal weight of evidence for and against a proposition
- 'Plausible' indicates that the weight of evidence supports the proposition
- 'Probable' means that there is at least one strong argument that the proposition is true and there are no arguments against it
- 'Inactive' indicates that the compound will not be mutagenic

For full details of the underlying methods and validation, please see the references above.

13.1 Derek Endpoint Descriptions

The following sections provide details on the derivation of the predictions for the major endpoints predicted by the Derek Nexus module in StarDrop.

13.1.1 Chromosome Damage

Derek Nexus v4.0 contains 98 alerts for chromosome damage, describing numerical (e.g. aneugenic) and structural (i.e. clastogenic) chromosomal aberrations *in vitro* and *in vivo*. Predictions are based on expert-derived structural alerts for chromosome damage (2D SARs), that take into account toxicological and mechanistic evidence, and where appropriate metabolism and physicochemical properties of compounds. External validations were carried out using public and proprietary data sets derived from *in vitro* and *in vivo* chromosome damage assays covering 3,361 and 1,802 unique compounds correspondingly.

Primary data used for alert development include:

- *in vitro* and *in vivo* chromosome aberration test.
- *in vitro* and *in vivo* micronucleus test.
- *in vitro* L5178Y TK+/- assay.

13.1.2 Mutagenicity

Derek Nexus v4.0 contains 111 alerts for bacterial mutagenicity. Predictions are based on expert-derived structural alerts for mutagenicity (2D SARs), that take into account toxicological and mechanistic evidence, and where appropriate metabolism and physicochemical properties of compounds. Following alert evaluation, Derek evaluates whether non-alerting query compounds contain any features that are either (i) also present in non-alerting mutagens in a large Ames test reference set (misclassified features) or (ii) not present in a large Ames test reference set (unclassified features). External validations were carried out using public data sets derived from Ames test assays covering 9,456 unique compounds.

Primary data used for alert development include:

- Ames test data in both Salmonella typhimurium and Escherichia coli.

Supporting data:

- *in vivo* lacZ-transgenic assay.
- *in vitro* L5178Y TK+/- assay.
- *in vitro* HGPRT gene mutation assay.
- *in vitro* Na+/K+ ATPase gene mutation assay.

13.1.3 Carcinogenicity

Derek Nexus v4.0 contains 77 alerts for carcinogenicity (both genotoxic and non-genotoxic). Predictions are based on expert-derived structural alerts for carcinogenicity (2D SARs), that take into account toxicological and mechanistic evidence, and where appropriate metabolism and physicochemical properties of compounds. External validations were carried out using public data sets derived from chronic carcinogenicity study data in rodents covering 2,181 unique compounds.

Primary data used for alert development include:

- chronic carcinogenicity assay data from studies conducted in rat and/or mouse.
- human data (cohort studies).

Secondary data sources:

- IARC classifications.

13.1.4 Skin Sensitisation

Derek Nexus v4.0 contains 73 alerts for skin sensitisation. Predictions are based on expert-derived structural alerts for skin sensitisation (2D SARs), that take into account toxicological and mechanistic evidence, and where appropriate metabolism and physicochemical properties of compounds. External validations were carried out using public data sets derived from LLNA and guinea pig assays covering 504 unique compounds.

Primary data used for alerts development include:

- guinea pig data, such as the Buehler and maximisation tests.
- human data from maximisation and patch tests.
- mouse data, mostly from the local lymph node assay.

Secondary data sources:

- BgVV categories.
- R43 classifications.

13.1.5 Hepatotoxicity

Derek Nexus v4.0 contains 84 alerts for hepatotoxicity. Predictions are based on expert-derived structural alerts for liver damage (2D SARs), that take into account toxicological and mechanistic evidence, and where appropriate metabolism and physicochemical properties of compounds.

Primary data used for alert development include:

- repeat dose toxicity studies in animals.
- clinical case reports on liver toxicity.
- hepatotoxicity epidemiological studies.

13.1.6 Teratogenicity

Derek Nexus v4.0 contains 63 alerts for teratogenicity. Predictions are based on expert-derived structural alerts for teratogenicity (2D SARs), that take into account toxicological and mechanistic evidence, and where appropriate metabolism and physicochemical properties of compounds.

Primary data used for alerts development include:

- teratogenicity studies in animals.
- *in vitro* data (embryo culture assays).
- human case reports on teratogenicity including FDA pregnancy categories.

Supporting data:

- *in vitro* assay data for specific enzymes known to disrupt pathways important for teratogenic effect of chemicals.
- *in vitro* data for known receptor mediated toxicity (endocrine disruption).

13.1.7 Irritation (of the skin)

Derek Nexus v4.0 contains 25 alerts for irritation of the skin, which includes both irritation and corrosion endpoints. Predictions are based on expert-derived structural alerts for skin irritants (2D SARs), that take into account toxicological and mechanistic evidence, and where appropriate metabolism and physicochemical properties of compounds.

Primary data used for alert development include:

- skin irritation studies in the rabbit.

Secondary data sources:

- R34, R35 and R38 classifications.

13.1.8 Irritation (of the eye)

Derek Nexus v4.0 contains 30 alerts for irritation of the eye. Predictions are based on expert-derived structural alerts for eye irritants (2D SARs), that take into account toxicological and mechanistic evidence, and where appropriate metabolism and physicochemical properties of compounds.

Primary data used for alert development include:

- eye irritation studies in the rabbit.

Secondary data sources:

- R34 and R36 classifications.

14 Example Applications

14.1 Example 1: Profiling Large Virtual Libraries to Identify Potential Liabilities within Chemical Series

14.1.1 Objective

The StarDrop *in silico* models may be used to identify potential ADME liabilities across large virtual libraries representing multiple chemotypes. This can be used to identify consistent liabilities within particular chemotype targeting synthesis on chemistries most likely to yield successful compounds. Downstream resources can also be focused to address potential ADME issues with a selected chemistry early in a project.

Product Profile: A chronic oral dose therapy against a CNS target.

History: Hits had been identified against the target and the chemistry expanded around these hits based on several core scaffolds.

14.1.2 Process

StarDrop's high throughput ADME models enable rapid profiling of virtual libraries containing up to millions of compounds. Consequently, only the breadth of chemistry ideas that can be conceived limits the size and chemical diversity of libraries that can be profiled.

To enable interpretation and visualisation of such a large data set, a simple analysis can be performed to show how many compounds (as a percentage) pass or fail the criteria for the predicted properties. Although, as discussed previously, simple filtering is not appropriate when prioritizing individual compounds for progression, this analysis enables chemotypes with consistently poor property values to be identified.

Figure 14.1 shows the profile of ADME properties from such an analysis across 5 ADME models, carried out on an array (structures centred on the same core scaffold, i.e. a single chemotype) of approximately 35,000 molecules, based on the following criteria:

Table 12 Example 1 – Selection criteria.

Property	Desired Value
CYP2C9 affinity (pK_i)	<6.0
CYP2D6 affinity	Low, Medium
HIA	+
BBB penetration category	+
Solubility (logS)	>1.0

Clearly, this particular array has potential problems in relation to the intended therapeutic target; with the majority of molecules having low aqueous solubility, which could potentially limit oral bioavailability, poor blood-brain barrier penetration, potentially reducing efficacy, and a high risk of interactions with CYP2D6, which is common for many CNS drugs, particularly those interacting with dopamine or 5-HT receptors.

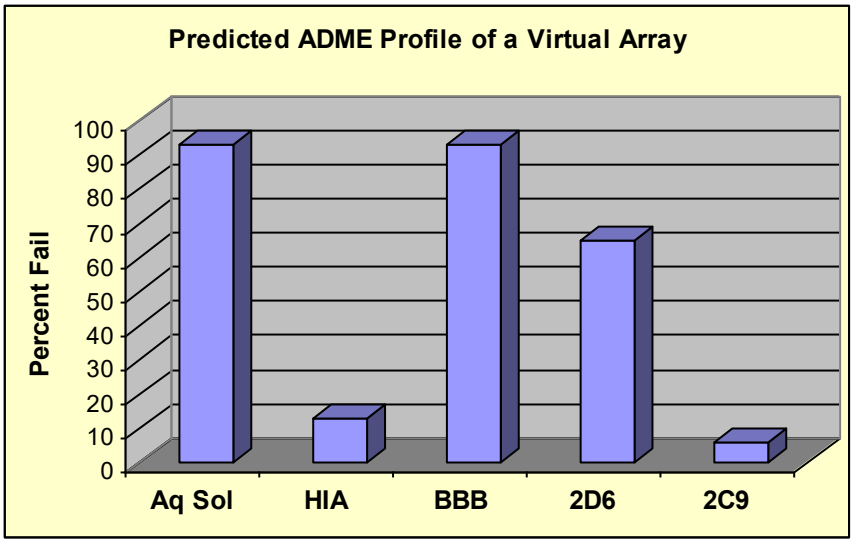


Figure 14.1 A graph of the failure rates for 5 ADME properties for a single chemotype. The higher bars indicate greater potential problems with the chemotype. The desired criteria for these properties are described in the text.

One of the values of such virtual profiling is that the chemist can subsequently 'drill down' into the data, comparing those molecules which *pass* the various criteria with those which *fail* and, using this data, try to fix the problems *in silico,* before expensive and time-consuming synthesis and testing have begun. Alternatively, if the chemistry ideas have already covered a wide range of potential chemotypes, it may be appropriate to simply focus resource on those arrays which show lower ADME risks.

The comparison of ADME risk across a number of different chemotypes can be performed easily by changing the graph format. **Error! Reference source not found.** illustrates the results of cumulative percentage failure rate of molecules for each of 23 chemotypes (or arrays) across the same five ADME models. Each coloured bar describes the percentage of compounds of that array that failed the respective property. Using this illustration, we can quickly identify high-risk arrays. Array 6, which is the chemotype previously illustrated in Figure 14.1, stands out as being one of the highest risk arrays. In such cases, where virtually all compounds in a chemotype fail to meet the required property criteria, it is likely that these properties arise from the core scaffold and not substituent groups. Conversely, we can also identify arrays with superior ADME properties, such as arrays 2 and 9, which may prove better starting points for lead generation, as there are no properties with consistent failures in these arrays.

In addition to helping guide chemistry to those chemotypes likely to have reduced risk, this analysis can also help to make better use of resources in downstream testing. For example, Array 18 illustrated here was of particular interest because of its chemical tractability and anticipated potency. Because of the potential risk for poor intestinal absorption predicted by the model, a small number of selected compounds from the library were synthesized and tested in an *in vitro* Caco-2 absorption assay. The data confirmed poor absorption for this chemotype and a key chemical modification to the core scaffold (removal of a peptidic bond) was subsequently made to improve the absorption properties of the entire array.

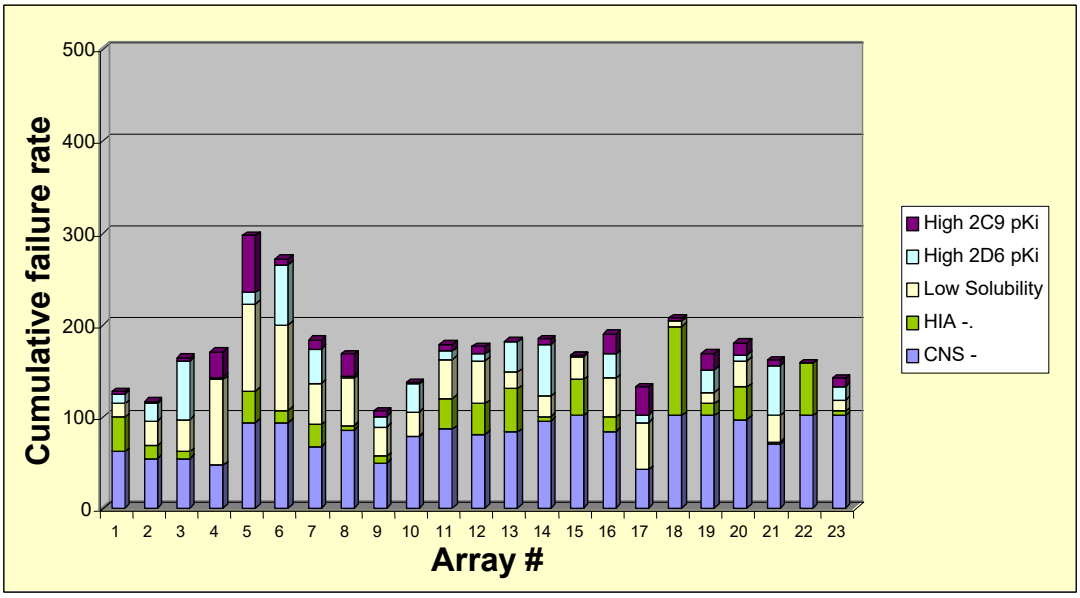


Figure 14.2 A graph of the cumulative failures for 5 ADME properties for 23 chemotypes containing a total of 8 million compounds. The arrays are shown on the x-axis and the cumulative percentage failure rates are shown on the y-axis. For example, if every compound in an array failed to meet the criterion for one property, that array would have a failure rate of 100%. If every compound failed to meet all five property criteria that array would have a cumulative failure rate of 500%.

14.2 Example 2: Prioritisation of Chemotypes Using Probabilistic Scoring

14.2.1 Objective

Probabilistic scoring may be used to assess the likelihood of success of all compounds in a chemotype. By analysing multiple chemotypes, these may be compared and prioritised; focusing resources on chemistries most likely to yield high value lead series.

Whilst the ADME risk 'Profiling' described in Example 1 can help to guide chemistry towards lower risk chemotypes, the process does not take into account the relative importance of the properties assessed in relation to the product profile, or the likely accuracy of individual models (or experimental results) across widely varying chemotypes. This is where the ability to input individual project property weightings, with which probabilistic scoring can be performed, enables more effective prioritisation.

Product Profile: A therapy against a peripheral oncology target. While oral bioavailability would be preferable, an IV formulation would be acceptable.

History: Hits against the project's target had been identified in four chemotypes via high-throughput screening.

14.2.2 Process

Virtual libraries were enumerated for each of the chemotypes, representing the range of accessible chemistries based on the core scaffold of each hit. The ADME properties of each compound were predicted using StarDrop and the compounds were scored using the following scoring profile (Figure 14.2).

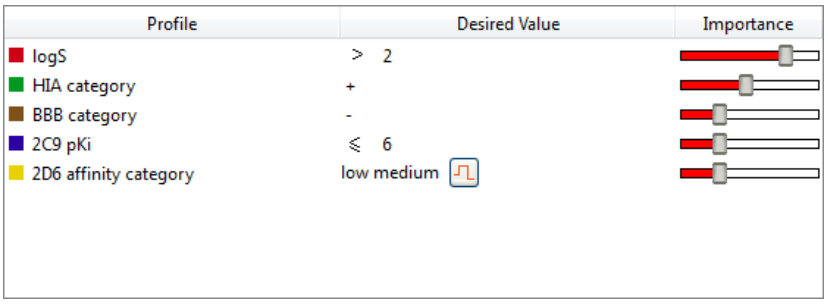


Figure 14.2 Example 2 – Scoring criteria. Scoring functions are one-threshold functions.

The distributions of scores for the libraries representing each chemotype were calculated and the results are presented in Figure 14.3.

From this, it can be seen that libraries 1 and 2 both contain compounds likely to have an excellent balance of properties, although there is a wide distribution of scores in these libraries. Library 4 has a significantly worse distribution of scores, although there is a subset of compounds with a higher likelihood of success than the majority in this library. The distribution of scores in library 3 suggests a very low probability that a compound with a good balance of properties could be derived from this chemotype. The conclusion drawn was that the majority of the resources in this project should be split between synthesis and testing of compounds in libraries 1 and 2, with a small effort spent exploring a small number of compounds from library 4, focusing on the subset of compounds in this library with higher scores.

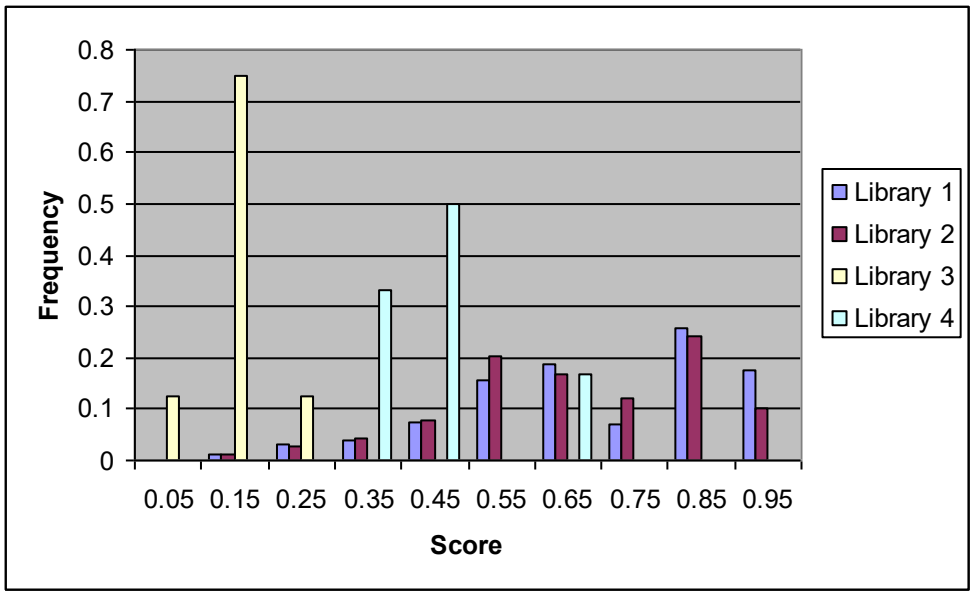


Figure 14.3 A comparison of score profiles for four virtual libraries. From this it can be clearly seen that both libraries 1 and 2 contain compounds likely to have an excellent balance of properties relative to the required profile for the project. Furthermore, it can be seen that libraries 1 and 2 are more likely to yield optimal compounds than 3 and 4, allowing further analysis, synthesis and testing to be prioritised accordingly.

14.3 Example 3: Focusing Resources in Hit-to-Lead

14.3.1 Objective

Following up on hits from primary screening campaigns can be a lengthy and resource intensive process. For programs where multiple hits are found across diverse chemistry, project teams may not have the resource to follow up on all potential chemotypes. This case study illustrates an example where downstream effort was directed towards those chemistries that have the highest overall chance of yielding successful drug candidates.

Product Profile Oral dose therapy for a non-CNS target

History: Following initial screening of over 3,000 compounds selected from a virtual library of over 13,000 molecules, the project identified multiple hits across 13 different chemistries. Due to resource limitations the project team had to decide where to focus their downstream effort in order to maximise the likelihood of identifying chemistries that would yield tractable lead series.

The chemical space plot in Figure 14.4 illustrates the diversity of the chemistries for which active compounds were identified in the context of the company's overall compound collection of 13,000 compounds.

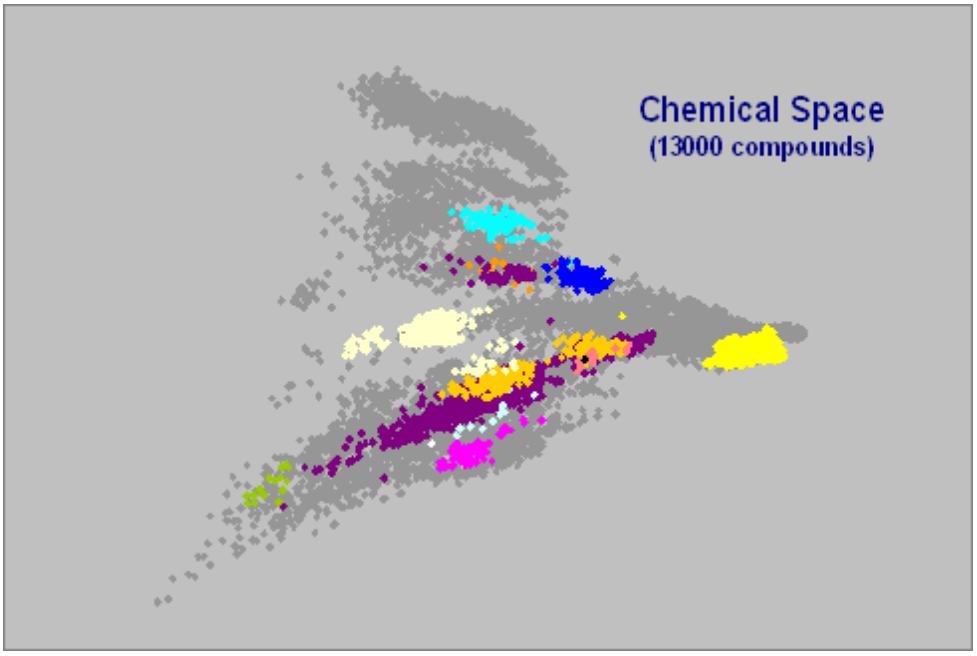


Figure 14.4 The chemical space of the compound collection from which active compounds were identified in a primary screen. The 13 chemotypes from which active molecules were identified are highlighted in colour.

14.3.2 Process

Figure 14.5 summarizes the scoring parameters for ten of models in the StarDrop ADME QSAR module in order of their relative importance with respect to the overall aims of the project.

Profile	Desired Value	Importance
logS	> 1	
HIA category	+	
logP	⩽ 3.5	
2D6 affinity category	low medium	
P-gp category	no	
PPB category (version 5.0)	low	
2C9 pKi	⩽ 6	
BBB category	-	
BBB log([brain]:[blood])	⩽ -0.5	
hERG pIC50	⩽ 5	

Figure 14.5 Example 3 – Scoring criteria

Each molecule was scored against these criteria and all compounds were then compared to assess which chemistries were at high risk of having ADME related problems and which chemistries could be considered to be at low risk from ADME problems.

The chemical space plot in Figure 14.6 shows that most areas of the chemistry space under consideration had the potential to yield compounds with the desired balance of properties, as designated by the lighter coloured dots. However, some chemical series that are unlikely to yield compounds with suitable ADME properties were immediately obvious, highlighted here as being 'high risk space'. Similarly, areas of chemistry that are predominantly 'low risk space' could be seen and these became the primary focus for further investigation.

Comparison of overall predicted ADME scores across different chemotypes enabled the project team to prioritise resource towards chemical ideas having the best overall likelihood of success against the project's defined criteria. Figure 14.7 shows four of the original 13 chemotypes along with the the distribution of scores (and associated uncertainties) for compounds within each series. From this it was clear that resource should be focused around the synthesis and testing of compounds from chemotype 3.

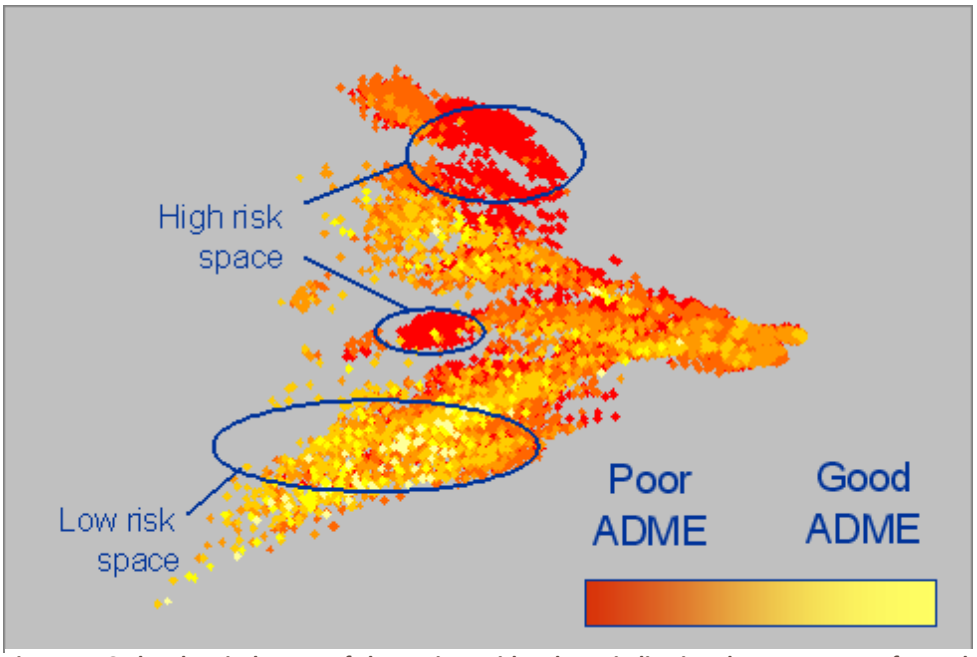


Figure 14.6 The chemical space of the project with colours indicating the ADME score for each compound from red (low) to yellow (high). Areas of particularly high and low risk are highlighted.

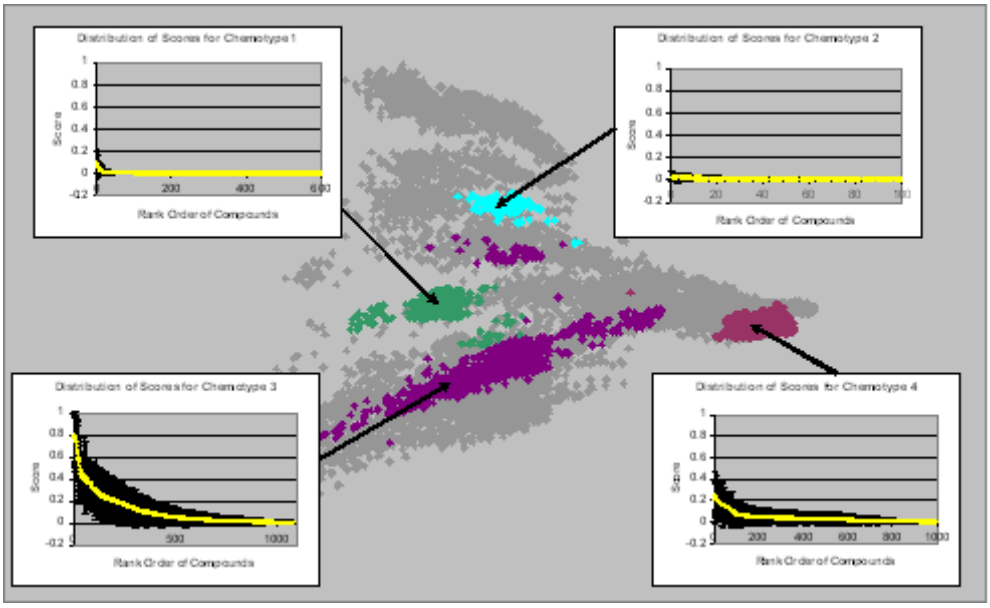


Figure 14.7 The chemical space of the project with four chemotypes highlighted from which active compounds were identified. The score distributions for these chemotypes are plotted, clearly indicating that Chemotype 3 offered the highest chance of containing good ADME properties.

14.4 Example 4: Reducing Synthesis Cycles in Lead Optimisation

14.4.1 Objective

When a project reaches an impasse, it is often useful to look back over past data to seek to identify potential opportunities that may not have been previously apparent. In this example, StarDrop was applied to data generated in a long-running lead optimisation project in order to identify any missed opportunities and to demonstrate how StarDrop could have identified suitable compounds more quickly.

Product Profile: An oral dose therapy for a CNS target

History: In the course of the project to date in excess of 3,000 compounds had been synthesized and screened, with 400 compounds being screened through a vigorous *in vitro* ADME cascade and 70 progressed to full *in vivo* pharmacokinetic analysis. In consultation with the project team, an analysis of the existing data was performed to identify the major decision criteria for compound progression. The following figures describe the chronological progress of the project towards its target project profile.

Figure 14.8 shows an analysis of the first 200 compounds progressed from *in vitro* potency to *in vitro* ADME profiling. These compounds were predominantly in one discreet area of chemical space. This was where the most potent compounds were located. However, as can be seen from the two examples shown in this chemical space plot, compounds in this area typically possessed either good bioavailability or good CNS penetration, but not both. In striving for greater potency, the project chemists had constrained themselves to an area of chemistry that was unlikely to yield successful compounds.

A compound scoring profile was generated to identify those compounds having the best overall balance of ADME properties consistent with the objectives of the target product profile. On the chemical space plot shown in Figure 67 compounds that scored highly against the criteria are coloured in light yellow, and those which had a poor predicted balance of ADME properties are coloured in red. Had this *in silico* analysis been carried out prior to compound synthesis it would have highlighted the difficulties in this area of chemical space, potentially saving "misdirected" resource. Analysis of the second 200 compounds progressed to *in vitro* ADME profiling, showed a marked change in synthetic focus as can be seen in the chemical space plot shown in Figure 14.9 (green dots). Compounds in the lower region of chemical space still maintained good levels of activity but had a much better balance between bioavailability and CNS penetration as can be seen by the two compounds highlighted. However, an enormous effort had been applied to reach this point.

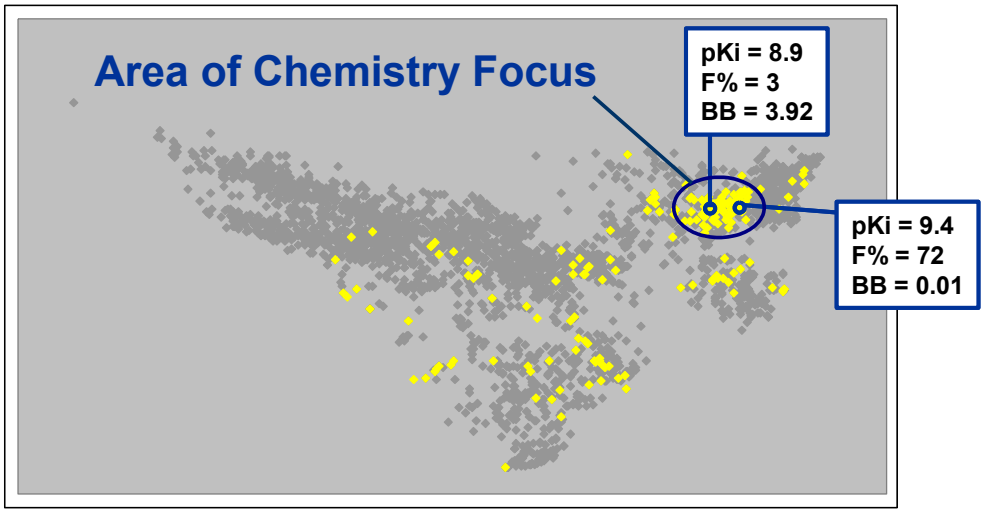


Figure 14.8 A chemical space plot illustrating the chemistry explored in Example 4. The first 200 compounds chosen to study *in vitro* ADME properties are shown along with two example compounds illustrating the typical *in vivo* pharmacokinetics achieved by the best compounds.

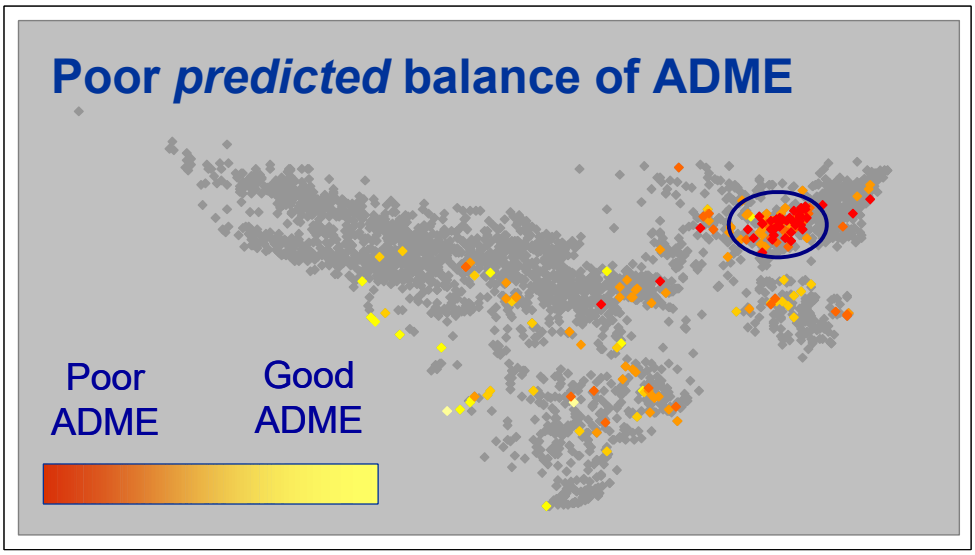


Figure 14.10 The chemical space explored in Example 4. *In silico* scores for the first 200 compounds chosen for progression are illustrated using the colour scale shown. This demonstrates that the compounds in the primary area of chemistry focus have a poor predicted ADME profile and, had this analysis been performed early in the project, would have been treated with caution.

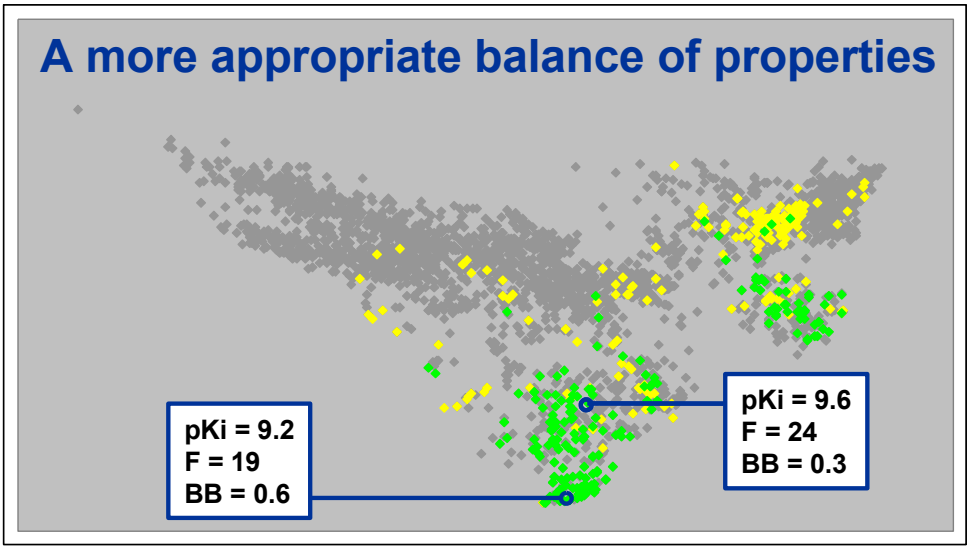


Figure 14.9 Chemical space plot of the chemistry explored in Example 4. Shown here (green dots) are the second set of 200 compounds chosen for progression to *in vitro* ADME studies. The resulting compounds, when tested *in vivo*, showed an improvement in pharmacokinetics as illustrated by the two highlighted examples.

14.4.2 Process

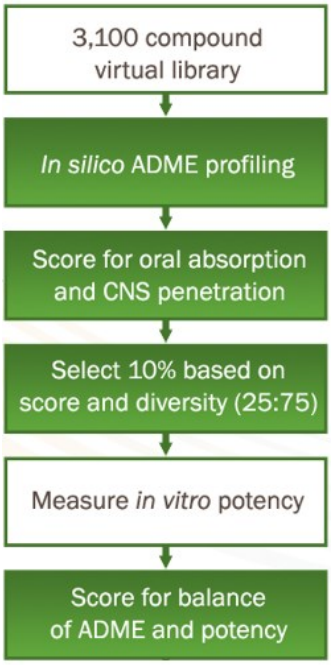


Figure 14.12 The process used in to select compounds from the virtual library for synthesis and testing

In order to demonstrate an alternative approach to the project's chemistry, all of the compounds synthesized to date were considered as a virtual library, for which no experimental data were available. The process illustrated in was applied in order to select an initial subset of 300 compounds from this virtual library. Although the library has been designed with target activity in mind, little is known about the potency SAR and initial compound selection typically needs to cover a wide chemical space in order to identify diverse hits. Therefore, the entire library was profiled for predicted ADME properties. This was used to bias the selection towards compounds likely to have the required balance of ADME properties, while maintaining the diversity of the selection. The resulting selection is shown in Figure 14.11.

The selected compounds would then be synthesized and screened for *in vitro* potency. Of course, in this retrospective analysis, we can identify the potencies for the selected compounds and these are shown in Figure 14.14. As can be seen, a good distribution of potencies was obtained in the selected 300 compounds, and therefore the compounds were then scored again; this time for an appropriate balance between good potency (measured) and good ADME (predicted). A further subset of 25 compounds was then selected for progression to *in vivo* pharmacokinetic studies. Here the bias in the selection is in favour of compounds having the best overall probability of success but with some degree of diversity factored-in to aid "back-up" or "second series" identification. The compounds selected in this way are illustrated in Figure 14.15 .

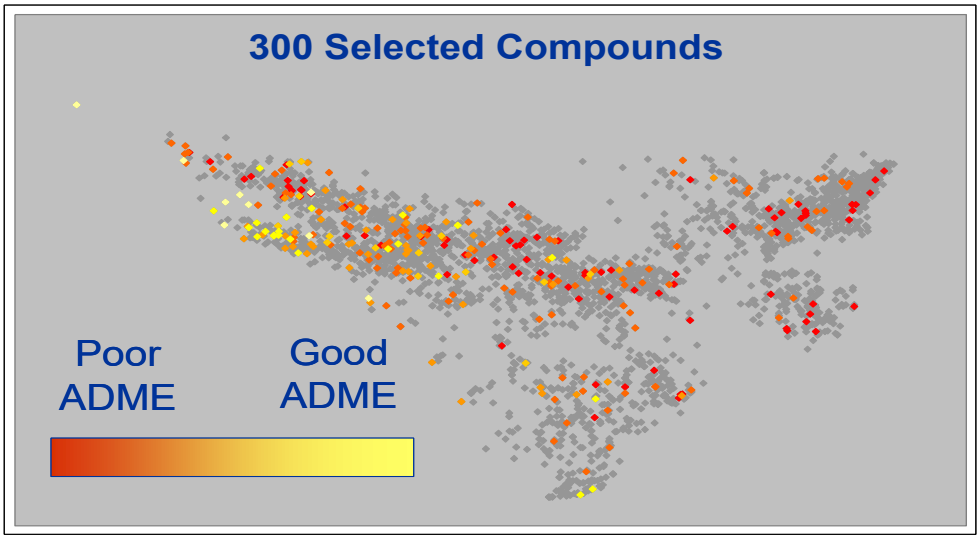


Figure 14.12 A chemical space plot illustrating the 300 compounds selected from the virtual library. These are coloured according to the predicted ADME score for these compounds from poor (red) to good (yellow).

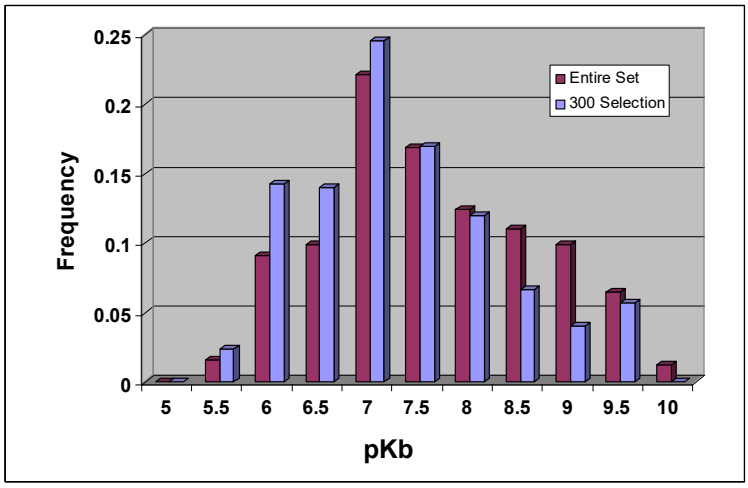


Figure 14.13 The distribution of potencies of the 300 compounds selected from the virtual library. This is compared with the distribution of potencies of all compounds which is known retrospectively. This indicates that a representative sample has been obtained despite a bias in the selection toward compounds with good ADME properties.

Whilst *in vivo* PK data are not available on all of the compounds selected using StarDrop, it can be seen in Figure 14.15 that the approach has selected key compounds that capture the progress of the project in relation to its target profile and also highlighted an area of space previously overlooked when viewed from a potency-biased standpoint.

This would have been achieved through the synthesis of only 10% of the compounds actually synthesized to date and with only 30% of the *in vivo* testing.

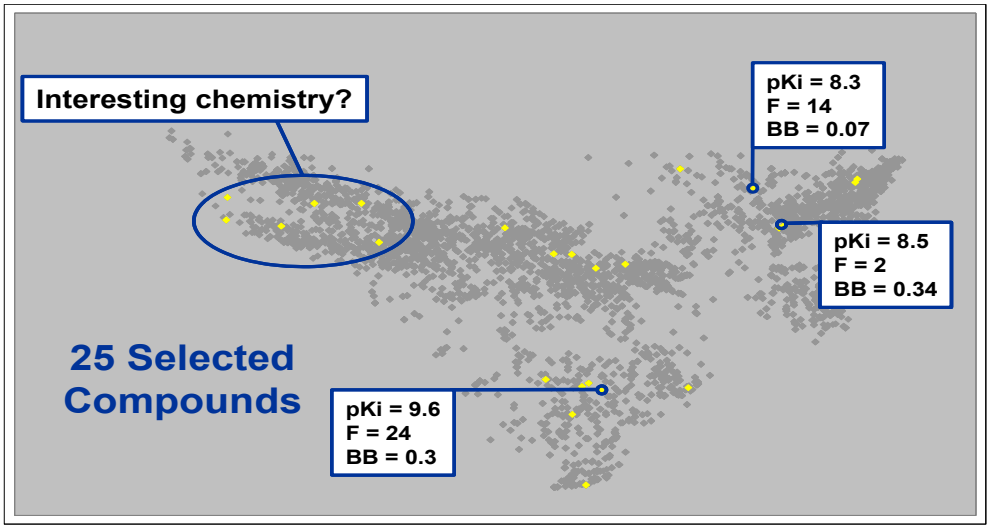


Figure 14.14 The compounds selected for progression to *in vivo* pharmacokinetic studies. Examples of the pharmacokinetic parameters of these compounds, where known, are show. In addition, compounds expected to have a good balance between potency and ADME properties that have not been investigated *in vivo* are highlighted.

14.5 Example 5: Prioritisation of Compounds in Lead Optimisation, based on *In Vitro* Data

14.5.1 Objective

Probabilistic scoring within StarDrop may be applied to *in vitro* data in the same manner as *in silico* predictions. Consideration of uncertainty and relevance of each measurement is just as important for experimentally-derived data. In this example, StarDrop was applied to an early-stage lead optimization project to identify compounds with an appropriate balance of potency, selectivity and ADME properties against the product profile, based on *in vitro* data.

