Structure generation and *de novo* design using reaction networks

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

at

The University Of Sheffield.

by

James Wallace

This work was sponsored by

Lilly

and

EPSRC

The University of Sheffield

Information School and Department of Chemistry

Faculty of Science

September 2016

Acknowledgements

First and foremost I would like to thank my supervisors Val Gillet and Beining Chen of The University of Sheffield, for their support and guidance throughout this work. From the industrial side of the project, thanks must go to Mike Bodkin, the original supervisor of this project when he was at Lilly UK, and the rest of the C3 team, in particular the two Davids (Thorner and Evans), who provided support and ultimate industrial supervision for the latter stages of this project.

Next, I would like to thank members of the CISRG, both past and present, for making my time in the group so enjoyable; so my thanks go to Edmund Duesbury, Simon Hand, John Holliday, Christos Kannas, Lucyantie Mazalan, Mira Omar, Nor Sani and Matthew Seddon. Special thanks must also go to Ben Allen and Richard Sherhod for helping me to get settled within the group and many interesting discussions at conferences, and Sonny Gan for many useful insights into the fundamentals of chemoinformatics. I would also like to thank the many members of the Beining Chen group past and present for their practical insight, in particular Harith, Calvin and Matthew who joined at the same time as I did and whose projects I have seen flourish, and Jenny Louth for some insightful discussions about chemistry as a whole. Finally, I would like to thank my parents for their tireless support and encouragement over the years, without which this may not have happened.

This work was made possible by funding from the EPSRC and Eli Lilly UK.

Disclaimer

All structures are either trivial small proof of concept examples, taken directly from the literature as cited, or *de novo* generated by the procedures outlined from literature SAR sets without any input of Lilly intellectual property.

Abstract

This project is concerned with *de novo* molecular design whereby novel molecules are built *in silico* and evaluated against properties relevant to biological activity, such as physicochemical properties and structural similarity to active compounds. The aim is to encourage cost-effective compound design by reducing the number of molecules requiring synthesis and analysis.

One of the main issues in de novo design is ensuring that the molecules generated are synthesisable. In this project, a method is developed that enables virtual synthesis using rules derived from reaction sequences. Individual reactions taken from reaction databases were connected to form reaction networks. Reaction sequences were then extracted by tracing paths through the network and used to create 'reaction sequence vectors' (RSVs) which encode the differences between the start and end points of the sequences. RSVs can be applied to molecules to generate virtual products which are based on literature precedents.

The RSVs were applied to structure-activity relationship (SAR) exploration using examples taken from the literature. They were shown to be effective in expanding the chemical space that is accessible from the given starting materials. Furthermore, each virtual product is associated with a potential synthetic route. They were then applied in *de novo* design scenarios with the aim of generating molecules that are predicted to be active using SAR models. Using a collection of RSVs with a set of small molecules as starting materials for *de novo* design proved that the method was capable of producing many useful, synthesisable compounds worthy of future study.

The RSV method was then compared with a previously published method that is based on individual reactions (reaction vectors or RVs). The RSV approach was shown to be considerably faster than de novo design using RVs, however, the diversity of products was more limited.

Table of Contents

List of Figures

List of Tables

List of Equations

Chapter 1:
Introduction

The field of drug discovery has, to some extent, used computational methods in a supporting role since the late 1950s (Willett, 2011, Leach and Gillet, 2003). Initially, this was in the form of simple substructure analyses of structure collections in databases, as in the work by Ray and Kirsch (1957). The benefits of this searching method over the previous manual efforts became apparent very quickly, leading to significant interest in the research field. This resulted in the formation in 1961 of the first journal for the field, the Journal of Chemical Documentation (this still exists as the Journal of Chemical Information and Modeling, having adopted this name in 2005). Much of the initial published work in this decade came from work from the Chemical Abstracts Service (CAS), as part of their efforts to computerise their existing collections (Weisgerber, 1997). This research produced a number of analysis and processing methods still used in some form today, such as the Morgan algorithm for producing canonical molecular representations (Morgan, 1965).

The success of the categorisation and analysis work soon led to an extension to studying the effect of structural features on activity, as in the work by Hansch and Fujita (1964). They introduced the concept of Quantitative Structure Activity Relationships (QSAR), whereby biological properties are related to structural parameters. During the 1970s and 1980s work on structure analysis methods continued as processing power improved; for example, the QSAR approach was enhanced and expanded and existing representations methods were extended to consider generic structures that represent multiple molecules in a single representation. Of particular interest during this period was the research work by Corey and Wipke (Corey et al., 1972), such as the OCCS (Organic Chemical Simulation of Synthesis) and LHASA (Logic and Heuristics Applied to Synthetic Analysis) programs. These tools were among the first to use computer graphics hardware to facilitate the

input and output of chemical structures, enabling drug design purely *in silico.* However, the calculations required for constructing and handling molecules were so computationally expensive that they severely limited the effectiveness of the modelling process.

The improvements in computational technology towards the end of the 1980s permitted the system limitations that had frustrated growth in this field to be overcome. This led to many new tools being developed to permit the design of molecular structures in this manner, including the first *de novo* (derived from the Latin for 'from new') design programs (Danziger and Dean, 1989). *De novo* tools are able to build novel molecules in virtual space with some element of evaluation of the generated results for suitability according to prescribed design constraints. These include shape complementarity with a target site, similarity to other known examples and other constraints based on molecular property calculations. With the twin influences of reducing costs of high performance computing, along with the steady increase in the costs of bringing a drug to market (over $US 1.8 billion per successful compound (Paul et al., 2010)), there has been significant effort in the development of these tools as a cost effective component of drug design. In particular, the use of these kinds of *in silico* methods has become ever more important as a way of suggesting new compounds that fit a particular model.

One common complaint with the first generation of structure design tools was that many did not take account of synthetic feasibility, i.e. the products generated by the tools were not necessarily capable of being created in the real world. More recent *de novo* methods are based around the concept of connecting together molecule fragments according to an established ruleset. These are often derived directly from examples in the published synthetic literature, restricting the transformation steps which can be applied to the input molecules to those that have a realistic synthesis route available. These tools have their own limitations in that many of the used rule sets are overly restrictive, with only limited capability to add new reactions to the collection.

Previous work in the Sheffield group adapted the so-called 'reaction vector' methods (Broughton et al., 2003) for classification into a *de novo* framework which enables any collection of reactions to be converted into a list of transformation rules that can be

applied to input molecules (Patel et al., 2008, Patel et al., 2009, Gillet et al., 2009). These reaction vectors are a simple method of encoding the changes that take place in chemical reactions based on the differences between the products and reactants. The vectors are very quick and simple to produce from a data set, and are obtained by a subtraction of the descriptors of the reactants from those for the products to give a list of differences. Reaction vectors can be used in a number of ways, such as categorisation of reactions, or as a simple text-based depiction of reaction transformations. However, the primary use by Gillet et al. was to apply reaction vectors to generate new structures from a given starting molecule, with this step forming the structure generation component of an evolutionary *de novo* design algorithm. Given that reaction vectors can be derived automatically from a database of reactions, this approach overcomes the limitation of working with a pre-defined set of transformations. However, the approach has a number of limitations including: execution time, especially when dealing with large numbers of starting materials or transformation rules, and issues with the optimisation especially when intermediates in a reaction sequence score poorly relative to the start and end point of a sequence.