Product Profile: An oral dose therapy.

History: *In vitro* potency, selectivity, solubility and microsomal stability data had been generated for a set of 150 compounds. Compounds had previously been filtered on the basis of selectivity and potency, which gave rise to solubility and metabolic stability problems *in vivo* in rats.

14.5.2 Process

The scoring criteria chosen for each of the *in vitro* properties are shown in Figure 14.15.

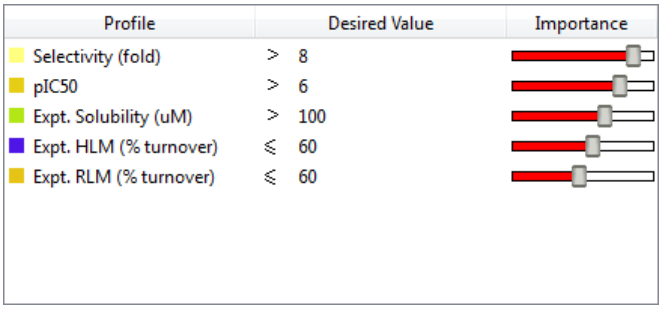


Figure 14.15 Example 5 – Scoring criteria.

14.5.3 Profile 1: Selectivity and Potency Only

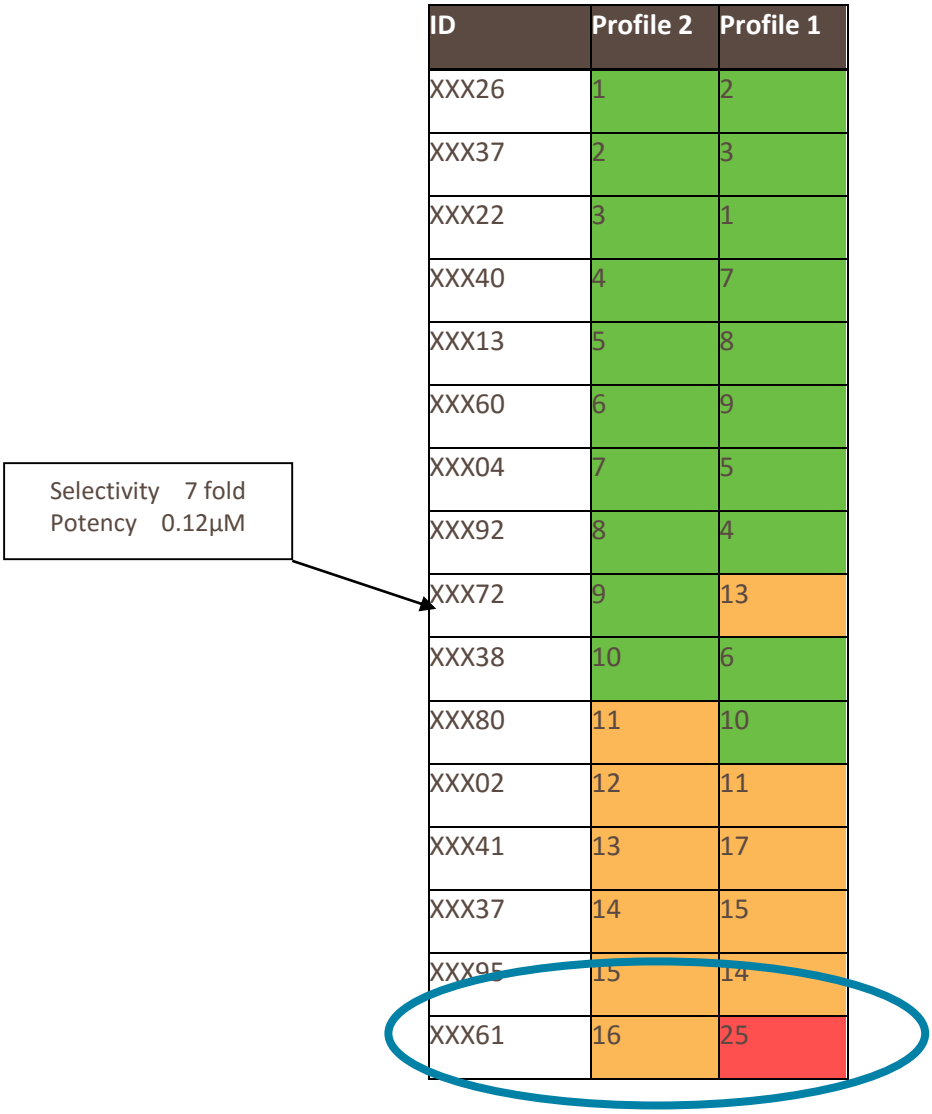
A 'traditional' method of selection had been applied previously, using selectivity and potency alone and filtering the compounds based on the threshold values listed above. The compounds were ordered by selectivity and then by potency, because selectivity was considered more important than potency in this case. The experimental uncertainties in the measurements were ignored. The results of this process are illustrated in Figure 14.16.

ID	Profile 1
XXX22	1
XXX26	2
XXX37	3
XXX92	4
XXX04	5
XXX38	6
XXX40	7
XXX13	8
XXX60	9
XXX80	10
XXX02	11
XXX82	12
XXX72	13
XXX95	14
XXX37	15
XXX81	16

Figure 14.16 The result of filtering and ordering the compounds based on their measured selectivity and potency. Experimental uncertainties were ignored. The top 10 compounds are coloured in green, compounds 10-20 in orange and >20 (not shown) in red. Compound XXX72 is highlighted for comparison with other profiles.

14.5.4 Profile 2: Selectivity and Potency with Uncertainty

Using StarDrop the compounds were scored on the basis of selectivity and potency alone, using criteria given above, and including the associated uncertainties in these properties. The results are shown in Figure 14.17.



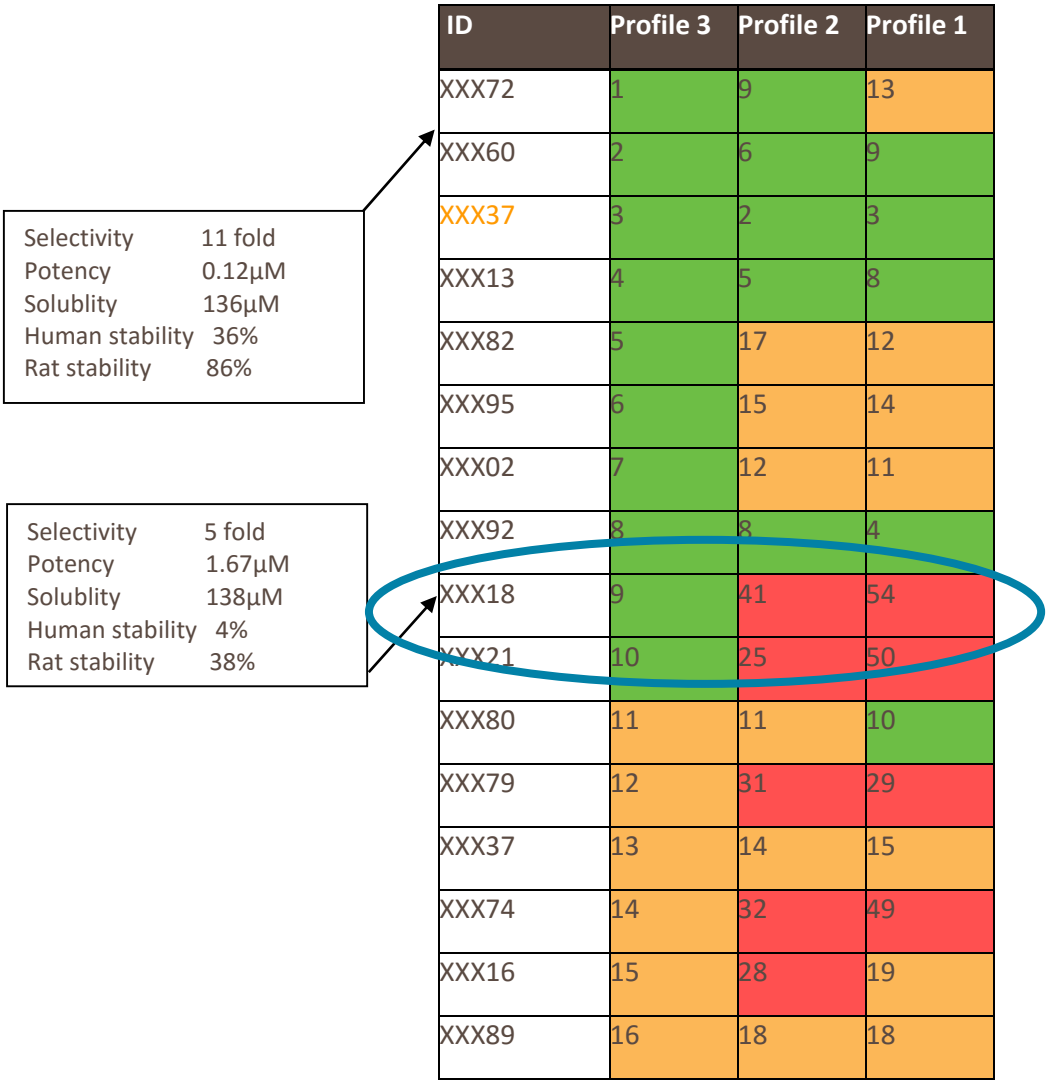
ID	Profile 2	Profile 1
XXX26	1	2
XXX37	2	3
XXX22	3	1
XXX40	4	7
XXX13	5	8
XXX60	6	9
XXX04	7	5
XXX92	8	4
XXX72	9	13
XXX38	10	6
XXX80	11	10
XXX02	12	11
XXX41	13	17
XXX37	14	15
XXX95	15	14
XXX61	16	25

Figure 14.17 The result of prioritising the compounds based on selectivity and potency alone, using the probabilistic scoring algorithm. Profile 2 shows this ordering and the previous position in profile 1 is shown for comparison. In each profile, the top 10 compounds are coloured in green, compounds 10-20 in orange and >20 in red. Compound XXX72 is highlighted for comparison with other profiles. Compound XXX61 is highlighted for further discussion in the main text.

From this, it can be seen that there was not a dramatic change in the compounds in the top 10, although the order was altered. However, the position of some compounds improved significantly. In particular, compound XXX61, previously number 25 in the list, was promoted to position 16. This is because, in the 'filtering' approach previously applied, this compound had been discarded due to its selectivity of 7-fold, below the threshold value of 8-fold. However, within the uncertainty of the experimental assay, there is a significant chance that the selectivity of this compound exceeds the required value. Therefore, due to its excellent potency, this compound achieved a significantly higher score.

14.5.5 Profile 3: All *In Vitro* Data with Uncertainty

The probabilistic scoring algorithm was finally applied to all of the experimental data and uncertainties, according to the scoring profile defined above. The results of this are shown in Figure 14.24.



ID	Profile 3	Profile 2	Profile 1
XXX72	1	9	13
XXX60	2	6	9
XXX37	3	2	3
XXX13	4	5	8
XXX82	5	17	12
XXX95	6	15	14
XXX02	7	12	11
XXX92	8	8	4
XXX18	9	41	54
XXX21	10	25	50
XXX80	11	11	10
XXX79	12	31	29
XXX37	13	14	15
XXX74	14	32	49
XXX16	15	28	19
XXX89	16	18	18

Figure 14.18 The result of prioritising the compounds based on selectivity, potency, solubility and microsomal stability, using the probabilistic scoring algorithm. Profile 3 shows this ordering. The previous positions in profiles 1 and 2 are shown for comparison. In each profile, the top 10 compounds are coloured in green, compounds 10-20 in orange and >20 in red. Compound XXX72 is highlighted for comparison with other profiles. Compound XXX18 is highlighted for further discussion in the main text.

The inclusion of ADME data, together with the uncertainty in the all measurements, had a significant influence over the top scoring compounds. Compound XXX72, originally scored in 13th place, is now the top scoring compound. This compound comfortably met the criteria for selectivity and potency and also had good solubility and stability in human liver microsomes. The only property for which this compound fell below the chosen threshold was stability in rat liver microsomes, which was considered to incur the lowest risk.

Four compounds were identified that had been originally overlooked due to the previous approach of filtering based on selectivity and potency. One of these compounds, Compound XXX18 had the second best PK profile when tested *in vivo* and was the only representative of a chemical series that had been discounted early in the project. This series was resurrected and new approaches to expanding its chemistry were explored, with a view to finding compounds with better all-round properties. This new chemistry would not otherwise have been considered.

14.6 Example 6: Developing Buspirone Analogues with Improved Metabolic Stability

The feasibility of pursuing a fast-follower for Buspirone, a 5HT1A ligand used as an anti-anxiolytic therapeutic, was explored in reference (Tandon, et al., 2004). Buspirone experiences rapid metabolism by CYP3A4 leading to low oral bioavailability and a short half-life in humans and this study aimed to identify analogues of Buspirone with greater metabolic stability whilst maintaining receptor affinity. The published study was guided by prospective application of an earlier version of the models described herein, but here we have repeated the calculations with the latest models.

The structure of Buspirone can be broken down into 3 regions:

- an arylpiperazine which is a protonatable recognition element important for receptor affinity and is metabolized via hydroxylation of pyrimidine C5
- a tetramethylene linker which is metabolized by N-dealkylation alpha to the piperazine N4
- a piperidinedione which is metabolized via oxidation of the spirocyclopentane ring

This study explored structural modifications with a view to improving metabolic stability. Here we will compare the experimentally observed changes in in vitro half-life with respect to metabolism by CYP3A4 to predictions from the latest models described in this paper. The predicted metabolic profile of Buspirone and its analogues are shown in Figure 14.19. The presence of two labile sites and a high CSL is consistent with the observed short half-life observed in vitro (4.6 minutes) and rapid metabolism *in vivo*.

Blocking the 5 position of the pyrimidine ring (predicted as labile with a regioselectivity of 58%) with fluorine led to compound 5, where activity at the target is maintained but the half-life increased to 52 minutes. In this case, one region of the molecule was modified but other labile and moderately labile sites remain, so only a small change in the overall CSL is observed even though this modification is beneficial. As noted above, other factors also influence the overall rate of metabolism and a direct correlation between the small changes to CSL and the CYP3A4 half-life is not necessarily expected. However, in this case the increase in half-life is reflected by a fall in CSL from 0.957 to 0.885.

An example of the complex relationship between structural changes and metabolic stability is demonstrated by molecule 10, which introduced a methyl substituent alpha to the piperazine in an attempt to hinder N-dealkylation from what is predicted to be a labile site. The models predict a further small decrease in CSL to 0.8458. However, the half-life falls to 14.8 minutes, indicating that the factors mentioned earlier are influential here. In this instance, the addition of the methyl group changes the lipophilicity and basicity of the compound which are likely to increase the binding affinity to CYP3A4 and hence offset the small decrease in CSL to increase the rate of metabolism.

Replacing the spirocyclopentane ring with a gem dimethyl to give compound 21 eliminated the predicted moderately labile site in the five-membered ring which was reflected by an increase in half-life to 78 minutes.

Overall, this example illustrates that the models can be used to guide development of a lead compound towards greater metabolic stability, where in this example compounds 5 and 21 show half-lives of 52 and 78 minutes respectively whilst maintaining activity of 0.2 µM or lower.

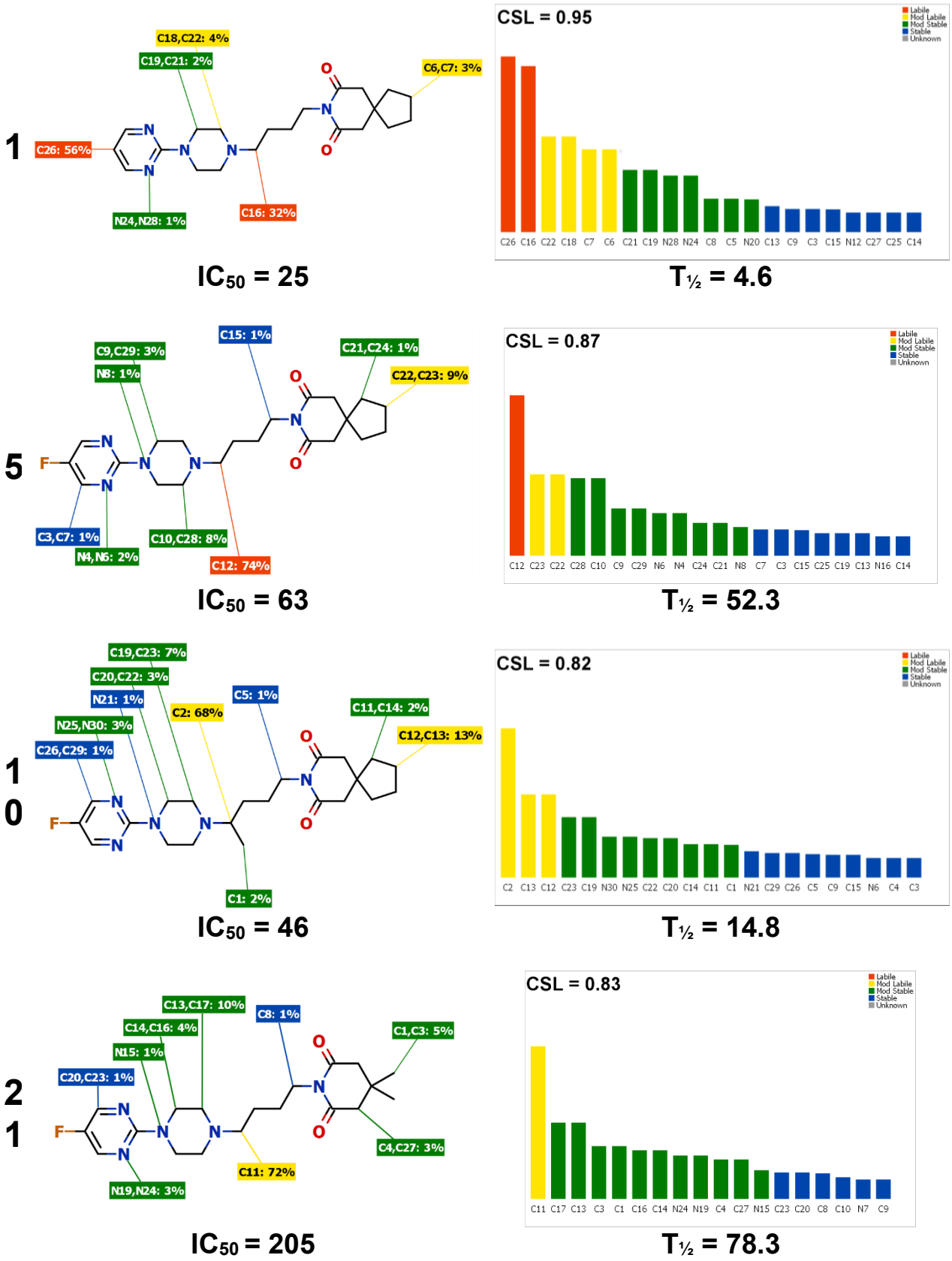


Figure 14.19 Cytochrome P450 metabolism predictions for example compounds from case study 1. The sites of metabolism and predicted regioselectivity are shown for each compound, along with a metabolic landscape illustrating the lability of each site with respect to metabolism by CYP3A4. For each compound the calculated CSL, and experimentally measured half-life ($T_{½}$ in minutes) with respect to *in vitro* metabolism by CYP3A4 and activity (IC_{50} in nM) against the target 5-HT_{1A} are shown (Tandon, et al., 2004).

14.7 Example 7: Developing HIV-1 Reverse Transcriptase Inhibitors with Improved Metabolic Stability

A series of N1-heterocyclic pyrimidinediones were investigated by Mitchell et al. (Mitchell, et al., 2010) for application as HIV-1 non-nucleoside reverse transcriptase inhibitors (NNRTIs) with the aim of improving the pharmacokinetic profile whilst maintaining activity.

Compound 1 showed the required target activity but the half-life of 45 minutes in human liver microsomes was a long way short of the target for once daily dosing. The models predicted the terminal amine group to be a labile site with the results for compound 1 and its analogues given in Figure 14.20. Compound 1 contains a fluorine substitution ortho to the pyridine N. Further substitution of the pyridine ring with fluorine to give compound 9 did not give a significant improvement in terms of prediction or measured data.

To significantly improve the half-life for this series it was necessary to address the metabolically labile terminal amine. Replacement with a hydrogen led to compound 13 where the CSL falls to 0.8309, corresponding to a large increase in half-life to 281 minutes. Further substitution of the pyridine to block both aromatic sites ortho to the pyridine N with fluorine, to give compound 10, led to a further reduction in the CSL and an increase in half-life to in excess of 395 minutes.

This relatively simple example shows how CYP models can quickly identify labile sites to focus on modifications that are likely to improve metabolic stability. In this case, the replacement of a labile terminal amine and blocking aromatic sites with fluorine, where compounds 10 and 13 showed improved metabolic stability and retained good antiviral potency.

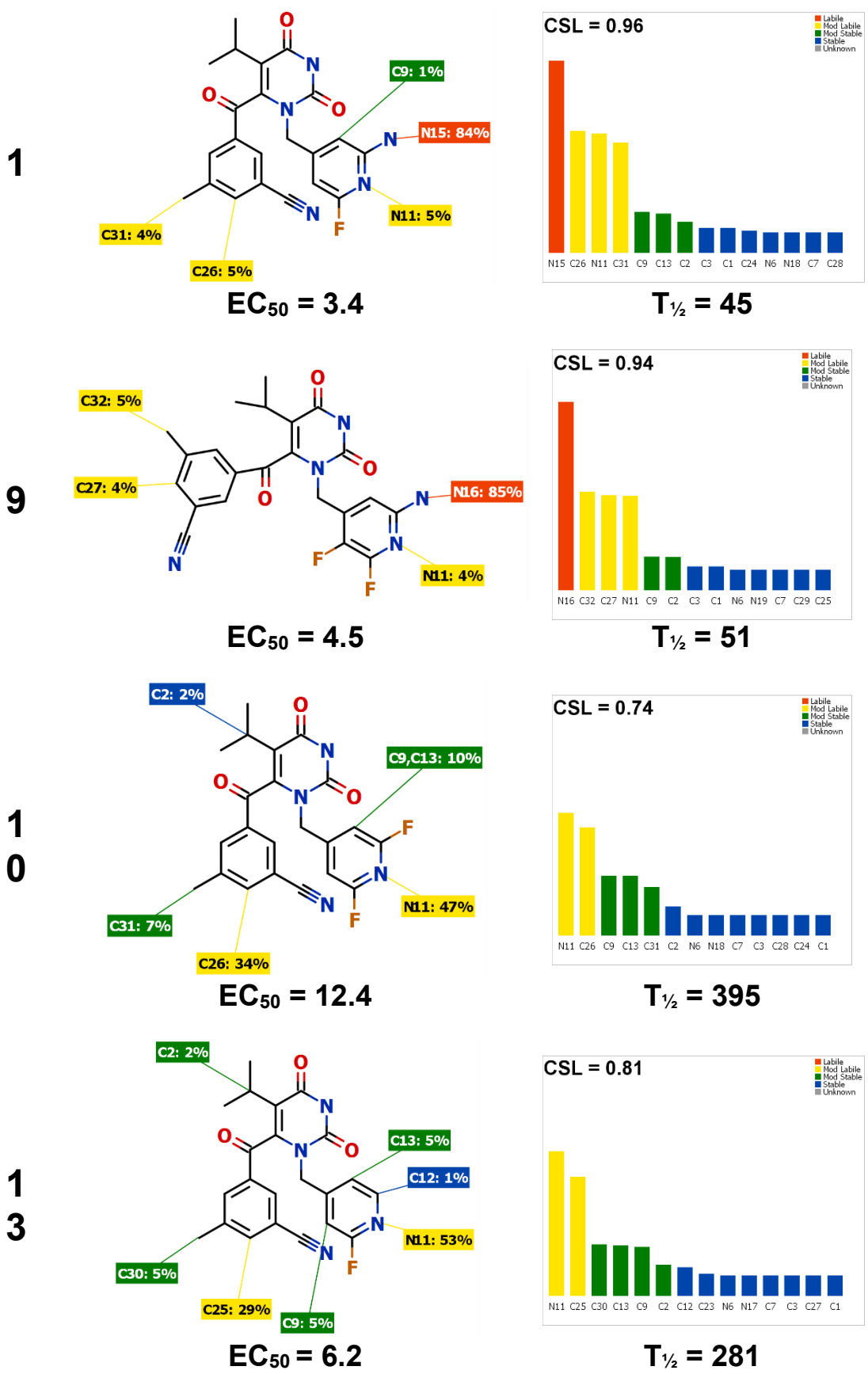


Figure 14.20 Cytochrome P450 metabolism predictions for example compounds from case study 2. The sites of metabolism and predicted regioselectivity are shown for each compound, along with a metabolic landscape illustrating the lability of each site with respect to metabolism by CYP3A4. For each compound the calculated CSL, and experimentally measured half-life ($T_{½}$ in minutes) in an *in vitro* human microsomal stability assay and activity (EC_{50} in nM) against the target HIV-1 reverse transcriptase are shown (Mitchell, et al., 2010).

14.8 Example 8: Novel Benzimidazoles as PDE10A Inhibitors with Improved Metabolic Stability

A series of novel benzimidazoles were developed by Chino et al. (Chino, et al., 2014) that show sub-micromolar activity as inhibitors of PDE10A, which is hypothesized to be effective in treating schizophrenia and a wide range of neurological, psychotic, anxiety and movement disorders by increasing levels of cAMP and cGMP in the brain.

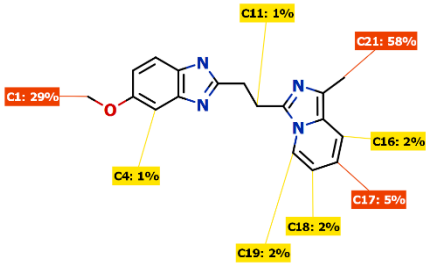
Compound 1 was identified from high throughput screening as a low micromolar PDE10A inhibitor where introduction of a phenyl ring to the N-1 position on benzimidazole was found to improve inhibitory activity. It was noted that compound 14a with a methyl at the 5-position on benzimidazole and a methyl in the 1-prime position of the imidazopyridine was approximately 3 times more active than compound 1, and removal of the methyl in the 1-prime position removed inhibitory activity, indicating the importance of this group (data not shown). This position is predicted by the CYP models to be metabolically labile, along with the 5-methyl on the benzimidazole, with further labile and moderately labile sites in the aromatic positions, as shown in Figure 14.21. This causes these compounds to have high risk of rapid metabolism by CYPs, as illustrated by the CSL values, and borne out in the experimental results with compound 14a exhibiting clearance of greater than 1000 mL/min/kg.

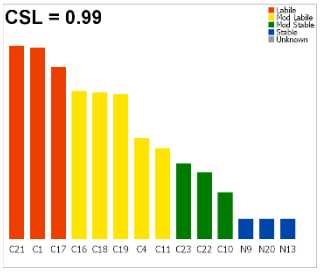
Introduction of another N into the fused ring system to give the imidazopyridazine in compound 16 gave improved inhibition but did not improve metabolic stability. The methyl substituents were shown to be important for activity so variations to the heterocycles that preserved these groups were made. The imidazopyridazine was replaced with an azabenzimidazolinone to give compound 10b, a change which gave improved metabolic stability, due to the loss of some moderately labile aromatic sites.

Focus then shifted to the benzimidazole part of the molecule to further improve metabolic stability. Variation of the 5 methyl and insertion of N into the benzimidazole ring system at the 7 position led to compound 24a showing improved metabolic stability and reflected by a lower CSL. Whilst these changes are not directly blocking a predicted labile site they do impact on the metabolic stability of sites elsewhere in the molecule and show that sometimes subtler longer range effects come into play.

This example highlights a situation where the lead series was developed by making larger changes to molecular fragments, not simply blocking a labile site, with non-intuitive changes to the metabolic stability. In this example the QM methodology employed by the models was able to capture these trends by considering each fragment in its actual environment and would allow a chemist developing a lead series to gain insights into the likely impact of even quite large changes to their molecules.

1

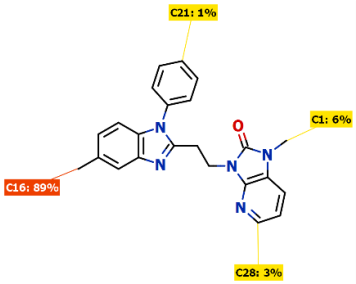


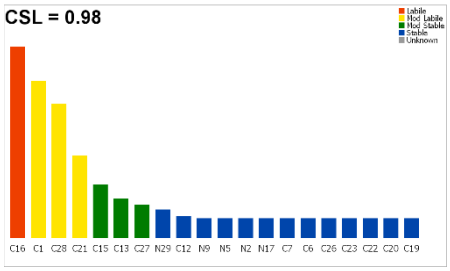


IC_{50} = N/A

Cl_{int} = N/A

10b

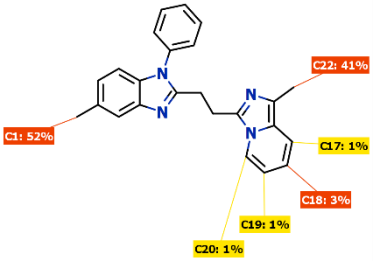


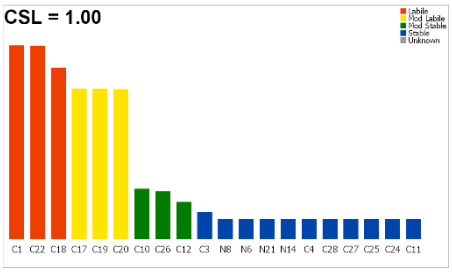


IC_{50} = 9.2

Cl_{int} = 330

14a

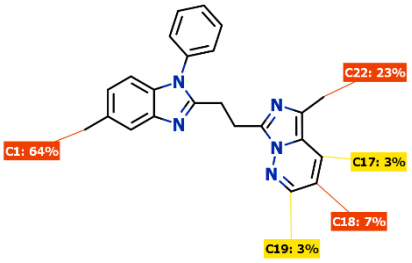


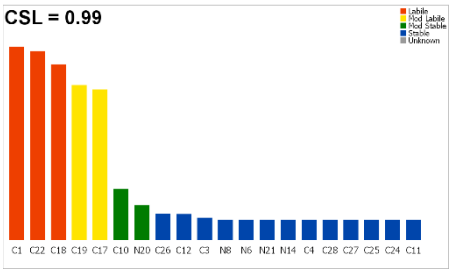


IC_{50} = 590

Cl_{int} = 1000

16

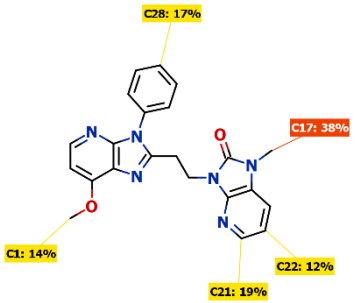


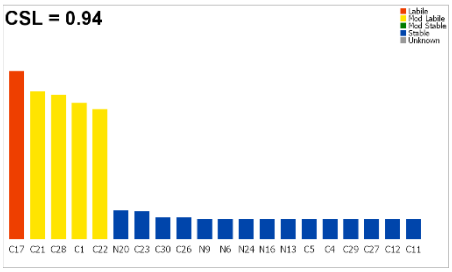


IC_{50} = 9.1

Cl_{int} = 1000

24a





IC_{50} = 29

Cl_{int} = 130

Figure 14.21 Cytochrome P450 metabolism predictions for example compounds from case study 3. The sites of metabolism and predicted regioselectivity are shown for each compound along with a metabolic landscape illustrating the lability of each site with respect to metabolism by CYP3A4. For each compound the calculated CSL, and experimentally measured intrinsic clearance (Cl_{int} in ml/min/kg) in an *in vitro* human microsomal stability assay and activity (IC50 in nM) against the target PDE_{10A} are shown (Chino, et al., 2014).

14.9 Example 9: Demonstrating the Use of Nova to Find Drugs from Leads

To illustrate the application of Nova to guide the search for optimised compounds based on an initial lead, we used the lead molecule that ultimately gave rise to the drug duloxetine as the parent molecule.

StarDrop's ADME QSAR models (See Chapter 6) and a model of the inhibitory constant K_i for the serotonin transporter, built with StarDrop's Auto-Modeller, were used to prioritise the compounds generated against the scoring profile, shown in Figure 14.22, which combines potency against the primary target with suitable ADME properties for an orally dosed compound against a CNS target.

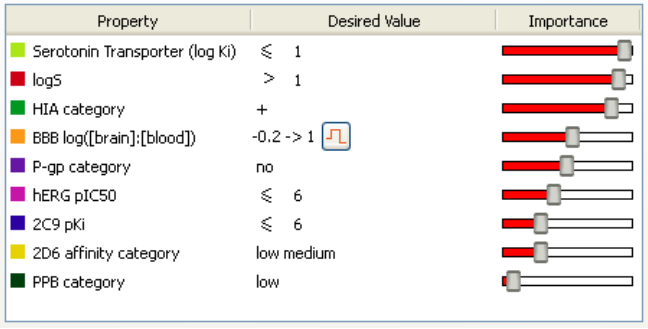


Figure 14.22 The scoring profile used to prioritise compounds generated from the duloxetine lead, showing the properties of interest, the desired value ranges and the importance of each criterion. For example, the most important property was inhibition of the serotonin transporter, for which a predicted K_i of less than 10 nM (log K_i <1) was required. This was followed by an aqueous solubility of greater than 10 µM (logS > 1) and positive prediction for human intestinal absorption.

The application of one generation of transformations produced 172 child compounds, which suggested that exhaustive enumeration of more than two generations would be intractable. Therefore, three generations were applied, but only the top-scoring 10% of the compounds in each of generations 1 and 2 were used as the basis for subsequent generations.

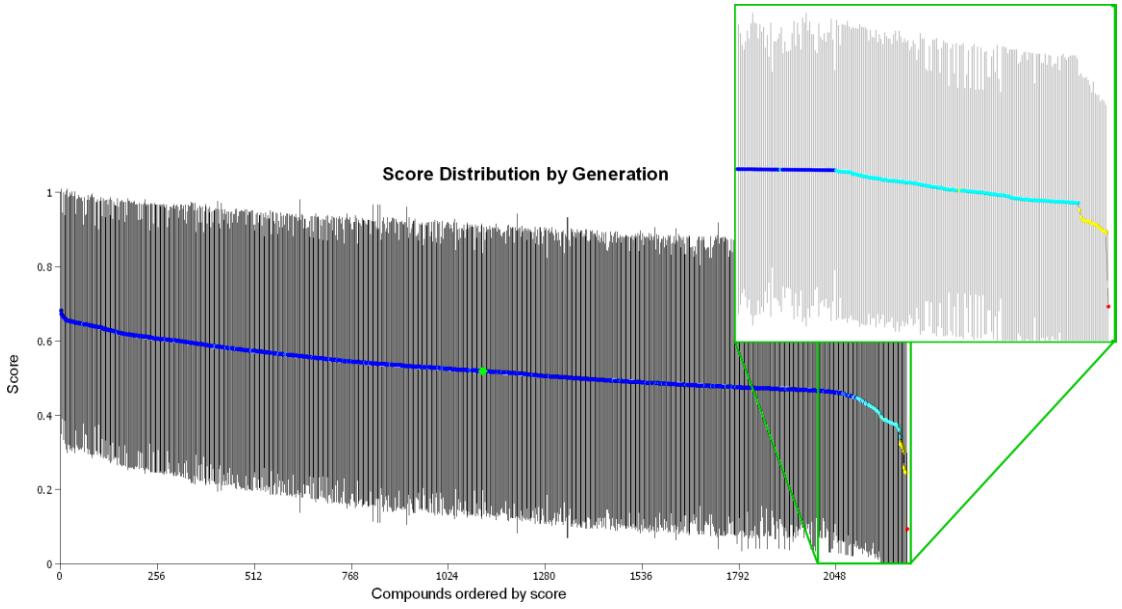


Figure 14.23 This graph shows the compounds generated by three generations of transformations starting with the lead compound for the project that yielded the drug duloxetine. Error bars show the uncertainty of the overall score for each compound due to the uncertainties in the underlying data. Only the top 10% of generations 1 and 2 were used as the basis for subsequent generations. The compounds are coloured by generation: Red is the parent, yellow generation 1, light blue generation 2 and dark blue generation 3. The drug Duloxetine was present in generation 3 and is shown by the green diamond.

The resulting data set contained 2,208 compounds (all of the compounds in the final generation were retained) and the scores for these compounds are plotted in Figure 14.23. From this, a number of observations may be made: The compounds in each generation typically show an increase in score over the previous generation; the score for the initial lead is 0.09 and the averages for the compounds in subsequent generations are 0.32, 0.44 and 0.53 respectively (note that only the top 10% of the first two generations are included). However, as the results from multiple uncertain predictions are combined to calculate the score, the uncertainties in the score are high, as shown by the error bars in Figure 14.23. Therefore, it is difficult to discriminate between compounds with confidence, particularly in the later generations. Finally, it is notable that duloxetine itself is present in the final generation, with a score that is significantly higher than the initial lead (with a probability of ~87%) and not significantly below that of the highest scoring compounds.

The structures and scores of the initial lead and duloxetine are shown in Figure 14.24 along with the three highest ranking molecules generated. The scores and uncertainties for the three top compounds indicate that they are significantly better than the initial leads with a confidence of ~94%. Although none of the top-three compounds could be identified in a search of PubChem (Bolton, Wang, Thiessen, & Bryant, 2008), the second-ranked compound bears a strong similarity (Tanimoto similarity >0.9) to litoxetine, shown in Figure 14.25, which was progressed to clinical trials and is active against the serotonin transporter with an IC50 of 6 nM (Andrews, et al., 2009).

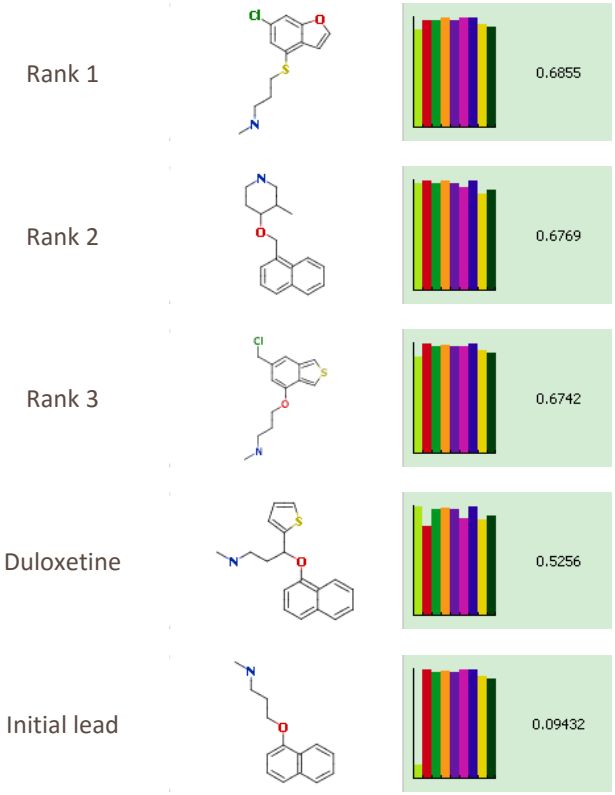


Figure 14.24 The initial lead that ultimately gave rise to duloxetine, the top three compounds generated from this lead and duloxetine, which was also generated by the algorithm. The score for each compound is show to the right along with a histogram indicating the contribution of each property to the overall score (the colour of each bar corresponds to the property key shown in Figure 14.21). All of these compounds are predicted to have good values for the predicted ADME properties. However, the initial lead has a much lower score due to a significantly poorer K_i predicted for the serotonin transporter.

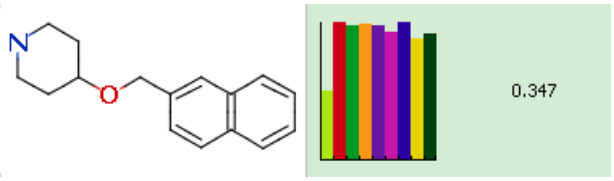


Figure 14.25 The structure and calculated score for litoxetine, a clinical candidate serotonin reuptake inhibitor. The predicted K_i for this compound is 10 nM, in line with the reported IC_{50} of 6 nM. Although this structure was not generated automatically in this example, it bears a strong similarity (Tanimoto similarity >0.9) with the second-ranked compound, which has a higher predicted affinity and hence a higher score.

The chemical space of the data set generated is shown in Figure 14.26. From this it is notable that a wide range of different chemical motifs has been explored and that there are multiple 'hot spots' containing high-scoring compounds; the best scoring compounds are not concentrated in one region, indicating that the algorithm has identified a number of different chemical strategies worthy of further consideration. The top three ranked molecules are structurally diverse, within the range of diversity explored around the initial lead, and are distinct from both the initial lead and duloxetine itself.

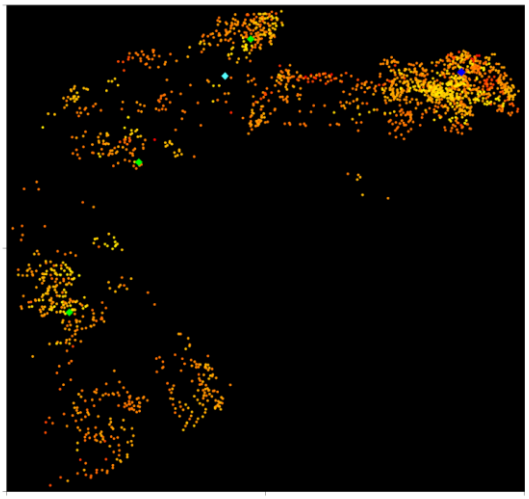


Figure 14.26 The chemical space of compounds generated from the initial lead that gave rise to duloxetine. The points corresponding to compounds are coloured by score, from the lowest (0.29) in red to the highest (0.69) in yellow. The initial lead is shown as a dark blue diamond, duloxetine as a light blue diamond. The top-three scoring compounds are shown as green diamonds.

In this example, the increase in score is driven primarily by the improvements in predicted target affinity between generations because the predicted ADME properties of the lead compound were good to begin with. However, the use of probabilistic scoring to select compounds with a good balance of properties was valuable as it eliminated compounds in early generations that were predicted to have high target affinity but were unlikely to have a good balance of ADME properties for the overall objective. Figure 14.27 shows the distribution of the scores for compounds in the first two generations with predicted K_i less than 10 nM, indicating that a significant number of compounds that were predicted to be active were rejected due to the predictions of poor values of other properties including solubility (184 compounds from generation 2 were used as the progenitors of generation 3).

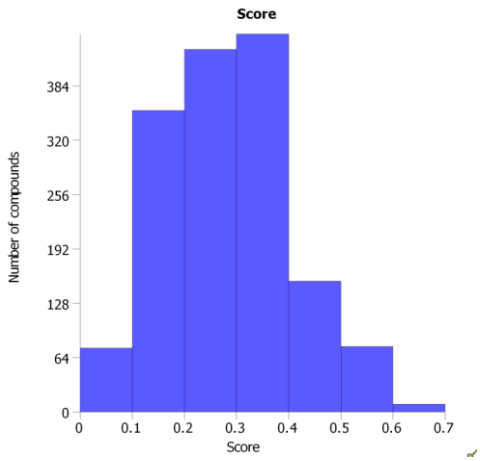
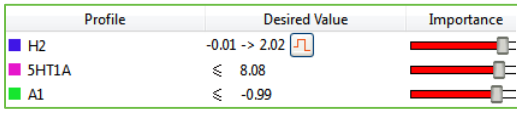


Figure 14.27 Score distribution for the compounds in generations 1 and 2 from the duloxetine lead compound with a predicted K_i of less than 10 nM. From this we can see that there are a significant number of compounds with poor scores, despite having high target affinity, indicating that they are likely to have poor values for other relevant properties.

14.10 Example 10: Using MPO Explorer to Identify Non-toxic Compounds

MPO Explorer's Profile Builder can be used to find rules for any therapeutic objective using any data, whether experimental or predicted. In this example we applied the Profile Builder to search for rules that help distinguish compounds with a low risk of *in vivo* toxicity, based on experimentally measured *in vitro* data. Here we explored data sets describing known drugs to determine their cardiotoxic and hepatotoxic potential in the clinic using a set of biochemical assays. These drugs were profiled in the CEREP Bioprint® assays panel (CEREP, n.d.) which offers biochemical assays against 185 targets including GPCR, kinase, nuclear hormone receptors and Cytochrome P450s, etc., and assesses the extent of off-target pharmacology of these compounds. The biochemical assays were run at a single concentration of 10μM and a reporting odds ratio (ROR) was used to detect a signal of potential drug-adverse event association using information from the FDA Adverse Event Reporting system (FAERS, formerly AERS) database (FDA, n.d.), which contains voluntary reports of adverse events submitted to the FDA by healthcare practitioners, manufacturers and consumers. The ROR signals for cardiotoxicity and hepatotoxicity were calculated for these drugs as reported elsewhere (Nadanaciva, et al., 2013). A ROR signal cut-off of 2.5 or above at the System Organ Class (SOC) level in the MedDRA Ontology was used to classify compounds as having cardiac or hepatic risks in the clinic, whereas a ROR signal of less than 2.5 was used to classify compounds as having no cardiotoxicity or hepatotoxicity. We split each of the two data sets into independent training, validation, and test sets comprising 70%, 15%, and 15% of the full set respectively.

The first data set consisted of 474 known drugs, 408 of which were labelled as 'cardiotoxic' and 66 as 'non-cardiotoxic' based on the ROR signal cut-off. It is worth noting that many of the drugs classified as cardiotoxic are in fact used in the treatment of cardiovascular diseases and so their 'toxicity' may result from either the underlying disease state or from the intended pharmacology of the drug in question. The Profile Builder generated rules comprising property criteria for increasing the probability of selecting non-cardiotoxic compounds from the training set, which are validated using compounds in the test and validation sets.



Set	Mean Improvement (%)	Support (%)
Training	233	9.3
Validation	173	10
Test	419	7.4

Figure 14.28 The rule found for identifying non-cardiotoxic compounds obtained by the Profile Builder when applied to a data set consisting of 408 cardiotoxic and 66 non-cardiotoxic compounds. The corresponding predictive performance of this rule over the training, validation, and test sets is also shown. The criteria correspond to percentage inhibition of the histamine 2 (H2), serotonin receptor 5-HT_{1A} and adenosine 1 (A1) receptors.

Figure 14.28 shows the rule obtained using a minimum support value of 8%. It is worth mentioning that the algorithm has only used the three most predictive properties (out of a total of 185) in order to prevent overtraining. One common way to assess the performance of a classifier is a Receiver Operating Characteristic (ROC) curve. The rule exhibits a large mean improvement of 419% over the test set, and the ROC curve (Figure 14.29) generated from the test set compounds shows that the rule performs well at selecting non-cardiotoxic compounds. Five of the 6 test set compounds selected by the rule are non-cardiotoxic, whereas only 13 of 81 compounds in the full test set are non-cardiotoxic. As such, over 83% of the compounds selected by the rule are non-cardiotoxic meaning that the rule offers a substantial improvement over chance because we would only expect approximately 16% of the selected compounds to be non-cardiotoxic if we had to guess. Furthermore, of the 20 test set compounds that fail every criterion in the rule, 19 are cardiotoxic, implying that any compound failing all the criteria comprising the rule has a very high chance of being cardiotoxic.

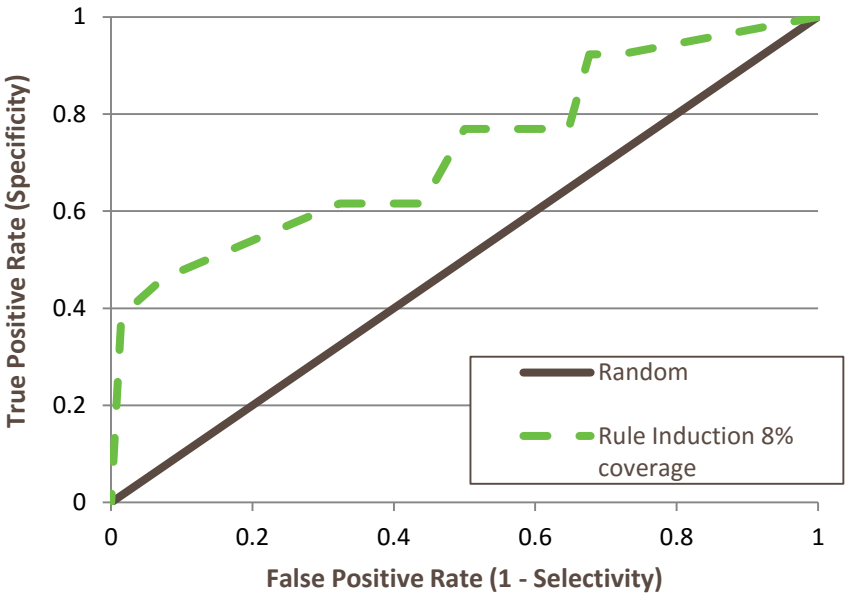
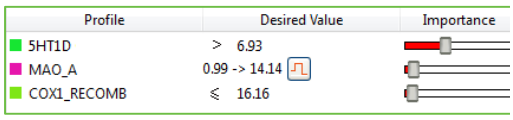


Figure 14.29 Classification of compounds as non-cardiotoxic or otherwise using the rule in Figure 14.28 derived with the Profile Builder. Here, a set of 66 non-cardiotoxic compounds was differentiated from 408 cardiotoxic compounds. A perfect classifier would be represented by the point in the top left and a performance below the identity line indicates worse performance than a random classification. However, in this case the area under the curve (AUC) is 0.72.

The specific values of the three property criteria identified are features of the measurements for specific compounds in the data. However, it is clear that they essentially correspond to absence of inhibition of the histamine 2 (H2), serotonin 5-Hydroxytryptamine (5-HT1A) and adenosine 1 (A1) receptors, which is biologically plausible, because interactions with these receptors have been previously associated with cardiotoxicity. For example, activation of 5-HT1A is known to cause a decrease in blood pressure and heart rate via modulation of sympathetic nerve activity (Dabiré, 1991) (Ramage, 1990). Likewise, stimulation of the adenosine receptor causes bradycardia and hypotension (Bonizzoni, Milani, Ongini, Casati, & Monopoli, 1995) and the activation of the H2 receptor is reported to cause vasorelaxation as reported by Jansen-Olesen et al. (Jansen-Olesen I, et al., 1997).

The second data set contained 470 compounds, 302 of which were labelled as ‘hepatotoxic’ and 168 as ‘non-hepatotoxic’. Here we searched for rules to increase the probability of selecting non-hepatotoxic compounds based on a minimum support value of 10%.



Set	Mean Improvement (%)	Support (%)
Train	51	12
Val	56	14
Test	39	11

Figure 14.30 A rule for identifying non-hepatotoxic compounds obtained by the Profile Builder when applied to a data set consisting of 168 hepatotoxic and 302 non-hepatotoxic compounds. The corresponding predictive performance of this rule over the training, validation, and test sets is also shown. The criteria correspond to percentage inhibition of the serotonin 5-HT_{1D} receptor and monoamine oxidase A (MAO_A) and cyclooxygenase 1 (COX1) enzymes.

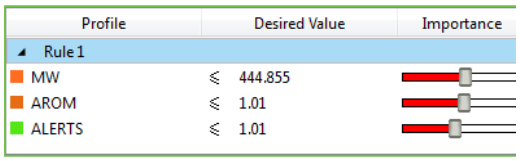
Figure 14.30 shows the rule obtained for selection of non-hepatotoxic compounds. The rule shows a reasonable mean improvement of 39% over the test set, with 9 of the 10 test set compounds being non-hepatotoxic versus 51 non-hepatotoxic compounds out of 80 in the full test set. However, the property criteria themselves do not appear to be biologically relevant. The rule relates to binding to the 5-hydroxytryptamine 1D (5-HT1D) receptor, mono-amine oxidase A (MAO-A) and cyclooxygenase 1 (COX1) enzymes. However, the criterion for 5-HT1D inhibition suggest that an increased inhibition of this enzyme reduces the risk of hepatotoxicity, while the criterion for inhibition of MAO-A suggest a narrow range of inhibition reduces hepatotoxicity risk, both of which are implausible. These statistically significant correlations may arise due to chance in a relatively small and noisy data set with many properties or may be due to correlation of a property with another causative relationship. This demonstrates the advantage of outputting rules as interpretable property criteria over a 'black-box' classifier; even if a rule appears to offer good predictive performance, we may still wish to discard or modify it based on an expert's understanding of the specific property criteria comprising the rule. In this case a plausible rule has not been found because the large majority of the targets for which data are present in the data set are not known to relate with hepatotoxicity. In the few examples of targets that are known to correlate with this toxic outcome, such as $PPAR_{\gamma}$ (Panasyuk, et al., 2012) (Rogue, et al., 2011), there are a statistically insignificant number of inhibitors in the data set and hence no correlation could be found. This reinforces the point that any method for finding rules will be limited by the quality of data available.

14.11 Example 11: Using MPO Explorer to Identify Drug-like Compounds

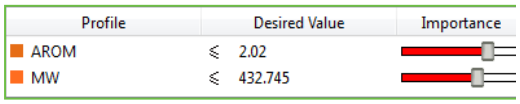
In this example we demonstrate how MPO explorer can be used to identify rules that identify drug-like compounds. A number of measures of 'drug-likeness' have been discussed in the literature, relating easily calculated molecular properties to outcomes such as oral activity or 'developability' (Lipinski, Lombardo, Dominy, & Feeney, 1997) (Veber, et al., 2002) (Ritchie & Macdonald, 2009) (Bickerton, Paolini, Besnard, Muresan, & Hopkins, 2012). Many of these take the form of simple rules, but Bickerton *et al.* describes a quantitative metric, QED, based on a combination of the outputs of desirability functions for logP, HBA, HBD, PSA, ROTB, AROM and the number of alerts for undesirable functionalities (ALERT) (Bickerton, Paolini, Besnard, Muresan, & Hopkins, 2012). Each desirability function corresponds to a single molecular property, and is derived empirically by fitting to this property's distribution over a set of 771 approved oral drugs. To compute the QED score for an individual compound, these desirability functions are combined by taking the geometric mean of all eight desirability scores, giving an overall QED score ranging from 0 (all properties are completely undesirable) to 1 (all properties are ideal).

An issue with this approach is that the QED score for a compound is based solely on the property distributions of a set of approved oral drugs; it does not take into account whether these distributions can *differentiate* the drugs from the 'non-drugs', i.e. the other compounds that might be synthesised. For this reason, an alternative approach, the Relative Drug Likelihood metric (RDL) (Yusof & Segall, 2013), defines a compound's desirability score for a property to be the relative probability of obtaining this compound's property value if it is a drug versus a 'non-drug'.

However, both of these approaches only consider the effect of one property at a time on the drug classification and combine these properties *post hoc*. Conversely, MPO Explorer's Profile Builder considers all property criteria simultaneously to find those criteria that, in combination, distinguish drugs from non-drugs. Furthermore, the Profile Builder will also tell us whether any of these eight properties are redundant for the objective of classifying a compound as a drug or non-drug.



Set	Mean Improvement (%)	Support (%)
Training	60	22
Validation	57	24



Set	Mean Improvement (%)	Support (%)
Training	51	23
Validation	46	23

Figure 14.31 Example multi-parameter scoring profile derived using the Profile Builder for identifying drug-like compounds when applied to a data set comprising 771 'positive' oral drugs and 1,000 'negative' non-drug compounds randomly selected from ChEMBL. The corresponding predictive performance of these rules over the training and validation sets is also shown. The rules were generated with a minimum support of 20%.

Profile	Desired Value	Importance
Rule 2		
ROTB	⩽ 4.04	
ALOGP	⩽ 2.727	

Set	Mean Improvement (%)	Support (%)
Training	35	57
Validation	35	58

Figure 14.32 Example multi-parameter scoring profile derived using the Profile Builder for identifying drug-like compounds when applied to a data set comprising 771 'positive' oral drugs and 1,000 "negative" non-drug compounds randomly selected from ChEMBL. The corresponding predictive performance of these rules over the training and validation sets is also shown. The rules were generated with a minimum support of 50%.

Figure 14.31 and Figure 14.32 show the rules obtained by the Profile Builder when applied to a data set comprising 771 'positive' oral drugs and 1,000 'negative' non-drug compounds randomly selected from ChEMBL for the objective of identifying oral drugs. The data set was randomly split into training and validation sets containing 70% and 30% of the compounds respectively, and we generated rules using minimum support values of 20% and 50%. Notice that the rules obtained only contain criteria for a subset of the eight properties used to train the algorithm, indicating that the excluded properties do not impart a significant amount of extra information about the objective compared with the subset of properties chosen by the Profile Builder.

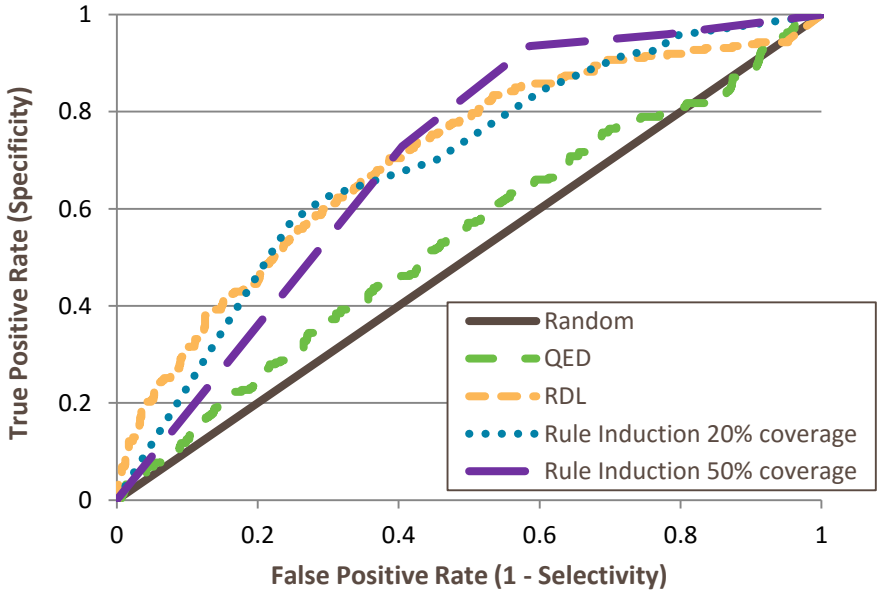


Figure 14.33 ROC plots of the true positive rate (TPR (sensitivity)) against the false positive rate (FPR (1 - specificity)) for the classification of compounds. A perfect classifier would be represented by the point in the top left and a performance below the identity line indicates worse performance than a random classification. A greater area under the curve (AUC) for a classifier indicates higher performance.

One common way to assess the performance of classifiers is a Receiver Operating Characteristic (ROC) curve. Figure 14.33 shows the performance of QED, RDL, and the rules generated by the Profile Builder on the task of differentiating an independent test set of 247 oral drugs from 1000 non-drugs randomly selected from ChEMBL (different from those used to find the rules). Although the Profile Builder rules only specify criteria for a subset of the original eight properties, they are able to match the performance of RDL on this benchmark. Note also that QED performs poorly in this instance with an AUC of just 0.52, showing that the choice of 'negative' set has a substantial impact on the effectiveness of this metric.

14.12 Example 12: Illustrative Application of Derek Nexus to Prioritise Compounds with Lower Potential for Toxicity

In early 'hit-to-lead' it is common to consider a library of compounds, representing multiple chemical series, with the objective to efficiently identify one or more high-quality lead series for progression. This example illustrates the application the Derek Nexus module in StarDrop to prioritise compounds from a library screened for inhibition of the Cyclooxygenase 2 (COX2) enzyme. For full details of this example, please see (Segall & Barber, 2014).

To illustrate one workflow for the practical application of these methods in the context of a hit-to-lead project, we have used a public domain data set, derived from the ChEMBL database (https://www.ebi.ac.uk/chembl/). This data set contains 152 compounds from multiple chemical series for which the inhibition of COX2 enzyme has been determined experimentally, including the drugs Celecoxib and Lumiracoxib. This is typical of a data set containing primary screening data in a hit-to-lead project targeting a fast-follower for an existing drug.

Figure 14.34(a) shows the 'chemical space' of this library, in which the colour of a point represents the score of each compound against the scoring profile shown in Figure 14.35(a), including the experimentally measured target inhibition and a range of predicted ADME properties, but not considering predicted toxicity. This illustrates the distribution of the compound scores across the chemical diversity of the library and indicates that there are three clusters of similar compounds that are likely to yield compounds with a good balance of potency and ADME properties. These high-scoring compounds include the drugs Celecoxib and Lumiracoxib.

The potential for these compounds to cause toxicities were then predicted using the Derek Nexus module for StarDrop for endpoints including mutagenicity, hepatotoxicity and genotoxicity. (Mutagens cause heritable changes to DNA whereas genotoxins damage a cell's genetic material but do not necessarily cause permanent damage to DNA sequences). Figure 14.34(b) shows the prediction of hepatotoxicity mapped onto the chemical space of the COX2 library, which clearly shows that several of the clusters have plausible evidence of hepatotoxicity and should be considered with care. Among those compounds with evidence of hepatotoxicity is Lumiracoxib, which was withdrawn from the market in several countries, mostly due to hepatotoxicity concerns, and has never been approved for use in the United States.

The toxicity predictions can be combined with the *in vitro* and *in silico* data for other properties in an overall scoring profile, shown in Figure 14.35(b), giving appropriate weight to the predictions of toxicity against the other factors. The resulting scores are plotted in the chemical space shown in Figure 14.34(c), in which one cluster clearly stands out as having several compounds with the highest likelihood of yielding a high quality lead series with good ADME properties and reduced chance of toxicity.

It is noteworthy that Celecoxib (the gold-standard COX2 inhibitor) (Moore, Derry, Makinson, & McQuay, 2005) is also identified as having plausible evidence of toxicity, illustrating the importance of balancing the potential for toxicity against the benefits. One advantage of Probabilistic Scoring is that it allows the contributions of each property and the uncertainty in the underlying data to be explicitly taken into account. Therefore, the series including Celecoxib and Lumiracoxib would not be rejected outright. For example, the score for Celecoxib (0.15 ± 0.08) is not statistically significantly different from the top-scoring compound (0.45 ± 0.30). This indicates that a rigorous strategy should select a small number of compounds from this series in order to experimentally confirm the required properties before making a final choice of lead series.

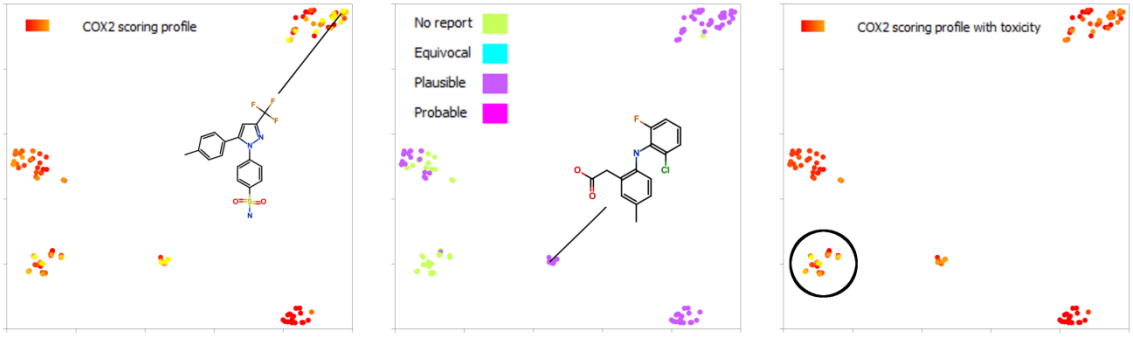


Figure 14.34 These chemical space plots (see Chapter 3) illustrate how predictions of the potential to cause toxicity can be combined with other experimental and predicted data to guide the selection of lead series in early drug discovery. (a) shows the compounds in a library of compounds with COX2 inhibition data containing 5 clusters of similar compounds, coloured by compound score from red (low) to yellow (high). The score was calculated using the profile shown in Figure 14.35(a), taking into account only potency and ADME properties. From this it can be seen that multiple clusters contain compounds with high-scoring compounds. For reference, the point corresponding to Celecoxib is identified. (b) shows the points coloured by predicted likelihood of hepatotoxicity, from which it can be seen that many regions of chemistry are predicted to have increased likelihood of exhibiting hepatotoxicity. The point corresponding to Lumiracoxib, a known hepatotoxin, is highlighted in this plot. In (c), this information is combined with the data for compound potency, predicted ADME properties and predictions for mutagenicity and genotoxicity using the scoring profile shown in Figure 14.35(b). The colours indicate low scoring compounds in red and high scoring compounds in yellow and the cluster containing the majority of high scoring compounds is circled.

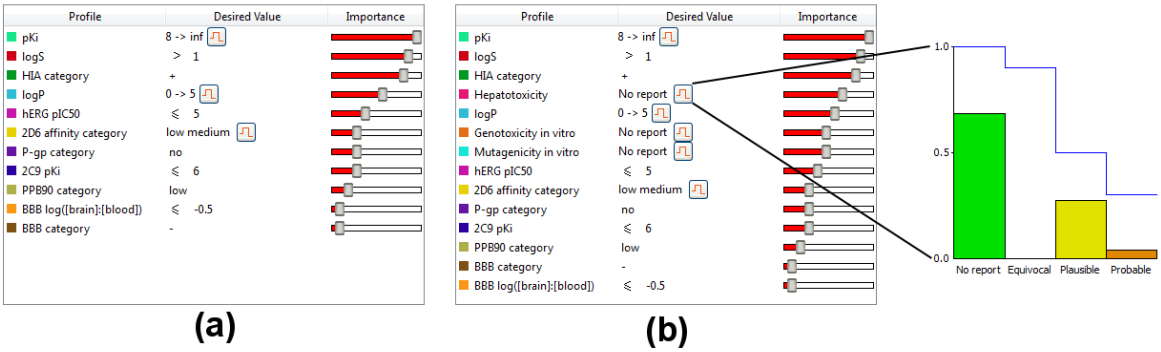


Figure 14.35 Example scoring profiles (see Chapter 2) for a range of experimental and predicted properties and the importance of each individual criterion to the overall objective of the project, specifically an orally dosed compound intended for a peripheral target. (a) shows an example of a profile includes experimental potency against the target and predicted ADME properties. (b) illustrates a profile combining these properties with knowledge-based predictions of toxicity endpoints. Also shown in (b) is an expansion of the criterion for hepatotoxicity, demonstrating how the impacts of different predicted likelihoods for this toxicity on the chance of a compound's success can be reflected by a 'desirability function' shown in blue. On this graph, the desirability of each outcome is shown by the blue line and the scale on the y-axis indicates the desirability on a scale of 0 to 1, where 1 indicates the ideal outcome. The histogram shows the distribution of the different predictions in the current data set.

Finally, considering the structure of Lumiracoxib in Figure 14.36, a single functionality is highlighted as the cause of the structural alert for increased hazard of hepatotoxicity, in common with all other members of this series. This suggests that approaches for reducing the associated risk, while retaining potency and other desirable properties, can be investigated at an early stage before rejecting this class of compounds.

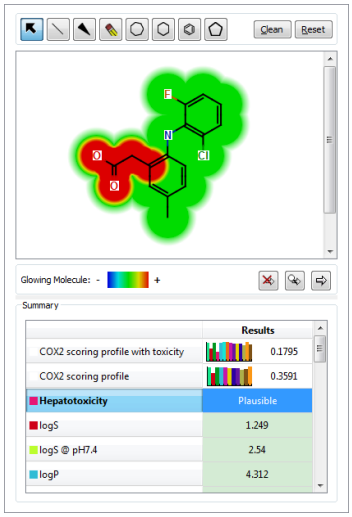


Figure 14.36 StarDrop's interactive designer, in which the structural alert giving rise to the prediction of an increased chance of hepatotoxicity for Lumiracoxib is highlighted by the Glowing Molecule (See Chapter 4). This enables exploration of strategies to reduce toxicity risk while providing instant feedback on the predicted impact of structural changes on multiple, relevant properties.

15 Appendices

15.1 ADME Models Reference

From a chemical structure, StarDrop generates estimates for:

- logP (Octanol/Water)
- $logD_{7.4}$ (Octanol/Water)
- Aqueous Solubility
- Intrinsic Aqueous Solubility (logS)
- Solubility at pH 7.4 ($logS_{7.4}$)
- Human Intestinal Absorption (HIA) Classification
- Blood-Brain Barrier Penetration
- Log([Brain]/[Blood]) (log(BB))
- Classification
- Cytochrome P450 Affinities
- CYP2C9 pKi
- CYP2D6 Classification
- P-gp Transport Classification
- hERG pIC_{50}
- Plasma Protein Binding Classification (90%)

A brief description of each model listed is given below in Table 13. For continuous models, R^2 gives the correlation between calculated and experimental values for the compounds in the external validation set. The RMSE (root mean squared error) is expressed in the same units as the predicted property values. When possible, the RMSE values are calculated for compounds within (IN), in close proximity (CLOSE) or outside (OUT) the chemical space of the model.

For classification models, the accuracy for each class is reported as the percentage of compounds correctly classified. The specificity refers to the percentage of correct classifications within the overall set of compounds predicted to be in that class.

Table 13 Model descriptions.

Model	Name used in StarDrop	Definition	Model statistics on test set
logP (Octanol/Water)	logP	Predicts the logarithm of the octanol/water partition coefficient for neutral compounds.	N = 2950 R^2 = 0.92 $RMSE_{IN}$ = 0.44 log units $RMSE_{OUT}$ = 0.63 log units
$logD_{7.4}$	logD	Predicts the logarithm of the octanol/water partition coefficient for ionised compounds at a fixed pH of 7.4	N = 257 R^2 = 0.88 $RMSE_{IN}$ = 0.58 log units $RMSE_{CLOSE}$ = 0.61 log units $RMSE_{OUT}$ = unknown
Intrinsic Aqueous Solubility (logS)	logs	Predicts the logarithm of the intrinsic aqueous solubility, S in µM, for neutral compounds.	N = 663 R^2 = 0.82 $RMSE_{IN}$ = 0.70 log units $RMSE_{OUT}$ = 1.03 log units
Solubility at pH 7.4 ($logS_{7.4}$)	logS@7.4	Predicts the logarithm of the apparent solubility at pH 7.4, S in µM, for ionised compounds.	N = 96 R^2 = 0.74 $RMSE_{IN}$ = 0.61 log units $RMSE_{OUT}$ = unknown
Blood-Brain Barrier Penetration log([Brain]/[Blood]) (log(BB))	BBB log([brain]:[blood])	Predicts the logarithm of the Brain/Blood ratio.	N = 75 R^2 = 0.72 $RMSE_{IN}$ = 0.36 log units $RMSE_{CLOSE}$ = 0.54 log units $RMSE_{OUT}$ = unknown
hERG pIC50	hERG pIC50	Predicts the pIC50 values for inhibition of hERG K+	N = 33 R^2 = 0.72

		channels expressed in mammalian cells.	RMSE = 0.64 log units
Cytochrome P450 CYP2C9 pKi	2C9 pKi	Predicts the pKi values for affinity with CYP2C9.	N = 25 R^2 = 0.64 $RMSE_{IN}$ = 0.6 log units $RMSE_{OUT}$ = unknown
Human Intestinal Absorption (HIA) Classification	HIA category	Predicts a classification of '+' for compounds which are ≥30% absorbed and '-' for compounds which are <30% absorbed.	N = 245 Accuracy Class '-' = 66% Accuracy Class '+' = 99% Specificity Class '-' = 91% Specificity Class '+' = 95%
Blood-Brain Barrier Penetration Classification	BBB category	Predicts a classification of '+' for compounds which have a log([brain]:[blood])≥-0.5 and '-' for compounds which have a ratio <-0.5.	N = 52 Accuracy Class '-' = 91% Accuracy Class '+' = 83% Specificity Class '-' = 91% Specificity Class '+' =83%
P-gp Transport Classification	P-gp category	Predicts a classification of 'yes' for substrates and 'no' for non-substrates.	N = 51 Accuracy Class 'yes' = 79% Accuracy Class 'no' = 82% Specificity Class 'yes' = 85% Specificity Class 'no' = 75%
Cytochrome P450 CYP2D6 Classification	2D6 affinity category	Predicts a classification of 'low' for compounds with a pKi<5, 'medium' for compounds with a pKi between 5 and 6, 'high' for compounds with a pKi between 6 and 7, and 'very high' for compounds with a pKi>7.	N = 45 Root mean class error = 0.87 classes
Plasma Protein Binding Classification (90%)	PPB90 category	Predicts a classification of 'low' for compounds which are <90% bound and 'high' for compounds which are >90% bound.	N=159 Accuracy Class 'high' = 81% Accuracy Class 'low' = 87% Specificity Class 'high' = 74% Specificity Class 'low' = 91%

15.1.1 logP (octanol/water)

logP, the logarithm of the octanol/water partition coefficient, gives a measure of the lipophilicity of a compound. Lipophilicity is an important property of a drug molecule as it influences a number of physiological properties including transport through cell membranes, rate of metabolism and interaction with receptor binding sites.

Data set

The StarDrop logP model is based on a large dataset of over 9,000 experimental octanol/water partition coefficient values obtained from the Medchem database (Daylight, n.d.) (Leo, 1993). The logP values are the most comprehensive and reliable source of logP data and most *in silico* models that predict logP are based on this data.

Model output

The logP model is based on over 100 2D-descriptors including atom and functionality counts. The model was trained on 6,887 compounds using the Radial Basis Function technique (Section 8.6), a widely used algorithm for supervised learning. The model was validated on a test set of 2950 compounds, on which it achieved an excellent R^2 value of 0.92 between observed and predicted values (see Figure 15.13).

As mentioned above, the model calculates the distance of each new compound from the descriptor-space of the training set, to gauge the validity of the results. Predictions for compounds within the chemical space of the model are reported with an RMSE value of 0.44 log units. Predicted logP values for compounds outside of the chemical space have an RMSE of 0.63 log units.

Table 14 further demonstrates the accuracy of the model for neutral compounds and compounds that accept or donate protons at physiological pH.

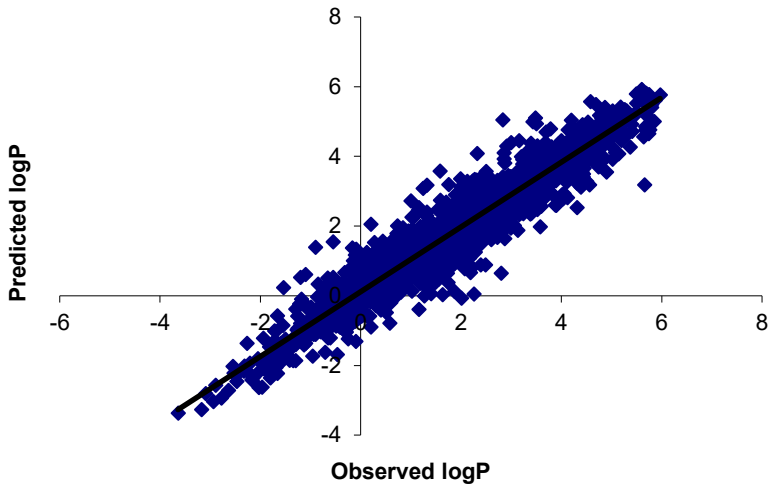


Figure 15.1 Plot of observed versus predicted logP for the independent test set.

Table 14 logP test set results broken down by compound type.

	Distribution of compounds within classes (%)		Statistical results for test compounds	
	Training set	Test set	R^2	RMSE
All compounds	100	100	0.92	0.44
Neutral compounds	64	63	0.91	0.45
Acidic compounds	5	5	0.89	0.43
Monobasic compounds	17	17	0.92	0.42
Polybasic compounds	5	6	0.88	0.46
Phenolic compounds	6	7	0.95	0.39
Zwitterionic compounds	2	2	0.91	0.54

Comparison with other predictive techniques

Models for logP are commonly used by pharmaceutical companies and there are several available commercially. It is difficult to rigorously assess the predictive power of competitors' models, as we do not know which compounds were used to train the various algorithms and which compounds represent true tests of their predictive power. This typically leads to an overestimate of a competitor models' performance. However, despite the limitations of this evaluation uncertainty, no competitor model has outperformed StarDrop's model in our tests.

15.1.2 $logD_{7.4}$

The distribution coefficient, D, provides a measure of lipophilicity at a physiologically relevant pH. In contrast to the partition coefficient P that refers to the concentration ratio of neutral species, D is defined as the sum of the concentration of all charge-state forms of a substance dissolved in the lipid phase, octanol, divided by the sum of those dissolved in water at a chosen pH. For this reason, logD is much more suitable parameter for correlating drug biological action, since it takes into account drug ionisation at a relevant pH.

This model predicts logD values at pH 7.4 at which many molecules exist in partially dissociated or ionised form.

A set of rules was defined to identify neutral or uncharged molecules at pH 7.4 for which $logD_{7.4}$ is equal to logP. In these cases, the prediction will be generated by the logP model described in Section 15.1.1.

Data set

The StarDrop $logD_{7.4}$ model is based on a dataset of 857 experimental octanol/water distribution coefficient values at pH 7.4, $logD_{7.4}$ (Avdeef, 2003). The majority of the logD values were obtained from StARLITe (Now ChEMBLdb https://www.ebi.ac.uk/chembldb/), a database containing medicinal chemistry and pharmacological data from two key peer-reviewed journals: Journal of Medicinal Chemistry (1980-2004) and Bio-organic Medicinal Chemical Letters (1991-2004).

Classification of the $logD_{7.4}$ set into acid, base, zwitterions and neutral categories was done according to proprietary SMARTS definitions of sub-structures likely to be ionised at pH 7.4. The $logD_{7.4}$ set was composed of 105 acidic, 684 basic and 68 zwitterionic molecules.

Model output

The $logD_{7.4}$ model was built by the automatic procedure implemented within the Auto-Modeller using standard settings. The initial dataset was split into three subsets using cluster analysis at Tanimoto level 0.7. The model was trained on 601 compounds and evaluated on validation set of 127 compounds and test set of 130 compounds. The $logD_{7.4}$ model was built using the Radial Basis Function technique (Section 8.6) using 173 2D-descriptors including atom and functionality counts. The logP descriptor was not used.

The predictive model for $logD_{7.4}$ was evaluated on the validation set, on which it achieved an excellent R^2 value of 0.88 and an RMSE value of 0.65 log units, and on the test set with $R^2 = 0.86$ and RMSE = 0.68 log units. On the combined validation and test sets the statistics were $R^2 = 0.88$ and RMSE = 0.67 log units. Figure 15.2 shows the observed versus predicted $logD_{7.4}$ values for the validation and test compounds.

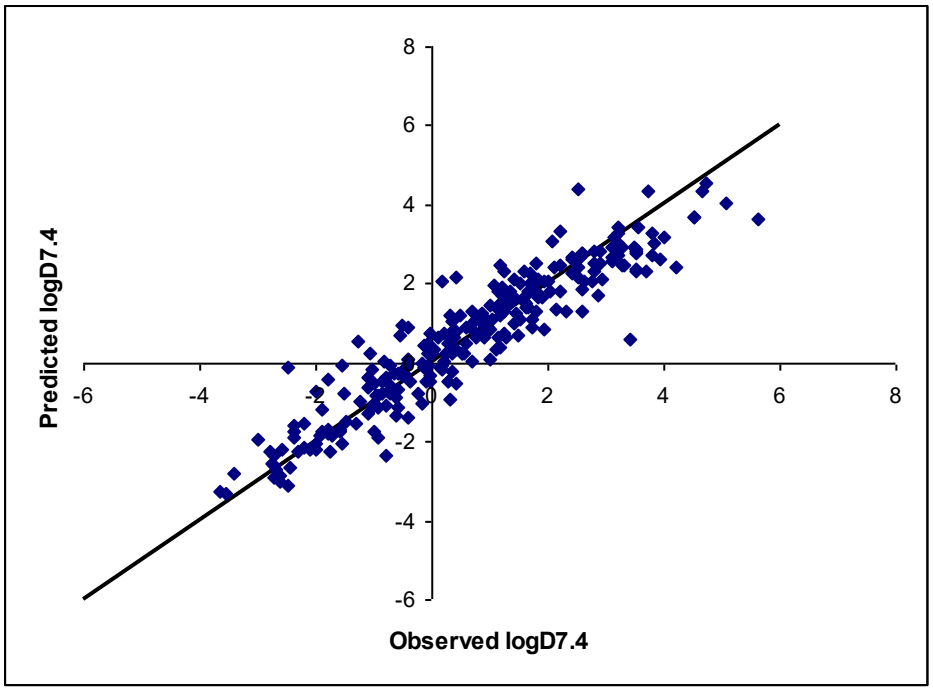


Figure 15.2 Plot of observed versus predicted $logD_{7.4}$ on the combined validation and test sets.

The performance of the model was further evaluated across several pre-defined chemical classes; the results are shown in Table 15.

The distance of each predicted compound from the descriptor-space of the training set, referred to as the chemical space of the model, is calculated in order to gauge the validity of the results. Predictions for compounds within the chemical space have an RMSE in prediction of 0.58 log units. Predictions for compounds outside, but in close proximity to, the chemical space have an RMSE of 0.61 log units. For compounds outside the chemical space the standard error in prediction is undefined (returned in the software as infinity) to indicate that the prediction must be treated with caution.

Table 15 Performance of the $logD_{7.4}$ model on groups of ionised compounds from combined validation and test sets of the model.

Class	Number of Compounds	RMSE
Acid	32	0.47
Base	200	0.68
Zwitterion	25	0.82

Comparison with other predictive techniques

The majority of pharmaceutical companies employ some form of model for octanol/water distribution coefficient at pH 7.4 and there are several available commercially. It is difficult to rigorously assess the predictive power of competitors' models because we do not know which compounds were used to train the various algorithms and which compounds represent true tests of their predictive power. This typically leads to an overestimate of a competitor models' performance. However, despite the limitations of this evaluation uncertainty, this model has significantly outperformed the predictive model for $logD_{7.4}$ values in SciTegic's Pipeline Pilot (version 6.1.1.0, now Accelrys). Poor statistics, R^2 = 0, r^2_{corr}=0.32 and RMSE = 2 log units, were obtained on the combined validation and test sets of 257 compounds.

Identification of neutral or uncharged compounds at pH 7.4

A set of proprietary SMARTS definitions was used to identify compounds likely to be neutral or uncharged at pH 7.4. Acidic and basic functionalities known to be fully or partially ionised at that pH

were coded up using SMARTS definition. If a compound contains none of the acidic or basic functionalities, it is then considered to be a neutral compound or uncharged at pH 7.4.

If a compound is identified as neutral or uncharged then logP will be used to predict $logD_{7.4}$.

Acidic functionalities:

Acid [OX2v2D1][C,P,B,S](=O)
Acid1 [OX2v2D1]C=CC=O
Acid2 [OX2v2D1][c][c][c]=O
diketone [CX4&!H0](C(=O))C=O
tetrazole [nH]1nnnc1
saccharin C(=O)[NX3&!H0]S(=O)=O
hydroxymicAcid [C](=O)[N&!H0][OX2v2D1]

Basic Functionalities:

nh2 [NH2][CX4,a]
nh1 [NH1]([CX4,a])[CX4]
nh0 [NH0]([CX4,a])([CX4])[CX4]
amidine NC=N
quatNitrogen [+]
hydrazine1 [N&!H0][N][CX4]
hydrazine2 [CX4][N][N][CX4]

15.1.3 Aqueous Solubility

Of all properties that determine a drug's ultimate *in vivo* ADME behaviour, solubility is one of the most important and deserves close attention in early discovery. Indeed, a drug's propensity to dissolve in aqueous media is a key property affecting its administration and absorption. Currently, most HTS *in vitro* solubility assays are in 2-5% DMSO/buffer, which does not necessarily correlate well with aqueous solubility. Therefore, the ability to predict aqueous solubility is important for early identification of compounds that are less likely to pose future difficulties in formulation and administration.

StarDrop has two models that predict aqueous solubility - a logS model to predict intrinsic water solubility and a $logS_{7.4}$ model to predict apparent solubility of charged compounds at pH 7.4.

15.1.4 logS

The logS model predicts intrinsic water solubility, i.e. solubility for uncharged compounds in water.

Data set

The StarDrop model is based on more than 3,300 aqueous solubility data points for intrinsic water solubility, S in μM, defined as the thermodynamic solubility of uncharged compound in water between 20-30°C. The data comes from the Syracuse database (Butina & Gola, Modeling aqueous solubility, 2003). It is noteworthy that most *in silico* models for prediction of intrinsic aqueous solubility are based on the same commercial database.