The focus of this thesis is to extend the concept of reaction vectors to encode reaction sequences as vectors and to investigate their use in *de novo* design. This involves reworking the reaction vector approach to encode complete reaction sequences as single vectors, avoiding the intermediate steps that cause the scoring issues. This also has the effect of significantly optimising the structure generation process, by reducing the amount of execution time required to produce a result set. To produce the sequence information, tools have been developed that take a source of reaction data (such as an electronic lab notebook database) and produce sequences via the use of interconnected reaction networks and graph theoretic methods. This involves the expression of reactions and sequences in network space, creating a knowledge base onto which various reaction properties can be mapped. In addition, the range of reactions that can be encoded and used in the process has been significantly extended through amendments to the processing code.

Chapter 2 presents a review of methods of representing molecular structures and reactions, along with specific drug design applications that utilise these approaches for structure development and reaction prediction. In Chapter 3, a more detailed study of

de novo design tools is presented. A summary of the existing reaction vector project work is presented in Chapter 4. This includes the two different methods used to generate and validate reaction vectors, as well as the means by which structures can be generated by combining a reaction vector with a given starting material.

Chapter 5 presents the methods developed for collecting reaction data and producing sequences, which are based on graph theoretic and networking methods. These sequences are then used with a new reaction sequence vector method capable of representing them in a single step. In Chapter 6 this new method is compared with the original reaction-led approach in terms of the novelty and number of product molecules generated, as well as analysing the distribution and nature of the reaction sequences used to create the products. While both approaches generate interesting structures, it is clear that the reaction-led approach produces larger populations, with a greater diversity. Using the vector methods to categorise the collections of reactions and sequences used show the inherent bias towards particular functionalities and structures that affect the types of compounds produced. Chapter 7 presents different applications of the reaction sequence method, including: using the reaction networks to identify multiple equivalent routes to products; using reaction sequences to build structure activity relationship (SAR) profiles from a given starting material. Chapter 8 uses the curated knowledge base with existing drug design case studies. This includes a direct comparison with the original reaction vector based multi-objective drug design method, highlighting the advantages and disadvantages of the new sequence-based approach for drug design.

Chapter 2:
Representations of reactions

2.1 Introduction

As chemical research has progressed, and new forms of collaboration have developed, there has been a need to develop different means of communicating chemical information to others. These can range from simple written formulae, to images of structures, to more complicated methods of storing atom and bond connection data for computerised tools and databases. This chapter presents an overview of these methods, as well as the ways in which these can be used for searching molecular databases, and assessing similarity of a given molecule to those already stored. As the tools produced in this thesis utilise their own method for storing reaction data internally, there is a need to consider the different methods for depicting and storing these. Some of these approaches are extensions of existing molecular depiction methods, whereas others were specifically developed with a view to facilitating drug design or reaction searching. The last two sections of this chapter discuss these different reaction representation methods, as well as the drug design tools that rely on the special features of the bespoke reaction depiction forms.

2.2 Molecular representation

Chemical structures require specialised representation methods in order to be stored in searchable databases. These methods fall largely into two groups, namely linear representations (such as text formulae) and less human-readable descriptions such as connection tables.

Text-based approaches such as SMILES (Simplified Molecular-Input Line-Entry Specification) (Weininger, 1988) and InChI (International Chemical Identifier) (McNaught, 2006) permit structural data to be represented using standard ASCII

characters. In the case of SMILES, capital letters represent the individual atomic symbols, lower case letters represent aromatic atoms, parentheses indicate molecular branching and a simple paired number system is used for assigning atoms in rings. Hydrogen atoms are usually omitted for SMILES strings, i.e. H_2O is represented as O. An example of a SMILES string for a molecule is given in Figure 2.1.

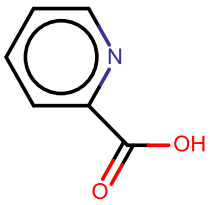
Molecule	SMILES
	OC(=O)c1ccccn1

Figure 2.1: Example of a SMILES String for 2-pyridinecarboxylic acid. (Wallace, 2015)

For substructures, an extension of this method known as SMARTS is available. This uses largely the same format as SMILES, with the addition of logical operators and wildcards to facilitate the specification of generic molecular queries. The InChI method, on the other hand, does not feature any specific substructure support. Instead, the advantage of InChI is the sheer amount of metadata that can be retained, with the string separated into a number of optional 'layers' containing structure, charge, stereochemistry and other relevant data, spaced by '\' characters. This enables more data to be encoded within the string, but at the expense of readability.

Alternatively, users may prefer to input structures using a chemical drawing package, in which case the storage of the converted data is commonly achieved through connection table formats such as the .MOL file (BIOVIA). The BIOVIA .MOL format carries a header, usually listing the molecule name and the generating program, followed by a non-redundant connection table (each atom is listed in turn, followed by each bond being recorded once only, as opposed to once for each atom in the redundant case), listing atomic properties, stereochemistry and any 'R group' designations in the case of generic 'Markush' structures.

Following the development of the World Wide Web, and XML (eXtensible Markup language) an XML schema has been developed for chemical information, the so-called

Chemical Markup Language (known as CML or ChemXML) (Murray-Rust and Rzepa, 1999). As with all XML documents, the CML schema establishes a series of rules for encoding information in a machine readable format that can be easily parsed, effectively embedding a form of connection table in a hierarchical structure. The schema can not only encode individual atoms and bonds within the molecule (Murray-Rust and Rzepa, 2003), but can also store spectral information (Kuhn et al., 2007) alongside, keeping all relevant data about a molecule together.

2.2.1 Molecular graph theory

From the earliest attempts to depict molecular structures, comparisons have been drawn between such structures and the basic forms used in graph theory (García-Domenech et al., 2008). Graph theory is best described as a field of mathematics that is concerned with the nature of connections between objects. A graph is a collection of entities, commonly referred to as nodes or vertices, alongside a set of pairs of these entities representing the connection between nodes, commonly known as edges. It should be noted that the nodes and edges define the graph, with the order of the entities being irrelevant. In a graph used in a chemical context, the nodes would represent the atoms, with the edges representing the bonds. If labelled with the relevant information, this form is sufficient for many chemical applications. An example of this form is shown in Figure 2.2.

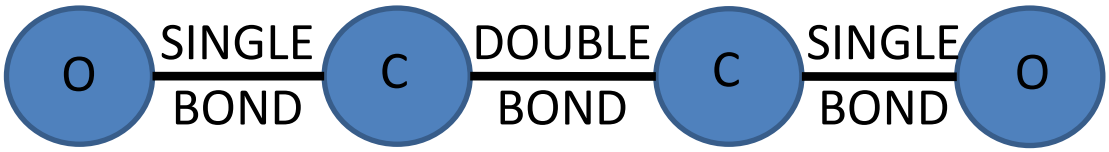


Figure 2.2: Molecular graph for HOC=COH. Hydrogens are omitted.

It should be noted that within this graph there is no ability to distinguish between isomers, such as the potential *cis* and *trans* forms around the double bond. However, these methods remain of significant value for representation of chemical structures in a mathematical context. Of particular interest is the concept of subgraph isomorphism, where the sets of nodes and edges of a particular graph are wholly contained within another. This is of particular interest in the context of molecular database searching, as

searching for a partial structure within a database is analogous to a search for a partial graph (subgraph) in a set of graphs.