Model output

The logS model uses over 100 2D-descriptors indicating compound size and counts of different atomic or functional groups or specific fragments. The model was trained on 2650 compounds using the Radial Basis Function technique (Section 8.6).

The model was validated on a test set of 663 compounds and observed and predicted values for this set are well correlated, with an R^2 value of 0.82 (see Figure 15.3). The performance of the model was further evaluated across several pre-defined chemical classes and, as can be seen from Table 16, the model gives consistently good R^2 and RMSE values throughout the classes.

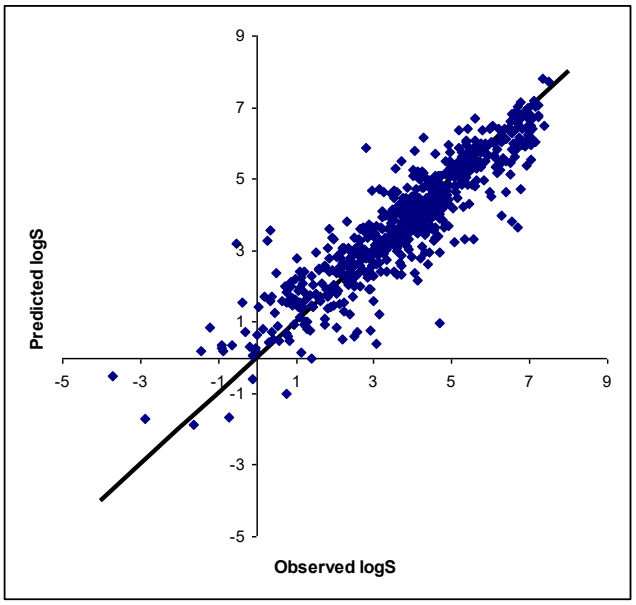


Figure 15.3 Plot of observed versus predicted logS (S in μM) for the test set.

As mentioned above, the distance of each new compound from the descriptor-space of the training set is calculated in order to gauge the validity of the results. Predictions for compounds within the chemical space of the model have an RMSE in prediction of 0.70 log units. The estimated logS values for compounds outside the chemical space have an RMSE of 1.03 log units.

Table 16 logS test set results broken down by compound type.

	Distribution of compounds within classes (%)		Statistical results on test set	
	Training set	Test set	R^2	RMSE
All compounds	100	100	0.81	0.80
Neutral compounds	60	64	0.82	0.78
Acidic compounds	14	12	0.77	0.70
Monobasic compounds	11	8	0.85	0.95
Polybasic compounds	4	3	0.73	1.06
Phenolic compounds	8	8	0.78	0.83
Zwitterionic compounds	3	4	0.83	0.65

Comparison with solubility assays

As previously mentioned, HTS *in vitro* solubility assays in 2-5% DMSO/buffer do not necessarily correlate well with intrinsic aqueous solubility. Figure 15.4 illustrates that there is no visible relationship between the two solubility measurements. Indeed, all of the compounds in the set chosen here with poor aqueous solubility, i.e. S<12.5 µM, have a very high 2% DMSO/buffer solubility, S> 100 µM.

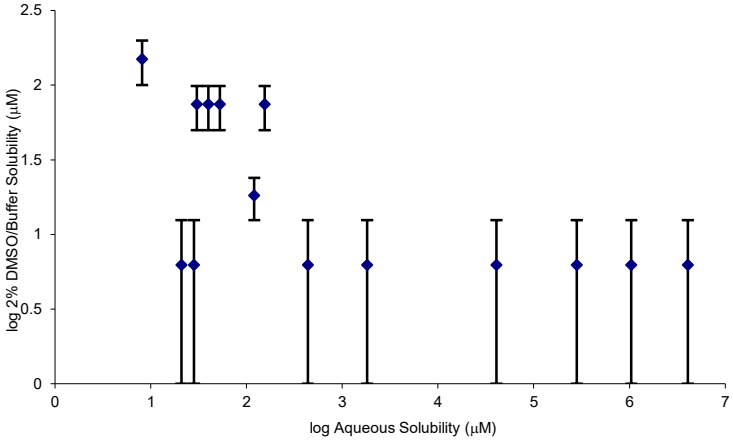


Figure 15.4 Comparison between observed intrinsic aqueous solubility and observed, 2% DMSO/buffer solubility. The error bars represent the values that are encompassed by this measurement.

15.1.5 Solubility at pH 7.4 ($logS_{7.4}$)

The $logS_{7.4}$ model predicts apparent solubility of ionised compounds at pH 7.4. At physiological pH many drug-like compounds exist in partially dissociated or ionised form.

A set of rules was defined to identify neutral or uncharged molecules at pH 7.4 for which $logS_{7.4}$ is equal to logS. In these cases, the prediction will be generated by the logS model described in Section 15.1.4.

Data set

A compilation of high quality solubility data measured in buffered solution at pH 7.4 ($logS_{7.4}$ with $S_{7.4}$ in µM) was gathered from BioFocus DPI's StARLITe database (Now ChEMBLdb https://www.ebi.ac.uk/chembldb/). Only those measurements that were determined between 25°C and 35°C were considered.

The StarDrop model is based on 322 charged drug-like compounds.

Model output

The $logS_{7.4}$ model was built by the automatic procedure implemented within the Auto-Modeller using standard settings. The initial dataset was split into three subsets using cluster analysis at Tanimoto level 0.7. The model was trained on 226 compounds and evaluated on validation and test sets of 48 compounds each. The best model was produced by the Radial Basis Function technique coupled with a genetic algorithm for descriptor selection (GA-RBF). The $logS_{7.4}$ model is based on 28 2D-descriptors measuring compound lipophilicity, negative charge and counts of different atomic and functional groups and specific fragments.

The predictive model was tested on the validation and test sets with a R^2 value of 0.74 and an RMSE of 0.61 log units on the combined sets (see Figure 15.5).

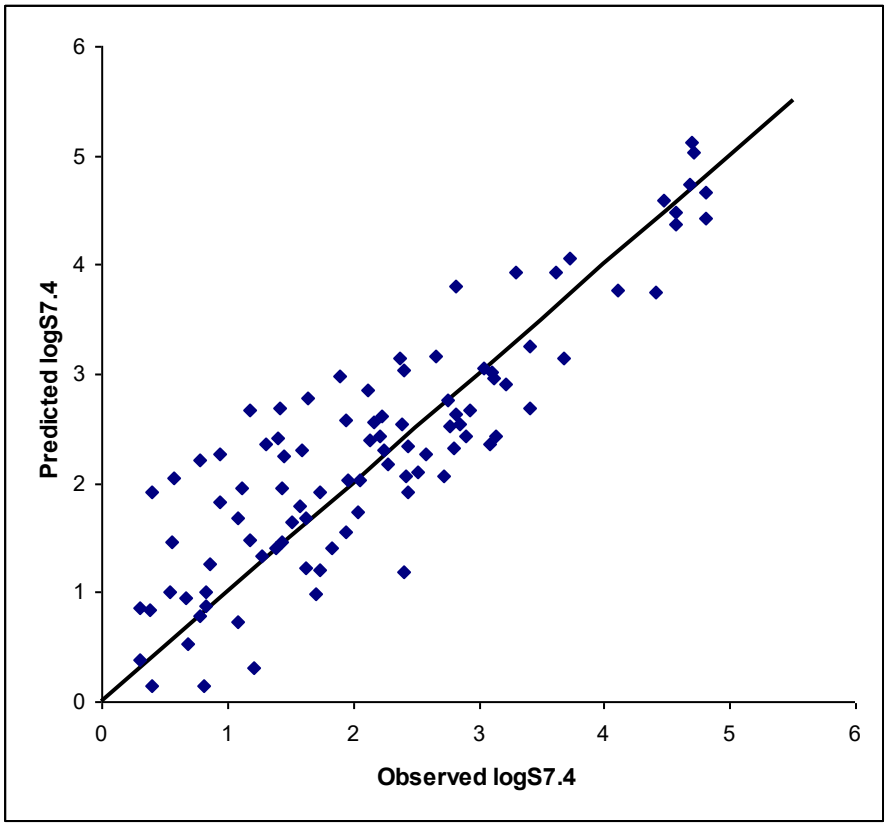


Figure 15.5 Plot of observed versus predicted $logS_{7.4}$ (S in µM) on combined validation and test sets.

The performance of the model was further evaluated across several pre-defined chemical classes and, as can be seen from
Table 17, the model gives consistently good R^2 and RMSE values throughout the classes.

Table 17 Performance of the $logS_{7.4}$ model on classes of ionized compounds from combined validation and test sets of the model.

	Distribution of compounds within classes (%)	R^2	r^2_{corr}	RMSE
All compounds	100	0.74	0.76	0.61
Acidic compounds	14	0.77	0.81	0.41
Basic compounds	54	0.74	0.75	0.60
Zwitterionic compounds	32	0.35	0.55	0.69

The distance of each predicted compound from the descriptor-space of the training set, referred to as the chemical space of the model, is calculated in order to gauge the accuracy of the results. Predictions

for compounds within the chemical space of the model have an RMSE in prediction of 0.62 log units. For compounds outside the chemical space the standard error in prediction is undefined (returned in the software as infinity) to indicate that the prediction must be treated with caution.

Comparison of logS and $logS_{7.4}$ with other predictive techniques

The majority of companies employ some form of models for aqueous solubility and there are several available commercially. It is difficult to rigorously assess the predictive power of competitors' models without knowing which compounds were used to train the algorithms and thus which compounds represent true tests of their predictive power. This typically leads to an overestimate of a competitor models' performance. Commercially-available solubility models and published literature models were recently reviewed and compared by Dearden (Dearden, 2006) and Schwaighofer *et al*. (Schwaighofer, et al., 2007). The latter paper also reports models predicting apparent solubility at pH 7.4 and pure aqueous solubility which were built by Gaussian Processes methods. The apparent solubility at pH 7.4 model achieved RMSE=0.77 log units and the pure solubility model achieved RMSE=0.61 log units (both measures were obtained by cross-validation).

15.1.6 Human Intestinal Absorption (HIA) Classification

The majority of high-value drugs on the market are orally administered. For this reason, a great deal of research has been carried out in attempts to predict human intestinal absorption (HIA) of compounds early in the drug discovery process. *In vivo* and *in vitro* models have been intensively used to estimate HIA, but these alternatives are costly, resource intensive and often difficult to interpret. Computational methods have been developed to overcome these hurdles and StarDrop has developed a classification model to identify compounds with good absorption based on a set of meaningful descriptors.

Data set

Percent Human Intestinal Absorption (%HIA) was used to build this particular *in silico* model. %HIA is defined as the percentage of orally administered drug reaching the hepatic portal vein. StarDrop's model is based on a dataset of over 250 compounds for which %HIA were reported in the literature. The major drawback with this data set is that it is highly biased towards well-absorbed compounds. However, most of the available *in silico* models for Human Intestinal Absorption prediction are based on the same dataset used at StarDrop (Zhao, et al., 2001).

The model was further tested on 245 proprietary data points.

Model output

By far the most common mechanism of absorption from the gastrointestinal tract is passive diffusion through the intestinal epithelial cells. This process depends heavily on the solute's ability to diffuse through the lipophilic phospholipids of the cellular membrane. In turn, the diffusion depends on the solvation/desolvation processes, on the surface interaction between solute and membrane, and on the H-bond potential of the solute. Based on this knowledge descriptors representing properties (e.g. hydrogen bond donors, hydrogen bond acceptors and size of the molecule) that favour passive transport process through the membrane were selected. The influence of the most important descriptors in the model are shown in Figure 15.7.

The current model for passive absorption across the intestinal wall provides either a '+' or '-' answer, indicating either >=30% passive absorption or <30% absorption respectively. The model had an overall classification rate of 96 % on the training set. Optimum results were obtained with 99% of '+' and 77% of '-' correctly predicted. The performance of this classifier was assessed on a test set, of which 99% of '+' and 66% of '-' were correctly classified.

A confidence for each prediction is reported, according to the strength of association of the compound's descriptor values with the predicted classification. Furthermore, the distance of the predicted compound from the chemical space of the training set is calculated in order to gauge our confidence in the result. As there are insufficient data points outside the chemical space of the training set to assess the confidence in predictions, no estimate regarding the confidence for such compounds can be made. In these cases, the probability that the result is correct is reported as 0.5, indicating an even distribution between the two possible classes.

Comparison with other predictive techniques

On a number of occasions we have compared the results from the model to those generated *in vitro*, using Caco-2 cells, and found it to be more predictive of human absorption. Because of the poor relationship between Caco-2 P_{app} values and human intestinal absorption and the variation in fit to the relationship, absolute absorption values are difficult to estimate from Caco-2 P_{app} data (see Figure 15.6) (Irvine, et al., 1999). Furthermore, this *in vitro* experiment is resource intensive and, of course, requires the synthesis and purification of the compounds to be tested, while the *in silico* model requires only the virtual structure.

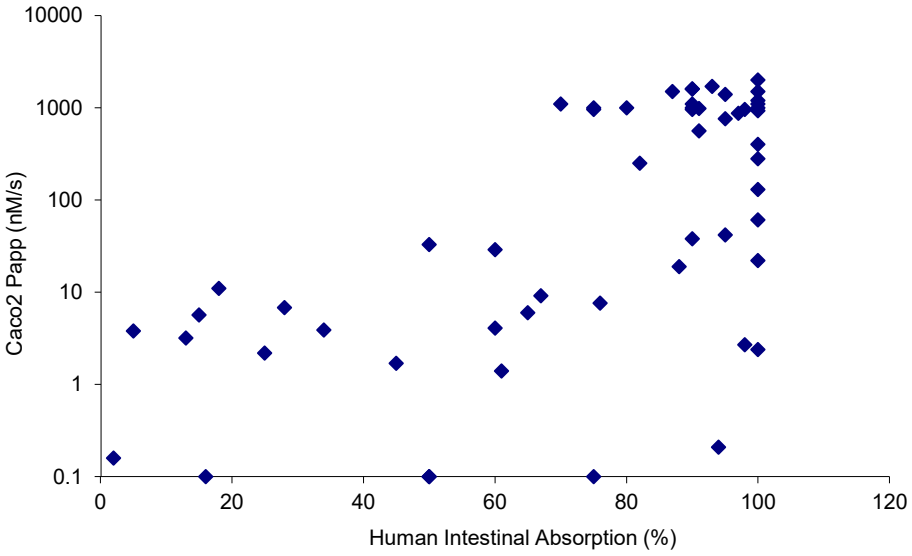


Figure 15.6 Correlation between Caco-2 P_{app} and %HIA.

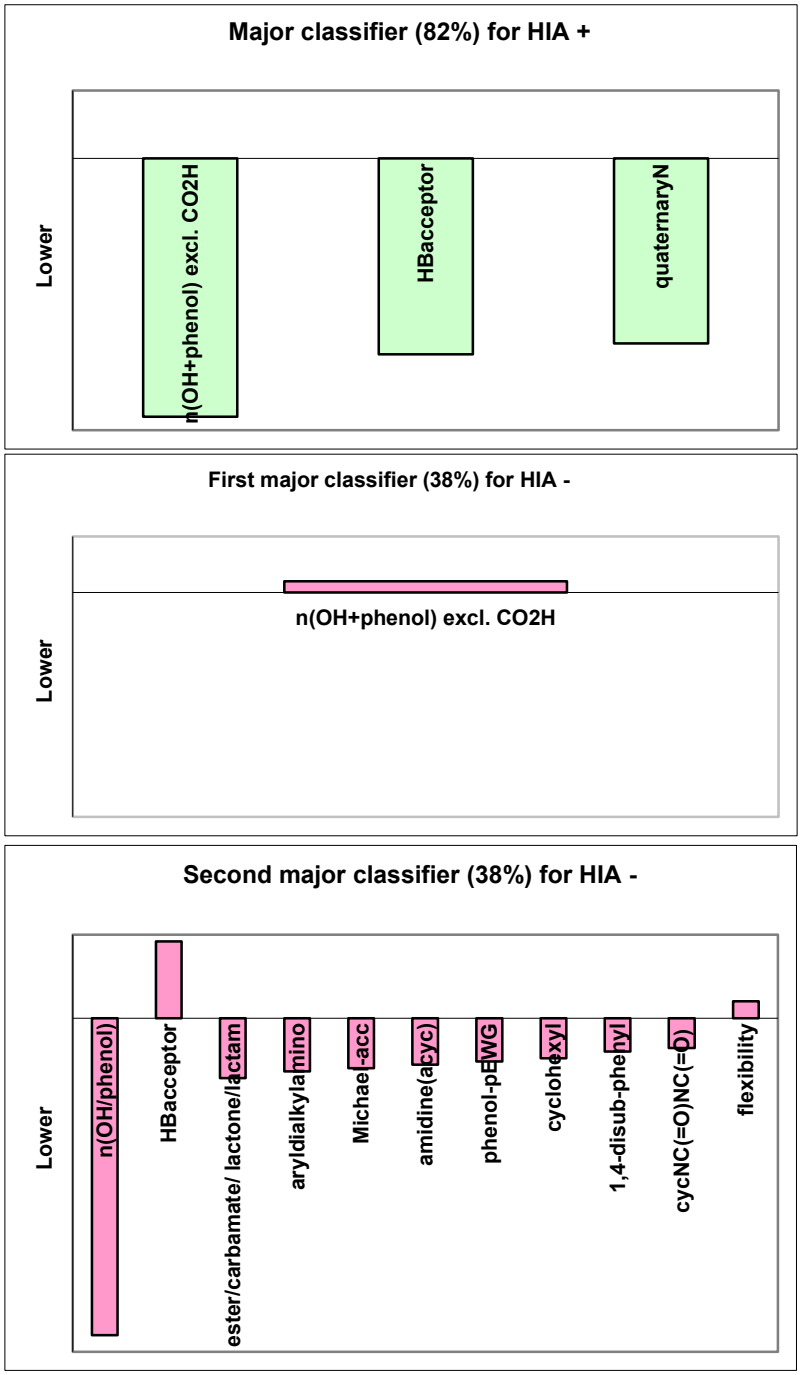


Figure 15.7 Histograms showing the influence of descriptors on the dominant rules of the HIA model. The directions of bars relative to the horizontal axis indicate whether the value of a descriptor must be higher or lower than a threshold. The length of a bar reflects the number of compounds retained if the condition is met. The vertical scale is uniform for all plots. Percentages in parentheses refer to the proportion of the class predictions made by the rule.

15.1.7 Blood-brain Barrier (BBB) Penetration

The ability to predict blood-brain barrier penetration is very important in drug development. For CNS therapeutic targets, good penetration is an absolute requirement, but for non-CNS targets blood-brain barrier penetration is undesirable, as it is a potential cause of side-effects.

StarDrop has two blood-brain barrier penetration models, log([brain]/[blood]) and a classification model, both of which have been included to allow determination of a consensus score. This approach provides a higher level of confidence because the models were developed independently. Hence,

compounds predicted to cross the blood-brain barrier by both models will have a higher consensus score than those only predicted to cross the blood-brain barrier by one model.

15.1.8 Log ([brain]/[blood])

Data set
The data set consists of 509 structures with a reported logarithm of the concentration ratio between brain tissue and plasma (log(BB)) which were derived from various literature sources: Abraham *et al* (Abraham, Ibrahim, Zhao, & Acree, 2006), Vilar *et al* (Vilar, Chakrabarti, & Costanzi, 2010) and Chico *et al* (Chico, Van Eldick, & Watterson, 2009). The model was trained on 70% of the compounds, with 15% saved for each of the validation and test sets.

Model output
The model was built by the automatic procedure implemented within the Auto-Modeller using the standard settings. The initial set was split into a training set (359), validation set (75) and test set (75) by using cluster analysis at Tanimoto level 0.7. The model was produced by the non-linear Radial Basis Function technique combined with a genetic algorithm to assist in descriptor selection (GA-RBF). The model uses 36 descriptors including logP, McGowan's volume, negative charge, polar surface area, hydrogen bond donors and counts of different atomic and functional groups.

The model predicts the log(BB) value for each compound, along with an estimate of the RMSE in prediction. The distance of each predicted compound from the descriptor-space of the training set, referred to as the chemical space of the model, is calculated in order to gauge the validity of the results. The model automatically determines whether or not a test compound lies within the chemical space. When a test compound lies outside the chemical space a prediction is returned, but the standard error in prediction is left as undefined (returned in the software as infinity) to indicate that the prediction must be treated with caution. The RMSE in prediction for compounds within the chemical space is 0.36 log units and the RMSE in prediction for compounds outside, but in close proximity to, the chemical space is 0.54 log units.

It is a feature of the RBF technique that it will usually provide a perfect fit for the training set. However, on the test set the model achieves an R^2 of 0.72 with an RMSE of prediction of 0.36 log units (see Figure 15.8).

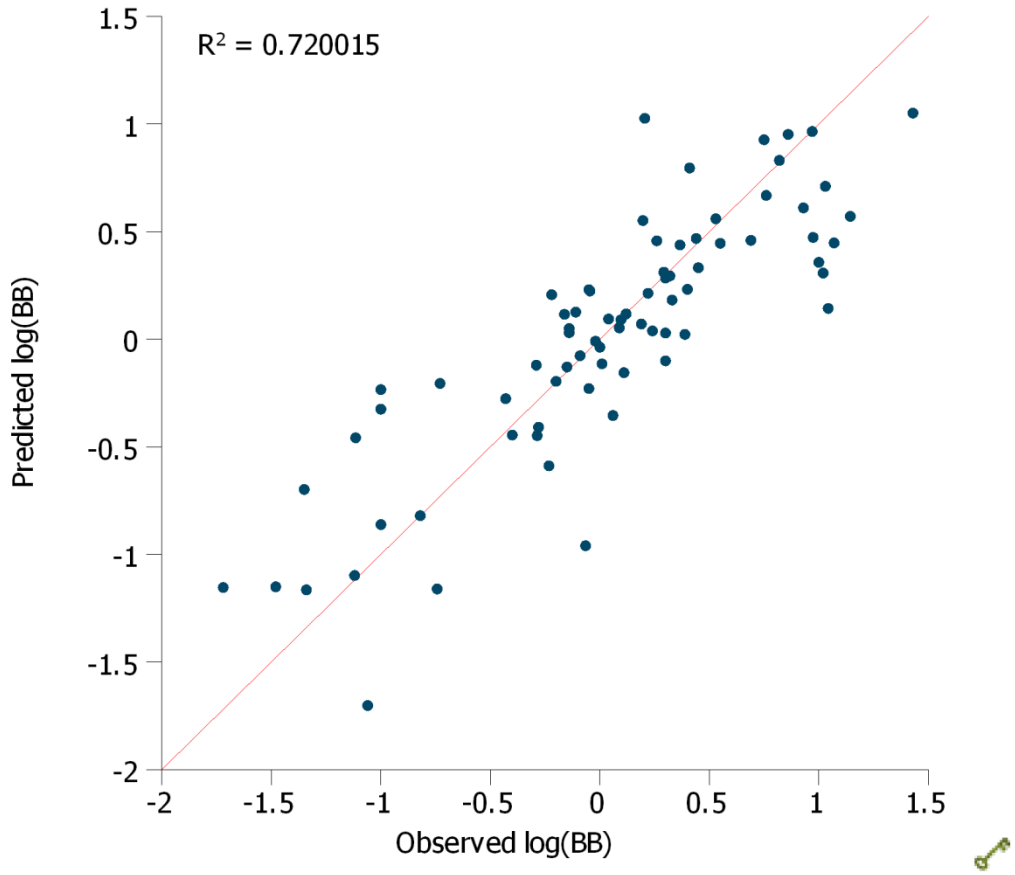


Figure 15.8 Plot of observed versus predicted log(BB) values for the combined validation and test sets

Comparison with other predictive techniques

Abraham and Hersey (Abraham, Hersey, Testa, & H., 2006) reviewed published continuous blood-brain barrier penetration models and concluded that a number of models can predict log(BB) values with an RMSE error of 0.3-0.35 log units, as also shown by Abraham *et al* (Abraham, Ibrahim, Zhao, & Acree, 2006) The estimated experimental error in log(BB) measurements is approximately 0.3 log units. Therefore, the RMSE of the StarDrop model compares well with published models.

15.1.9 BBB Classification

Data set

The data set consists of 201 structures classified as BBB+ and BBB- that are reported in literature models. This data was divided into a training set containing 101 compounds with an even distribution between BBB+ and BBB- compounds and an internal evaluation set of 48 compounds, with a 3.5:1 ratio between BBB+ and BBB- compounds, which was used to monitor the training of the model. The remaining 52 structures were utilized as an independent test set with a 1:2 ratio of BBB+ and BBB- compounds.

Model output

The model is a random forest classification model which uses descriptors that are consistent with the general observations that neutral molecules tend to penetrate the CNS better than charged compounds and that cations generally penetrate the CNS better than anions.

The model generates a prediction for each compound as BBB crossing (BBB+) or non-crossing (BBB-). This is based on a nominal classification boundary of log(BB)=-0.5 between BBB- and BBB+ compounds. For the independent test set, 91% of BBB- predictions were correct in relation to the known category, whereas BBB+ predictions were correct in 83% of cases.

A confidence for each prediction is reported, according to the strength of association of the compound's descriptor values with the predicted classification. Furthermore, the distance of the predicted compound from the chemical space of the training set is calculated to gauge the confidence in the result. As there are insufficient data points outside the chemical space of the training set to assess the confidence in predictions, no estimate regarding the confidence for such compounds can be made. In these cases, the probability that the result is correct is reported as 0.5, indicating an even distribution between the two possible classes.

Comparison with other predictive techniques

The model statistics compare well to recent literature BBB classification models (Crivori, Cruciani, Carrupt, & Testa, 2000) (Ajay, Bemis, & Murcko, 1999) (Engkvist, Wrede, & Rester, 2003) (Keseru, Molnar, & Greiner, 2000) (Doniger, Hofmann, & Yeh, 2002) where BBB+ prediction accuracy ranges from 80% to 100% and BBB- prediction accuracies lie between 65% and 87%.

15.1.10 Cytochrome P450 Affinities

The Cytochromes P450 are a superfamily of metabolic enzyme present in a wide range of organs, and cells (Danielson, 2002). In particular, phase I metabolism by P450s in the liver is a major route of clearance for many drug compounds and, in some cases, may result in bioactivation, forming toxic metabolites. There are a large number of isoforms, each with different substrate specificities, distributions in the body and rates of metabolism.

StarDrop contains models of affinity for P450 isoforms CYP2D6 and CYP2C9, two of the three most significant drug-metabolising enzymes, along with CYP3A4. Although metabolism of drugs via these enzymes is not as common as for CYP3A4, interaction with CYP2D6 or CYP2C9 is a significant cause of drug-drug interactions, due to inhibition of clearance of another drug primarily metabolised by the same enzyme. In addition, interaction with CYP2D6 is also considered to be unfavourable as it is the subject of a well-known genetic polymorphism, resulting in approximately 10% of the Caucasian population having a 'poor metabolise' phenotype in which activity of CYP2D6 is dramatically reduced.

The affinity of a ligand is defined by the K_i, the molar concentration required to occupy half the binding sites available to a competitor ligand, in the absence of radioligand or competitors (if the K_i value is low then the affinity is high). This is commonly reported as a pK_i value (i.e. $\log_{10}(1/Ki)$ with K_i in M). In this case, the greater the pK_i value, the higher the affinity. It should be noted that a high affinity does not necessarily indicate that a compound will be metabolised by the particular P450. Conversely, very low affinity compounds are unlikely to be significantly turned over by the enzyme.

15.1.11 CYP2C9 pKi

Data set

The data for this model were generated in house, due to the high inter-laboratory variation observed in reported P450 affinities in the literature. The data consist of accurate K_i values generated for competitive inhibitors using a multi-point K_i protocol. Data for a total of 130 compounds were generated in this data set covering a wide range of chemical diversity.

Model output

A continuous random forest model for CYP2C9 inhibition was developed. The model predicts a compound's pK_i and also produces an estimate of the RMSE in prediction. The model automatically determines whether or not a test set compound lies within the chemical space formed by the training set. As there are insufficient compounds available outside the training set chemical space, no rigorous estimates regarding the confidence for such compounds can be made. In these cases, a prediction is returned, but the standard error in prediction is undefined (shown as infinity).

The observed R^2 for the training set of 105 compounds was 0.92 and the RMSE in fit was 0.33 log units. The R^2 value for the independent test set of 25 compounds was 0.64 (see Figure 15.9) and the standard error in prediction was 0.60 log units.

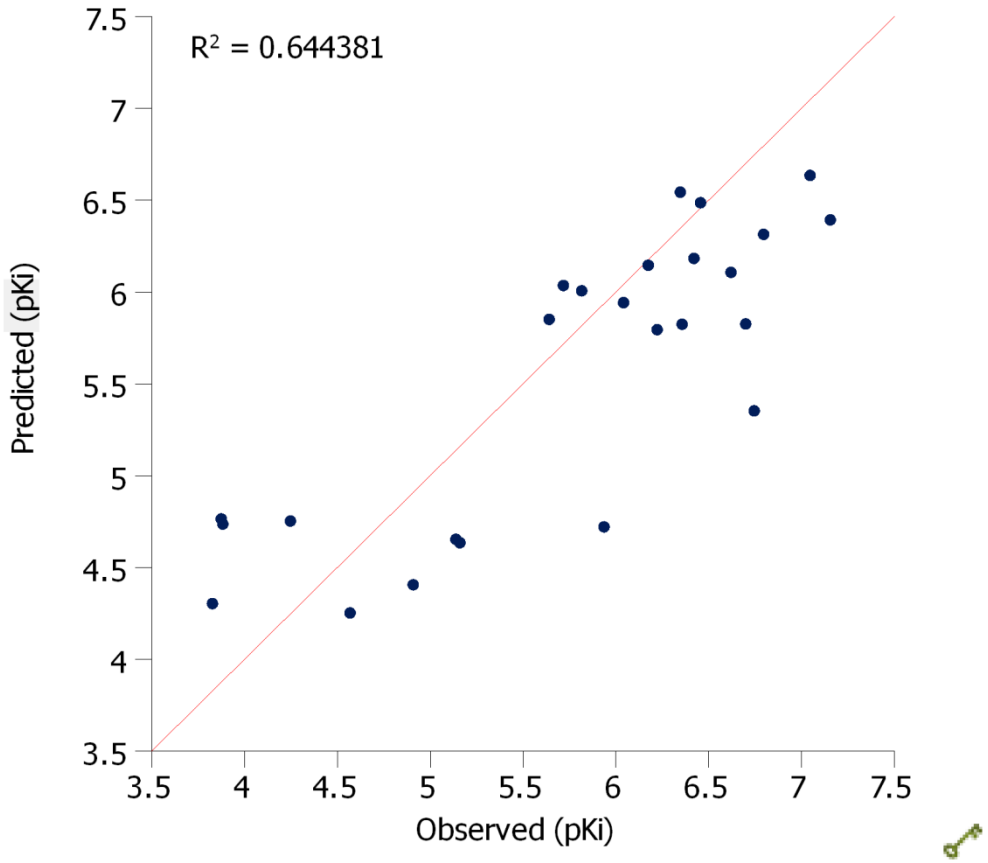


Figure 15.9 Observed versus predicted pK_i for CYP2C9 Affinity.

Comparison with other predictive techniques

There has been significant published work on quantitative structure-activity relationships for affinity to CYP2C9; in particular by David Lewis *et al*. (Lewis D. F., Essential requirements for substrate binding affinity and selectivity toward human CYP2 family enzymes, 2003) (Lewis D. F., On the recognition of mammalian microsomal cytochrome P450 substrates and their characteristics: towards the prediction of human p450 substrate specificity and metabolism, 2000) (Lewis, Modi, & Dickins, Structure-activity relationship for human cytochrome P450 substrates and inhibitors, 2002) and Sean Ekins *et al*. (Ekins, de Groot, & Jones, Pharmacophore and three-dimensional quantitative structure activity relationship methods for modeling cytochrome p450 active sites, 2001). The majority of the *in silico* models proposed for identifying compounds with high CYP2C9 affinity are based on few training cases and require 3D structures (Afzelius, et al., 2004).

15.1.12 CYP2D6 Classification

Data set

The data for this model were generated in house, due to the high inter-laboratory variation observed in reported P450 affinities in the literature. The data consist of accurate K_i values generated for competitive inhibitors using a multi-point K_i protocol. A total of 213 data points were generated in this data set.

Due to an uneven distribution of K_i values in the data set, and uncertainty regarding the purity of some compounds, the CYP2D6 affinity data were classified into 4 categories; low (pK_i<5), medium (5=<pK_i<6), high (6=<pK_i<7) and very high (pK_i>=7).

Model output

The current model used thirteen 2D descriptors including the lipophilicity term, logP, a flexibility index and the molecular weight. The remaining descriptors are related to more specific functionalities.

The major discriminator of the model is logP. Compounds with poor lipophilicity are unlikely to inhibit CYP2D6, likewise for more lipophilic molecules bearing more than two electron-donating groups or having the potential to form intramolecular hydrogen bonds. Compounds with high affinity for CYP2D6 are large and flexible with hydrogen bond donor groups (see Figure 15.10).

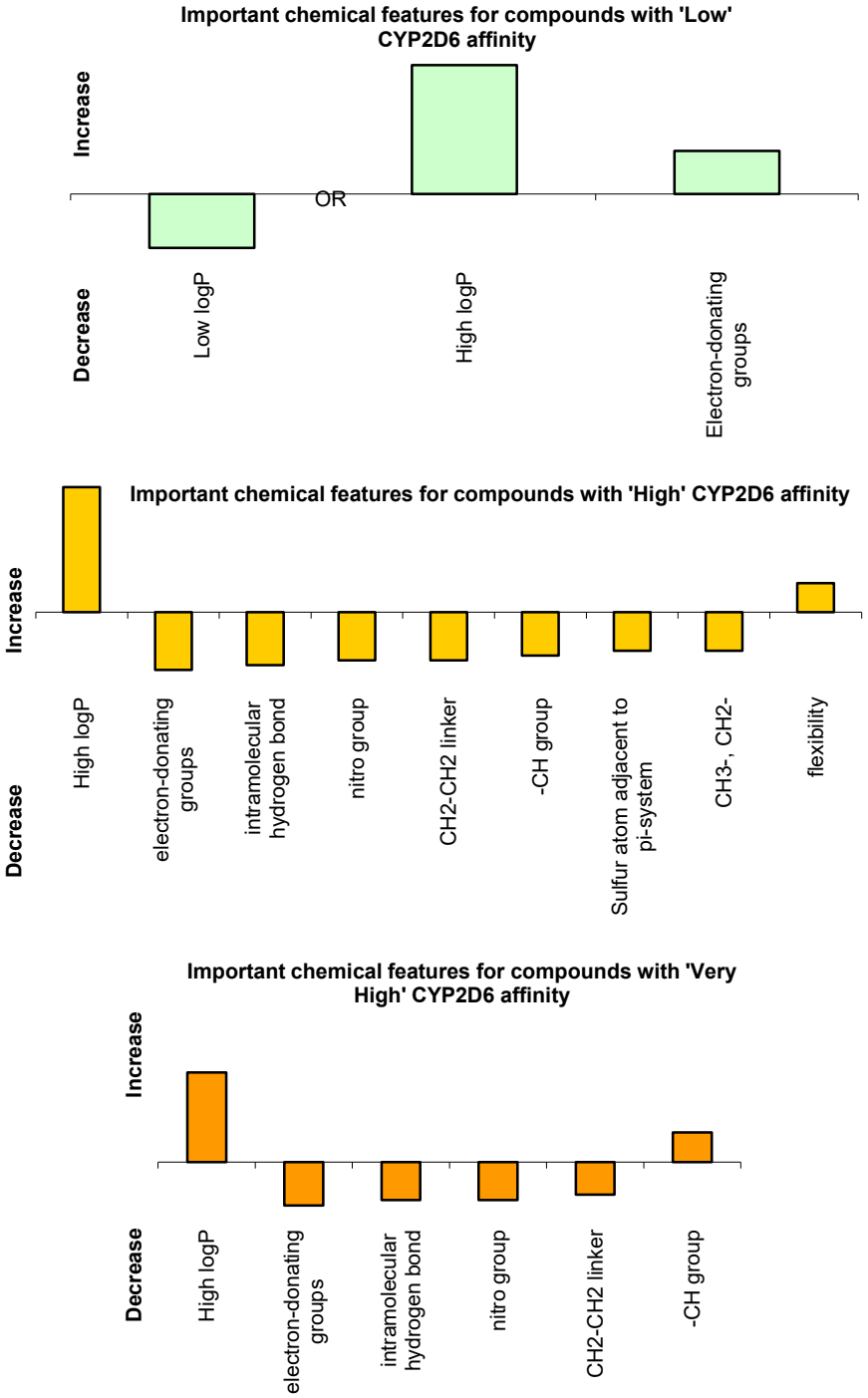


Figure 15.10 Histograms showing the influence of descriptors on the dominant rules of the CYP2D6 affinity classification model.

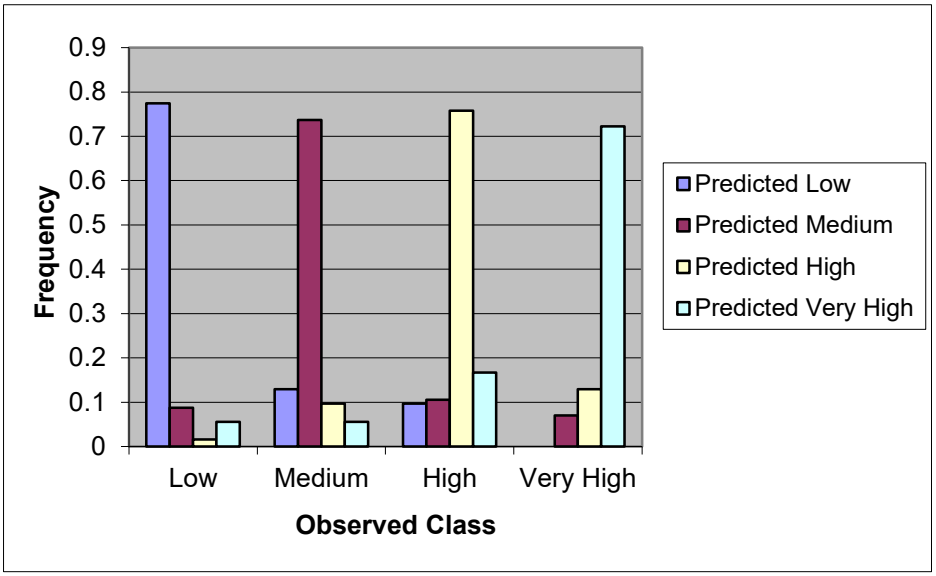


Figure 15.11 Training set predictions for the CYP2D6 affinity model.

The current model for CYP2D6 affinity classifies compounds as 'low', 'medium', 'high' or 'very high' affinity, according to the class boundaries given above. A confidence for each prediction is reported, according to the strength of association of the compound's descriptor values with the predicted classification. Furthermore, the distance of the predicted compound from the chemical space of the training set is calculated to gauge the confidence in the result. As there are insufficient data points outside the chemical space of the training set to assess the confidence in predictions, no estimate regarding the confidence for such compounds can be made. In these cases, the probability that the result is correct is reported as 0.25, indicating an even distribution between the four possible classes.

The results for the training and test set are shown in Figure 15.11 and Figure 15.12 respectively. These show that the model can identify compounds with high/very high affinity for CYP2D6.

Comparison with other predictive techniques

There has been significant published work on quantitative structure activity relationships for affinity to CYP2D6 (Ekins, Berbaum, & Harrison, Generation and validation of rapid computational filters for cyp2d6 and cyp3a4, 2003) (Hutzler, Walker, & Wienkers, 2003) (Koymans, et al., 1992) (Langdon, Barret, & Buxton, 2003) (Lewis, Eddershaw, Goldfarb, & Tarbit, 1996) (Modi, et al., 1996) (Strobl, von Kruedener, Stockigt, Guengerich, & Wolff, 1993) (Susnow & Dixon, 2003). The majority of these are pharmacophore models requiring 3D structures and are based on data points obtained from literature survey.

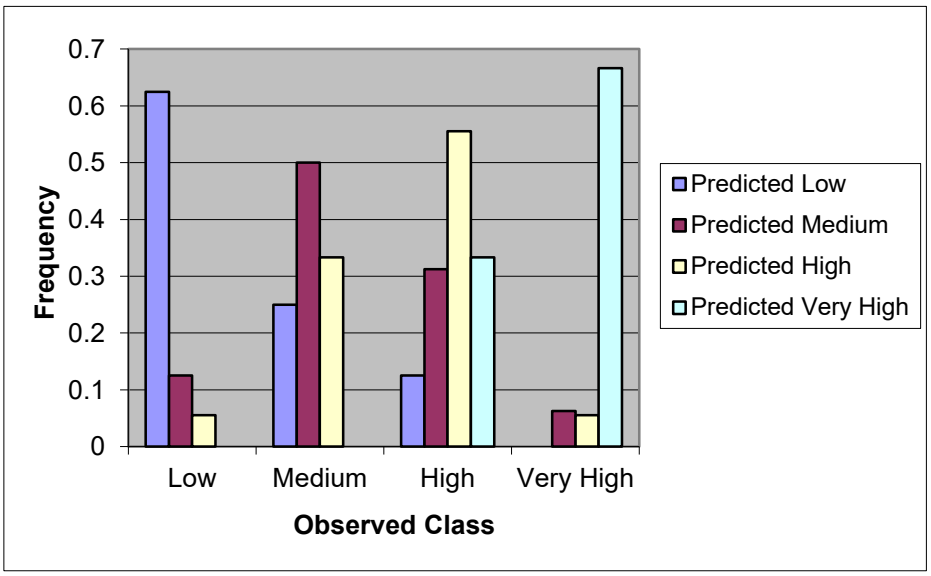


Figure 15.12 Test set predictions for the CYP2D6 affinity model.