2.2.2 Molecular database searching

The search algorithms used within a molecular database differ depending on the user requirements, for example, the user may want to look at the properties of a complete structure, for compounds with a particular substructure element, or indeed, those structures that are similar to a given example. In the case of an 'exact search' query, this may seem to be a straightforward operation at first, but the problem is that the structure connection table or SMILES data string may be produced in a number of equivalent ways, depending on how one considers the order of the atoms. This can lead to duplication in storage, or possibly false negatives when searching. Testing every possible numbering method during a search is computationally expensive, as for a table of N atoms there are $N!$ numberings to consider. Therefore it is necessary to ensure all structures stored and submitted use a standardised approach (a so-called 'canonical' representation), whereby no matter what the input order of atoms the output representation is always the same.

The first reported example of an automated algorithm for generating such representations was by Morgan, in his work on behalf of the Chemical Abstracts Service (Morgan, 1965). At first, the 'connectivity values' of each atom are set to the number of non-hydrogen atoms directly connected to it, thereafter a new value is calculated (the sum of the values of the neighbouring atoms). This process is repeated iteratively until the number of different connectivity values the molecule possesses is at a maximum. Once this state is reached, the atom with the highest extended connectivity value is chosen to be first in the connection table for the molecule, then moving in succession through its neighbours in descending order of extended connectivity and so on through the molecule. In the event of two identical connectivity values, properties such as atomic number and bond order are used to break the tie. Similar approaches have been used to create canonical SMILES strings and InChI representations for storage and searching purposes.

Once the structures are canonical, both the submitted query and the stored data can be directly compared. This can be done with direct reference to the connection table or string, or through 'hashing' the query and the stored data. Hashing is the generation of a new alphanumerical string based on the data according to a given algorithm, such as in the Freeland approach, a continuation of Morgan's work (Freeland et al., 1979). For large structures, the hash string is easier to process than the original structure and speeds up data retrieval. However, there is a risk of hash 'collisions' i.e. two molecules having the same calculated hash, which must be resolved if the database is to remain fully searchable.

When considering substructure searching, one approach is to utilise graph theory as discussed above. The substructure problem can be considered as a form of subgraph isomorphism – whether the structure subgraph from the query can be found wholly contained within the graph of the stored database structure. The earliest attempts to use computational methods for substructure searching were developed by Ray and Kirsch (1957). In their method, all of the molecular graphs representing the database entries are compared one at a time. However, such approaches are relatively slow over a data set the size of a typical chemical database (Barnard and Downs, 1992). To speed this up, a screening step (Dittmar et al., 1983) is used to remove all structures that cannot possibly match, before the subgraph search is performed on the remainder. The screening process relies on the creation of individual bit strings for each molecule in the database and the query based on a given set of rules. These bits are set (1) or unset (0) depending on the presence or absence of given substructural features, such as augmented atoms, linear sequences and other structural features such as rings. An augmented atom feature is a representation of a substructure, defined in terms of the atoms attached to a given central atom. For greater search precision, the augmented atoms also specify bonding information, for example, ring bonds (marked with '*') or chain bonds (marked with '–') and bond types (single, double, triple or aromatic) by adding the appropriate numerical bond index (1, 2, 3 or 4 respectively). Examples of augmented atom descriptions are shown in Figure 2.3. Depending on the degree of precision, a number of different fragment types can be used, excluding or including the bond type, and the relative atom numbers. In the figure, three different fragment types are listed in order of complexity, from simple lists of atoms in the top row, to full descriptions of the bond type in the bottom row.

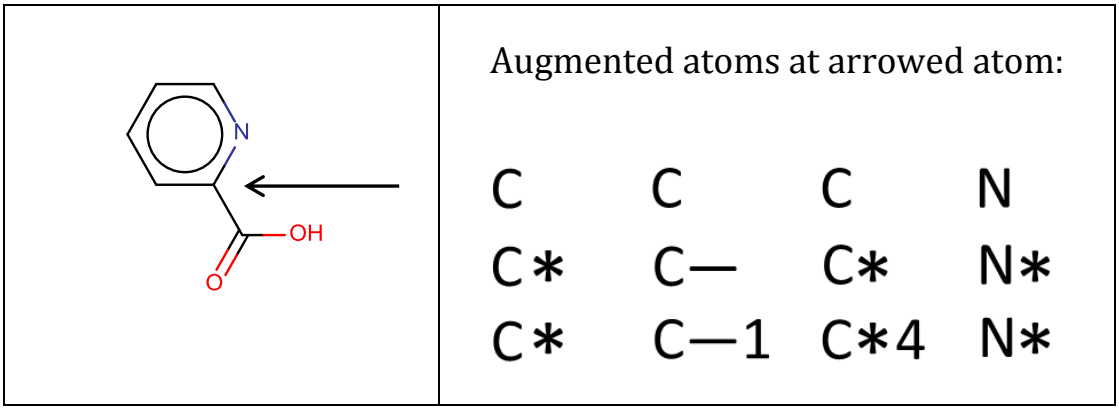


Figure 2.3: Illustration of potential augmented atoms for 2-pyridinecarboxylic acid. Each line represents an augmented atom term for the arrowed atom, at differing levels of detail. The top line is a simple list of the elements involved, before the addition of the bond type in each case ('*' indicates a ring bond, and '–' indicates a standard chain bond). The bottom line further adds the nature of the bonds, such as single, double etc. (1= single bond, 4 = aromatic bond). (Wallace, 2015)

The linear sequence screen is very similar to the augmented atom screen, but relates to a chain of between four and six interconnected non-hydrogen atoms rather than radiating from a central atom. This works as an effective substructure screen when combined with the bond type designation seen in the augmented atom method. Examples of these fragments are shown in Figure 2.4. As can be seen, because the sequences are processed in order, both the forward and reverse sequences need to be stored to ensure that the search operates correctly.

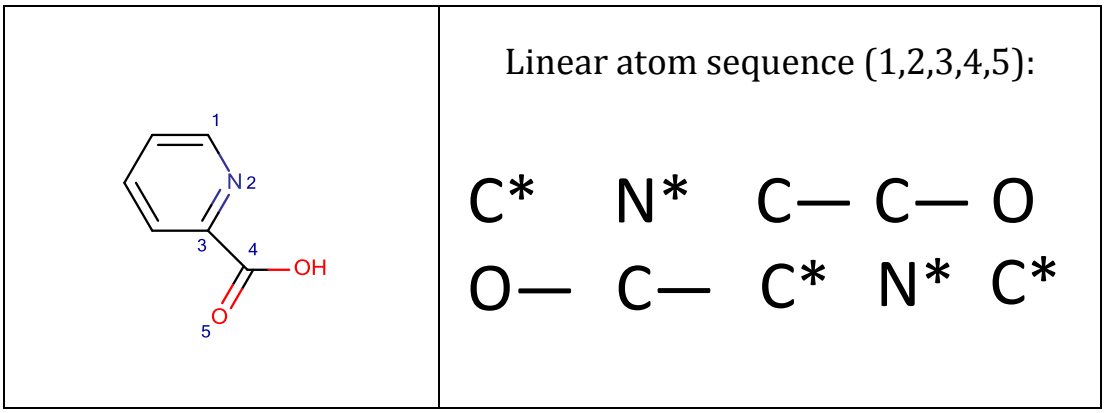


Figure 2.4: Illustration of linear sequence screens for a given substructure of 2-pyridinecarboxylic acid. (Wallace, 2015)

In Dittmar's approach, combinations of 12 categories of screens similar to those described above are used to generate a 2048 bit string for each molecule (in contrast, similar screen approaches used in modern systems are typically limited to 1024 bits for simplicity). The strings of the query and the molecule in question are then compared, with results only returned where every bit set in the query is also set in the molecule. To ensure the system is efficient, it is necessary to carefully select the structural elements used for the bits, ensuring that they occur sufficiently often to be useful, while remaining independent of one another to ensure maximum effectiveness.