15.1.13 P-gp Transport Classification

P-gp is an ATP-driven efflux pump encoded by the MDR1 gene, capable of transporting a wide spectrum of chemical structures as well as different classes of drugs (Selwood, et al., 1990). Active transport by P-glycoprotein (P-gp) can represent a serious hurdle for pharmaceuticals as transport by P-gp has been associated with reduced bioavailability of orally administered drugs and with decreased ability of drug candidates to cross blood-tissue barriers such as the blood-brain barrier (Ayrton & Morgan, 2001). In addition, if a drug is subject to significant P-gp efflux, its distribution, absorption and elimination could be altered by potent P-gp inhibitors. Evidence for drug-drug interactions due to inhibition of P-gp have been reported in human clinical studies (Schwab, Fischer, Tabatabaei, Poli, & Huwyler, 2003). This is best documented for quinidine-digoxin interactions in which decreased renal and intestinal clearance of digoxin and increased plasma drug levels have been reported when quinidine is administered to patients taking digoxin (Hochman, Yamazaki, Ohe, & Lin, 2002). These changes have been attributed to inhibition of P-gp by quinidine where a significant portion of digoxin elimination is mediated by P-gp (Hochman, Yamazaki, Ohe, & Lin, 2002). Therefore, from the drug discovery and development perspective, knowledge of the transport of drug candidates by P-gp is desirable at an early stage of the drug design process.

Data set

A database of 256 chemically diverse compounds with P-gp transport properties was assembled from the literature. The P-gp transport of each compound was assigned “yes” if transported by the protein and “no” if not transported. There is no single experimental method to conclusively identify a compound as a substrate for P-gp. Therefore, identification of the transport classification was based on at least two concurrent literature values from different assays, for example bi-directional Caco-2 measurements, ATPase activity or inhibition of transport of marker substrates.

Model output

The model is a random forest classification model which classifies molecules as likely to be substrates for P-gp (yes) or not likely (no). The performance of this model was assessed on an independent test set of 51 compounds, of which 82% of the non-substrates and 79% of the substrates were correctly classified.

A confidence for each prediction is reported, according to the strength of association of the compound’s descriptor values with the predicted classification. Furthermore, the distance of the predicted compound from the chemical space of the training set is calculated to gauge the confidence in the result. As there are insufficient data points outside the chemical space of the training set to assess the confidence in predictions, no estimate regarding the confidence for such compounds can be made. In these cases, the probability that the result is correct is reported as 0.5, indicating an even distribution between the two possible classes.

Comparison with other predictive techniques

The model statistics compare well to recent literature P-gp classification models where P-gp substrate prediction accuracy on independent test sets ranges from 53% to 72% and P-gp non-substrate prediction accuracies lie between 79% and 80% (Penzotti, Lamb, Evensen, & Grootenhuis, 2002) (Stouch, Gudmunson, & Ge, 2002) (Didziapetris, Japertas, & Petrauskas, 2004).

15.1.14 hERG pIC50

Inhibition of the human Ether-a-go-go-Related Gene (hERG) potassium channel by medications appears to be the most common mechanism of acquired QT interval prolongation. QT interval prolongation is a side effect induced by structurally-diverse drugs that has been linked to life threatening ventricular arrhythmias including Toursade de Pointes. Because more and more non-antiarrhythmic drugs are being shown to have the potential to prolong QT interval, it is important that all new chemical entities (NCE) are thoroughly investigated for this potential early in their preclinical development. Therefore, *in silico* prediction of the hERG screening plays an important role in understanding the hERG-drug binding. Such predictive hERG models are highly valuable as *in vitro* and *in vivo* measurements are costly, labour intensive and not widely available. In recent years, structure-activity studies on the growing number of marketed drugs and investigational compounds exhibiting inadvertent hERG channel blockade have been reported.

Data set

Data on hERG K+ channel blockers were derived from various literature sources. 168 structures with patch-clamp IC_{50} values for inhibition of hERG K+ channels expressed in mammalian cells were selected, as this is the 'gold-standard' experimental technique for determining hERG inhibition. Other higher-throughput approaches to measuring hERG inhibition show poor correlation with patch-clamp measurements in mammalian cells and with each other and were therefore not used in the development of this model.

Model output

A model was built using the non-linear Gaussian Processes technique (GP2DSearch) implemented within the Auto-Modeller. The model was trained on 135 compounds and tested on 33 compounds. The model uses 158 descriptors measuring compound lipophilicity, McGowan's volume, negative and positive charges, polar surface area and counts of different atomic or functional groups or specific fragments.

The initial set of 168 compounds was split into the training set (135 compounds), validation set (17 compounds) and test set (16 compounds) using cluster analysis at Tanimoto level 0.7. The R^2 value for the training set is 0.78 and the RMSE is 0.66 log units. On the combined validation and test sets (33 compounds) the model achieved an R^2 value of 0.72 (r^2_{corr}=0.74) and an RMSE in prediction of 0.64 log units (see Figure 15.13).

Together with each prediction, Gaussian Processes modelling techniques are able to calculate a standard deviation in prediction. The model output combines a prediction of the compound's pIC_{50} ($-\log_{10} IC_{50}$) along with a standard deviation of prediction. A large standard deviation means that the compound is outside of the descriptor space of the model, and such a prediction should be treated with caution.

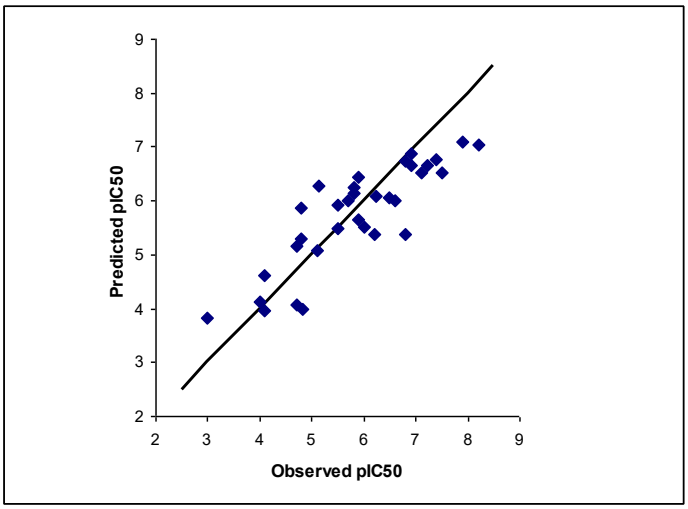


Figure 15.13 Observed versus predicted pIC_{50} values for hERG inhibition for combined validation and test sets.

Testing the model on external data

The model was further tested on external data from three separate sources.

- Cadwell *et al.* reported the potassium channel activity of 22 fluoropyrrolidine amides as measured by hERG binding reporting Ki values in nM (Cadwell, et al., 2004). They used a displacement binding assay of [^{35}S]-radiolabeled MK-499 in membranes derived from HEK293 cells stably transfected with the hERG gene and expressing the Ikr channel.
- Fletcher *et al.* evaluated the displacement of [^{3}H]-dofetilide binding to HEK cells stably expressing hERG (Fletcher, et al., 2002). The authors reported the Ki values in nM for 19 4-(phenylsulfonyl)piperidines.
- Bell *et al.* studied the affinity of 20 3-aminopyrrolidinone farnesyltransferase inhibitors (Bell, et al., 2001). They used a radioligand competition assay. The hERG channel was stably expressed in HEK-293 cells and plasma membrane fractions prepared from these cells were used for competition experiments with [^{35}S]-MK499. Results were reported as inflection points.

The experimental conditions for the above 61 compounds are different from ones for the initial set used for building a model which must be taken into consideration when making a comparison. However, the results of prediction for these compounds are shown in Figure 15.14. On this 61-compound set the model achieved R^2 = 0.53, RMSE=0.64 log units and the squared correlation coefficient r^2_{corr}= 0.72.

Comparison with other predictive techniques

The predictive power of this model is as high as other QSARs reported in the literature. Recent reviews of predictive *in silico* models for hERG inhibition were given by Gola *et al.* and by Song and Clark (Gola, Obrezanova, Champness, & Segall, 2006) (Song & Clark, 2006). The predictive power of 2D QSAR regression models evaluated on test sets ranges from R^2 = 0.52, RMSE=0.68 log units (training set of 439 compounds) (Seierstad & Agrafiotis, 2006) to R^2 = 0.85, RMSE=0.60 log units (training set of 71 compounds) (Song & Clark, 2006).

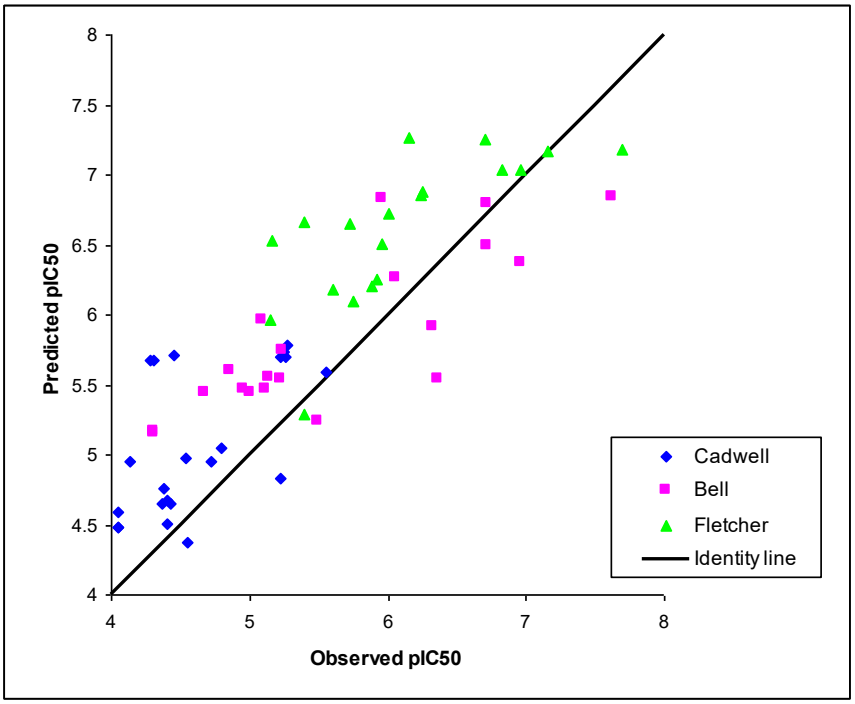


Figure 15.14 Observed versus predicted pIC_{50} values for hERG inhibition for an external test set of 61 compounds. Compounds reported by Cadwell *et al.* are shown in blue, compounds reported by Bell *et al.* are shown in pink and the compounds reported by Fletcher *et al.* are shown in green.

15.1.15 Plasma Protein Binding

The degree of binding to plasma proteins significantly influences the pharmacokinetic and pharmacodynamic properties of a drug. The efficacy of the drug will be related to the exposure to the amount of unbound drug in plasma, i.e. the proportion free to penetrate into surrounding tissues. The bound drug in plasma can also serve as a reservoir for free drug removed by various elimination processes thus prolonging the duration of action. The bulk of experimental data available relates to percent or fraction of drug bound to all plasma proteins (mainly serum albumin and α1-acid glycoprotein).

15.1.16 Plasma Protein Binding Classification (90% threshold)

Data set

Commercial and proprietary databases, compendia and drug monographs were searched for experimentally measured values for % drug bound to human plasma protein the data were incorporated into a database, which, after rigorous quality control, led to data set of 1,107 compounds. This dataset is somewhat biased toward high percentage values with 33% of the compounds reported as ≥ 90% bound. Values of % bound < 90% were classified as low, values ≥ 90% were classified as high. The structures were assigned randomly to training (n = 775), internal evaluation (n = 166) and independent test (n = 166) sets. The latter was excluded from the model development process.

Model output

The model is a random forest that predicts the extent of test set compounds' plasma protein binding as either "high" or "low" in relation to the threshold described above. Calculated logP plus a further thirteen 2D descriptors, relating to the occurrence of certain functional groups and structural fragments and the protonation state of certain groups, are used in the model.

The model classifies molecules as having high or low affinity based on a classification boundary of 90%. For the independent test set predictions of high and low plasma protein binding were correct on 81% and 87% of occasions respectively.

A confidence for each prediction is reported, according to the strength of association of the compound's descriptor values with the predicted classification. Furthermore, the distance of the predicted compound from the chemical space of the training set is calculated to gauge the confidence in the result. As there are insufficient data points outside the chemical space of the training set to assess the confidence in predictions, no estimate regarding the confidence for such compounds can be made. In these cases, the probability that the result is correct is reported as 0.5, indicating an even distribution between the two possible classes.

Comparison with other predictive techniques

Comparison with recent literature models is difficult as they are based on binding to human serum albumin only and with data obtained via chromatographic, rather than older-established methods (Kratochwil, Huber, Muller, Kansy, & Gerber, 2002) (Colmenarejo, Alvarez-Pedraglio, & Lavandera, 2001).

15.2 Descriptors

Both 1D descriptors and whole molecule properties are used within StarDrop.

The whole molecule properties available by default when building models with the StarDrop Auto-Modeller are:

- logP - the lipophilicity of a compound, calculated as described in Section 15.1.1
- TPSA - Topological Polar Surface Area is calculated based on the method described by Ertl and co-workers (Ertl, Rhodes, & Selzer, 2000). Using SMARTS atom type definitions, the authors proposed the calculation of two PSA values. ERTLNoTPSA reports the polar surface area for Nitrogen and Oxygen atoms only. ERTLSPNoTPSA reports tha polar surface area for Nitrogen, Oxygen, Sulfur and Phosphorus atoms
- MWT - the molecular weight of the molecule
- Vx - the McGowan volume (Abraham & McGowan, The use of characteristic volumes to measure cavity terms in reversed-phase liquid-chromatography., 1987)
- Flex - the flexibility index
- Number of positive, negative and overall charges. Overall charge is the number of positive charges minus the number of negative charges
- Number of aromatic rings

The SMARTS based patterns are described in Table 27.

Table 27 SMARTS based patterns in StarDrop

SMARTS definition	Name	Definition	Reference
[N,n,O,o]	HBA-lip	Number of hydrogen bond acceptors according to the Lipinski definition	StarDrop
[$([O,S;H1][!#1;!$(*=[O,N,P,S])]),$(N#C),$([O;H0]-[!#1]),$([S;H0&X2]),$([O,S]=[C,S,N,P;!$([C,S,N,P]-[O;H1])]),$([o,s]),$([n;H0&r5]),$([O,S]=a)]	HBA-prof	Number of hydrogen bond acceptors including sulfur, oxygen and nitrogen atoms	StarDrop
[NH1,NH2,nH,OH]	HBD-lip	Number of hydrogen bond donors according to the Lipinski definition	StarDrop
[$([N;!H0][C,S]=[O,S]),$([n;H1;+0]),$([O,S;H1][!#1;!$(*=[O,N,P,S])])]	HBD-prof	Number of hydrogen bond donors including sulfur, oxygen and nitrogen atoms with higher specificity	StarDrop
[$([N;!H0][C,S]=[O,S]),$([n;H1;+0]),$([O,S;H1][CX4,a])]	HBD-cam	Number of hydrogen bond donors including sulfur, oxygen and nitrogen atoms with lower specificity	StarDrop
[$([N+](*)(*)(*)*),$([n+](*)(*)*)]	quatN	Number of quaternary nitrogens	StarDrop
[r;!r3;!r4;!r5;!r6;!r7;!r8;!r9]	Macrocyclic	Number of rings with more than nine atoms	StarDrop
[$([O](=[C&!R]([#6])[NH1,NH2]))]	ACamideO-nh-nh2	Number of carboxamide groups with at least one hydrogen on the amide nitrogen	StarDrop
[$([O](=[C&!R]([#6])[NH0]))]	ACamideO-nh0	Number of carboxamide groups with no hydrogen on the amide nitrogen	StarDrop
[$([S&!R](=O)(=O)([#6])[NH1,NH2])]	ASamideO-nh-nh2	Number of sulfonamide groups with at least one hydrogen on the amide nitrogen	StarDrop
[$([S&!R](=O)(=O)([#6])[NH0])]	ASamideO-nh0	Number of sulfonamide groups with no hydrogen on the amide nitrogen	StarDrop
[$(N[C&!R]=[N&!R])]	Aamidine	Number of non cyclic amidine groups	StarDrop
[$([NH0&!R]([CX4])([CX4])[CX4])&!$(([NH0&!R]([CX4])([CX4])[CX4].[OH1][C,S,P](=O)))]	AbasicNH0	Number of tertiary nitrogens not in a ring linked to three sp3 carbons. The molecule should not be zwitterionic	StarDrop
[$([NH1&!R]([CX4,a])[CX4])&!$(([NH1&!R]([CX4,a])[CX4].[OH1][C,S,P](=O)))]	AbasicNH1	Number of secondary nitrogens not in a ring linked to three sp3 carbons. The molecule should not be zwitterionic	StarDrop

[$(C(Br)(Br)Br)]	CBr	Number of tribromomethyl groups	StarDrop
[$(C(F)(F)F)]	CF3	Number of trifluoromethyl groups	StarDrop
[$([CH0]([CX4,a])([CX4,a])([CX4,a])[CX4,a])]	CH0Aa	Number of aliphatic sp3 carbons with no hydrogen linked to four saturated aliphatic carbons or aromatic atoms	StarDrop
[$([CH1]([CX4,a])([CX4,a])[CX4,a])]	CH1Aa	Number of aliphatic sp3 carbons with exactly one hydrogen linked to three saturated aliphatic carbons or aromatic atoms	StarDrop
[$([CH2]([CX4,a])[CX4,a])]	CH2Aa	Number of aliphatic sp3 carbons with exactly two hydrogens linked to two saturated aliphatic carbons or aromatic atoms	StarDrop
[$([CH2][O,S,N])]	CH2hetero	Number of aliphatic sp3 carbons with exactly two hydrogens linked to an non-aromatic oxygen, sulfur or nitrogen	StarDrop
[$([CH2][CH2])]	CH2link	Number of methylene-methylene groups	StarDrop
[$([CH2][CH2][CH2][CH2][CH2])]	CH2long	Number of aliphatic chains with five methylene groups	StarDrop
[$([CH3][CX4,a])]	CH3Aa	Number of methyl groups linked to either aliphatic sp3 carbons or aromatic carbons	StarDrop
[$([CH3][O,S,N])]	CH3hetero	Number of methyl groups linked to an aliphatic oxygen, nitrogen or sulfur	StarDrop
[SX2]([#6])[SX2][#6]	CSSC	Number of disulfanyl groups	StarDrop
[$([NH0][C](=O)[#6])]	CamideNH0	Number of tertiary nitrogens in carbo-amide groups	StarDrop
[$([O,S&R]([CX4,a])[C,S&!R]=O)]	Ester	Number of ester and thioester groups	StarDrop
[$([CX4]([F,Cl,Br,I])[CX4,a])]	HaloC	Number of carbons bearing halogen atoms	StarDrop
[$(C(=O)C=[CH])]	Michael-accept	Number of Michael acceptor type groups	StarDrop
[a]	NBA	Number of aromatic atoms	StarDrop
[NH1](C(=O)[#6])C(=O)[#6]	NH1and2CdO	Number of diacetamide groups	StarDrop
[NX3]-[O]	NO	Number of nitroso groups	StarDrop
[*]-[*]	NRB	Number of single bonds to heavy atoms	StarDrop
[$([OH1][CH2,CH1]C(=O))]	OHCHCdO	Number of hydroxyl groups in the beta position of an amide	StarDrop
[$([OH0][C,S](=O)(N))]	Ocarbamate	Number of carbamate groups	StarDrop
P(=[S,O])[NX3]	Pamide	Number of Pamide groups	StarDrop
P(=[O,S])[O,S]	Pester	Number of Pester groups	StarDrop
[$([O](=[C&R]([#6])[NH1,NH2]))]	RCamideO-nh-nh2	Number of primary and secondary carboxamide groups with cyclic sp2 carbons	StarDrop
[$([O](=[C&R]([#6])[NH0]))]	RCamideO-nh0	Number of tertiary carboxamide groups with cyclic sp2 carbons	StarDrop
[SX2;v2]	RSR	Number of thioethers, thioesters and sulpfurs from thiocarbamates	StarDrop
[$([S&R](=O)(=O)([#6])[NH1,NH2])]	RSamideO-nh-nh2	Number of primary and secondary carbo-sulfonamides.	StarDrop
[$([S&R](=O)(=O)([#6])[NH0])]	RSamideO-nh0	Number of tertiary carbo-sulfonamides	StarDrop
[$(N[C&R]=[N&R])]	Ramidine	Number of cyclic amidine groups	StarDrop
[$([NH0&R]([CX4])([CX4])[CX4])&!$(([NH0&R]([CX4])([CX4])[CX4].[OH1][C,S,P](=O)))]	RbasicNH0	Number of cyclic sp3 nitrogens with no hydrogen connected to three sp3 carbons in a molecule with no acidic groups	StarDrop
[$([NH1&R]([CX4,a])[CX4])]	RbasicNH1	Number of cyclic sp3 nitrogens connected to at least one sp3 carbon	StarDrop

[$([NH2][S](=O)[#6]),$([NH1][S](=O)[#6])]	Samide-NH	Number of aliphatic sp3 nitrogens with at least one hydrogen in a sulfonamide group	StarDrop
[$([NH0][S](=O)[#6])]	SamideNH0	Number of tertiary nitrogens in carbo-sulfonamide groups	StarDrop
[$([CH2,CH1]([C,N,S](=O))[C,N,S](=O))]	activatedCH	Number of aliphatic sp3 carbons with at least one hydrogen and connected to either carbon, sulfur or nitrogen connected to an sp2 oxygen	StarDrop
[$(O=[CH1][CX4,a])]	aldehydes	Number of aldehydes	StarDrop
[$([OH][CX4])&!$([OH]C(C)(C)C)]	aliphOH-t6	Number of hydroxyl groups connected to an aliphatic sp3 carbon but not tert-butyl	StarDrop
[$(C=C[CX4&!H0])]	allylic-oxyd-t10	Number of aliphatic sp2 carbons connected to an sp2 carbon bearing an sp3 carbon with at least one hydrogen	StarDrop
[$([NH1](C(=O))C(=O))]	amide-dicarbonyl	Number of aliphatic sp3 nitrogens connected to two carboxy groups	StarDrop
[$([NH0]CC[N,O])]	aminoethanol0	Number of aliphatic sp3 nitrogens with no hydrogen in the beta position of heteroatoms (O or N)	StarDrop
[$([NH]CC[N,O])]	aminoethanol1	Number of aminoethanol side chains	StarDrop
[$([O]=[*])]	anycarbonyl	Number of carbonyl groups	StarDrop
[$([Br][a])]	aromBr	Number of bromine atoms linked to an aromatic atom	StarDrop
[$([Cl][a])]	aromCl	Number of chlorine atoms linked to an aromatic atom	StarDrop
[$([F][a])]	aromF	Number of fluorine atoms linked to an aromatic atom	StarDrop
[$([I][a])]	aromI	Number of bromine atoms linked to an aromatic atom	StarDrop
o	aromO	Number of aromatic sp2 oxygens	StarDrop
[NX3;H1]([a])-[C;!R]=O	arylNHCO	Number of secondary amides with nitrogen connected to an aromatic atom	StarDrop
[$([NH2][CX4])&!$(([NH2][CX4].[OH1][C,S,P](=O)))]	basic-NH2	Number of primary amines (not in zwitterionic compounds)	StarDrop
[$(C1CNc2ccccc2C=N1)]	benzdiaz-t18	Number of benzodiazepine rings with no additional fused ring	StarDrop
[$(N1aaC(a)=NCC1)]	benzdiazepine-ring	Number of benzdiazepine rings	StarDrop
[OH][CX4]a	benzylicOH	Number of benzyl groups	StarDrop
[$([CX4;H1&!R]),$([CX4;H0&!R])]	branchedCnotRing	Number of sp3 carbons with exactly one or no hydrogens and not in a ring	StarDrop
[S,O]=C([#7])[S,O]	carbamate-and-thio	Number of carbamate and thiocarbamate groups	StarDrop
[CX3](O)([O,N])=O	carbonate-carbamate	Number of carbonate or carbamate group	StarDrop
[$([CH2]([CH2])[a])&!$([CH2]([CH2][O,N])[a])]	ch2-lipo-t9	Number of methylene groups connected to exactly one aromatic atom	(Yoshida & Topliss, 2000)
[$(N(=[*])[O&!R])]	dNO	Number of sp2 nitrogens connected to a non-cyclic oxygen and double-bonded to any atom	(Hall, Kier, & Brown.B.B, 1995)
[$(N(=O)=O),$(C#N),$([F,Cl,Br,I])][CX4][$(N(=O)=O),$(C#N),$([F,Cl,Br,I])]	di-withdraw-cx4	Number of withdrawing electron groups on sp3 carbons	StarDrop
N(=N[a])[a]	diazo-aryl	Number of diazo groups with nitrogen connected to aromatic atoms	StarDrop
[NX2]=[NX2]	diazo	Number of diazo groups	StarDrop
O=CC=CC=O	dione-1-4	Number of 1,4-dione groups	StarDrop
[$(CN1C=CC=CC1),$([SH][CX4])]	easy-oxy-t13	Number of sulfhydryl and dihydropyridyl groups	(Yoshida & Topliss, 2000)

[S](-*)-*	ertl-33	Number of sulfur atoms with at least two single bonds	(Ertl, Rhodes, & Selzer, 2000)
[S](-*)(-*)=*	ertl-35	Number of sulfur atoms with two single bonds and at least one double bond	(Ertl, Rhodes, & Selzer, 2000)
[SH]-*	ertl-37	Number of sulfhydryl groups	(Ertl, Rhodes, & Selzer, 2000)
[s](=*)(:*):*	ertl-39	Number of aromatic sulfur atoms	(Ertl, Rhodes, & Selzer, 2000)
[P](-*)=*	ertl-41	Number of phosphorus atoms with at least one double bond and one single bond. Low specificity	(Ertl, Rhodes, & Selzer, 2000)
[PH](-*)(-*)=*	ertl-43	Number of phosphorus atoms with at least one hydrogen and connected to other atoms by at least two single bonds and one double bond	(Ertl, Rhodes, & Selzer, 2000)
[$(C(=O)([C&R])O[C&R]),$([#6]C(=O)O[#6]),$(N1C(=O)CC1),$(CNC(=O)OC)]	est-lact-latm-carbm-t7	Number of esters, lactones, beta-lactams and alkylcarbamates	(Yoshida & Topliss, 2000)
[$([O,S]([CX4,a])[CX4,a])&!$(O1CCOC1)]	ether	Number of oxygen and sulfur atoms linked to aliphatic or aromatic carbons but not 5-membered cyclic ketals	StarDrop
[F,Cl][CX4][CX4][F,Cl]	halosp3sp3halo	Number of 1,2-fluoro/chloro ethyl groups linked to another 1,2-fluoro/chloro ethyl group	StarDrop
[cH0](:[nX2])(:[nX2])[#8,#7,#16,#15,#34,#9,#17,#35,#53]	hetero-halo-di-n-arom	Number of aromatic carbons connected to exactly two aromatic nitrogens and one heteroatom	StarDrop
[OH]c1c([CX4])cccc1[CX4]	hindred-phenol	Number of phenolic groups with two sp3 carbons in the ortho position to the hydroxyl	StarDrop
[$([OH1][CX4])]	hydroxyA	Number of hydroxyl groups linked to an sp3 carbon	StarDrop
[$(c1(O([CX4]))c[cH][cH][cH][cH]1),$(c1(N([#6])([#6]))c[cH][cH][cH][cH]1),$(c1([NH](C(=O)[#6]))c[cH][cH][cH][cH]1)]	hydroxylation-t8	Number of hydrogens in the para position of an activating group	(Yoshida & Topliss, 2000)
[$([OH1,NH1,nH]~[*]~[*](=O))]	intraHbond5	Number of hydrogen bond donors three bonds away from an sp2 oxygen	StarDrop
[$([OH1,NH1,nH]~[*]~[*]~[*](=O))]	intraHbond6	Number of hydrogen bond donors four bonds away from an sp2 oxygen	StarDrop
[$(C1OCCO1)]	ketal	Number of cyclic carbons connected to two cyclic alkyl oxygens	StarDrop
[$(C(=O)([#6,CX4])([CX4]))]	ketone-t14	Number of sp2 carbons in a carboxy group whose sp2 carbon is attached to two sp3 carbons or one sp3 carbon and one carbon	(Yoshida & Topliss, 2000)
[$(O=C([CX4])[CX4,a])]	ketones	Number of sp2 oxygens in a carboxy group whose sp2 carbon is attached to two sp3 carbons or one sp3 carbon and one aromatic atom	StarDrop

[$([CH2]([$([#6])&!$([#6]~[#7,#8,S,P])])[$([#6])&!$([#6]~[#7,#8,S,P])]),$([CH3][$([#6])&!$([#6]~[#7,#8,S,P])]),$([cH]([$([#6])&!$([#6]~[#7,#8,S,P])])[$([#6])&!$([#6]~[#7,#8,S,P])]),$([CH1]([$([#6])&!$([#6]~[O,N])])=[CH1][$([#6])&!$([#6]~[O,N])])]	lipovolume	Number of lipophilic atoms	StarDrop
[$([nH0]1([*&!R])aaa2aaaaa21)]	nH0indole-like	Number of indole nitrogens with no hydrogen.	StarDrop
[$([nH]1aaa2aaaaa21)]	nHindole-like	Number of indole nitrogens with one hydrogen	StarDrop
na(=O)na(=O)	nc(do)n	Number of 1,3-diazinane-2,4-diones	StarDrop
[$([C,c]([O]N(=O)(=O)))]	nitro-O	Number of nitrooxy groups	StarDrop
[$(c([cH])([cH])N(=O)(=O))]	nitro-no-ortho-t15	Number of nitro groups on a benzene ring with no substituent in the ortho position	(Yoshida & Topliss, 2000)
[$(N(=O)(=O)[#6])]	nitro	Number of nitro groups connected to a carbon	StarDrop
[$([*&!R])]	nonring-at	Number of non-cyclic atoms	StarDrop
[OH][CX4;!R][CX4;!R][OH]	not-ring-diol	Number of linear and aliphatic diols	StarDrop
[$([OH]CC[NH]C(C)(C)C),$([OH]C[C&!R][N&R])]	ohccn-t17	Number of hydroxyl groups in the beta position of a secondary amine	(Yoshida & Topliss, 2000)
[S,O,N,F,Cl,Br,I]-c:a:a:c-[S,O,N,F,Cl,Br,I]	p-hetero-or-halo	Number of heteroatoms (S,N,O and the first four halogens) para to heteroatoms (S,O,N and the first four halogens)	StarDrop
[OH]c1ccc(cc1)[$([F,Cl,Br,I]),$(N(=O)=O),$(C(F)(F)F)]	p-withdraw-phenol	Number of hydroxyls on a benzene ring para to electron withdrawing groups (the first four halogens, the nitro group and the trifluoromethyl group)	StarDrop
[$(C(F)(F)C(F)(F)C(F)(F)C(F)(F))]	perfluoro	Number of di-trifluoromethyl side chains	StarDrop
[$([OH1]a(a)a(a)n)]	phenol-pyr2r	Number of hydroxyl groups on a pyrrole ring	StarDrop
[$([OH1][a])]	phenol	Number of hydroxyl groups connected to an aromatic atom	StarDrop
[O]=[c]	phenolic-tautomer	Number of aromatic carbonyls	StarDrop
[$(O1CCCCC1.O1CCCCC1)]	poly-sugars	Number of polysugar groups	StarDrop
[$([CH1]([OH1])[CH1]([OH1])[CH1]([OH1]))]	polyOH	Number of side chains with three sp3 carbons with one hydrogen and connected to one hydroxyl group	StarDrop
[$([n&X2]1ccccc1)]	pyridine	Number of nitrogens in a pyridine ring	StarDrop
[$([O]=a1aanaa1),$([O]=a1naaaa1)]	pyridones	Number of sp2 oxygens in a pyridone ring	StarDrop
O=[C,c]1[C,O,c]~[C,c][C,c](=O)[C,c]~[C,c]1	quinone-type	Number of quinone type rings	StarDrop
[$([c](:a)(:a):a)]	ring-join	Number of aromatic carbons at an aromatic/aromatic boundary	StarDrop
[$([nH0&r5])]	ring5-nH0	Number of aromatic nitrogens with no hydrogens in a 5-membered ring	StarDrop
[$([nH&r5])]	ring5nH	Number of aromatic nitrogens with one hydrogen in a 5-membered ring	StarDrop
[$([C,S&R]([*&R])(=O)[*&R])]	ringOdouble	Number of aliphatic sp2 carbons or sp2 sulfurs in a ring connected to one sp2 oxygen and to two other atoms in rings	StarDrop
[$([*&R])]	ringat	Number of cyclic atoms	StarDrop
[OH][CX4;R][CX4;R][OH]	ringdiol	Number of diol groups on an aliphatic ring	StarDrop
[$([CX2])]	sp-carbons	Number of sp carbons	(Hall, Kier, & Brown.B.B, 1995)

[$([CX3])]	sp2-carbons	Number of non-aromatic sp2 carbons	(Hall, Kier, & Brown.B.B, 1995)
[$(C([*&R])([*&R])([*&R])[*&R])]	spiroC	Number of aliphatic spiro carbons	(Hall, Kier, & Brown.B.B, 1995)
[$([C,c](S(=O)(=O)[OH]))]	sulfonicacid	Number of aliphatic and aromatic carbons connected to a sulfo group	StarDrop
[SX4](=O)(=O)([#6])[#6]	sulfates	Number of sulfonyl groups with sulfur connected to two carbons	StarDrop
S(=O)(=O)[NH2]	sulphonamide-t5	Number of sulfonamido groups	(Yoshida & Topliss, 2000)
[$([NH2][a])]	t-16-1	Number of anilines or amino-heteroaromatics	(Yoshida & Topliss, 2000)
[$([NH2][NH1][a])]	t-16-2	Number of hydrazino groups	(Yoshida & Topliss, 2000)
[$([NH2]C(=[NH1])[a])]	t-16-3	Number of amidino groups.	(Yoshida & Topliss, 2000)
[$(N([CX4])([CX4])[CX4])]	tert-amine-t11	Number of tertiary nitrogen non-anilines	StarDrop
[$(C(=S)[SH])]	thio-acid	Number of sp2 carbons in a thioacid group	StarDrop
[S]=C	thio-keto	Number of thiocarboxyl groups	StarDrop
[$(NC(=[O,S])N)]	urea-thio	Number of urea and thiourea groups	StarDrop
O=C([NX3])[NX3]	urea	Number of non-aromatic urea groups.	StarDrop
[$([N,O,a]CCN[CH3]),$([N,O,a]CCN[CH2][CH3])]	xccn-t12	Number of secondary amines connected to either methyl or ethyl groups on one side and in the beta position of an oxygen or nitrogen atom on the other side	(Yoshida & Topliss, 2000)
([NH2][CX4].C(=O)[OH1])	zw1	Number of primary amines connected to one sp3 carbon	StarDrop
([NH1]([CX4])[CX4].C(=O)[OH1])	zw2	Number of secondary amines connected to two sp3 carbons	StarDrop
([NH0]([CX4])([CX4])[CX4].C(=O)[OH1])	zw3	Number of tertiary amines connected to three sp3 carbons	StarDrop
[$([CX3,c])&!$([#6]=[O])]	nC(sp2)	Number of sp2 carbons not connected to an sp2 oxygen	StarDrop
[$([CX4])]	nC(sp3)	Number of sp3 carbons	StarDrop
[$([OH1][C](=O)([C,c]))]	nCOOH	Number of carboxylic acid groups connected to any carbon	StarDrop
[$([O;H1][C,c]);!$([O;H1][C]=O)]	nOH	Number of hydroxyl groups connected to any carbon not in a carboxylic acid	StarDrop
[$([O]=[C,c]);!$([O]=[C][OH1])]	nCO	Number of carbonyl groups not in a carboxylic acid	StarDrop
[$([O,o,S,s]([*])[*])]	nOS	Number of sulfurs and oxygens attached to two heavy atoms	StarDrop
[F,Cl,Br,I]	nX	Number of fluorines, chlorines, bromines and iodines	StarDrop

[$([NX3&!$(*c)&!$(*([C,c]=[O,o,P,S]))&!$(*[C,c]([N,n])=[N,n])&!$(*[c,C]=[n,N][a])&!$(*~[O,N,o,P,S])&!$(*[C,c][Cl,Br,F,I])&!$(*[C,c]O[CH3])&!$(*[C,c]C(F)(F)F)]),$([NH1&!R]C=[NH1;!R]),$([NH1;R][CX2;R&!$(*([Cl,Br,F,N,I,O]))]=[NH0;R][CX4;R&!$(*([Cl,Br,F,I,N,O,S]))][CX4;R&!$(*([Cl,Br,F,I,N,O,S]))]),$([NH1]1[CX4;!$(*([Cl,Br,F,I,O,S]))][CX4;!$(*([Cl,Br,F,I,O,S]))][NH0]=[CH0]1),$([NH1;!R][C;R&!$(*([Cl,Br,F,I,O]))&!$(*([NH1&R]))]=[NH0;R]),$([NH1&R][C&R](=[N&!R])[NH1&R]),$(n1([CX4])cncc1),$([nH]1[cH]n[cH][c]1[CX4]),$([nH]1[cH][n]c([CX4])c1[CX4]),$([n&H0][c&!$(*([Cl,Br,F,I,O]))][n][c&!$(*([Cl,Br,F,I,O]))]([N])[c]);!$([NX3&R][C&R]~[C&R][C](=O)O)]	nNprot	Number of protonated nitrogens at pH 7.4	StarDrop
[$([CX2&H2](=*))]	dCH2	Number of sp2 carbons with exactly two hydrogens	StarDrop
[$([CX4&H2](-*)-*)]	ssCH2	Number of aliphatic sp3 carbons with exactly two hydrogens	(Hall, Kier, & Brown.B.B, 1995)
[$([CH](#*))]	tCH	Number of sp1 carbons with exactly one hydrogen	(Hall, Kier, & Brown.B.B, 1995)
[$([CH](=*)-*)]	dsCH	Number of sp2 carbons with exactly one hydrogen	StarDrop
[$([cH]([a])[a])]	aaCH	Number of aromatic carbons with exactly one hydrogen	(Ertl, Rhodes, & Selzer, 2000)
[$([CX4&H1](-*)(-*)-*)]	sssCH	Number of sp3 carbons with exactly one hydrogen	(Hall, Kier, & Brown.B.B, 1995)
[$([CX2&H0](=A)=A)]	ddC	Number of sp2 carbons with exactly no hydrogens and two double bonds	StarDrop
[$([CX2&H0](#*)-*)]	tsC	Number of sp1 carbons with no hydrogens	(Hall, Kier, & Brown.B.B, 1995)
[$([C&H0](=*)([A])([A]))]	dssC	Number of sp2 carbons with no hydrogens	StarDrop
[$([cH0]([a])([a])-*)]	aasC	Number of aromatic carbons with no hydrogens and connected to at least two aromatic atoms	(Ertl, Rhodes, & Selzer, 2000)
[$([cH0]([a])([a])[a])]	aaaC	Number of aromatic carbons with no hydrogens and connected to three aromatic atoms	(Ertl, Rhodes, & Selzer, 2000)
[$([CX4&H0](-*)(-*)(-*)-*)]	ssssC	Number of aliphatic sp3 carbons with no hydrogens	(Hall, Kier, & Brown.B.B, 1995)
[$([NX4&H3]-*)]	sNH3+	Number of aliphatic quaternary nitrogens with exactly three hydrogens	(Hall, Kier, & Brown.B.B, 1995)
[$([NX3&H2]-*)]	sNH2	Number of aliphatic sp3 nitrogens with exactly two hydrogens	(Hall, Kier, & Brown.B.B, 1995)
[$([NX4&H2](-*)-*)]	ssNH2+	Number of quaternary nitrogens with exactly two hydrogens	(Hall, Kier, & Brown.B.B, 1995)
[$([NX2&H1]=*)]	dNH	Number of sp2 nitrogens with exactly one hydrogen and one double bond	StarDrop
[$([NX3&H1](-*)-*)]	ssNH	Number of secondary amides and aniline nitrogens	(Hall, Kier, & Brown.B.B, 1995)

[$([nH]([a])[a])]	aaNH	Number of aromatic nitrogens with one hydrogen	(Ertl, Rhodes, & Selzer, 2000)
[$([NX1]#*)]	tN	Number of sp1 nitrogens	(Hall, Kier, & Brown.B.B, 1995)
[$([NX4&H1](-*)(-*)-*)]	sssNH+	Number of quaternary nitrogens with exactly one hydrogen.	(Hall, Kier, & Brown.B.B, 1995)
[$([NX2&H0](-*)=*)]	dsN	Number of sp2 nitrogens with no hydrogens	StarDrop
[$([nh0]([a])[a])]	aaN	Number of aromatic nitrogens with no hydrogens	(Ertl, Rhodes, & Selzer, 2000)
[$([N&H0](-*)(-*)-*)]	sssN	Number of amide and sulphonamide nitrogens	(Hall, Kier, & Brown.B.B, 1995)
[$([NX3](=*)(=*)-*)]	ddsN	Number of nitro groups	StarDrop
[$([NX3]([a])([a])-*)]	aasN	Number of sp3 nitrogens connected to at least two aromatic atoms	(Ertl, Rhodes, & Selzer, 2000)
[$([#7+](-*)(-*)(-*)-*)]	ssssN+	Number of quaternary nitrogens with no hydrogens	(Hall, Kier, & Brown.B.B, 1995)
[$([OX2H]-*)]	sOH	Number of aliphatic sp3 oxygens connected to one hydrogen	(Hall, Kier, & Brown.B.B, 1995)
[$([O&X2&H0](-*)-*)]	ssO	Number of aliphatic sp3 oxygens with no hydrogens	(Hall, Kier, & Brown.B.B, 1995)
[$([F]-*)]	sF	Number of fluorine atoms	(Hall, Kier, & Brown.B.B, 1995)
[$([SiX4&H3]-*)]	sSiH3	Number of sp3 silicons with exactly three hydrogens	(Hall, Kier, & Brown.B.B, 1995)
[$([SiX4&H2](-*)-*)]	ssSiH2	Number of sp3 silicons with exactly two hydrogens	(Hall, Kier, & Brown.B.B, 1995)
[$([SiX4&H1](-*)(-*)-*)]	sssSiH	Number of sp3 silicons with exactly one hydrogen	(Hall, Kier, & Brown.B.B, 1995)
[$([SiX4&H0](-*)(-*)(-*)-*)]	ssssSi	Number of silicons with four neighbouring atoms	(Hall, Kier, & Brown.B.B, 1995)
[$([P&H2](-*))]	sPH2	Number of phosphorus atoms connected to exactly two hydrogens	(Hall, Kier, & Brown.B.B, 1995)
[$([P&H1](-*)-*)]	ssPH	Number of phosphorus atoms connected to exactly one hydrogen	(Hall, Kier, & Brown.B.B, 1995)
[$([P&H0](-*)(-*)-*)]	sssP	Number of phosphorus atoms with no hydrogen	(Hall, Kier, & Brown.B.B, 1995)
[$([P](=*)(-*)(-*)-*)]	dsssP	Number of phosphorus atoms with four neighbours	(Hall, Kier, & Brown.B.B, 1995)
[$([P](-*)(-*)(-*)(-*)-*)]	sssssP	Number of phosphorus atoms with five neighbours	(Hall, Kier, & Brown.B.B, 1995)
[$([S&H1]-*)]	sSH	Number aliphatic sp3 sulfurs with exactly one hydrogen	(Hall, Kier, & Brown.B.B, 1995)

[$([S&H0]=*)]	dS	Number of sulfurs linked to by a double bond	StarDrop
[$([S&H0](-*)-*)]	ssS	Number of aliphatic sp3 sulfurs with no hydrogen	(Hall, Kier, & Brown.B.B, 1995)
[$([S&H0](-[a])-[a])]	aaS	Number of sp3 sulfurs with no hydrogens and connected to two aromatic atoms	(Ertl, Rhodes, & Selzer, 2000)
[$([S&H0](=*)(-*)-*)]	dssS	Number of sulfur atoms	(Ertl, Rhodes, & Selzer, 2000)
[$([S&H0](=*)(=*)(-*)-*)]	ddssS	Number of sulfonyl and sulfo groups	StarDrop
[$([Cl]-*)]	sCl	Number of chlorine atoms	(Hall, Kier, & Brown.B.B, 1995)
[$([Br]-*)]	sBr	Number of bromine atoms	(Hall, Kier, & Brown.B.B, 1995)
[$([I]-*)]	sI	Number of iodine atoms	(Hall, Kier, & Brown.B.B, 1995)
[$([N,n])&!$([NX3&!$(*c)&!$(*([C,c]=[O,o,P,S]))&!$(*[C,c]([N,n])=[N,n])&!$(*[c,C]=[n,N][a])&!$(*~[O,N,o,P,S])&!$(*[C,c][Cl,Br,F,I])&!$(*[C,c]O[CH3])&!$(*[C,c]C(F)(F)F)])&!$([NH1&!R]C=[NH1;!R])&!$([NH1;R][CX2;R&!$(*([Cl,Br,F,N,I,O]))]=[NH0;R][CX4;R&!$(*([Cl,Br,F,I,N,O,S]))][CX4;R&!$(*([Cl,Br,F,I,N,O,S]))])&!$([NH1]1[CX4;!$(*([Cl,Br,F,I,O,S]))][CX4;!$(*([Cl,Br,F,I,O,S]))][NH0]=[CH0]1)&!$([NH1;!R][C;R&!$(*([Cl,Br,F,I,O]))&!$(*([NH1&R]))]=[NH0;R])&!$([n&H0][c&!$(*([Cl,Br,F,I,O]))][n][c&!$(*([Cl,Br,F,I,O]))]([N])[c])&!$([NH1&R][C&R](=[N&!R])[NH1&R])]	nNneutral	Number of neutral nitrogens	StarDrop
[N,n;!H0]	NnH	Number of nitrogens with hydrogen	StarDrop
[$([NH1+0](A)a),$([NH1+0](a)a)]	N4	Number of non-aromatic, uncharged, nitrogens with exactly one hydrogen, connected to an aromatic atom	(Ghose & Crippen, 1987)
[#7]	NbN	Number of nitrogens	StarDrop
[$(N1([CX4])CCC(CC1)*)]	fg5	Number of substituted piperidines with a non-aromatic carbon	StarDrop
[$([NH2]C([#6])=O),$([NH1]C([#6])=O)]	CamideNH	Number of nitrogen amides with one or two hydrogens	StarDrop
[$([NH0&R]([CX4])([CX4])[CX4])&!$(([NH0]([CX4])([CX4])[CX4].[OH1][C,S,P](=O)))].(c1c[c,n]ccc1)	BasicNH0R2AroRings	Number of cyclic sp3 nitrogens with three sp3 carbon connections and no hydrogens when there is a phenyl or pyridyl ring and no carboxylic, sulforic or phosphoric acid groups	StarDrop
[$([NH0]([CX4])([CX4])[CX4])&!$(([NH0]([CX4])([CX4])[CX4].[OH1][C,S,P](=O)))].(c1c[c,n]ccc1)	BasicNH02AroRings	Number of sp3 nitrogens with three sp3 carbon connections and no hydrogens when there is a phenyl or pyridyl ring and no carboxylic, sulforic or phosphoric acid groups	StarDrop
[$([NH1&R]([CX4])([CX4]))&!$(([NH1]([CX4])[CX4].[OH1][C,S,P](=O)))].(c1c[c,n]ccc1)	BasicNH1R2AroRings	Number of cyclic sp3 nitrogens with two sp3 carbon connections and one hydrogen when there is a phenyl or pyridyl ring and no carboxylic, sulforic or phosphoric acid groups	StarDrop
[$([NH1]([CX4])([CX4]))&!$(([NH1]([CX4])[CX4].[OH1][C,S,P](=O)))].(c1c[c,n]ccc1)	BasicNH12AroRings	Number of sp3 nitrogens with two sp3 carbon connections and one hydrogen when there is a phenyl or pyridyl ring	StarDrop

		and no carboxylic, sulforic or phosphoric acid groups	
[!#1;!#6;!#7;!#8;!#9;!#16;!#17;!#35;!#53]	NonOrganicAtom	Number of non-organic atoms	StarDrop
[#7,#8]-[#6,#15,#16]-[#7,#8]	PRX-time1	Number of nitrogens or oxygens connected to carbon, sulfur or phosphorus atoms and also connected to either nitrogen or oxygen	StarDrop
[#7]-[#6,#16]=[#8]	PRX-time-1	Number of amide and sulfonamide side chains	StarDrop
[$(*=,#,:*);!$(N(=O)=O)]	UB	Number of triple, aromatic and double bonds. Bonds in nitro groups are not counted	StarDrop
[$([#7;!H0;!$(*(S(=O)=O)C(F)(F)F);!$(n1nnnc1);!$(n1nncn1)]),$([#7;-])]	HDN	Number of non-negatively charged nitrogens with at least one hydrogen and in trifluorosulfonamide groups and not in tetrazole ring systems	StarDrop
[$([#7;v3&!$([nH])&!$([#7](-a)-a)])]	HAN	Number of nitrogens with total bond order equal to three other than H-pyrrole nitrogens and nitrogens connected to aromatic atoms	StarDrop
[$([#8]-[#7])]	PRX-time2	Number of oxygens connected to nitrogen	StarDrop
[$([$([#16]);!$(*=N~O);!$(*~N=O);X1,X2])]	HAS	Number of sulfur atoms not connected to nitroso or nitro groups	StarDrop
[$([$([#8,#16]);!$(*=N~O);!$(*~N=O);X1,X2]),$([#7;v3;!$([nH]);!$(*(-a)-a)])]	HAT	Number of HAN, HAO and HAS as described	StarDrop
[$([$([#8]);!$(*=N~O);!$(*~N=O);X1,X2])]	HAO	Number of oxygens not in nitro groups	StarDrop
[$([A;X4&H1,X3&H0,X5&H2,X6&H3](@*)(@*)~[!#1])]	AliRingAttachment	Number of cyclic atoms with three explicit bonds	StarDrop
[$([CH0X4]a)]	C12	Number of sp3 carbons with no hydrogens and connected to at least one aromatic atom	(Ghose & Crippen, 1987)
[$([CH1X4][N,O,P,S,F,Cl,Br,I]),$([CH0X4][N,O,P,S,F,Cl,Br,I])]	C4	Number of sp3 carbons with either no or one hydrogen connected to any of the following atoms: N, O, F, S, Cl, Br and I	(Ghose & Crippen, 1987)
[$([CH2X4]a)]	C10	Number of sp3 carbons with exactly two hydrogens and connected to at least one aromatic atom	(Ghose & Crippen, 1987)
[$([CH2]=C),$([CH1](=C)A),$([CH0](=C)(A)A)]	C6	Number of aliphatic sp2 carbons connected to one aliphatic sp2 carbon substituted by aliphatic atoms	(Ghose & Crippen, 1987)
[$([CH3][N,O,P,S,F,Cl,Br,I]),$([CH2X4][N,O,P,S,F,Cl,Br,I])]	C3	Number of sp3 carbons with two or three hydrogens and connected to any of the following atoms: N, O, F, S, Cl, Br and I	(Ghose & Crippen, 1987)
[$([CH3][a;!c])]	C9	Number of sp3 carbons with exactly three hydrogens and connected to an aromatic atom that is not carbon	(Ghose & Crippen, 1987)
[$([CH3]c)]	C8	Number of sp3 carbons with exactly three hydrogens and connected to an aromatic carbon	(Ghose & Crippen, 1987)
[$([CH4]),$([CH3]C),$([CH2](C)C)]	C1	Number of methyl groups with only one, or methylene groups with only two, carbon substituents	(Ghose & Crippen, 1987)
[$([CHX4]a)]	C11	Number of sp3 carbons with exactly one hydrogen and connected to at least one aromatic atom	(Ghose & Crippen, 1987)
[$([CH](C)(C)C),$([C](C)(C)(C)C)]	C2	Number of sp3 carbons with no or exactly one hydrogen and connected to aliphatic carbons only	(Ghose & Crippen, 1987)

[$([CX4][#5]),$([CX4][#14]),$([CX4][#15]),$([CX4][#33]),$([CX4][#34]),$([CX4][#50]),$([CX4][#80])]	C27	Number of sp3 carbons connected to any of the following atoms: Si, B, P, As, Se, Sn and Hg	(Ghose & Crippen, 1987)
[$([C](=C)(a)A),$([C](=C)(c)a),$([CH](=C)a),$([C]=c)]	C26	Number of aliphatic sp2 carbons connected to one aliphatic sp2 carbon and at least one aromatic atom	(Ghose & Crippen, 1987)
[$([N+0](=A)A),$([N+0](=A)a),$([N+0](=a)A),$([N+0](=a)a)]	N6	Number of neutral sp2 nitrogens connected to either aromatic or aliphatic atoms	(Ghose & Crippen, 1987)
[$([N+0](A)(A)A)]	N7	Number of neutral aliphatic sp3 nitrogens connected to exactly three aliphatic atoms	(Ghose & Crippen, 1987)
[$([N+0](a)(A)A),$([N+0](a)(a)A),$([N+0](a)(a)a)]	N8	Number of neutral aliphatic sp3 nitrogens connected to at least one aromatic atom	(Ghose & Crippen, 1987)
[$([N+]#A),$([N-]),$([N+](=[N-])=N)]	N14	Number of charged sp3 or sp2 aliphatic nitrogens	(Ghose & Crippen, 1987)
[$([NH+0](A)A)]	N2	Number of uncharged aliphatic nitrogens with one hydrogen	(Ghose & Crippen, 1987)
[$([NH0+](A)(A)(A)A),$([NH0+](=A)(A)A),$([NH0+](=A)(A)a),$([NH0+](=[#6])=[#7])]	N13	Number of positively charged aliphatic nitrogens with no hydrogens	(Ghose & Crippen, 1987)
[$([NH1]),$([OH1][#7])]	H3	Number of secondary amine nitrogens and hydroxyl groups connected to a nitrogen atom	(Hall, Kier, & Brown.B.B, 1995)
[$([NH2+0]A)]	N1	Number of primary amine nitrogens connected to one aliphatic atom	(Ghose & Crippen, 1987)
[$([NH2]-[CX4]),$([NH](-[CX4])-[CX4]),$(N(-[CX4])(-[CX4])-[CX4]),$([*;+;!$(*~[*;-])]),$(N=C-N),$(N-C=N)]	BasicGroup	Number of basic nitrogens	StarDrop
[$([NH3+]),$([NH2+]),$([NH+])]	N10	Number of charged aliphatic nitrogens with at least one hydrogen	(Ghose & Crippen, 1987)
[$([O;H1,-&!$(*-N=O)]),$([S;H1&X2,-&X1]),$([#7;!H0;!$(*(S(=O)=O)C(F)(F)F);!$(n1nnnc1);!$(n1nncn1)]),$([#7;-])]	HDT	Number of HDO and HDN and the number of sulfhydryl groups	StarDrop
[$([O;H1,-&!$(*-N=O)])]	HDO	Number of hydroxyl groups not in a nitro group	StarDrop
[$([O;H1]-[C,S,P]=O),$([*;-;!$(*~[*;+])]),$([NH](S(=O)=O)C(F)(F)F),$(n1nnnc1),$(n1nncn1)]	AcidGroup	Number of acidic groups	StarDrop
[$([OH1]C=[#6]),$([OH1]C=[#7]),$([OH1]C=O),$([OH1]C=S),$([OH1][O]),$([OH1][S])]	H4	Number of hydroxyl groups connected to sp2 carbons	(Hall, Kier, & Brown.B.B, 1995)
[$([OH1][CX4]),$([OH1]c),$([OH1][#5]),$([OH1][#14]),$([OH1][#15]),$([OH1][#33]),$([OH1][#33]),$([OH1][#50]),$([BH1]),$([SiH1]),$([PH1]),$([SH1]),$([SnH1])]	H2	Number of hydroxyl groups	(Hall, Kier, & Brown.B.B, 1995)
[$([OX1-;!N,S])]	O7	Number of negatively charged sp2 oxygens	(Ghose & Crippen, 1987)
[$([OX1-][#16])]	O6	Number of negatively charged sp2 oxygens connected to sulfur	(Ghose & Crippen, 1987)
[$([O](C)C),$([O](C)[A;!#6]),$([O]([A;!C])[A;!C])]	O3	Number of aliphatic sp3 oxygens connected to two aliphatic atoms	(Ghose & Crippen, 1987)
[$([O]=C([A;!C])[A;!C]),$([O]=C([A;!C])[a;!c]),$([O]=C([a;!c])[a;!c])]	O11	Number of carbonyl groups connected to no carbons	(Ghose & Crippen, 1987)

[$([O]=[#8]),$([O]=[#7]),$([OX1-][#7])]	O5	Number of sp2 oxygen connected to oxygen or nitrogen atoms	(Ghose & Crippen, 1987)
[$([O]=[CH]C),$([O]=C(C)C),$([O]=C(C)[A;!C]),$([O]=[CH]O),$([O]=[CH2]),$([O]=[CX2]=O),$([O]=[CH]N)]	O9	Number of carbonyl groups connected to at least one carbon describing aldehyde, ketone, ester, carboxylic acid or amide groups	(Ghose & Crippen, 1987)
[$([O]=[CH]c),$([O]=C(C)c),$([O]=C(c)c),$([O]=C(c)[a;!#6]),$([O]=C(c)[A;!C]),$([O]=C(C)[a;!c])]	O10	Number of carbonyl groups attached to an aromatic system	(Ghose & Crippen, 1987)
[$([S+]),$([S-])]	S2	Number of charged aliphatic sulfurs	(Ghose & Crippen, 1987)
[$([a;X4&H1,X3&H0,X5&H2,X6&H3](@*)(@*)([!#1]))]	AroRingAttachment	Number of aromatic atoms with three explicit bonds	StarDrop
[$([c](:a)(:a)=C),$([c](:a)(:a)=N),$([c](:a)(:a)=O)]	C25	Number of aromatic atoms connected to an aliphatic sp2 carbon, nitrogen or oxygen	(Ghose & Crippen, 1987)
[$([c][#5]),$([c][#14]),$([c][#15]),$([c][#33]),$([c][#34]),$([c][#50]),$([c][#80])]	C13	Number of aromatic carbons connected to any of the following atoms: Si, B, P, As, Se, Sn and Hg	(Ghose & Crippen, 1987)
[$([n+0])]	N11	Number of neutral aromatic nitrogens	(Ghose & Crippen, 1987)
[$([n+])]	N12	Number of positively charged aromatic nitrogens	(Ghose & Crippen, 1987)
[C;!$(C-[!C]);!$(C-C-[!C]);!$(C=,#*);!$(C-C=,#*)]	HydrophobicGroup	Number of aliphatic hydrophobic groups	StarDrop
[C;H1]	H1a	Number of aliphatic carbons with one hydrogen	(Hall, Kier, & Brown.B.B, 1995)
[C]=[N,O,P,S,F,Cl,Br,I]	C5	Number of aliphatic sp2 carbons connected to any of the following atoms: N, O, F, S, Cl, Br and I	(Ghose & Crippen, 1987)
[c](:a)(:a)-C	C21	Number of aromatic carbons connected to two aromatic atoms and one aliphatic carbon	(Ghose & Crippen, 1987)
[c](:a)(:a)-N	C22	Number of aromatic carbons connected to two aromatic atoms and one aliphatic nitrogen	(Ghose & Crippen, 1987)
[c](:a)(:a)-O	C23	Number of aromatic carbons connected to two aromatic atoms and one aliphatic oxygen	(Ghose & Crippen, 1987)
[c](:a)(:a)-S	C24	Number of aromatic carbons connected to two aromatic atoms and one aliphatic sulfur	(Ghose & Crippen, 1987)
[c](:a)(:a)-a	C20	Number of aromatic carbons connected to exactly three aromatic atoms	(Ghose & Crippen, 1987)
[s]	S3	Number of aromatic sulfurs	(Ghose & Crippen, 1987)
[$([CX4&!$(*~[Cl,Br,I,F,O,N])][a])]	ed70	Number of sp3 carbons connected to at least one aromatic atom and not connected to any of the first four halogens, oxygen or nitrogen	StarDrop
[$([N&H0][a])]	ed20	Number of nitrogens with no hydrogens and connected to an aromatic atom	StarDrop
[$([N&H1](C(=O))([CX4])[a])]	ed50	Number of amide nitrogens with exactly one hydrogen connected to one aromatic atom and one sp3 carbon	StarDrop

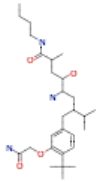
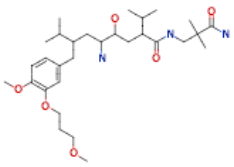
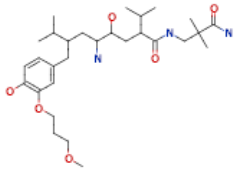
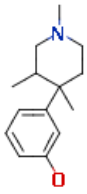
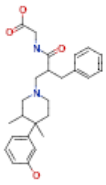
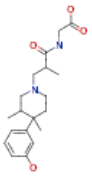
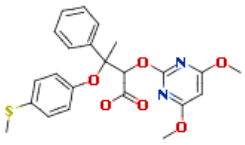
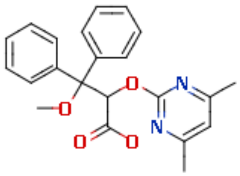
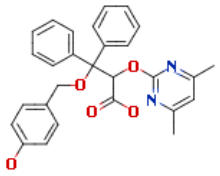
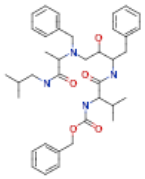
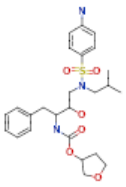
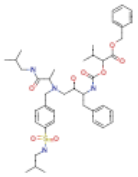
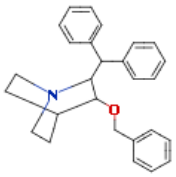
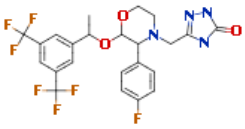
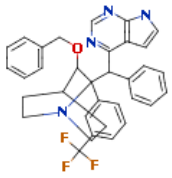
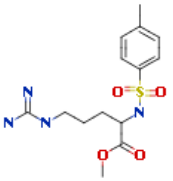
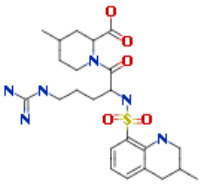
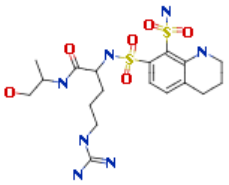
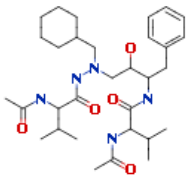
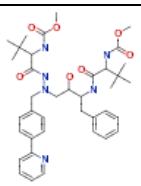
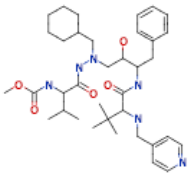
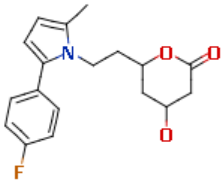
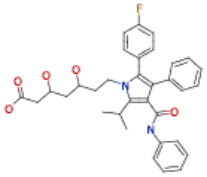
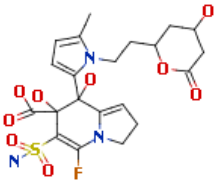
[$([O&H0](C(=O))([CX4])[a])]	ed60	Number of esteric oxygens with exactly no hydrogens connected to one aromatic atom and one sp3 carbon	StarDrop
[$([O&H0][a])]	ed40	Number of sp3 oxygens with no hydrogens connected to an aromatic atom	StarDrop
[$([c]-[a])]	ed80	Number of aromatic carbons connected to at least one aromatic atom	StarDrop
[$(C(#N)[a])]	ew70	Number of nitrile groups connected to one aromatic atom	StarDrop
[$(C(F)(F)(F)[a])]	ew60	Number of trifluoromethyl groups connected to one aromatic atom	StarDrop
[$(N(=O)(=O)[a])]	ew90	Number of nitro groups connected to one aromatic atom	StarDrop
[$(S(=O)(=O)([OH1])[a])]	ew80	Number of sulfo groups connected to one aromatic atom	StarDrop
[$(S(=O)(=O)N)]	ew75	Number of sulfonamide groups	StarDrop
[$([CX2&H0](=O)([CX4])[a])]	ew30	Number of ketone groups connected to one aromatic atom and one sp3 carbon	StarDrop
[$([CX2&H0](=O)([Cl])[a])]	ew50	Number of chloroformyl groups connected to one aromatic atom	StarDrop
[$([CX2&H0](=O)([OH1])[a])]	ew40	Number of carboxyl groups connected to one aromatic atom	StarDrop
[$([CX2&H1](=O)[a])]	ew20	Number of aldehyde groups connected to one aromatic atom	StarDrop
[$([Cl,I,F,Br][a])]	ew10	Number of fluorine, chlorine, bromine and iodine atoms connected to one aromatic atom	StarDrop
c1[c,n,s,o][c,o,n,s]n[c,n,s,o]1	ew100	Number of 5-member aromatic rings with at least one aromatic carbon and one aromatic nitrogen	StarDrop
a(!:*):*:*:a!:*	f004	Number of para interactions between substituents singly or doubly bonded to the aromatic atoms	(Olah, Bologa, & Oprea, 2004)
a(-&!:*):*:*:a-&!:*	f005	Number of para interactions between substituents singly bonded to the aromatic atoms	(Olah, Bologa, & Oprea, 2004)
a(-&!:*):*:a-&!:*	f007	Number of meta interactions between substituents singly bonded to the aromatic atoms	(Olah, Bologa, & Oprea, 2004)
[#8](!@*)!@*	f015	Number of sp3 non cyclic oxygens	(Olah, Bologa, & Oprea, 2004)
[CX4H3]-[!#6]	f147	Number of sp3 carbons with exactly three hydrogens and connected to an atom other than carbon	(Olah, Bologa, & Oprea, 2004)
[#6!H0]~&@[#6!H0]~&@[#6!H0]~&@[#6!H0]	f244	Number of cyclic chains of four carbon atoms which have at least one hydrogen	(Olah, Bologa, & Oprea, 2004)
[#6!H0]~&@[#6!H0]~&@[#6!H0]~&@[#6!H0]~&@[#6!H0]	f245	Number of cyclic chains of five carbon atoms which have at least one hydrogen	(Olah, Bologa, & Oprea, 2004)
[#7]~*~*~*~*~*~*~*~*~[#8]	f301	Number of groups containing a nitrogen atom separated from an oxygen atom by nine bonds	(Olah, Bologa, & Oprea, 2004)
[#9,#17,#35,#53]!@*@*	f390	Number of fluorine, chlorine, bromine and iodine atoms connected to an atom in the alpha position of a cyclic atom	(Olah, Bologa, & Oprea, 2004)

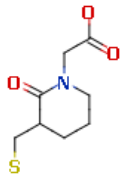
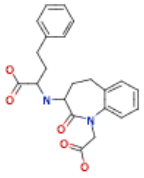
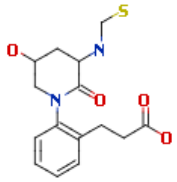
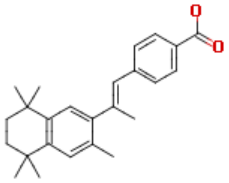
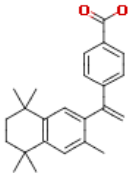
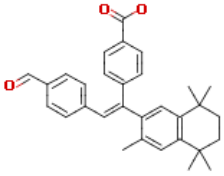
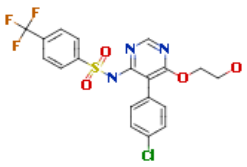
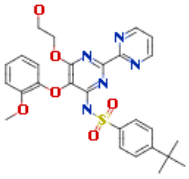
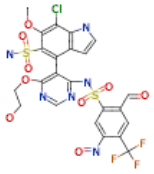
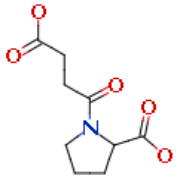
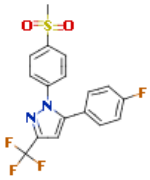
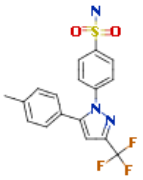
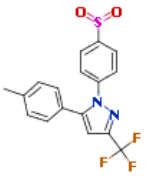
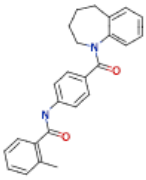
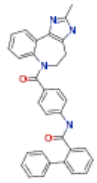
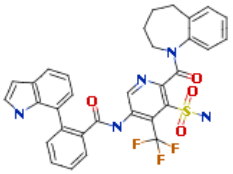
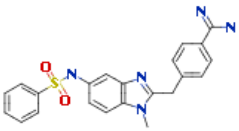
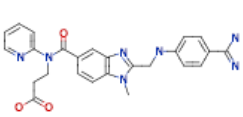
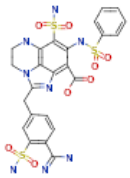
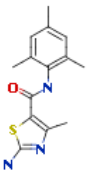
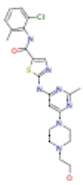
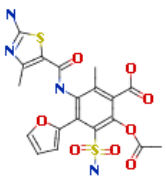
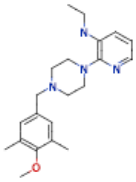
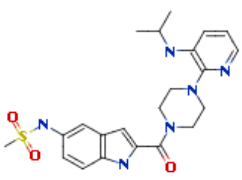
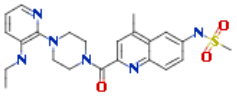
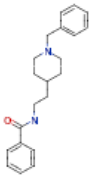
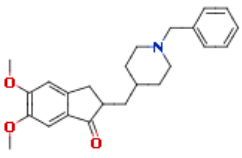
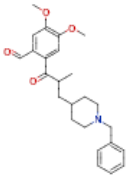
[#9,#17,#35,#53]-[R]	f392	Number of fluorine, chlorine, bromine and iodine atoms connected to a cyclic atom	(Olah, Bologa, & Oprea, 2004)
[#9,#17,#35,#53]~*(~*)~*	f393	Number of fluorine, chlorine, bromine and iodine atoms connected to a disubstituted atom	(Olah, Bologa, & Oprea, 2004)
[$([#6]-&!@[A!C!H1]-&!@[#6])]	f407	Number of carbons connected to one non-cyclic carbon atom and to one non-cyclic atom other than carbon	(Olah, Bologa, & Oprea, 2004)
[$([#8X2v2]([#6])[#6])]	f413	Number of non-aromatic oxygens connected to two carbons	(Olah, Bologa, & Oprea, 2004)
[CX4H3]-*~*~a	f440	Number of methyl groups separated by three bonds from an aromatic atom	(Olah, Bologa, & Oprea, 2004)
[CX4H3]-*~*~[R]	f441	Number of methyl groups separated by three bonds from a ring atom	(Olah, Bologa, & Oprea, 2004)
[CX4H3]-*~a	f443	Number of methyl groups separated by two bonds from an aromatic atom	(Olah, Bologa, & Oprea, 2004)
[CX4H3]-*~[R]	f444	Number of methyl groups separated by two bonds from a ring atom	(Olah, Bologa, & Oprea, 2004)
[#9,#17,#35,#53]~[X4&H1,X3&H0,X5&H2,X6&H3]	f456	Number of fluorine, chlorine, bromine and iodine atoms connected to an atom with three explicit connections	(Olah, Bologa, & Oprea, 2004)
(!@)!@*	q017	Number of non-cyclic side chains consisting of at least three atoms	(Olah, Bologa, & Oprea, 2004)
[R](-*-*):a	q039	Number of aromatic atoms with a two atom substitution where the atoms are singly bonded	(Olah, Bologa, & Oprea, 2004)
[R](-*-*)~*:a	q040	Number of atoms in a ring separated from an aromatic atom by two bonds with at least one aromatic bond	(Olah, Bologa, & Oprea, 2004)
[R](-*-*)~*~*:a	q041	Number of atoms in a ring separated from an aromatic atom by three bonds with at least one aromatic bond.	(Olah, Bologa, & Oprea, 2004)
[R]~*~*~[!#6]	q137	Number of atoms in a ring separated from a non-carbon atom by three bonds	(Olah, Bologa, & Oprea, 2004)
[R]~*~[!#6]	q139	Number of atoms in a ring separated from a non-carbon atom by two bonds	(Olah, Bologa, & Oprea, 2004)
[!$(*#*);!$([X6&H5,X5&H4,X4&H3,X3&H2,X2&H1,X1])]-&!@[!$(*#*);!$([X6&H5,X5&H4,X4&H3,X3&H2,X2&H1,X1])]	q155	Number of rotatable bonds	(Olah, Bologa, & Oprea, 2004)
[!a](=*)~*~*~*~[R]	q192	Number of sp2 non-aromatic atoms separated from another ring atom by four bonds	(Olah, Bologa, & Oprea, 2004)
[#6]~&!@[#8]	q257	Number of carbons connected to oxygen by a non-cyclic bond	(Olah, Bologa, & Oprea, 2004)
[#6X3v3-,#6X3v4+0,#6X3v3+]	q277	Number of sp2 carbons connected to at least three heteroatoms	(Olah, Bologa, & Oprea, 2004)
[#7]~*(~*)~*	q300	Number of nitrogens in the alpha position of a branched substituent	(Olah, Bologa, & Oprea, 2004)
[#8,#16]	q358	Number of sulfurs and oxygens	(Olah, Bologa, & Oprea, 2004)

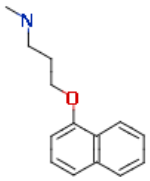
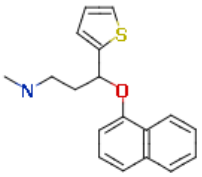
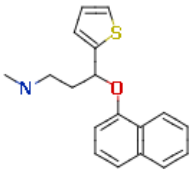
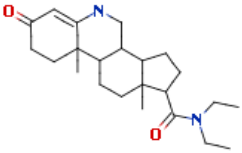
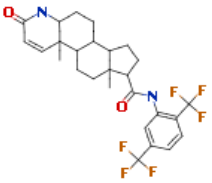
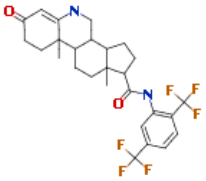
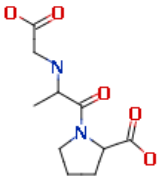
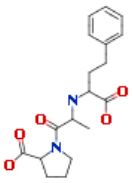
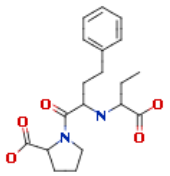
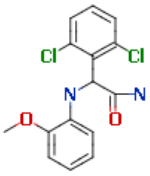
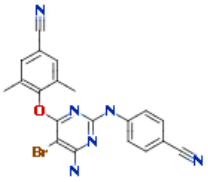
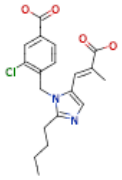
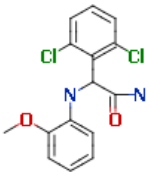
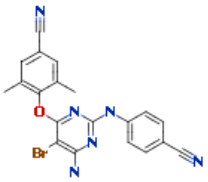
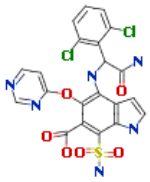
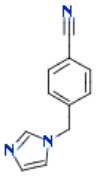
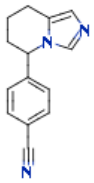
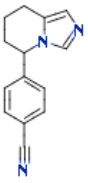
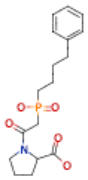
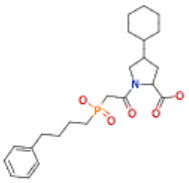
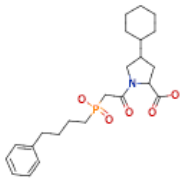
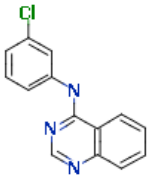
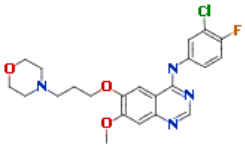
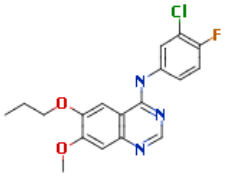
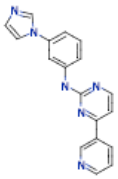
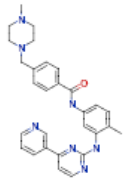
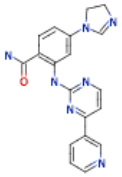
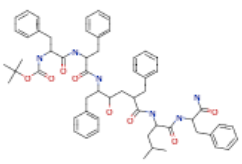
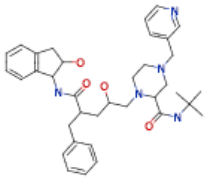
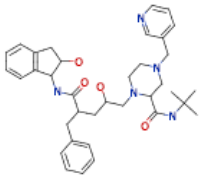
[X4&H1,X3&H0,X5&H2,X6&H3]~*~*~*~[!#6]	q453	Number of atoms with three explicit connections separated by four bonds from an atom other than carbon	(Olah, Bologa, & Oprea, 2004)
[#7]~[X4&H1,X3&H0,X5&H2,X6&H3]	q457	Number of atoms with three explicit connections connected to a carbon atom	(Olah, Bologa, & Oprea, 2004)
[X4&H1,X3&H0,X5&H2,X6&H3]~[R]	q458	Number of atoms with three explicit connections connected to a cyclic atom	(Olah, Bologa, & Oprea, 2004)
[R](-*-*)~*~*~*~a	q481	Number of aliphatic atoms in a ring with two singly bonded atoms connected and separated from an aromatic atom by four bonds	(Olah, Bologa, & Oprea, 2004)
[R](-*-*)~*~*~a	q483	Number of aliphatic atoms in a ring with two singly bonded atoms connected and separated from an aromatic atom by three bonds	(Olah, Bologa, & Oprea, 2004)
[R](-*-*)~*~a	q485	Number of aliphatic atoms in a ring with two singly bonded atoms connected and separated from aromatic atoms by two bonds	(Olah, Bologa, & Oprea, 2004)
c1(*)[c;H1][c;H1]c(*)[c;H1][c;H1]1	frg-8	Number of para interactions in a disubstituted benzene ring with no other substituents	(Olah, Bologa, & Oprea, 2004)
[$([NH0;X3;+0]([CX4])([CX4])[a])]	frg-26	Number of sp3 nitrogens with no hydrogens connected to two sp3 carbons and one aromatic sp2 carbon	(Olah, Bologa, & Oprea, 2004)
[$([CX4;H1](*)1[CX4;H2][CX4;H2][CX4;H2][CX4;H2][CX4;H2]1)]	frg-54	Number of six member aliphatic rings with sp3 carbons	(Olah, Bologa, & Oprea, 2004)
[N,n&H0]	Nn	Number of nitrogens with no hydrogens	StarDrop

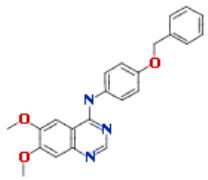
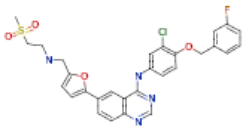
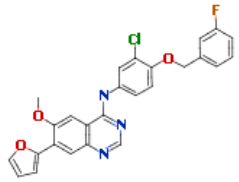
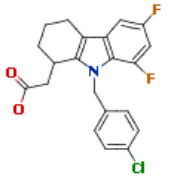
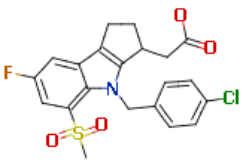
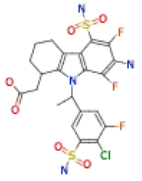
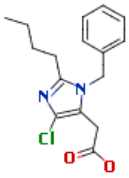
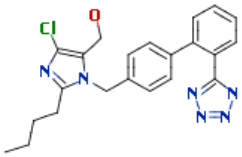
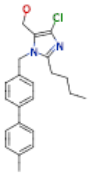
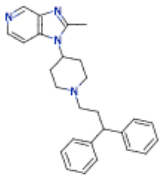
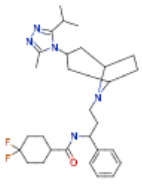
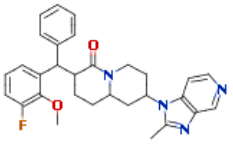
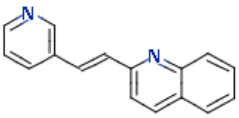
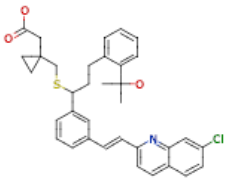
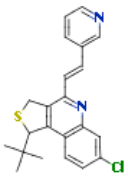
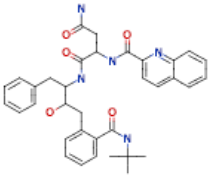
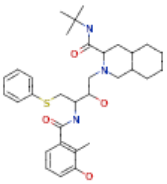
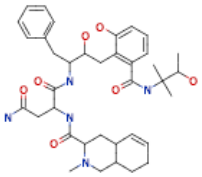
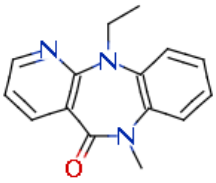
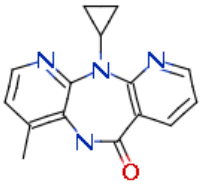
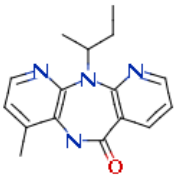
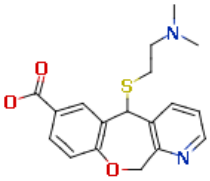
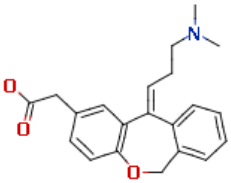
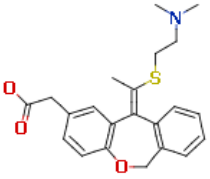
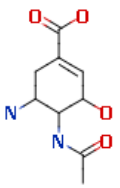
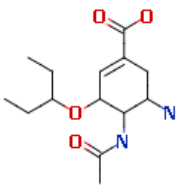
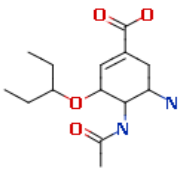
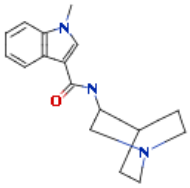
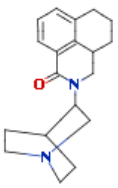
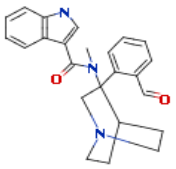
15.3 Results of Lead to Drug Transformations for Nova Validation

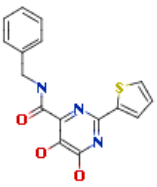
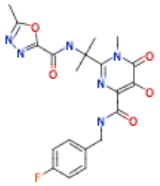
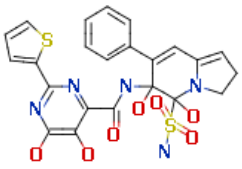
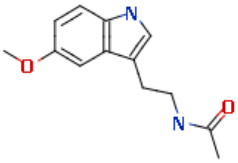
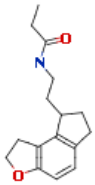
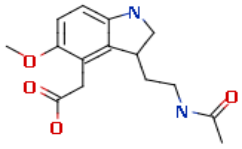
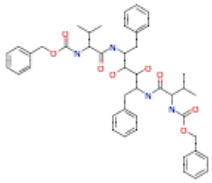
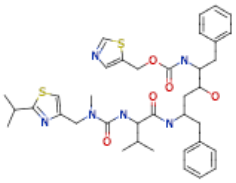
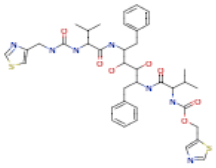
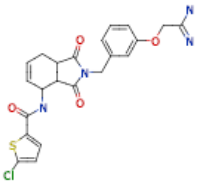
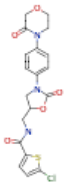
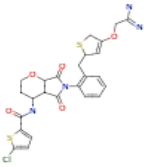
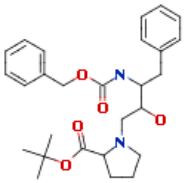
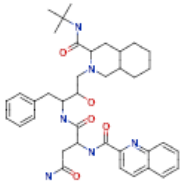
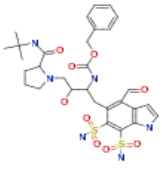
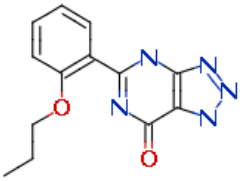
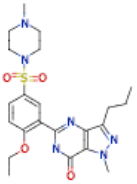
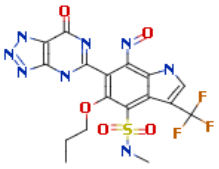
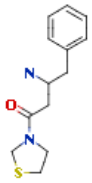
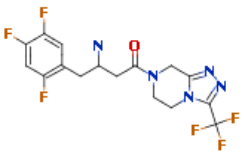
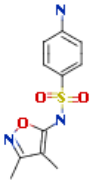
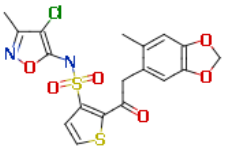
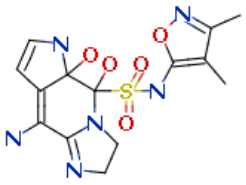
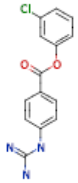
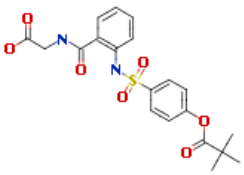
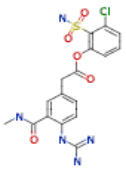
Table 28 Initial leads, marketed drugs and closest child compounds generated in 5 generations for Perola Set