These same bit strings can be used to provide a means for searching based on molecular similarity, such as with the Tanimoto coefficient (Willett et al., 1998, Willett and Winterman, 1986). The Tanimoto coefficient between two molecules A and B (S_{AB}) is:

$$S_{AB} = \frac{c}{a + b - c}$$

Equation 2.1: Tanimoto coefficient for molecular similarity.

where a represents those bits set for molecule A, b represents those set for B and c represents the bits set in both (the 'common' elements). A similarity value of '1' indicates that the two molecules have identical bit strings (they are not necessarily identical themselves) while '0' indicates that there is no commonality in terms of the bit strings. By calculating the Tanimoto coefficient for each entry relative to the query molecule, a ranking of molecules by similarity can be obtained.

While the Tanimoto coefficient is one of the more commonly used measures of molecular similarity, other measures exist that can be used in the same manner. These include the Dice and Cosine coefficients (Equation 2.2).

$$S_{AB} = \frac{2c}{[a + b]} \qquad S_{AB} = \frac{c}{\sqrt{[ab]}}$$

Dice Cosine

Equation 2.2: Equations for the Dice and Cosine coefficients of molecular similarity.

Alternatively, measurements exist that give an indication of how dissimilar two molecules are that work as inverses of the similarity measures (so called 'distance' measurements). Examples of these coefficients are shown in Equation 2.3.

$$D_{AB} = [a + b - 2c] \qquad D_{AB} = \sqrt{[a + b - 2c]} \qquad D_{AB} = 1 - \frac{c}{[a + b - c]}$$

Hamming Euclidean Soergel

Equation 2.3: Equations for the Hamming, Euclidean and Soergel molecular distance measurements.

2.3 Reaction representation

Storage and representation of reactions can be considered in a similar manner to individual molecular; however the methods have to be adapted to cope with multiple substances and roles within the same entry.

In reaction SMILES, reactants and products are separated from one another with '>' signs, and any catalysts or other agents are specified between the two groups. In cases where there are multiple reagents or products a '.' character separates them. Figure 2.5 shows a reaction SMILES string for a typical reaction.

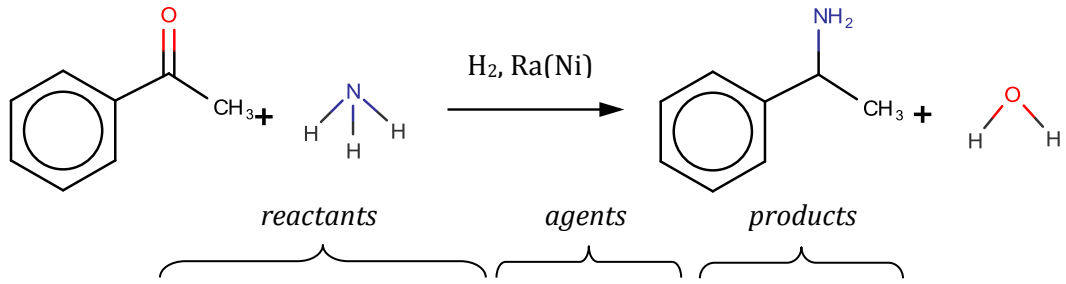
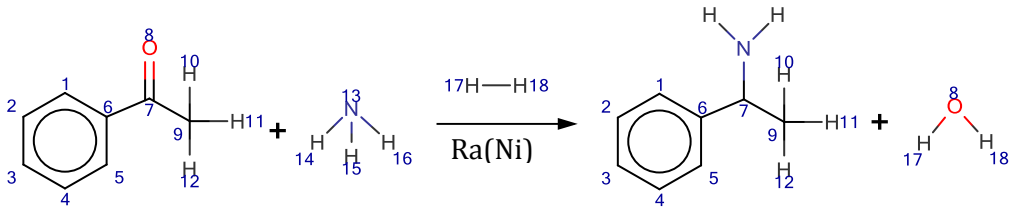


Figure 2.5: Reaction SMILES string for the Mignonac amination reaction. (Wang, 2010, Wallace, 2015)

The MDL .RXN format serves the same purpose for reactions as the .MOL file does for molecules. The format is related to .MOL in the sense that each structure is represented as a connection table, but additional information is included, such as whether a structure represents a reactant or product (via tags and positioning within the file

record, depending on the precise revision), and potentially, atom mapping information. In atom mapping, each atom within a reaction is tagged, usually using a numerical system. The idea is that the same atom receives the same tag throughout the reaction, thus linking the reactant structures to the products, and indeed linking one reaction in a scheme to another, as necessary. An example of a mapped reaction is shown in Figure 2.6. A comprehensive overview of atom mapping in its various forms can be found in the review by Chen, et al. (2013).



Reactants

[H:11][C:9]([H:10])([H:12])[C:7](=[O:8])[c:6]1[cH:5][cH:4][cH:3][cH:2][cH:1]1.[H:14][N:13]([H:15])[H:16]>

Agents

([H:17][H:18]).RaNi>

Products

[H:14][N:13]([H:15])[CH:7]([c:6]1[cH:1][cH:2][cH:3][cH:4][cH:5]1)[C:9]([H:12])([H:10])[H:11].[H:18][O:8][H:17]

Figure 2.6: Example of atom mapping for the Mignonac amination reaction, including Reaction SMILES (split for legibility).(Wang, 2010, Wallace, 2015)

The process of generating atom maps for reactions can be entirely automated, and is a key element in the storage and recall of reactions in databases. The first fully automatic mapping method was reported by Lynch and Willett (Lynch and Willett, 1978b, Lynch and Willett, 1978a). Initially, the method relied on comparison of the two sides of a reaction written in Wiswesser Line Notation (an early text-based molecule representation), but this was followed up with a more extensible approach based on matching the maximal common substructure (MCS) between the two sides (McGregor and Willett, 1981), (Funatsu et al., 1988). As the name suggests, the MCS method compares the molecule graphs of the product and reactant sides to find the largest

common element to both, and uses this to ensure any relevant mapping or comparisons are canonical. While all of the automated algorithms are relatively quick and highly effective, if the initial reaction is imbalanced the chances of failure are high.

A recent development of the MCS method (Apostolakis et al., 2008) for reaction mapping adds additional weight values to bonds based on the atoms that form them (for example, C-C σ-bonds are assigned a weight of 1.5, while C-N amine bonds are weighted as 0.48). These weights correspond to the likelihood of a bond being broken in a transformation (a lower weight indicates an easier breakage), and therefore also represents the cost of not matching particular bonds between the two sides of a reaction, the unmatched bonds representing the reaction centre. Alternatively, a general representation of a given reaction can be produced from superimposition of the two sides of the reaction onto one another to identify the reaction centre (de Luca et al., 2012). This condensed reaction graph, as illustrated in Figure 2.7, can be treated in the same manner as any normal molecule, and therefore molecule similarity measures can be used to compare reactions with given queries. In the graph, the changing bonds are colour coded to highlight the differences, with red lines indicating lost bonds, and green lines indicating created bonds.