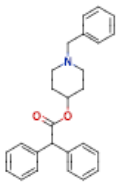
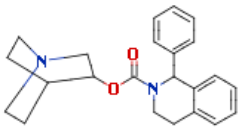
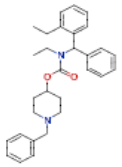
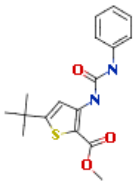
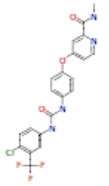
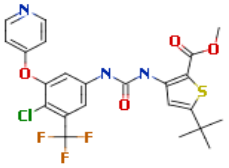
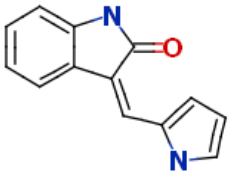
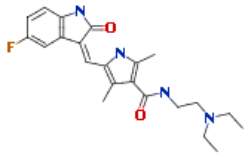
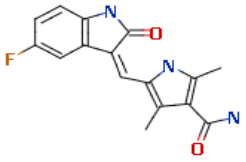
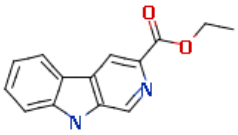
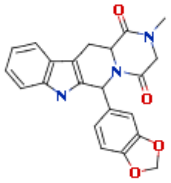
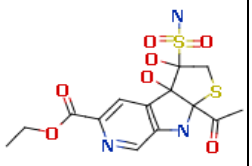
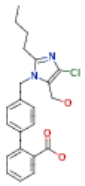
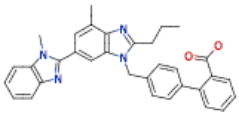
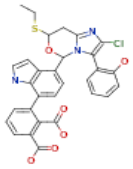
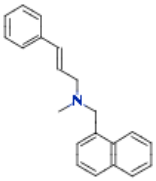
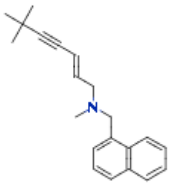
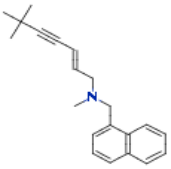
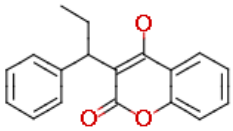
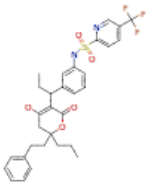
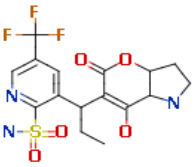
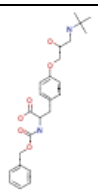
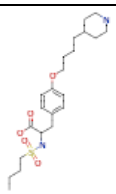
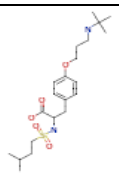
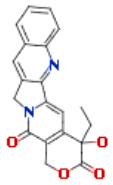
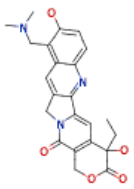
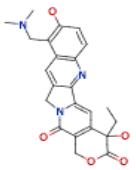
Drug Name	Initial Lead Structure	Drug Structure	Closest Child Generated	Lead-Drug Similarity	Child-Drug Similarity
ALISKIREN				0.711	0.985
ALVIMOPAN				0.744	0.937
AMBRISENTAN				0.659	0.931
AMPRENAVIR				0.621	0.737
APREPITANT				0.530	0.720
ARGATROBAN				0.540	0.862
ATAZANAVIR				0.755	0.842
ATORVASTATIN				0.644	0.925

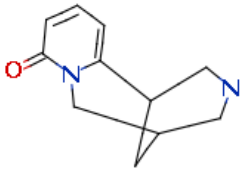
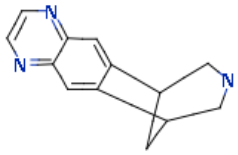
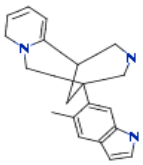
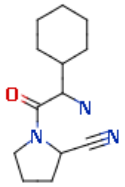
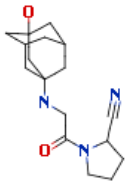
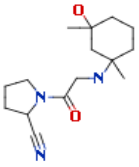
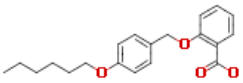
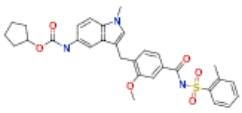
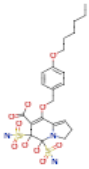
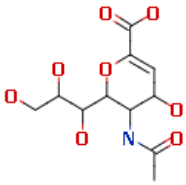
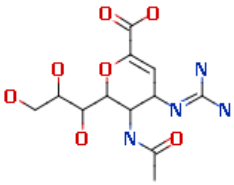
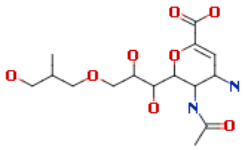
BENAZEPRILAT				0.468	0.786
BEXAROTENE				0.743	0.911
BOSENTAN				0.724	0.825
CAPTOPRIL				0.791	1.000
CELECOXIB				0.905	0.967
CONIVAPTAN				0.643	0.851
DABIGATRAN				0.639	0.794
DASATINIB				0.624	0.793
DELAVIRDINE				0.650	0.911
DONEPEZIL				0.493	0.920

DULOXETINE				0.497	1.000
DUTASTERIDE				0.549	0.838
ENALAPRILAT				0.787	0.921
EPROSARTAN				0.659	0.821
ETRAVIRINE				0.473	0.660
FADROZOLE				0.488	1.000
FOSINOPRILAT				0.878	1.000
GEFITINIB				0.651	0.952
IMATINIB				0.650	0.786
INDINAVIR				0.675	0.839

LAPATINIB				0.674	0.833
LAROPIPRANT				0.893	0.938
LOSARTAN				0.733	0.866
MARAVIROC				0.604	0.729
MONTELUKAST				0.461	0.732
NELFINAVIR				0.676	0.768
NEVIRAPINE				0.823	0.990
OLOPATADINE				0.634	0.846
OSELTAMIVIR CARBOXYLATE				0.835	1.000
PALONOSETRON				0.631	0.771

RALTEGRAVIR				0.682	0.859
RAMELTEON				0.416	0.645
RITONAVIR				0.714	0.884
RIVAROXABAN				0.619	0.688
SAQUINAVIR				0.614	0.807
SILDENAFIL				0.653	0.882
SITAGLIPTIN				0.402	0.709
SITAXENTAN				0.570	0.842
SIVELESTAT				0.380	0.651

SOLIFENACIN				0.446	0.795
SORAFENIB				0.479	0.634
SUNITINIB				0.721	0.957
TADALAFIL				0.502	0.919
TELMISARTAN				0.718	0.845
TERBINAFINE				0.757	1.000
TIPRANAVIR				0.601	0.879
TIROFIBAN				0.528	0.883
TOPOTECAN				0.985	1.000

VARENICLINE				0.455	0.604
VILDAGLIPTIN				0.592	0.970
ZAFIRLUKAST				0.373	0.838
ZANAMIVIR				0.797	0.908
Average				0.636	0.853

15.4 File Formats

StarDrop provides a number of opportunities for customisation where it is possible to import your own data to be used within its algorithms. The following sections describe the file formats used in each case.

15.4.1 Descriptors/Filters

Descriptors for the Auto-Modeller and filters to use when tidying a data set or running Nova can be imported as SMARTS patterns (see Section 8.2.1 for more information about SMARTS). To define your own you must create a text file containing SMARTS and their associated names. The SMARTS patterns must not contain any spaces and there should be a space to separate the pattern from the name, with one pattern and name on each line of the file.

Example:

```
[S,C](=[O,S])[F,Br,Cl,I] acid halide
[Cl]C([C&R0])=N chloramidine
[P,S][F,Cl,Br,I] P/S halide
```

15.4.2 Transformations

Custom Nova transformations can be imported as SMIRKS patterns (Weininger, 1998). The name of the file will be used to provide the name of the new group created in the tree displayed in the StarDrop client. The file must be a text file and each line should contain either two or three tab delimited entries. The first entry must be the SMIRKS pattern and the second entry should be a name of this pattern (this will be displayed in the tree). The third entry is an optional reference for this pattern.

Example:

```
[C:1][CH2][C:2]>>[C:1]O[C:2]      Secondary carbon to ether  Reference1
[C:1][CH2][C:2]>>[C:1]N[C:2]      Secondary carbon to amine  Reference2
[C:1][CH2][C;!R:2]>>[C:1][C;!R:2] Remove secondary carbon
```

15.4.3 Fragments

Custom fragments to use during library enumeration can be imported in SD file format. The SD file format is designed to define a complete structure, whereas a fragment is specifically a sub-structure that can be connected. As such, appropriate data elements must be provided to indicate which atoms and bonds describe the actual fragment that will be used and which atoms and bonds are placeholders for the attachment point.

The <FragmentAtomIds> tag must be used to contain a semi-colon delimited list of indices of the atoms in the fragment.

The <FragmentBondIds> tag must be used to contain a semi-colon delimited list of indices of the bonds in the fragment.

The <CollectionName> tag is optional. Where used it provides the name of the group in which this fragment will be displayed within the StarDrop client. Where this is not provided, the filename will be used as the name of the group.

Example:

```
trifluoro
StarDropFragmentManager_1

  8 7 0 0 0 0 0 0 0 0999 V2000
     3.732 0.5 0 C 0 0 0 0 0 0 0 0 0 0 0 0
     2.866 0 0 C 0 0 0 0 0 0 0 0 0 0 0 0
     2.366 0.866 0 F 0 0 0 0 0 0 0 0 0 0 0 0
         2 -0.5 0 F 0 0 0 0 0 0 0 0 0 0 0 0
     3.366 -0.866 0 F 0 0 0 0 0 0 0 0 0 0 0 0
     4.042 -0.03694 0 H 0 0 0 0 0 0 0 0 0 0 0 0
     4.269 0.81 0 H 0 0 0 0 0 0 0 0 0 0 0 0
     3.422 1.037 0 H 0 0 0 0 0 0 0 0 0 0 0 0
  1 2 1 0 0 0 0
```

```
  2  3  1  0  0  0  0
  2  4  1  0  0  0  0
  2  5  1  0  0  0  0
  1  6  1  0  0  0  0
  1  7  1  0  0  0  0
  1  8  1  0  0  0  0
M  END
> <FragmentAtomIds>
2;3;4;5;
>
> <FragmentBondIds>
2;3;4;
>
> <CollectionName>
Halogens
>
$$$$
```

15.5 Legacy Reference

The following models have been superseded by newer, improved models but are still available.

15.5.1 Log ([brain]/[blood]) (version 5.2)

Data set

The data set consists of 292 structures with a reported logarithm of the concentration ratio between brain tissue and plasma (log(BB)) which were derived from various literature sources. The data set is largely the same as the set used by Abraham *et al* (Abraham, Ibrahim, Zhao, & Acree, 2006). The data set contains 86 volatile compounds and 206 non-volatile compounds. The model was trained on 205 compounds and tested on 87 compounds.

Model output

The model was built by the automatic procedure implemented within the Auto-Modeller using the standard settings. The initial set was split into a training set (205 compounds), validation set (44 compounds) and test set (43 compounds) by using cluster analysis at Tanimoto level 0.7. The model was produced by the non-linear Radial Basis Function technique combined with a genetic algorithm to assist in descriptor selection (GA-RBF). The model uses 29 descriptors including logP, McGowan's volume, negative charge, polar surface area, hydrogen bond terms and counts of different atomic and functional groups.

The model predicts the log(BB) value for each compound, along with an estimate of the RMSE in prediction. The distance of each predicted compound from the descriptor-space of the training set, referred to as the chemical space of the model, is calculated in order to gauge the validity of the results. The model automatically determines whether or not a test compound lies within the chemical space. There are insufficient compounds in the validation and test sets available outside the chemical space to obtain a rigorous estimate of the confidence for such compounds. In these cases, a prediction is returned, but the standard error in prediction is left as undefined (returned in the software as infinity) to indicate that the prediction must be treated with caution. The RMSE in prediction for compounds within the chemical space is 0.27 log units and the RMSE in prediction for compounds outside, but in close proximity to, the chemical space is 0.47 log units.

It is a feature of the RBF technique that it will always provides a perfect fit for the training set. However, on the combined validation and test sets the model achieves an R^2 of 0.74 with an RMSE of prediction of 0.32 log units (see Figure 15.8). The model performance was also evaluated on separated subsets of volatile and non-volatile compounds from the combined validation and test sets. For the 56 non-volatile compounds the R^2 is 0.69 and the RMSE is 0.38 log units; for the 31 volatile compounds the R^2 is 0.88 and the RMSE is 0.14 log units.

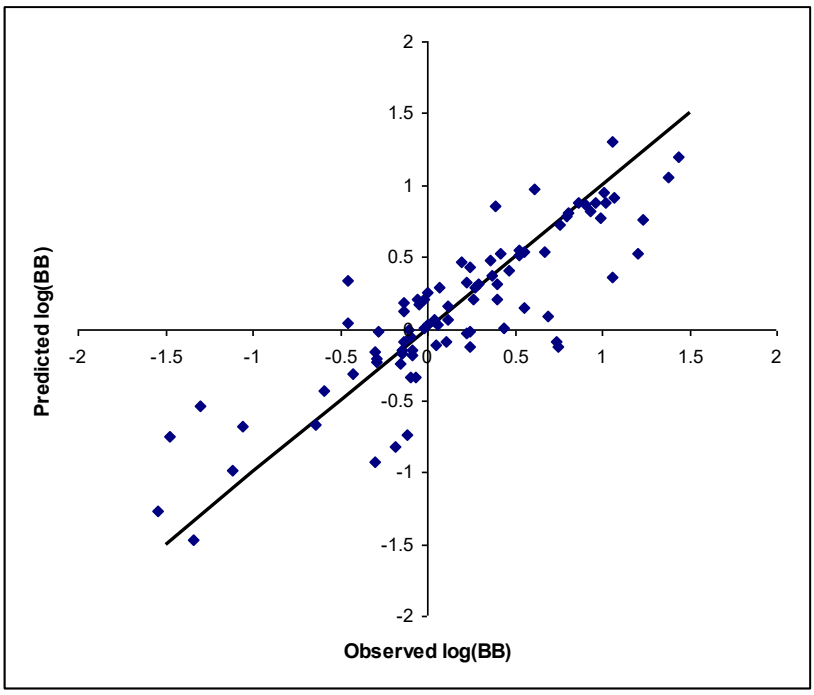


Figure 15.15 Plot of observed versus predicted log(BB) values for the combined validation and test sets

The model was also tested against an independent set of 1599 structures categorised as BBB+/BBB-, but which had no log(BB) values. This dataset was recently used by Zhao *et al.* (Zhao, et al., 2007)to develop classification models for BBB permeation. The results of this test can be seen in Figure 15.16, which illustrate that the predicted log(BB) distributions for the BBB+ and BBB- are clearly different. There is significant overlap of these distributions between log(BB) values of –1 and 0, where there is a moderate degree of BBB penetration, and activity will be dependent on the potency of the compound. Mispredictions of BBB- compounds as penetrating the blood-brain barrier are believed to be examples of active efflux of compounds by transport proteins such as P-gp, which would otherwise penetrate the blood-brain barrier by passive diffusion.

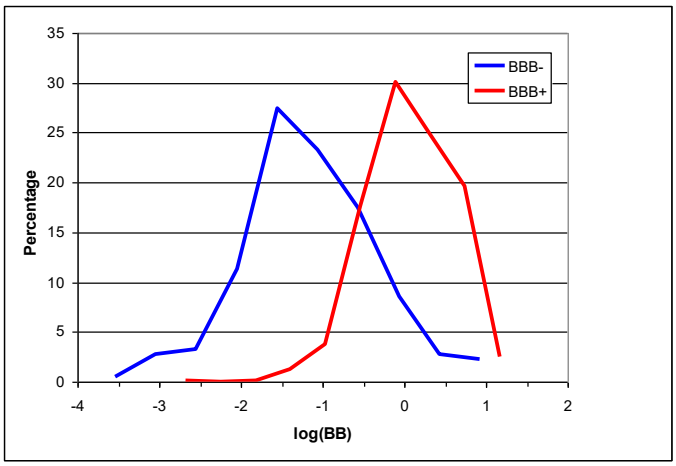


Figure 15.16 Frequency distribution of predicted log(BB) values for 360 compounds observed as BBB- and 1239 compounds observed as BBB+. Percentages are calculated within BBB+ and BBB- subsets.

Comparison with other predictive techniques

Recently, Abraham and Hersey (Abraham, Hersey, Testa, & H., 2006) reviewed published continuous blood-brain barrier penetration models and concluded that a number of models can predict log(BB) values with an RMSE error of 0.3-0.35 log units, as also shown by Abraham *et al* (Abraham, Ibrahim,

Zhao, & Acree, 2006) The estimated experimental error in log(BB) measurements is approximately 0.3 log units. Therefore, the RMSE of the StarDrop model compares well with published models.

15.5.2 BBB classification (version 5.1)

The data set consists of 201 structures classified as BBB+ and BBB- that are reported in literature models. This data was divided into a training set containing 101 compounds with an even distribution between BBB+ and BBB- compounds and an internal evaluation set of 48 compounds, with a 3.5:1 ratio between BBB+ and BBB- compounds, which was used to monitor the training of the model. The remaining 52 structures were utilized as an independent test set with a 1:2 ratio of BBB+ and BBB- compounds.

Model output

The model uses McGowan volume plus eight 2D descriptors relating to the numbers of hydrogen bond donors and acceptors, potential ionisation and overall polarity of the compounds to produce a decision tree. The descriptors used are consistent with the general observations that neutral molecules tend to penetrate the CNS better than charged compounds and that cations generally penetrate the CNS better than anions.

The model generates a prediction for each compound as BBB crossing (BBB+) or non-crossing (BBB-). This is based on a nominal classification boundary of log(BB)=-0.5 between BBB- and BBB+ compounds.

For the independent test set, 91% of BBB- predictions were correct in relation to the known category, whereas BBB+ predictions were correct in 83% of cases. Overall training and test set classifications were 96% and 93% correct respectively. The model also correctly predicted the BBB+ category for 18 of the top 20 best-selling drugs in 2003. The only incorrect predictions were for compounds identified as substrates for active uptake or efflux transporter proteins.

A confidence for each prediction is reported, according to the strength of association of the compound's descriptor values with the predicted classification. Furthermore, the distance of the predicted compound from the chemical space of the training set is calculated to gauge the confidence in the result. As there are insufficient data points outside the chemical space of the training set to assess the confidence in predictions, no estimate regarding the confidence for such compounds can be made. In these cases, the probability that the result is correct is reported as 0.5, indicating an even distribution between the two possible classes.

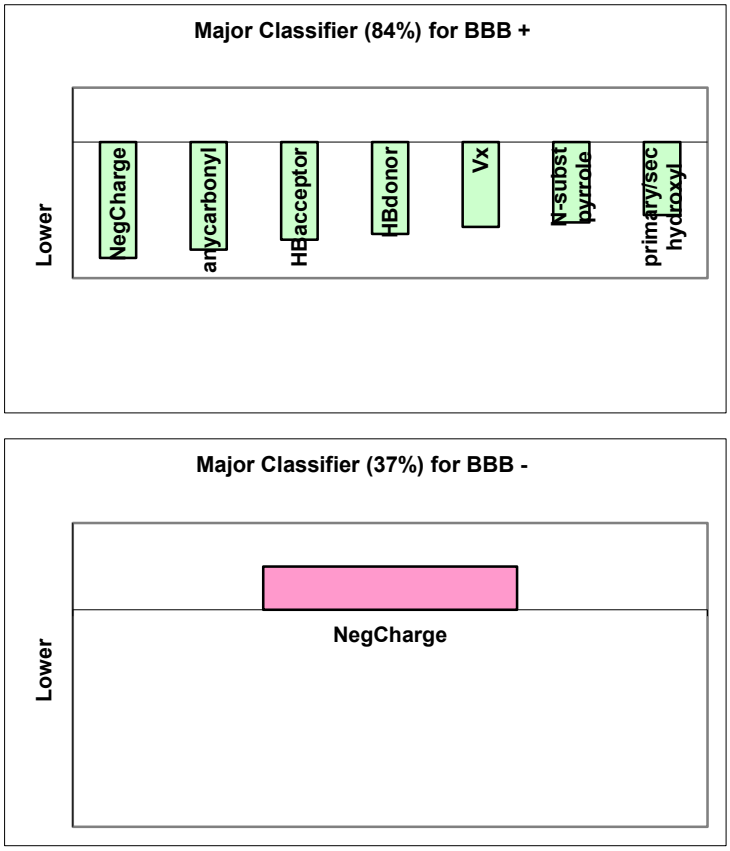


Figure 15.17 Histograms indicating the influence of descriptors on the dominant rules of the categorized BBB+/- model. The directions of bars relative to the horizontal axis indicate whether the value of a descriptor must be higher or lower than a threshold. The length of a bar reflects the number of compounds retained if the condition is met. The vertical scale is uniform for both plots. Percentages in parentheses refer to the proportion of the class predictions made by the rule.

Comparison with other predictive techniques

The model statistics compare well to recent literature BBB classification models (Crivori, Cruciani, Carrupt, & Testa, 2000) (Ajay, Bemis, & Murcko, 1999) (Engkvist, Wrede, & Rester, 2003) (Keseru, Molnar, & Greiner, 2000) (Doniger, Hofmann, & Yeh, 2002) where BBB+ prediction accuracy ranges from 80% to 100% and BBB- prediction accuracies lie between 65% and 87%.

15.5.3 CYP2C9 pKi (version 5.1)

Data set

The data for this model were generated in-house, due to the high inter-laboratory variation observed in reported P450 affinities in the literature. The data consist of accurate K_i values generated for competitive inhibitors using a multi-point K_i protocol. Data for a total of 130 compounds were generated in this data set covering a wide range of chemical diversity.

Model output

A rule-based continuous model for CYP2C9 inhibition was developed using nine 2D descriptors, including compound lipophilicity, size and aromaticity, as well as presence of certain types of nitrogen atom. Three rules were defined by the presence of cationic charges on nitrogen atoms and the number of sp^2 carbons. A partial least square equation (SIMCA 8, Umetrics) was built for each rule. The influence of each descriptor in each rule is displayed in Figure 15.18. For instance, the pK_i values for compounds in rule 1 are positively influenced by the number of sp2 carbons and the size of the molecule. On the other hand, the presence of carbonyl groups tends to lessen pK_i values and hence affinity of the compound for CYP2C9.

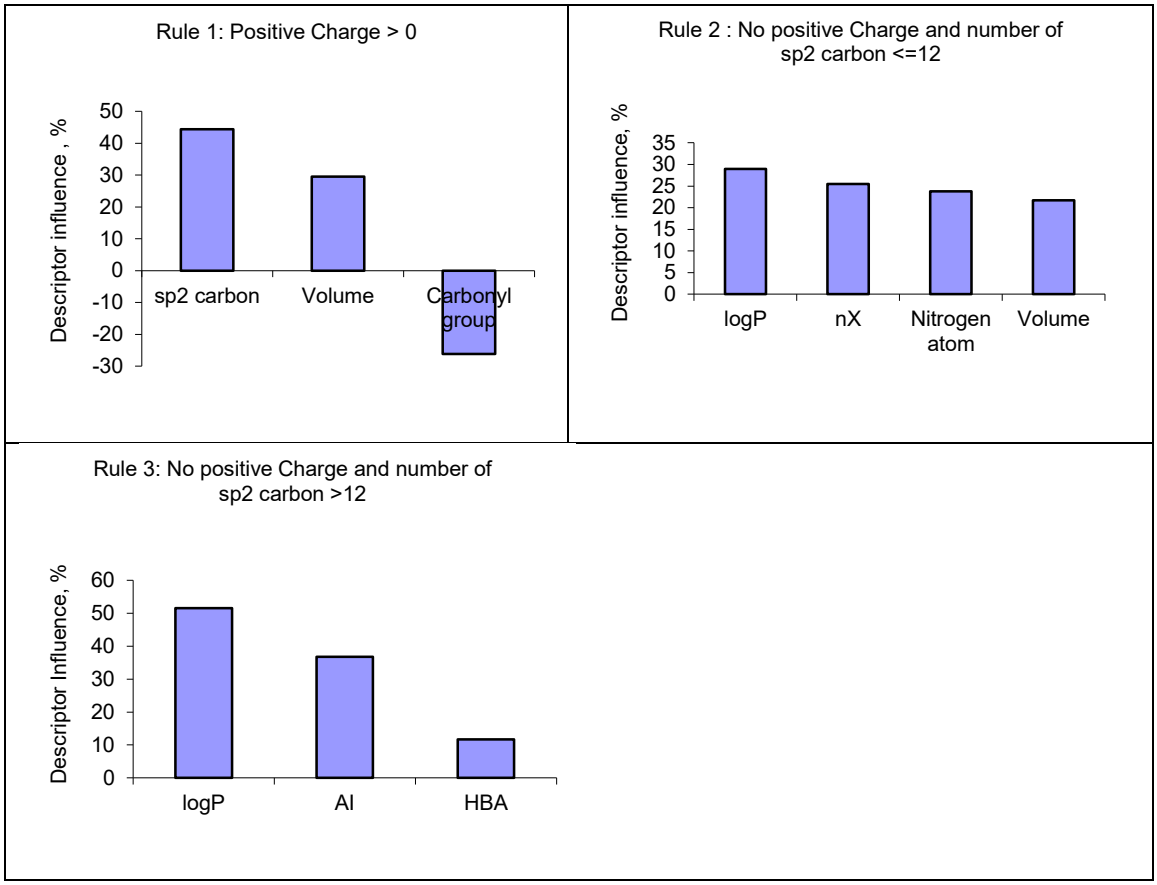


Figure 15.18 Descriptors' importance in each rule where nX is the number of halogen atoms, AI represents an aromaticity index and HBA the number of hydrogen bond acceptors.

The model outputs a prediction of a compound's pK_i along with an estimate of the RMSE in prediction. The model automatically determines whether or not a test set compound lies within the chemical space formed by the training set. As there are insufficient compounds available outside the training set chemical space, no rigorous estimate regarding the confidence for such compounds can be made. In these cases, a prediction is returned, but the standard error in prediction is undefined (shown as infinity).

The observed R^2 for the training set of 105 compounds was 0.78 and the RMSE in fit was 0.536 log units. The R^2 value for the independent test set of 25 compounds was 0.62 (see Figure 15.19) and the standard error in prediction was 0.625 log units.

Comparison with other predictive techniques

There has been significant published work on quantitative structure activity relationships for affinity to CYP2C9; in particular by David Lewis *et al*. (Lewis D. F., Essential requirements for substrate binding affinity and selectivity toward human CYP2 family enzymes, 2003) (Lewis D. F., On the recognition of mammalian microsomal cytochrome P450 substrates and their characteristics: towards the prediction of human p450 substrate specificity and metabolism, 2000) (Lewis, Modi, & Dickins, Structure-activity relationship for human cytochrome P450 substrates and inhibitors, 2002) and Sean Ekins *et al*. (Ekins, de Groot, & Jones, Pharmacophore and three-dimensional quantitative structure activity relationship methods for modeling cytochrome p450 active sites, 2001) The majority of the *in silico* models proposed for identifying compounds with high CYP2C9 affinity are based on few training cases and require 3D structures (Afzelius, et al., 2004).

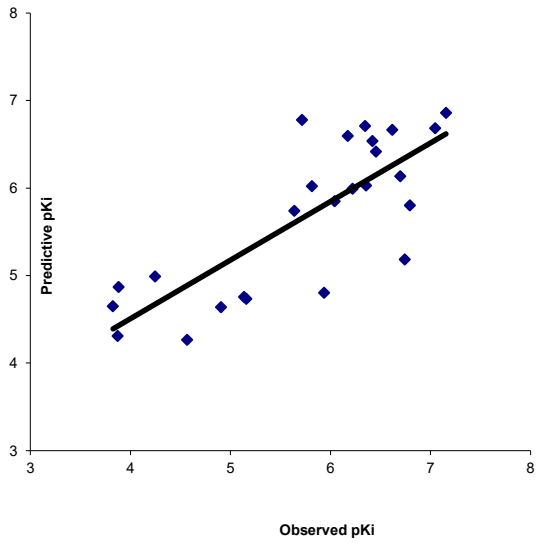


Figure 15.19 Observed versus predicted pK_i for CYP2C9 Affinity.

15.5.4 P-gp transport classification (version 5.1)

P-gp is an ATP driven efflux pump encoded by the MDR1 gene, capable of transporting a wide spectrum of chemical structures as well as different classes of drugs (Selwood, et al., 1990). Active transport by P-glycoprotein (P-gp) can represent a serious hurdle for pharmaceuticals as transport by P-gp has been associated with reduced bioavailability of orally administered drugs and with decreased ability of drug candidates to cross blood-tissue barriers such as the blood-brain barrier (Ayrton & Morgan, 2001). In addition, if a drug is subject to significant P-gp efflux, its distribution, absorption and elimination could be altered by potent P-gp inhibitors. Evidence for drug-drug interactions due to inhibition of P-gp have been reported in human clinical studies (Schwab, Fischer, Tabatabaei, Poli, & Huwyler, 2003). This is best documented for quinidine-digoxin interactions in which decreased renal and intestinal clearance of digoxin and increased plasma drug levels have been reported when quinidine is administered to patients taking digoxin (Hochman, Yamazaki, Ohe, & Lin, 2002). These changes have been attributed to inhibition of P-gp by quinidine where a significant portion of digoxin elimination is mediated by P-gp (Hochman, Yamazaki, Ohe, & Lin, 2002). Therefore, from the drug discovery and development perspective, knowledge of the transport of drug candidates by P-gp is desirable at an early stage of the drug design process.

Data set

A database of 256 chemically diverse compounds with P-gp transport properties was assembled from the literature. The P-gp transport of each compound was assigned "yes" if transported by the protein and "no" if not transported. There is no single experimental method to conclusively identify a compound as a substrate for P-gp. Therefore, identification of the transport classification was based on at least two concurrent literature values from different assays, for example bi-directional Caco-2 measurements, ATPase activity or inhibition of transport of marker substrates.

Model output

The model is based on eleven 2D structural descriptors including molecular size, flexibility, planarity, aromaticity and polarity. Presence of hydrogen bond acceptor and donor groups is also an important feature of the model. The McGowan's volume, Vx, is the major discriminator of the model; the bigger the molecules, the more likely they are to be transported. Indeed, more than 75% of the compounds with large Vx values were found to be P-gp substrates and 72% of compounds of smaller size were not transported by P-gp. Figure 15.20 also illustrates important rules discriminating between substrates and non-substrates. For instance, large compounds with primary or secondary amines and few sp^2 carbons are identified as substrates. In addition, small but flexible compounds with no carboxylic groups and few hydrogen bond acceptors are defined as non-substrates.

The current model classifies molecules as likely to be substrates for P-gp (yes) or not likely (no) with an overall correct classification rate of 94 % on the training set. Optimum results were obtained with 95% of P-gp substrates correctly classified and 93% of the non-substrates correctly predicted. The

performance of this classifier was assessed on an independent test set of 51 compounds, of which 68% of the non-substrates and 86% of the substrates were correctly classified.

A confidence for each prediction is reported, according to the strength of association of the compound's descriptor values with the predicted classification. Furthermore, the distance of the predicted compound from the chemical space of the training set is calculated to gauge the confidence in the result. As there are insufficient data points outside the chemical space of the training set to assess the confidence in predictions, no estimate regarding the confidence for such compounds can be made. In these cases, the probability that the result is correct is reported as 0.5, indicating an even distribution between the two possible classes.

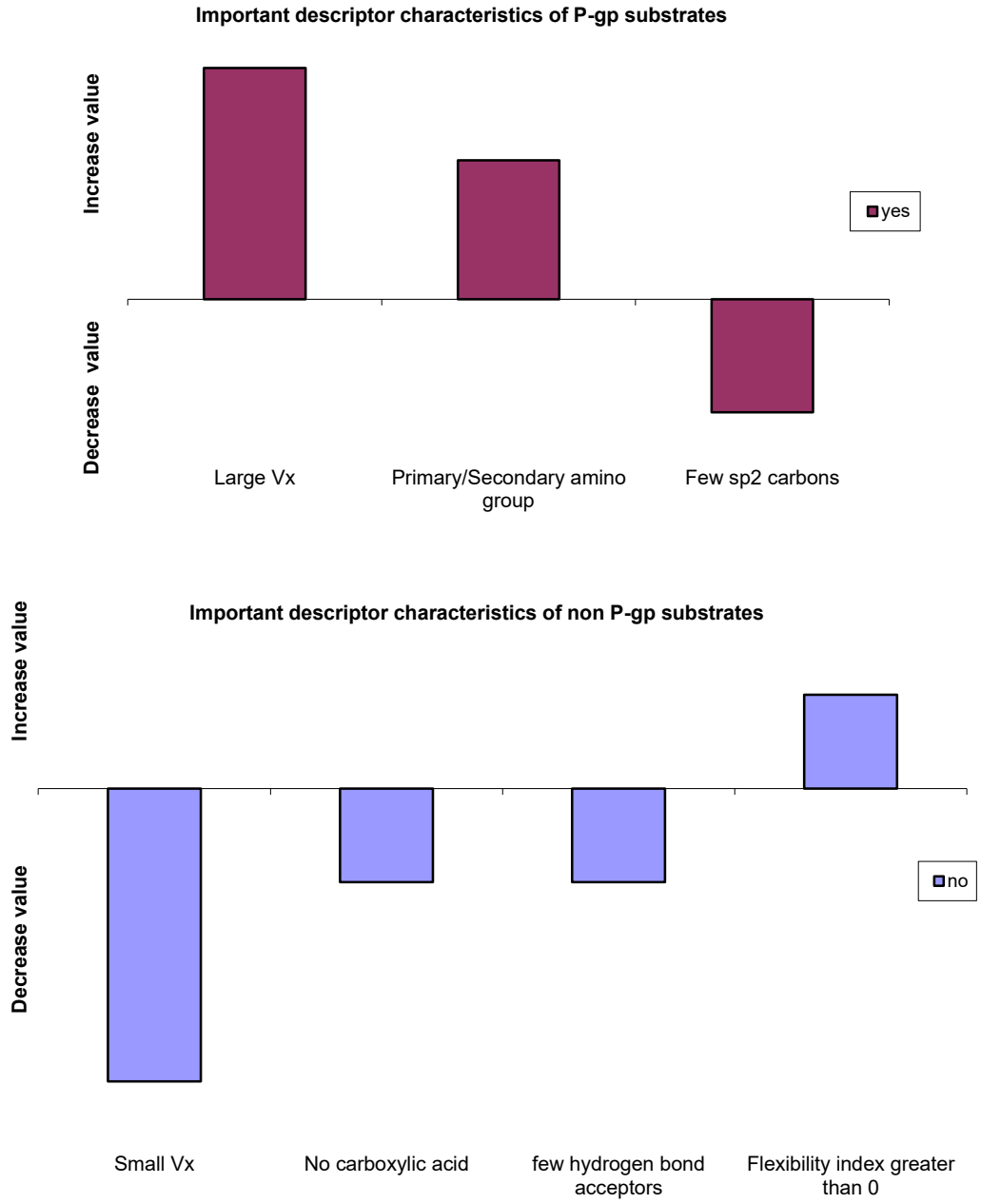


Figure 15.20 Histograms showing the influence of descriptors on the dominant rules of the P-gp classification model.

Comparison with other predictive techniques

The model statistics compare well to recent literature P-gp classification models where P-gp substrate, yes, prediction accuracy on independent test sets ranges from 53% to 72% and P-gp non-substrate prediction accuracies lie between 79% and 80% (Penzotti, Lamb, Evensen, & Grootenhuis, 2002) (Stouch, Gudmunson, & Ge, 2002) (Didziapetris, Japertas, & Petrauskas, 2004).

15.5.5 Plasma protein binding classification (80% threshold)

Data set

Commercial and proprietary databases, compendia and drug monographs were searched for % drug bound to plasma protein values and the data incorporated into a database of 888 structures. Rigorous quality control led to the elimination of 92 compounds for which the veracity of the data could not be confirmed. The resultant 796-compound dataset is highly biased toward high percentage values with 41% of the compounds reported as ≥ 90% bound. Values of % bound < 80% were classified as low,

values ≥ 80% were classified as high. This threshold corresponds approximately to a log K_a of 3.8 (binding affinity) and a K_D of 150 μM (dissociation constant). The structures were assigned randomly to training (n = 478), internal evaluation (n = 159) and independent test (n = 159) sets. The latter was excluded from the model development process.

Model output

The model is a decision tree that predicts the extent of test set compounds' plasma protein binding as either "high" or "low" in relation to the threshold described above. Calculated logP plus a further thirteen 2D descriptors, relating to the occurrence of certain functional groups and structural fragments and the protonation state of certain groups, are used in the model.

The model classifies molecules as having high or low affinity based on a classification boundary of 80%. Overall, training set classifications were 89% correct. For the internal evaluation set, used to monitor the performance of the model during training, predictions of high and low plasma protein binding were correct on 80% and 82% of occasions respectively. The corresponding figures for the independent test set were 77% and 78%.

A confidence for each prediction is reported, according to the strength of association of the compound's descriptor values with the predicted classification. Furthermore, the distance of the predicted compound from the chemical space of the training set is calculated to gauge the confidence in the result. As there are insufficient data points outside the chemical space of the training set to assess the confidence in predictions, no estimate regarding the confidence for such compounds can be made. In these cases, the probability that the result is correct is reported as 0.5, indicating an even distribution between the two possible classes.

Comparison with other predictive techniques

Comparison with recent literature models is difficult as they are based on binding to human serum albumin only and with data obtained via chromatographic, rather than older-established methods (Kratochwil, Huber, Muller, Kansy, & Gerber, 2002) (Colmenarejo, Alvarez-Pedraglio, & Lavandera, 2001).

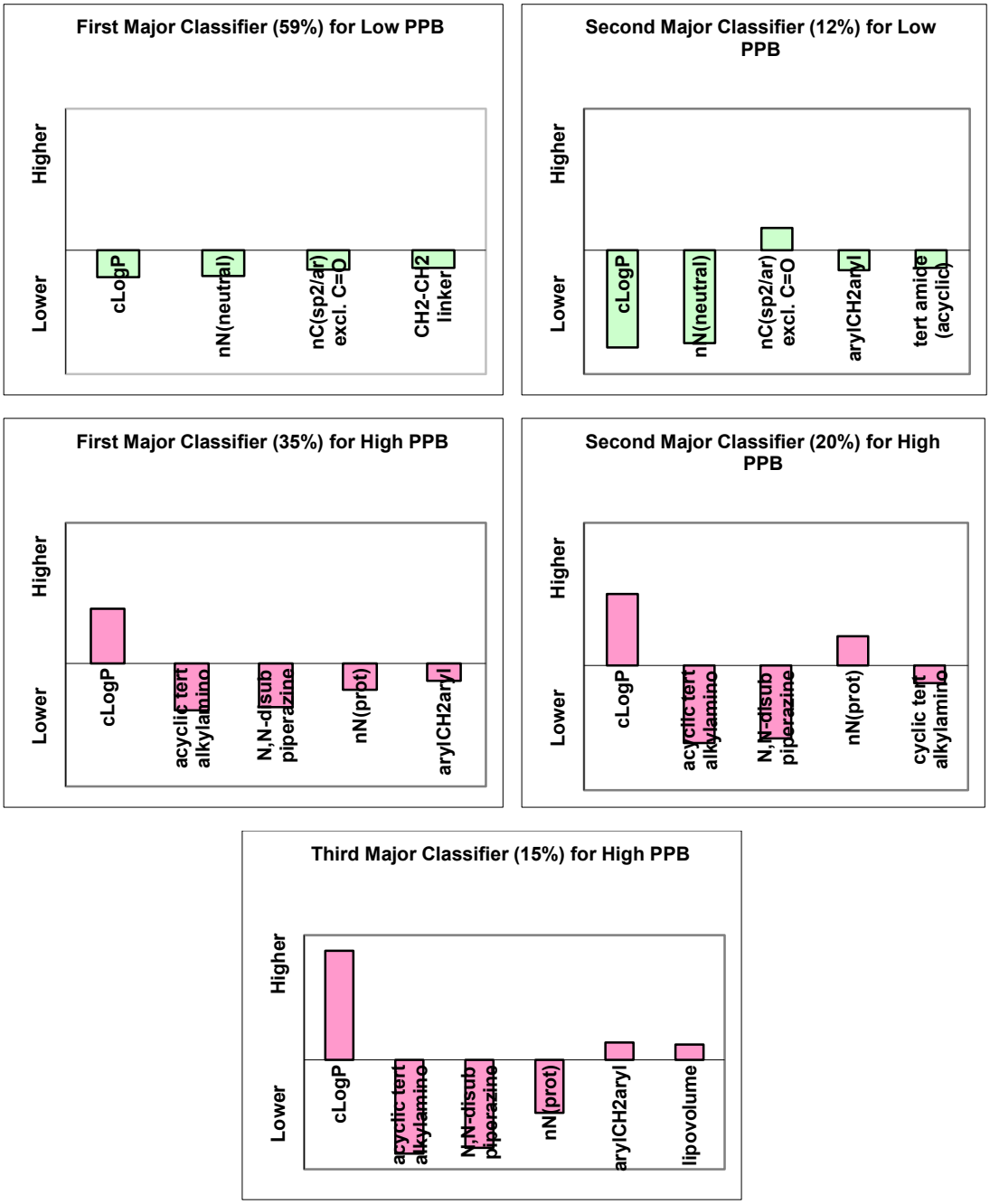


Figure 15.21 Histograms showing the influence of descriptors on the dominant rules of the PPB model. The directions of bars relative to the horizontal axis indicate whether the value of a descriptor must be higher or lower than a threshold. The length of a bar reflects the number of compounds retained if the condition is met. The vertical scale is uniform for all plots. Percentages in parentheses refer to the proportion of the class predictions made by the rule.

16 References

Abraham, M. H., & McGowan, J. C. (1987). The use of characteristic volumes to measure cavity terms in reversed-phase liquid-chromatography. *Chromatographia, 23*, 243-246.

Abraham, M. H., Hersey, A., Testa, B., & H., v. d. (2006). In B. Testa, & H. van de Waterbeemd, *ADMET/Property Based Approaches* (2nd Ed. ed., Vol. 5). Elsevier.

Abraham, M. H., Ibrahim, A., Zhao, Y., & Acree, W. E. (2006, 10). A data base for partition of volatile organic compounds and drugs from blood/plasma/serum to brain, and an LFER analysis of the data. *J. Pharm. Sci., 95*(10), 2091-2100.

Afzelius, L., Masimirembwa, C. M., Andersson, T., Karlen, A., & Zamora, I. (2003). An Almond model for identification and quantitative prediction of CYP2C9 inhibitors.

Afzelius, L., Masimirembwa, C. M., Karlen, A., Andersson, T. B., & Zamora, I. (2002, 07). Discriminant and quantitative PLS analysis of competitive CYP2C9 inhibitors versus non-inhibitors using alignment independent GRIND descriptors. *16*(7), pp. 443-458.

Afzelius, L., Zamora, I., Masimirembwa, C. M., Karlen, A., Andersson, T. B., Mecucci, S., . . . Cruciani, G. (2004, 02 12). Conformer- and alignment-independent model for predicting structurally diverse competitive CYP2C9 inhibitors. *J. Med. Chem., 47*(4), 907-914.

Agrafiotis, D. K. (2001). Multiobjective optimization of combinatorial libraries. *IBM J. Res. & Dev., 45*(3/4), 545-566.

Ajay, Bemis, G. W., & Murcko, M. A. (1999, 12 02). Designing libraries with CNS activity. *J. Med. Chem., 42*(24), 4942-4951.

Americ, S., Sullivan, J., Briggs, C., Donnelly-Roberts, D., Anderson, D., Raszkiewicz, J., . . . al., e. (1994). (S)-3-methyl-5-(1-methyl-2-pyrrolidinyl) isoxazole (ABT 418): a novel cholinergic ligand with cognition-enhancing and anxiolytic activities: I. In vitro characterization. *J. Pharmacol. Exp. Ther., 270*, 310-318.

Andrews, M., Brown, A., Chiva, J., Fradet, D., Gordon, D., Lansdell, M., & MacKenny, M. (2009). Design and optimisation of selective serotonin re-uptake inhibitors with high synthetic accessibility: part 2. *Bioorg. Med. Chem. Lett., 19*, 5893-5897.

Avdeef, A. (2003). *Absorption and Drug Development. Solubility, Permeability and Charge State.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Ayrton, a., & Morgan, P. (2001). The role of transport proteins in drug absorption, distribution and excretion. *Xenobiotica, 31*(8-9), 469-497.

Bathelt, C., Ridder, L., Mulholland, A., & Harvey, J. (2003). Aromatic hydroxylation by cytochrome P450: model calculations of mechanism and substituent effects. *J. Am. Chem. Soc., 125*(49), 15004-15005.

Bell, I. M., Gallicchio, S. N., Abrams, M., Beshore, D. C., Buser, C. A., Culberson, J. C., . . . Williams, T. M. (2001, 08 30). Design and biological activity of (S)-4-(5-([1-(3-chlorobenzyl)-2-oxopyrrolidin-3-ylamino]methyl)imidazol-1- ylmethyl)benzonitrile, a 3-aminopyrrolidinone farnesyltransferase inhibitor with excellent cell potency. *J. Med. Chem., 44*(18), 2933-2949.

Bickerton, G., Paolini, G., Besnard, J., Muresan, S., & Hopkins, A. (2012). Quantifying the chemical beauty of drugs. *Nature Chemistry, 4*(2), 90-98.

Binder, D., Hromatka, O., Geissler, F., Schmied, K., Noe, C., Burri, K., . . . Zeller, P. (1987). Analogues and derivatives of tenoxicam. 1. Synthesis and antiinflammatory activity of analogues with different residues on the ring nitrogen and the amide nitrogen. *J. Med. Chem., 30*, 678-682.

Black, J., Duncan, W., & Shanks, R. (1965). Comparison of some properties of pronethalol and propranolol. *Br. J. Pharmacol. Chemother., 25*, 577-591.

Bolton, E., Wang, Y., Thiessen, P., & Bryant, S. (2008). PubChem: Integrated Platform of Small Molecules and Biological Activities. In *Annual Reports in Computational Chemistry* (Vol. 4, pp. 217-241). Washington DC: American Chemical Society.

Bonizzoni, E., Milani, S., Ongini, E., Casati, C., & Monopoli, A. (1995). Modeling hemodynamic profiles by telemetry in the rat. A study with A1 and A2a adenosine agonists. *Hypertension, 25*(4 Pt 1), 564-569.

Bonnet, P., & Robins, R. (1993). Modulation of leukocyte genetic expression by novel purine nucleoside analogues. A new approach to antitumor and antiviral agents. *J. Med. Chem., 36*, 635-653.

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32.

Brown, N. (Ed.). (2012). *Bioisosteres in Medicinal Chemistry* (Vol. 54). Weinheim, Germany: Wiley-VCH.

Buhman, M. D. (2003). *Radial basis functions: theory and implementations.* Cambridge: Cambridge University Press.

Burden, F. R. (2001, 05). Quantitative structure-activity relationship studies using Gaussian processes. *41*(3), pp. 830-835.

Burger, A. (1970). *Medicinal Chemistry* (3rd ed.). San Francisco: John Wiley & Sons Inc.

Butina, D. (1999). Unsupervised Data Base Clustering Based on Daylight's fingerprint and Tanimoto Similarity: A fast and automated way to cluster small and large data set. *J. Chem. Inf. Comput. Sci, 39*, 747-750.

Butina, D., & Gola, J. M. (2003, 05). Modeling aqueous solubility. *J. Chem. Inf. Comput. Sci., 43*(3), 837-841.

Cadwell, G. W., Chen, P., He, J., Parmee, E. R., Leiting, B., Marsilio, F., . . . Weber, A. E. (2004). Fluoropyrrolidine amides as dipeptidyl peptidase IV inhibitors. *Bioorg. Med. Chem. Lett., 14*, 1265-1268.

Campagna-Slater, V., Pottel, J., Therrien, E., Cantin, L., & Moitessier, N. (2012). Development of a computational tool to rival experts in the prediction of sites of metabolism of xenobiotics by P450s. *J. Chem. Inf. Model., 52*(9), 2471-2483.

CEREP. (n.d.). *CEREP Bioprint*. Retrieved from see http://www.cerep.fr/Cerep/Users/pages/productsservices/bioprintservices.asp),

Cheeseright T, M. M. (2006). Molecular Field Extrema as Descriptors of Biological Activity: Definition and Validation. *J Chem Inf Model, 46*(2), 665-676.

ChemAxon. (n.d.). *JChem Base*. Retrieved from http://www.chemaxon.com/products/jchem-base

ChEMBL. (n.d.). Retrieved from https://www.ebi.ac.uk/chembldb/

Chico, L., Van Eldick, L., & Watterson, D. (2009). Targeting protein kinases in central nervous system disorders. *Nature Rev. Drug Discov., 8*(11), 892-909.

Chino, A., Masuda, N., Amano, Y., Honbou, K., Mihara, T., Yamazaki, M., & Tomishima, M. (2014). Novel benzimidazole derivatives as phosphodiesterase 10A (PDE10A) inhibitors with improved metabolic stability. *Bioorg. Med. Chem., 22*(13), 3515-3526.

Colmenarejo, G., Alvarez-Pedraglio, A., & Lavandera, J. L. (2001, 12 06). Cheminformatic models to predict binding affinities to human serum albumin. *J. Med. Chem., 44*(251), 4370-4378.

Cox, C., Breslin, M., Mariano, B., Coleman, P., Buser, C., Walsh, E., . . . Hartman, G. (2005). Kinesin spindle protein (KSP) inhibitors. Part 1: The discovery of 3,5-diaryl-4,5-dihydropyrazoles as potent and selective inhibitors of the mitotic kinesin KSP. *Bioorg. Med. Chem. Lett., 15*, 2041-2045.

Cresset. (n.d.). *FieldAlign*. Retrieved from http://www.cresset-group.com/

Crivori, P., Cruciani, G., Carrupt, P. A., & Testa, B. (2000, 06 01). Predicting blood-brain barrier permeation from three-dimensional molecular structure. *J. Med. Chem., 43*(11), 2204-2216.

Dabiré, H. (1991). Central 5-hydroxytryptamine (5-HT) receptors in blood pressure regulation. *Therapie, 46*(6), 421-429.

Danielson, P. B. (2002, 12). The cytochrome P450 superfamily: biochemistry, evolution and drug metabolism in humans. *Curr. Drug Metab., 3*(6), 561-597.

Daylight Chemical Information Systems Inc. (n.d.). *SMARTS Tutorial*. Retrieved 7 2, 2012, from http://www.daylight.com/dayhtml tutorials/languages/smarts/index.html

Daylight. (n.d.). *Daylight Chemical Information Systems Inc.* Retrieved from www.daylight.com

Daylight Software Release 4.4. (2002). *(1)*.

de Visser, S., & Shaik, S. (2003). {A proton-shuttle mechanism mediated by the porphyrin in benzene hydroxylation by cytochrome p450 enzymes. *J. Am. Chem. Soc., 125*(24), 7413-7424.

Dearden, J. C. (2006). In silico prediction of aqueous solubility. *Expert Opin. Drug Discov., 1*(1), 31-52.

Didziapetris, R., Japertas, P., & Petrauskas, A. (2004). *Critical Compilation of P-gp transport and inhibition data. Development of predictive algorithms.* Retrieved from http://www.ap-algorithms.com/presentation.htm

Digital Chemistry. (n.d.). *MOLSMART*. Retrieved from http://www.digitalchemistry.co.uk/prod_chemicalquery.html

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

Dossetter, A., Griffen, E., & Leach, A. (2013). Matched Molecular Pair Analysis in drug discovery. *Drug Discov. Today, 18*(15-16), 724-731.

Draper, N. R., & Smith, H. (1981). *Applied Regression Analysis* (Vol. Second). New York: Wiley.

Ekins, S., Boulanger, B., Swaan, P. W., & Hupcey, M. A. (2002). Towards a new age of virtual ADME/TOX and multidimensional drug discovery. *J. Comput. Aided Mol. Des., 16*, 381-401.

Ekins, S., Andreyev, S., Ryabov, A., Kirillov, E., Rakhmatulin, E., Bugrim, A., & Nikolskaya, T. (2005). Computational prediction of human drug metabolism. *Expert Opin. Drug Metab. Toxicol., 1*(2), 303-324.

Ekins, S., Berbaum, J., & Harrison, R. K. (2003, 09). Generation and validation of rapid computational filters for cyp2d6 and cyp3a4. *Drug Metab. Dispos., 31*(9), 1077-1080.

Ekins, S., de Groot, M. J., & Jones, J. P. (2001, 07). Pharmacophore and three-dimensional quantitative structure activity relationship methods for modeling cytochrome p450 active sites. *Drug Metab. Dispos., 29*(7), 936-944.

Ekins, S., Honeycutt, J., & Metz, J. (2010). Evolving molecules using multi-objective optimization: applying to ADME/Tox. *Drug Discov. Today, 15*, 451-60.

Engkvist, O., Wrede, P., & Rester, U. (2003, 01). Prediction of CNS activity of compound libraries using substructure analysis. *J. Chem. Inf. Comput. Sci., 43*(1), 155-160.

Ertl, P., Rhodes, B., & Selzer, P. (2000). Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its application to the Prediction of Drug Transport. *J. Med. Chem., 43*, 3714-3717.

Everitt, B. S., & Dunn, G. (2001). *Applied Multivariate Data Analysis* (Vol. 2nd). London: Arnold.

FDA. (n.d.). *FDA Adverse Event Reporting system*. Retrieved from http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm

Feng, J., Zhang, Z., Wallace, M., Stafford, J., Kaldor, S., Kassel, D. B., . . . Gwaltney II, S. (2007). Discovery of Alogliptin: A Potent, Selective, Bioavailable, and Efficacious Inhibitor of Dipeptidyl Peptidase IV. *J. Med. Chem., 50*(10), 2297-2300.

Fletcher, S. R., Burkamp, F., Blurton, P., Cheng, S. K., Clarkson, R., O'Connor, D., . . . MacLeod, A. M. (2002, 01 17). 4-(Phenylsulfonyl)piperidines: novel, selective, and bioavailable 5-HT(2A) receptor antagonists. *J. Med. Chem., 45*(2), 492-503.

Fournié-Zaluski, M., Coric, P., Turcaud, S., Rousselet, N., Gonzalez, W., Barbe, B., . . . Roques, B. (1994). New dual inhibitors of neutral endopeptidase and angiotensin-converting enzyme: rational design, bioavailability, and pharmacological responses in experimental hypertension. *J. Med. Chem., 37*, 1070-1083.

Friedman, J., & Fisher, N. (1999). Bump hunting in high-dimensional data. *Statistics and Computing, 9*(2), 123-143.

Gaussian Processes website. (2007).

Geladi, P. (1992). Wold, Herman, the father of PLS. *Chemometrics and Intelligent Laboratory Systems, 15*(1), pp. R7-R8.

Ghose, A. K., & Crippen, G. M. (1987, 02). Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. *J. Chem. Inf. Comput. Sci., 27*(1), 21-35.

Gillet, V. J., Khatib, W., Willett, P., Fleming, P. J., & Green, D. V. (2002, 03). Combinatorial library design using a multiobjective genetic algorithm. *J. Chem. Inf. Comput. Sci., 42*(2), 375-385.

Gola, J. M., Obrezanova, O., Champness, E., & Segall, M. D. (2006). ADMET property prediction: The state of the art and current challenges. *QSAR Comb. Sci., 25*(12), 1172-1180.

Goldberg, D. E. (1988). *Genetic Algorithms in Search, Optimisation and Machine Learning.* Reading, MA: Addison-Wesley.

Greene, N., Judson, P., Langowski, J., & Marchant, C. (1999). Knowledge-Based Expert Systems for Toxicity and Metabolism Prediction: DEREK, StAR and METEOR. *SAR and QSAR in Environmental Research, 10*(2-3), 299-314.

Guengerich, F. (2004, 05). Cytochrome P450: what have we learned and what are the future issues? *Drug. Metab. Rev., 36*(2), 159-197.

Guengerich, P. (2006). Cytochrome P450s and other enzymes in drug metabolism and toxicity. *The AAPS Journal, 8*(1), E101-11.

Guha, R., & Van Drie, J. (2008). Structure−Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model., 48*(3), 646-658.

Gupta-Ostermann, D., Wawer, M., Wassermann, A., & Bajorath, J. (2012). Graph Mining for SAR Transfer Series. *J. Chem. Inf. Model., 52*(4), 935-942.

Hajduk, P., & Sauer, D. (2006). Statistical Analysis of the Effects of Common Chemical Substituents on Ligand Potency. *J. Med. Chem., 51*(3), 553-564.

Hall, L. H., Kier, L. B., & Brown.B.B. (1995). Molecular Similarity Based on Novel Atom-Type Electrotopological State Indices. *J. Chem. Inf. Comput. Sci., 35*, 1074-1080.

Harris, D., & Loew, G. (1995, 03 15). Prediction of regiospecific hydroxylation of camphor analogs by cytochrome-P450(cam). *J. Am. Chem. Soc., 117*(10), 2738-2746.

Harris, D., & Loew, G. (1998). Theoretical investigation of the proton assisted pathway to formation of cytochrome P450 compound I. *J. Am. Chem. Soc., 120*(35), 8941-8948.

Hochman, J. H., Yamazaki, M., Ohe, T., & Lin, J. H. (2002, 06). Evaluation of drug interactions with P-glycoprotein in drug discovery: in vitro assessment of the potential for drug-drug interactions with P-glycoprotein. *Curr. Drug Metab., 3*(3), 257-273.

Hughes, J., Blagg, J., Price, D., Bailey, S., Decrescenzo, G., Devraj, R., . . . Zhang, Y. (2008). Physicochemical drug properties associated with in vivo toxicological outcomes. *Bioorg. Med. Chem. Lett., 18*(17), 4872-4875.

Hutzler, J. M., Walker, G. S., & Wienkers, L. C. (2003, 04). Inhibition of cytochrome P450 2D6: structure-activity studies using a series of quinidine and quinine analogues. *Chem. Res. Toxicol., 16*(4), 450-459.

Hynes Jr., J., AJ, D., Lin, S., Wrobleski, S., Wu, H., Gillooly, K., . . . al., e. (2008). Design, Synthesis, and Anti-inflammatory Properties of Orally Active 4-(Phenylamino)-pyrrolo[2,1-f][1,2,4]triazine p38α Mitogen-Activated Protein Kinase Inhibitors. *J. Med. Chem., 51*, 4-16.

Ihlenfeldt, W., Takahashi, Y., Abe, H., & Sasaki, S. (1994). Computation and Management of Chemical Properties in CACTVS: An extensible Networked Approach toward Modularity and Flexibility. *J. Chem. Inf. Comp. Sci., 34*, 109-116.

Irvine, J. D., Takahashi, L., Lockhart, K., Cheong, J., Tolan, J. W., Selick, H. E., & Grove, J. R. (1999, 01). MDCK (Madin-Darby canine kidney) cells: A tool for membrane permeability screening. *J. Pharm. Sci., 88*(1), 28-33.

Jansen-Olesen I, O. A., Strunk, S., Lassen, L., Olesen, J., Mortensen, A., Engel, U., & L., E. (1997). Role of endothelium and nitric oxide in histamine-induced responses in

human cranial arteries and detection of mRNA encoding H1- and H2-receptors by RT-PCR. *Br. J. Pharmacol., 121*(1), 41-48.

Janssens, F., Leenaerts, J., Diels, D., De Boeck, B., Megens, A., Langlois, X., . . . Borgers, M. (2005). Norpiperidine Imidazoazepines as a New Class of Potent, Selective, and Nonsedative H1 Antihistamines. *J. Med. Chem., 48*(6), 2154-2166.

Jones, J. P., Mysinger, M., & Korzekwa, K. R. (2002, 01). Computational models for cytochrome P450: a predictive electronic model for aromatic oxidation and hydrogen atom abstraction. *Drug Metab. Dispos., 30*(1), 7-12.

Jones, J. P., Rettie, A. E., & Trager, W. F. (1990, 04). Intrinsic isotope effects suggest that the reaction coordinate symmetry for the cytochrome P-450 catalyzed hydroxylation of octane is isozyme independent. *J. Med. Chem., 33*(4), 1242-1246.

Jones, J., Mysinger, M., & Korzekwa, K. (2002). Computational models for cytochrome P450: a predictive electronic model for aromatic oxidation and hydrogen atom abstraction. *Drug Metab. Dispos., 30*(1), 7-12.

Judson, P., Stalford, S., & Vessey, J. (2013). Assessing confidence in predictions made by knowledge-based systems. *Toxicol. Res., 2*, 70-79.

Keseru, G. M., Molnar, L., & Greiner, I. (2000, 12). A neural network based virtual high throughput screening test for the prediction of CNS activity. *Comb. Chem. High Throughput Screen., 3*(6), 535-540.

Kirchmair, J., Williamson, M., Tyzack, J., Tan, L., Bond, P., Bender, A., & Glen, R. (2012). Computational prediction of metabolism: sites, products, SAR, P450 enzyme dynamics, and mechanisms. *J. Chem. Inf. Model., 52*(3), 617-648.

Korzekwa, K. R., Swinney, D. C., & Trager, W. F. (1989). Isotopically labeled chlorobenzenes as probes for the mechanism of cytochrome P-450 catalysed aromatic hydroxylation. *Biochemistry, 28*, 9019-9027.

Korzekwa, K. R., Trager, W. F., Gouterman, M., Spangler, D., & Loew, G. (1985). Cytochrome P450 mediated aromatic oxidation: A theoritical study. *J. Am. Chem. Soc., 107*, 4273-4279.

Korzekwa, K., Jones, J., & Gillette, J. (1990). Theoretical studies on cytochrome P-450 mediated hydroxylation: a predictive model for hydrogen atom abstractions. *J. Am. Chem. Soc., 112*(19), 7042-7046.

Korzekwa, K., Jones, J., & Gillette, J. (1990). Theoretical Studies on Cytochrome P-450 Mediated Hydroxylation: A predictive model for Hydrogen Atom Abstractions. *J. Am. Chem. Soc., 112*, 7042-7046.

Korzekwa, K., Trager, W., & Gillette, J. (1989). Theory for the observed isotope effects from enzymatic systems that form multiple products via branched reaction pathways: cytochrome P-450. *Biochemistry, 28*(23), 9012-9018.

Koymans, L., Vermeulen, N. P., van Acker, S. A., te Koppele, J. M., Heykants, J. J., Lavrijsen, K., . . . GM, D.-O. d. (1992, 03). A predictive model for substrates of cytochrome P450-debrisoquine (2D6). *Chem. Res. Toxicol., 5*(2), 211-219.

Kratochwil, N. A., Huber, W., Muller, F., Kansy, M., & Gerber, P. R. (2002, 11 01). Predicting plasma protein binding of drugs: a new approach. *Biochem. Pharmacol., 64*(9), 1355-1374.

Kulkarni, S., Zhu, J., & Blechinger, S. (2005). In silico techniques for the study and prediction of xenobiotic metabolism: a review. *Xenobiotica, 35*(10-11), 955-973.

Kumar, D., Karamzadeh, B., Sastry, G., & de Visser, S. (2010). What Factors Influence the Rate Constant of Substrate Epoxidation by Compound I of Cytochrome P450 and Analogous Iron(IV)-Oxo Oxidants? *J. Am. Chem. Soc., 132*(22), 7656-7667.

Langdon, W., Barret, S., & Buxton, B. (2003, 2003/12/08/). Predicting Biochemical Interactions - Human P450 2D6 Enzyme Inhibition. *COngress on Evolutionary Computation. Congress on Evolutionar Computation 2003, Canberra, 8-12 Dec 2003*, pp. 8-12. Canberra: IEEE Press.

Larsen, A., & Lish, P. (1964). A NEW BIO-ISOSTERE: ALKYLSULPHONAMIDOPHENETHANOLAMINES. *Nature, 203*, 1283-1284.

Leach, A., Jones, H., Cosgrove, D., Kenny, P., Ruston, L., MacFaul, P., . . . Law, B. (2006). Matched molecular pairs as a guide in the optimization of pharmaceutical properties; a study of aqueous solubility, plasma protein binding and oral exposure. *J. Med. Chem., 49*(23), 6672-6682.

Leo, A. (1993). Calculating log Poct from Structures. *Chem. Reviews, 93*(4), 1281-1306.

Lewis, D. (2004). 57 varieties: the human cytochromes P450. *Pharmacogenomics, 5*(3), 305-318.

Lewis, D. F. (2000, 08 01). On the recognition of mammalian microsomal cytochrome P450 substrates and their characteristics: towards the prediction of human p450 substrate specificity and metabolism. *Biochem. Pharmacol., 60*(3), 293-306.

Lewis, D. F. (2003, 01 01). Essential requirements for substrate binding affinity and selectivity toward human CYP2 family enzymes. *Arch. Biochem. Biophys., 409*(1), 32-44.

Lewis, D. F., Eddershaw, P. J., Goldfarb, P. S., & Tarbit, M. H. (1996, 10). Molecular modelling of CYP3A4 from an alignment with CYP102: identification of key interactions between putative active site residues and CYP3A-specific chemicals. *Xenobiotica, 26*(10), 1067-1086.

Lewis, D. F., Modi, S., & Dickins, M. (2002, 02). Structure-activity relationship for human cytochrome P450 substrates and inhibitors. *Drug Metab. Rev., 34*(1-2), 69-82.

Lhasa. (n.d.). *Lhasa Limited*. Retrieved from http://www.lhasalimited.org/

Lightfoot, T., Ellis, S. W., Mahling, J., Ackland, M. J., Blaney, F. E., Bijloo, G. J., . . . Tucker, G. T. (2000, 03). Regioselective hydroxylation of debrisoquine by cytochrome P4502D6: implications for active site modelling. *Xenobiotica, 30*(3), 219-233.

Lipinski, C., Lombardo, F., Dominy, B., & Feeney, P. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev., 23*, 3-25.

MacKay, D. J. (2003). *Information Theory, Inference, and Learning Algorithms.* Cambridge: Cambridge University Press.

Marti-Renom, M., Stuart, A., Fiser, A., Sanchez, R., Melo, F., & Sali, A. (1985). Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc., 107*(13), 3902-3909.

McGregor, M., & Pallai, P. V. (1997). Clustering of Large Databases of Compounds: Using the MDL "Keys" as structural Descriptors. *J. Chem. Inf. Comput. Sci., 37*, 443-448.

Mitchell, M., Son, J., Lee, I., Ching, K., Kim, H., Guo, H., . . . Kim, C. (2010). N1-Heterocyclic pyrimidinediones as non-nucleoside inhibitors of HIV-1 reverse transcriptase. *Bioorg. Med. Chem. Lett., 20*(5), 1585-1588.

Modi, S., Paine, M. J., Sutcliffe, M. J., Lian, L. Y., Primrose, W. U., Wolf, C. R., & Roberts, G. C. (1996, 04 09). A model for human cytochrome P450 2D6 based on homology modeling and NMR studies of substrate binding. *Biochemistry, 35*(14), 4540-4550.

Moore, R., Derry, S., Makinson, G., & McQuay, H. (2005). Tolerability and adverse events in clinical trials of celecoxib in osteoarthritis and rheumatoid arthritis: systematic review and meta-analysis of information from company clinical trial reports. *Arthritis Res. Ther., 7*(3), R644-R665.

Nadanaciva, S., Aleo, M., Strock, C., Stedman, D., Wang, H., & Will, Y. (2013). Toxicity assessments of nonsteroidal anti-inflammatory drugs in isolated mitochondria, rat hepatocytes, and zebrafish show good concordance across chemical classes. *Toxicol. Appl. Pharmacol., 272*(2), 272-280.

O'Boyle, N., Bostrom, J., Sayle, R., & Gill, A. (2014). Using Matched Molecular Series as a Predictive Tool To Optimize Biological Activity. *J. Med. Chem., 57*(6), 2704-2713.

Obrezanova, O., Csanyi, G., Gola, J., & Segall, M. (2007). Gaussian processes: a method for automatic QSAR modeling of ADME properties. *J. Chem. Inf. Model., 47*(5), 1847-1857.

Obrezanova, O., Gola, J., Champness, E., & Segall, M. (2009). Automatic QSAR modeling of ADME properties: blood-brain barrier penetration and aqueous solubility. *J. Comput. Aided Mol. Des., 22*, 431-440.

Ogliaro, F., Harris, N., Cohen, S., Filatov, M., de Visser, S., & Shaik, S. (2000). A Model Rebound Mechanism of Hydroxylation by Cytochrome P450: Stepwise and

Effectively Concerted Pathways and their Reactivity Patterns. *J. Am. Chem. Soc., 122*(37), 8977-8989.

Olah, M., Bologa, C., & Oprea, T. I. (2004, 07). An automated PLS search for biologically relevant QSAR descriptors. *J. Comput. Aided Mol. Des., 18*(7-9), 437-449.

Optibrium. (n.d.). Retrieved January 8, 2015, from http://www.optibrium.com/stardrop

Optibrium. (n.d.). *StarDrop*. Retrieved March 3, 2011, from http://www.optibrium.com/stardrop

Overington, J. (2009). ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). Interview by Wendy A. Warr. *J. Comput. Aided Mol. Des., 23*, 195-198.

Panasyuk, G., Espeillac, C., Chauvin, C., Pradelli, L., Horie, Y., Suzuki, A., . . . Pende, M. (2012). PPARγ contributes to PKM2 and HK2 expression in fatty liver. *Nat. Commun., 3*(672), 14.

Parks, D., Lafrance, L., Calvo, R., Milkiewicz, K., Gupta, V., Lattanze, J., . . . Lu, T. (2005). 1,4-Benzodiazepine-2,5-diones as small molecule antagonists of the HDM2-p53 interaction: discovery and SAR. *Bioorg. Med. Chem. Lett., 15*, 765-70.

Patani, G., & LaVoie, E. (1996). Bioisosterism: A Rational Approach in Drug Design. *Chem. Rev., 96*, 3147-3176.

Pearlman, R. S. (1998). Novel software tools for chemical diversity. *Perspect. Drug Discov. Des., 9*, 339-353.

Penzotti, J. E., Lamb, M. L., Evensen, E., & Grootenhuis, P. D. (2002, 04 25). A computational ensemble pharmacophore model for identifying substrates of p-glycoprotein. *J. Med. Chem., 45*(9), 1737-1740.

Perola, E. (2010). An analysis of the binding efficiencies of drugs and their leads in successful drug discovery programs. *J. Med. Chem., 53*, 2986-2997.

Press, W. H. (1988). *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge: Cambridge University Press.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning.* San Mateo: Morgan Kaufmann Publishers, Inc.

Quinlan, R. (1986). Induction of Decision Trees. *Machine Learning, 1*, 81-106.

Quinlan, R. (1996). Bagging, Boosting and C4.5. *Proceedings of the Thirteenth National Conference on Artificial Intelligence* (pp. 725-730). AAAI Press.

Ramage, A. (1990). Influence of 5-HT1A receptor agonists on sympathetic and parasympathetic nerve activity. *J. Cardiovasc. Pharmacol., 15*(Suppl 7), S75-85.

Rasmussen, C. E., & Williams, C. K. (2006). *Gaussian Processes for Machine Learning.* Cambridge, Massachusetts: The MIT Press.

Raymond, J., Watson, I., & Mahoui, A. (2009). Rationalizing lead optimization by associating quantitative relevance with molecular structure modification. *J. Chem. Inf. Model., 49*, 1952-62.

RDKit: Cheminformatics and Machine Learning Software. (n.d.). Retrieved 2011 йил 2-March from http://www.rdkit.org/

Ridings, J., Barratt, M., Cary, R., Earnshaw, C., Eggington, C., Ellis, M., . . . al., e. (1996, January 8). Computer prediction of possible toxic action from chemical structure: an update on the DEREK system. *Toxicology, 106*(1-3), 267-279.

Ritchie, T., & Macdonald, S. (2009). The impact of aromatic ring count on compound developability – are too many aromatic rings a liability in drug design? *Drug Discov. Today, 14*(21/22), 1011-1020.

Rocheblave, L., Bihel, F., De Michelis, C., Priem, G., Courcambeck, J., Bonnet, B., . . . Kraus, J. (2002). Synthesis and antiviral activity of new anti-HIV amprenavir bioisosteres. *J. Med. Chem., 45*, 3321-3324.

Roehrig, S., Straub, A., Pohlmann, J., Lampe, T., Pernerstorfer, J., Schlemmer, K., . . . Perzborn, E. (2005). Discovery of the novel antithrombotic agent 5-chloro-N-({(5S)-2-oxo-3- [4-(3-oxomorpholin-4-yl)phenyl]-1,3-oxazolidin-5-yl}methyl)thiophene- 2-carboxamide (BAY 59-7939): an oral, direct factor Xa inhibitor. *J. Med. Chem., 48*, 5900-5908.

Rogers, D., & Tanimoto, T. (1960). A computer program for classifying plants. *Science, 132*(10), 1115-1118.

Rogue, A., Lambert, C., Jossé, R., Antherieu, S., Spire, C., Claude, N., & Guillouzo, A. (2011). Comparative gene expression profiles induced by PPARγ and PPARα/γ agonists in human hepatocytes. *PLoS One, 6*(4), e18816.

Roweis, S. (1997). *Neural Information Processing Systems 10.* Denver, CO, USA: NIPS.

Rydberg, P., & Olsen, L. (2010). SMARTCyp: A 2D Method for Prediction of Cytochrome P450-Mediated Drug Metabolism. *ACS Medicinal Chemistry Letters, 1*(3), 96-100.

Rydberg, P., & Olsen, L. (2012). Ligand-Based Site of Metabolism Prediction for Cytochrome P450 2D6. *ACS Med. Chem. Lett., 3*(1), 69-73.

Rydberg, P., Gloriam, D., Zaretzki, J., Breneman, C., & Olsen, L. (2010). SMARTCyp: A 2D method for prediction of cytochrome P450-mediated drug metabolism. *ACS Med. Chem. Lett., 1*(3), 96-100.

Rydberg, P., Ryde, U., & Olsen, L. (2008). Sulfoxide, Sulfur, and Nitrogen Oxidation and Dealkylation by Cytochrome P450. *J. Chem. Theo. Comp., 4*(8), 1369-1377.

Sanderson, D., & Earnshaw, C. (1991, July). Computer prediction of possible toxic action from chemical structure; the DEREK system. *Hum. Exp. Toxicol., 10*(4), 261-273.

Schwab, D., Fischer, H., Tabatabaei, A., Poli, S., & Huwyler, J. (2003, 04 24). Comparison of in vitro P-glycoprotein screening assays: recommendations for their use in drug discovery. *J. Med. Chem., 46*(9), 1716-1725.

Schwaighofer, A., Schroeter, T., Mika, S., Laub, J., Laak, A. T., Sulzle, D., . . . Muller, K. R. (2007). Accurate solubility prediction with error bars for electrolytes: A machine learning approach. *J. Chem. Inf. Model., 47*, 407-424.

Segall, M., & Barber, C. (2014). Addressing toxicity risk when designing and selecting compounds in early drug discovery. *Drug Discov. Tod., 19*(5), 688-693.

Segall, M., Beresford, A., Gola, J., Hawksley, D., & Tarbit, M. (2006). Focus on Success: Using in silico optimisation to achieve an optimal balance of properties. *Expert Opin. Drug Metab. Toxicol., 2*.

Segall, M., Champness, E., Obrezanova, O., & C, L. (2009). Beyond Profiling: Using ADMET models to guide decisions. *Chemistry & Biodiversity, 6*, 2144 - 2151.

Seierstad, M., & Agrafiotis, D. K. (2006, 04). A QSAR model of HERG binding using a large, diverse, and internally consistent training set. *Chem. Bio. Drug Des., 67*(4), 284-296.

Selwood, D. L., Livingstone, D. J., Comley, J. C., O'Dowd, A. B., Hudson, A. T., Jackson, P., . . . Stables, J. N. (1990, 01). Structure-activity relationships of antifilarial antimycin analogues: a multivariate pattern recognition study. *J. Med. Chem., 33*(1), 136-142.

Shaik, S., Cohen, S., Wang, Y., Chen, H., Kumar, D., & Thiel, W. (2010). P450 enzymes: their structure, reactivity, and selectivity-modeled by QM/MM calculations. *Chem. Rev., 110*(2), 949-1017.

Shaik, S., de Visser, S., Ogliaro, F., Schwarz, H., & Schroder, D. (2002). Two-state reactivity mechanisms of hydroxylation and epoxidation by cytochrome P-450 revealed by theory. *Curr. Opin. Chem. Biol., 6*(5), 556-567.

Sharma, P., de Visser, S., & Shail, S. (2003). Can a single oxidant with two spin states masquerade as two different oxidants? A study of the sulfoxidation mechanism by cytochrome p450. *J. Am. Chem. Soc., 125*(29), 8698-8699.

Song, M., & Clark, M. (2006, 01). Development and evaluation of an in silico model for hERG binding. *J. Chem. Inf. Comput. Sci., 46*(1), 392-400.

Stewart, K., Shiroda, M., & James, C. (2006). Drug Guru: a computer software program for drug design using medicinal chemistry rules. *Bioorg. Med. Chem., 14*, 7011-22.

Stouch, T. R., Gudmunson, O., & Ge, S. E. (2002, 04 01). Prediction of PGP transporter activity using calculated molecular properties. *BTEC/CINF Abstracts.*

Strobl, G. R., von Kruedener, S., Stockigt, J., Guengerich, F. P., & Wolff, T. (1993, 04 30). Development of a pharmacophore for inhibition of human liver cytochrome P-450 2D6: molecular modeling and inhibition studies. *J. Med. Chem., 36*(9), 1136-1145.

Stumpfe, D., & Bajorath, J. (2012). Exploring Activity Cliffs in Medicinal Chemistry. *J. Med. Chem., 55*(7), 2932-42.

Sun, Q., Gatto, B., Yu, C., Liu, A., Liu, L., & LaVoie, E. (1995). Synthesis and evaluation of terbenzimidazoles as topoisomerase I inhibitors. *J. Med. Chem., 38*, 3638-44.

Susnow, R. G., & Dixon, S. L. (2003, 07). Use of robust classification techniques for the prediction of human cytochrome P450 2D6 inhibition. *J. Chem. Inf. Comput. Sci., 43*(4), 1308-1315.

Tandon, M., O'Donnel, M., Porte, A., Vensel, D., Yang, D., Palma, R., . . . Ashwell, M. (2004). The design and preparation of metabolically protected new arylpiperazine 5-HT1A ligands. *Bioorg. Med. Chem. Lett., 14*(7), 1709-1712.

Tarcsay, A., & Keseru, G. (2011). In silico site of metabolism prediction of cytochrome P450-mediated biotransformations. *Expert Opin. Drug Metab. Toxicol., 7*(3), 299-312.

Ujváry, I., & Hayward, J. (2012). BIOSTER: A Database of Bioisosteres and Bioanalogues. In N. Brown, *Bioisosteres in Medicinal Chemistry.*

Uno, T., Kondo, H., Inoue, Y., Kawahata, Y., Sotomura, M., Iuchi, K., & Tsukamoto, G. (1990). Synthesis of antimicrobial agents. 3. Syntheses and antibacterial activities of 7-(4-hydroxypiperazin-1-yl)quinolones. *J. Med. Chem., 33*, 2929-2932.

Van de Waterbeemd, H., & Gifford, E. (2003). ADMET in silico modelling: towards prediction paradise? *Nat. Rev. Drug Discovery, 2*, 192-204.

van der Maaten, L., & Hinton, G. (2008). Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research, 9*, 2579-2605.

Vaz, R., Zamora, I., Li, Y., Reiling, S., Shen, J., & Cruciani, G. (2010). The challenges of in silico contributions to drug metabolism in lead optimization. *Expert Opin. Drug Metab. Toxicol., 6*(7), 851-861.

Veber, D., Johnson, S., Cheng, H., Smith, B., Ward, K., & Kopple, K. (2002). Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem, 45*(12), 2615-2623.

Vilar, S., Chakrabarti, M., & Costanzi, S. (2010). Prediction of passive blood-brain partitioning: straightforward and effective classification models based on in silico derived physicochemical descriptors. *J. Mol. Graph., 28*(8), 899-903.

Wagener, W., & Lommerse, J. (2006). The Quest for Bioisosteric Replacements. *J. Chem. Inf. Model., 46*(2), 677-685.

Walsh, D., Franzyshen, S., & Yanni, J. (1989). Synthesis and antiallergy activity of 4-(diarylhydroxymethyl)-1-[3-(aryloxy)propyl]piperidines and structurally related compounds. *J. Med. Chem., 32*, 105-118.

Warner, D., Griffen, E., & St-Gallay, S. (2010). WizePairZ: A Novel Algorithm to Identify, Encode, and Exploit Matched Molecular Pairs with Unspecified Cores in Medicinal Chemistry. *J. Chem. Inf. Model., 50*(8), 1350-1357.

Wasserman, A., & Bajorath, J. (2011). A Data Mining Method to Facilitate SAR Transfer. *J. Chem. Inf. Model., 51*(8), 1857-1866.

Wawer, M., & Bajorath, J. (2011). Local Structural Changes, Global Data Views: Graphical Substructure−Activity Relationship Trailing. *J. Med. Chem., 54*(8), 4944-2951.

Wehrens, R., & Mevik, B. (2007). The pls Package: Principal Component and Partial Least Squares Regression in R. *J. Stat. Soft., 18*(2), 1-24.

Weininger, D. (1998). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci., 28*, 31–36.

Wishart, D., Knox, C., Guo, A., Cheng, D., Shrivastava, S., Tzur, D., . . . Hassanali, M. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res., 36(Database issue)*, D901-6.

Wold, S., Sjostrom, M., & Eriksson, L. (1999). Partial Least Squares Projectoins to Latent Structures (PLS) in Chemistry. In P. Schleyer, N. Allinger, T. Clark, J. Gasteiger, P. Kollman, H. Schaefer III, & P. Schreiner, *The Encyclopedia od Computational Chemistry* (pp. 1-16). Chichester, UK: John Wiley and Sons.

Yoshida, F., & Topliss, J. G. (2000, 06 29). QSAR model for drug human oral bioavailability. *J. Med. Chem., 43*(13), 2575-2585.

Yoshino, K., Kohno, T., Morita, T., & Tsukamoto, G. (1989). Organic phosphorus compounds. 2. Synthesis and coronary vasodilator activity of (benzothiazolylbenzyl) phosphonate derivatives. *J. Med. Chem., 32*, 1528-1532.

Yusof, I., & Segall, M. (2013). Considering the impact drug-like properties have on the chance of success. *Drug Discovery Today, 18*(13-14), 659-666.

Zaretzki, J., Bergeron, C., Rydberg, P., Huang, T.-W., Bennet, K., & Breneman, C. (2011). RS-Predictor: A New Tool for Predicting Sites of Cytochrome P450-Mediated Metabolism Applied to CYP 3A4. *J. Chem. Inf. Model., 51*(7), 1667-1689.

Zaretzki, J., Rydberg, P., Bergeron, C., Bennett, K., Olsen, L., & Breneman, C. (2012). RS-Predictor models augmented with SMARTCyp reactivities: robust metabolic regioselectivity predictions for nine CYP isozymes. *J. Chem. Inf. Model., 52*(6), 1637-1659.

Zhang, B., Wasserman, A. M., Vogt, M., & Bajorath, J. (2012). Systematic Assessment of Compound Series with SAR Transfer Potential. *J. Chem. Inf. Model., 52*(12), 3138-3143.

Zhao, Y. H., Abraham, M. H., Ibrahim, A., Fish, P. V., Cole, S., Lewis, M. L., . . . Reynolds, D. P. (2007, 01). Predicting penetration across the blood-brain barrier from simple descriptors and fragmentation schemes. *J. Chem. Inf. Comput. Sci., 47*(1), 170-175.

Zhao, Y. H., Le, J., Abraham, M. H., Hersey, A., Eddershaw, P. J., Luscombe, C. N., . . . Platts, J. A. (2001, 06). Evaluation of human intestinal absorption data and subsequent derivation of a quantitative structure-activity relationship (QSAR) with the Abraham descriptors. *J. Pharm. Sci., 90*(6), 749-784.