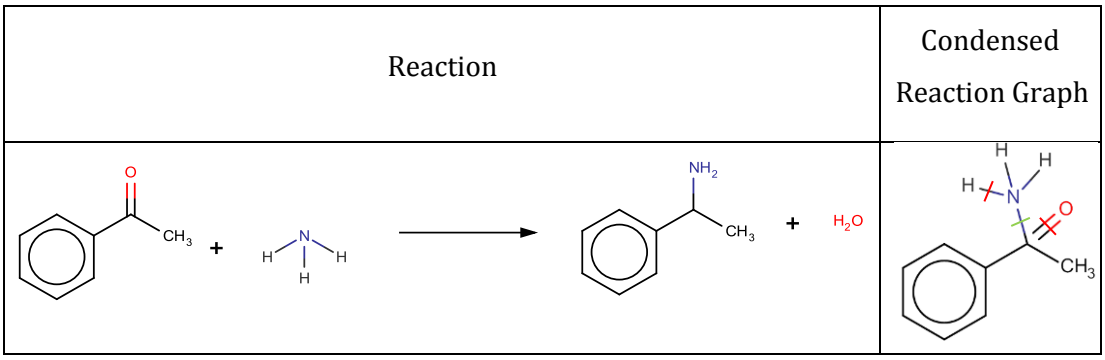


Figure 2.7: Illustration of the condensed reaction graph approach for the Mignonac amination reaction. (Wang, 2010, Wallace, 2015)

The atom map assigned to the reaction can be used to identify the reaction centre via direct comparison of the two sides, forgoing the need for the user to specifically identify the region in the query.

The ChemXML format also has a reaction subset, CMLReact (Holliday et al., 2005). This uses the hierarchical structure of XML to collect the molecular representation and other data for reactants and products within a parent 'reaction' element. As before, the relevant connection tables to encode the individual molecules are stored. However, the reaction subset also permits tagging and mapping of individual atoms, bonds and electrons across molecules, allowing the encoding of mechanistic details that are useful for synthesis. By further exploitation of the parent and child principle, a reaction 'step list' can be produced, encoding an entire linear reaction sequence in the same manner within the same file.

2.3.1 Chemical reaction databases

In order to enable coherent processing of published literature, many chemical databases exist. While there are databases that deal exclusively with compounds and their properties such as the NIST Webbook (NIST) and the CAS registry (Chemical Abstract Services), many include some element of reaction data. These can be divided into two groups (Boiten et al., 1995):

- Those that aim to cover the entire literature base within certain set boundary conditions such as CASREACT (Blake and Dana, 1990), and CrossFire Beilstein/Gmelin (Hicks, 1990) (now the Reaxys database).
- Subsets of useful reactions without any claim to completeness, some of which are freely available and some are licensed for in-house use such as SYNLIB (Distributed Chemical Graphics). Many of the original examples of these such as ORAC, REACCS (Mills et al., 1988), and IRDAS have been discontinued as a result of consolidation between suppliers, but tools such as BIOVIA DiscoveryGate (BIOVIA) offer access to similar data collections.

Some of the organic reaction databases currently available are given in Table 2.1 and Table 2.2, representing free and commercial databases respectively.

Database name	Source	Approximate number of reactions	Rate of expansion
Organic Syntheses (Organic Syntheses Inc.)	Independently confirmed reaction routes	>5000	~40 new routes a year
Boston CMLD Synthesis protocols (Boston University)	Boston University Chemical Methodology and Library Development	133	No further expansion noted
The Chemical Thesaurus (Leach)	Open access submission to editor	4000	No new reactions for some time
Webreactions (openmolecules.org)	Extraction from ChemSynth and ChemReact	~400,000	Unknown
ChemSpider Synthetic Pages (Royal Society of Chemistry)	Open access submission to editor (Public Domain)	~250,000	Dependent on submissions and review
USPTO Collection via NextMove Software (Lowe and Sayle, 2014)	Extraction of reactions from US patent applications (2001-2013) and grants (1976-2013)	1,000,000	n/a

Table 2.1: A list of freely available organic reaction databases and their properties.

Database name	Source	Approximate number of reactions	Rate of expansion
Reaxsys (Elsevier)	Journals and patents (Formerly Beilstein, Gmelin and Patent Chemistry Databases)	>22,000,000	200,000 new reactions annually
CASREACT (Chemical Abstract Services)	Journals and patents	>60,000,000	30,000-50,000 new reactions weekly
ChemInform Reaction Library (CIRX) (Wiley/FIZ CHEMIE Berlin)	Journals	>1,200,000	Monthly updates of varying sizes
Current Chemical Reactions (Thomson Reuters)	Journals and some US patents	>1,000,000	Monthly updates, no figures given
Derwent Journal of Synthetic Methods (Reuters)	Journals and patents	>75,000	No longer updated or supported.
e-EROS (Encyclopaedia of Reagents for Organic Synthesis) (John Wiley & Sons)	Submissions to editor	~70,000	Twice annual updates, 150 reagents and articles updated in last release
Science of Synthesis (Thieme Chemistry Publishing)	Submissions to editor	>300,000	~14,000 new reactions annually (estimated)
SPRESI (InfoChem)	Journals and Patents	~4,400,000	100,000 new reactions annually

Table 2.2: A list of commercially available reaction databases and their properties.

2.3.2 Reaction database searching

The same approaches used to search for an individual molecule in a database (Section 2.2.2) can be used in a reaction context, allowing for users to find reactions that can be performed with a given starting material or that give a particular product. However, these methods will not necessarily work for finding reactions that retain a particular substructure, or for searching by reaction type. These additional mapping approaches are necessary to ensure a reaction search is effective. Conducting a standard substructure search in this context (looking for reactions that use a particular functional group, for example) may produce a large number of unsuitable hits. This is due to the fact that the search will lead to a collection of structures that have the structural features within the reaction centre, but not necessarily in a reactive position, due to steric effects or other considerations. The usual approach to these types of searches is to instead consider only the portions of a reaction that change (known as the reaction centre), and return results based on this.

When performing a search within a reaction centre, additional input methods are required to indicate which structural features are parts of the query. For example, the BIOVIA Draw program (BIOVIA) allows this through special bond indicators that indicate the changes, placing an 'X' through any bonds that do not change and using bespoke symbols on those that do. A table of these indicators is included below in Table 2.3, along with a sample of a suitably labelled reaction in Figure 2.8.

Bond Symbol	Effect
#	Bond change unspecified, can change type, be broken or be formed (same as no symbol)
|	Bond changes in type (single to double, double to triple etc.)
||	Bond is formed (if used on product) or broken (if used on reactant)
|||	Combination of | and ||, bond is formed/broken and changes type
X	Bond remains unchanged

Table 2.3: The bond symbols used in BIOVIA Draw and their meanings.

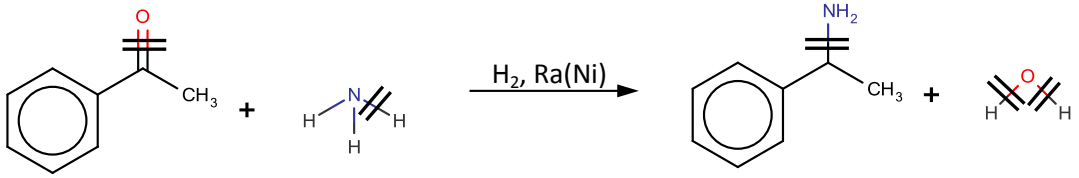


Figure 2.8: Bond change marking of the amination reaction in Figure 2.5. (Wang, 2010, Wallace, 2015)

In addition to the bond indicators, some of these programs also use the atom mapping approach described previously, enabling the user to select individual atoms and assign numerical tags to indicate their role and position in the final structure.

To input a reaction query in a SMILES text form, another subset of the SMILES language, SMIRKS, is required. This relies on five simple rules to generate a compatible string for searching (Daylight Chemical Information Systems) :

- The reactant and product sides of the reaction have to have the same numbers and types of mapped atoms i.e. each mapped atom in the reactant should have a counterpart in the product. If need be, atoms can be added or removed during the reaction as necessary (if an agent or catalyst has to be specified, for example), but these atoms cannot carry a mapping. To define a mapping, the atom symbol must be followed by ':N', where 'N' is the mapping number.
- Stoichiometry within the string is assumed to be 1:1 for all atoms on both sides, if additional equivalents of reactant or product are required, they must be entered an appropriate number of times.
- If hydrogen atoms are stated explicitly on one side, the equivalent hydrogen atoms on the other side must also be stated. They must also be atom mapped in both cases.
- Bond expressions must be valid SMILES strings; it is not possible to use wildcards to represent multiple bond types. Atom queries on the other hand can use the '*' or '?' wild cards as appropriate.
- Atoms where the connectivity and bond order remain identical in both cases can be specified in the SMARTS molecular pattern format; otherwise they must be SMILES strings.

These criteria result in an expanded string, similar to the SMILES string seen previously, where atom numbers are specified by values after the colon, as seen in Figure 2.9.

[*:6]([C:7](=[O:8])[C:9]).([N:13][H:14][H:15][H:16)>([H:17][H:18]).Ni>
[*:6][C:7]([N:13][H:14][H:15])[C:9][H:10][H:11][H:12].[H:17][O:8][H:18]

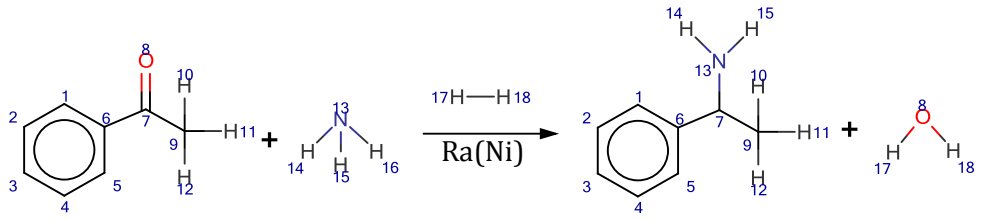


Figure 2.9: SMIRKS string for the reaction shown in Figure 2.5 (split at product portion for clarity). For reference, the mapped reaction is included. (Wang, 2010, Wallace, 2015)

2.4 Reaction classification methods

A number of different approaches to the storage and processing of reaction centres have been created to permit effective searching and reaction classification (Chen, 2008). Many of these rely on some form of atom mapping approach (Section 2.3), with the REACCS database (Grethe and Moock, 1990) combining this with a substructure fingerprint approach for searching, similar to that used in standard molecule database searching methods.

Reaction classification methods are also based on reaction centres. One such example is the appropriately named Classify method by InfoChem (InfoChem), which is used by a number of commercial databases. This has three different levels of search depending on how many of the atoms immediately connected to the reaction centre are included, with the narrower searches including more of the specific environment surrounding the reaction centre. Hash codes are generated for reaction centres at each level and can then be summed to give an overall value for the reaction, which can be compared with others in the database to assess similarity of reaction type.

There also exists a subset of reaction classification methods that are specifically designed to be used with a particular tool or for a specialised purpose. The Dugundji–Ugi reaction model (1973) is one such method, used in the IGOR synthesis method that generates novel reactions from first principles (Ugi et al., 1993). The model states that reactions of ensembles of molecules may be treated in the same manner as a graph isomorphism problem, effectively creating a graph with the molecules on the nodes and reactions on the edges. Using this model, the reaction transformation operation (the exchange of atoms and electrons that occurs during the reaction itself) is represented as a bond-electron matrix, as illustrated in Figure 2.10. There are three matrices in total, one representing the reactants (B - beginning), one representing the products (E - end) and an overall reaction matrix R (effectively E-B). By consideration of what the numbers represent, and the basic rules of chemistry, several conclusions can be drawn about the properties of the matrix:

- Since charge is conserved over the reaction, the sum of all the individual entries of the reaction matrix must equal zero (no electrons can be created or destroyed).
- It therefore follows that, if the formal charge does not change for any atom, then this also applies to the individual row and column pairings corresponding to the atoms (if charges do migrate then numerical imbalances occur at the points of migration, in accordance with the first rule).

As a result of these properties, it becomes possible to predict reactions via determining a solution for the matrix set that satisfies the charge rules, given either the reactant matrix or the reaction transformation (B or R in the nomenclature). This strict mathematical treatment can and has led to previously unknown reactions being presented as solutions, and exploration of new solution space (Herges and Ugi, 1985).

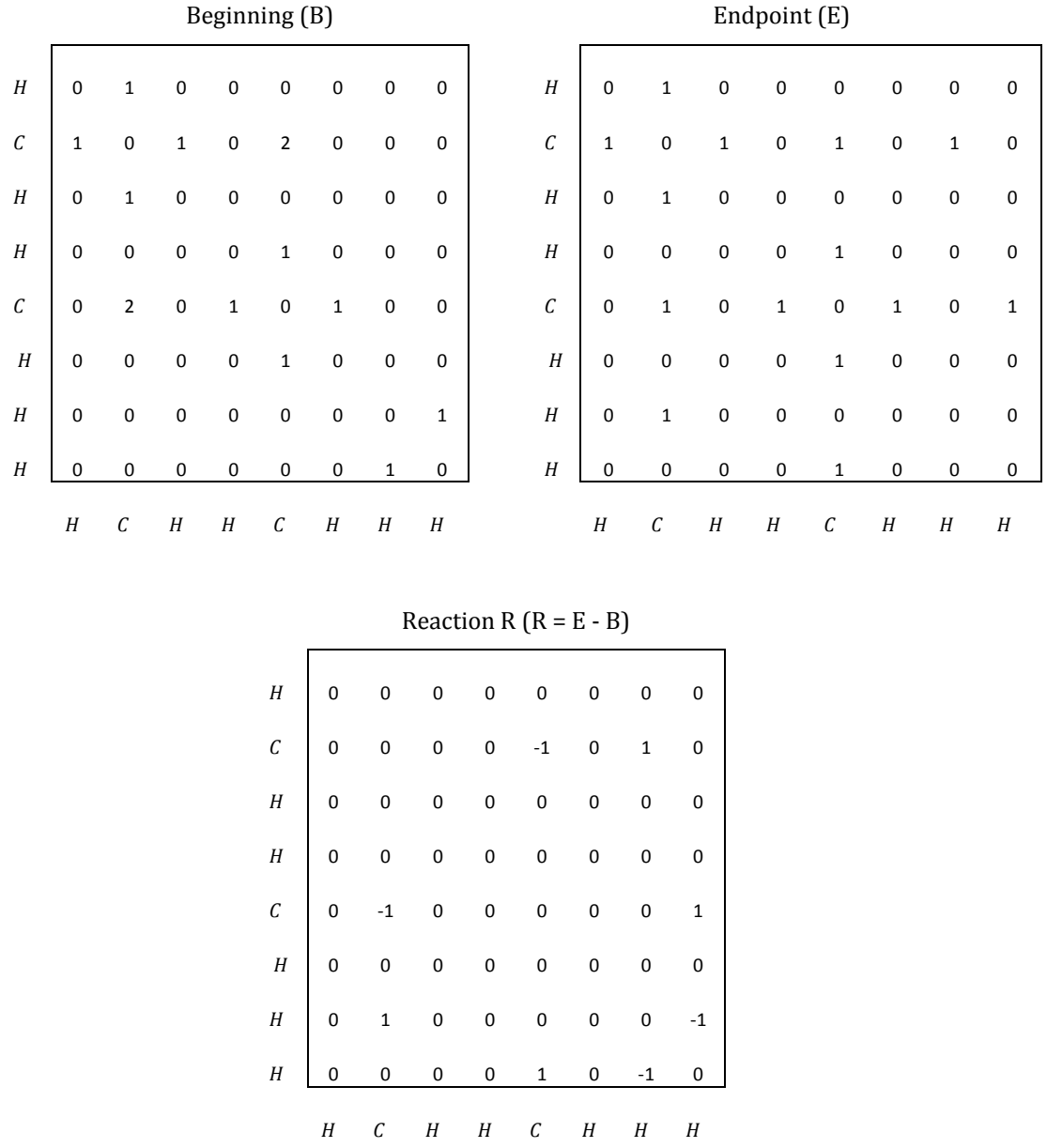


Figure 2.10: Dugundji-Ugi matrices for the reaction $CH_2CH_2 + H_2 \rightarrow CH_3CH_3$.

The Hendrickson classification method (Hendrickson, 1997b) attempts to represent chemical reactions in terms of the bonds broken and formed, extending similar visual methods proposed by Fujita, Vladutz and Balaban (Chen, 2008). These methods all share the idea of representing the bonds and atoms of the reaction centre as a cycle, but the Hendrickson case is greatly simplified, with the broken bonds shown as solid lines and the newly made bonds as dashed lines. This effectively condenses the information at the reaction centre into an 'imaginary transition state', as seen in Figure 2.11. In

these centres any bonds that are unchanged in the reaction but are useful for classification can be designated as shell bonds (such as the σ-bond in σ, π double bond systems), marked as a bolder line.

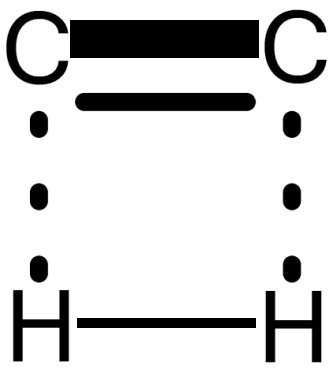


Figure 2.11: Transition state logo for the reaction shown in Figure 2.10. The σ-bond in the double bond system appears as a bolder shell bond (Wallace, 2015).

While initially complicated to set up, the idea of an imaginary reaction centre has significant advantages when describing the intermediary phases of a reaction scheme, and can depict the change in a molecule without the need for atom mapping, as would be the case for connection table-based formats where the centre is not immediately obvious. However, this system as designed is intended for pictorial representation, and for computational purposes a linear, text-based 'Synthesis Tree' approach using the same data is implemented instead. This is effectively a branched retrosynthesis diagram showing all possible disconnects, and can be generated easily as an aid to synthetic planning for a given target compound.

A further variant (Hendrickson and Miller, 1990) is more specifically designed for *in silico* operation, focussing on rapid retrieval of reaction data from models and databases. In this form, classification is limited to carbon atoms in the reaction centre, with other atoms represented simply in terms of electronegativity relative to carbon. As a consequence, four types of bonding can be defined, namely carbon-carbon σ-bonding (R), carbon-carbon π-bonding (Π), carbon to electronegative atom (Z) and carbon to electropositive atom (H). From this method, it can be seen that a relative change in the total of R would indicate a structural change, with the other values indicating functionality changes.

One graph theory approach to handling reactions (Crabtree and Mehta, 2009) is to consider that the sum of the graphs of the reactants is effectively transformed into the sum of the graphs of the products during a reaction. The key is to find therefore a mapping for both sides that minimises the number of bonds formed or broken, attempting to match the relevant subgraphs. Many approaches to solving this problem exist, either as atom mapping functions in their own right such as the NAUTY (McKay, 1981) and Faulon (Faulon et al., 2004) algorithms, or as attempts to provide canonical names and structural representations of molecules, such as the Maximal Common Substructure/Subgraph method (MCS) (McGregor and Willett, 1981), (Funatsu et al., 1988), as discussed in Section 2.3.

The reaction vector approach (Broughton et al., 2003) involves determining a difference vector between the reactants and products. It was originally used as a means of comparing reactions for classification purposes, but the generic representation it generates has since found use within *de novo* design (Patel et al., 2008). The idea of categorising reactions in this manner dates back to the work of Vléduts (1963) (Willett, 1980). Vléduts' method relies on the logical premise that the bonds produced as the result of a given reaction transformation will be different in nature to those destroyed in the reactant, and as a consequence the reaction centre can be identified by tracking these changes. At the time of the Vléduts work, limits in computational power prohibited automatic processing for all but the most simple connection tables, but by the time of Willett's work, technology enabled the principle to be extended to the vast majority of reactions. In the case of the reaction vector method, no connection tables are required; features such as atom pairs are used to encode the individual components of a reaction either using the number of occurrences of a particular descriptor or its presence or absence as the vector elements and the difference is calculated by subtracting the reactant vectors from the product vectors.

Atom pairs are essentially substructure representations, encoding characteristics about the properties of each atom (type, connectivity, etc.) relative to all others in a molecule, usually expressed in the form '*atom type a – (distance in bonds) – atom type b'*. In this case, as with the bond-electron matrices, an overall reaction vector can be obtained via subtraction of the product pair list from the reactant pair list, with the result giving an

indication of the changes within the reaction centre without reference to the rest of the structure.

2.5 Forward synthetic planning methods

A number of approaches have been developed to predict products that can be synthesised from given starting materials. For example, CAMEO (Computer Assisted Mechanistic Evaluation of Organic Reactions) (Salatin and Jorgensen, 1980) analyses the atom types present in a given molecule before choosing an appropriate mechanism and simulating the likely synthetic products. A similar concept is used with the previously mentioned IGOR (Interactive Generation of Organic Reactions) (Ugi et al., 1993). The main difference between the two approaches is that, rather than using an existing knowledge base of mechanisms, IGOR is capable of generating novel reactions, based on the Dugundji–Ugi model of reaction representation. Following the Dugundji–Ugi electron redistribution rules, every generation step in the process minimises the reallocation of valence electrons while forcing a structural change, in some cases creating new reaction classes for study and predicting likely new products.

The EROS (Elaboration of Reactions for Organic Synthesis) tool (Gasteiger and Jochum, 1978, Höllering et al., 2000) is another example of this type of methodology. This is a rule based system that attempts to predict the products of a reaction. The process of rule creation is complex, and attempts to recreate the mixing conditions of the real world system. This requires the definition of the system in terms of the processes used (defined as 'reactors' in the program), details of the phases of the mixture in the system, and any other information about the reaction behaviour that is necessary. This information includes definitions of what combinations of starting materials are possible (such as whether dimers can form), while also being used to mark particular reactors and phases in terms of reaction behaviour (described as the reaction 'mode'). If reactions do not occur in a particular phase, the mode must be set to 'inert' to ensure the kinetic profile for the predicted reaction is correct. With these rules defined, the EROS program can then simulate the reaction products from a given set of starting materials, as well as calculating relevant physicochemical properties to assist in building a kinetic profile for the process.

2.6 Retrosynthetic approaches

The storage of structural information for a given reaction can also allow for automated retrosynthesis of molecules, in much the same way an individual would with a drawn structure (Cook et al., 2012). In retrosynthesis, an established end molecule is deconstructed at key attachment points, with each disconnection responding to an existing chemical reaction in the forward synthesis. The usual approach when performing retrosynthesis *in silico* is to use a pre-established rule set to provide a 'knowledge base' for the system and identify the points of disconnection. The first full program to perform a retrosynthetic role, LHASA (Logic and Heuristics Applied to Synthetic Analysis) (Corey et al., 1972) relied on manual encoding of reaction rules in a bespoke language and so was reliant on the operators to provide the disconnection logic and add functionality. A number of revisions were made to LHASA during its lifetime to add additional features and improve the quality of the retrosynthesis (Corey and Jorgensen, 1976). However, as research progressed, a more automated approach was sought leading to alternative programs such as the SEEDS tool (Honma, 2003), which fragments the molecules, before checking commercial databases for suitable derivatives and starting materials.

Combinations of retrosynthesis and reaction prediction have also been reported, such as Hendrickson's SYNGEN (Hendrickson, 1997a). This uses his reported system of *in silico* reaction classification (Section 2.4.1) to identify bonds in the target molecule that are suitable for retrosynthetic disconnection. These are then used to fragment the molecule, generating a suitable synthesis reaction. In order to ensure feasibility, an additional filtration step removes any reactions that are too complicated, or that contain a starting material not present in the SYNGEN knowledge base. A similar approach is used by WODCA (Workbench for the Organisation of Data for Chemical Applications) (Gasteiger et al., 2000). This is used to retrosynthesise a target molecule which can then be passed to the EROS program to generate a reaction sequence for synthesis.

2.7 Reaction networks

Graph-theoretic approaches can be used to depict reactions and sequences (Temkin and Bonchev, 1992). In reaction networks, each vertex represents a molecule (reactant, product or intermediate), with each edge representing an individual reaction step. In order to correctly represent the flow of the reaction, the edges can have a direction assigned, indicating a transition that only occurs in one direction (reversible reaction steps are represented using undirected edges). An example of a reaction graph is shown in Figure 2.12.

Figure 2.12: Example of a reaction graph (left) for a simple reaction sequence (right).

This method can also be used to represent multiple step sequences in the same manner, with separate reaction graphs linked together to show the reaction progression. These graphs can be used for mechanism elucidation, by generating all possible reaction combinations that fit a given set of criteria (Sinanoglu, 1975). In this approach, an exhaustive set of graphs are generated, representing all possible reaction mechanism combinations for a given number of reactions. A list of potential mechanisms for a simple two step sequence from the hypothetical starting material MOL1 is illustrated in Figure 2.13.

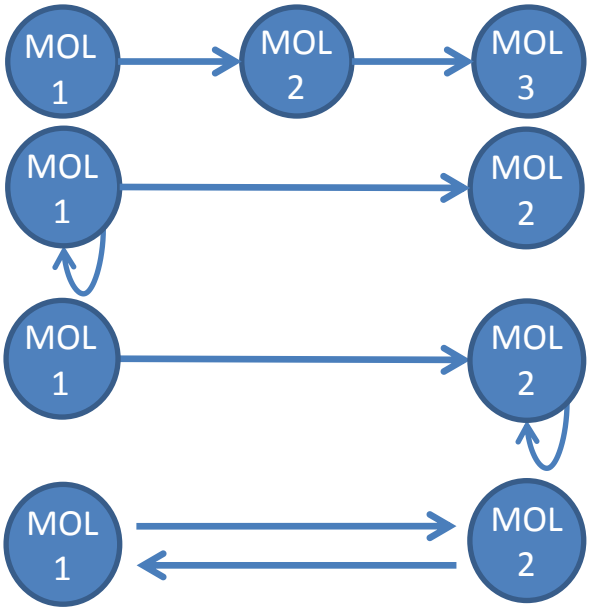


Figure 2.13: The reaction graphs for a two step reaction sequence, including loops where a loop represents a rearrangement.The four graphs represent: a simple two step progression; a rearrangement of molecule 1 followed by a reaction to molecule 2; a reaction to molecule 2 followed by a rearrangement; and a dynamic equilibrium. With comparison to the observed behaviour of the reaction system, logical conclusions can be drawn as to which of these mechanisms are valid, facilitating assignment.

By mapping the known reactant and product molecules of the sequence onto the nodes, it is possible to ascertain which mechanism is correct. Comparing observed experiments with these graphs makes it clear if certain mechanisms are impossible in given circumstances, and indicates which candidates are most likely. In this example, if an intermediate compound is detected as well as the starting material and product, this indicates that the linear mechanism (the top of Figure 2.13) must be correct, as both of the other options only lead to the generation of a single molecule. This is of particular benefit for more complicated sequences where mechanisms and in some cases, reaction rates can be identified via application of this approach in conjunction with the observation of reaction kinetics (Temkin and Bonchev, 1992).

By combining collections of reaction graphs, it is possible to build a network illustrating the interactions between synthetic pathways. Initial literature referred to these forms as chemical networks (Sellers, 1967), but to avoid confusion with property networks

Oster and Perelson (1974), as well as others, propose the term 'reaction networks'. By utilising graphs in this manner, it is possible to gain an insight into a reaction collection. A number of metrics can be calculated to profile the nature of the connections within a reaction data set. One example of particular interest is the node degree value, which represents the number of edges that include a particular node (either inbound or outbound). This can be used to highlight areas of over and underrepresentation in a reaction network context, such as the k_{in} and k_{out} parameters reported by Grzybowski et al. (2009). In this context, the identification of nodes with particularly low degree values would indicate regions in the network space that are underrepresented. Another, more straightforward visualisation of reaction sequences for comparison was developed by Proudfoot (2013). This takes a sequence and represents it as a road map, colour coding each edge based on reaction yield and increasing the edge weights based on the reaction scale. This means that high yield, large scale reactions are represented by heavy, green edges. The nodes themselves also vary in size, so molecules with high atom counts are larger. By generating these networks for multiple reaction pathways, the relative properties are made immediately obvious for comparison.

A more detailed approach to navigated reaction sequences is the ChemGPS project by the Grzybowski group (Fialkowski et al., 2005, Bishop et al., 2006). This project uses a reaction network derived from the CrossFire database (Elsevier) to create navigable views of reaction databases. These can be effectively crawled and mined for information in a similar way to the process used to index files within a search engine crawler. As a consequence, ChemGPS can be used as a means of mapping and searching large collections of reactions for relevant information. The network creation process for ChemGPS is the same as with other reaction network methods, with molecules as nodes, and edges connecting nodes where a reaction between the two molecules is reported. Initially, the tool was used to highlight areas of chemical space where there are gaps in synthetic knowledge that can be resolved by developing new syntheses (Grzybowski et al., 2009). In this case those areas where there is a low density of edges (low node degree), or a gap within the network would be highlighted as regions of interest.

Later work (Kowalik et al., 2012, Gothard et al., 2012) saw ChemGPS extended to consider the pathways contained within the network for sequence optimisation.

Kowalik reports a method for suggesting and optimising syntheses from a given starting material. In this approach, one of the many algorithms for computing the shortest path between nodes in a network (Dijkstra, 1959, Hart et al., 1968, Cherkassky et al., 1996), is used to present all possible reaction pathways between two points. However, as part of the optimisation, each edge in the network is assigned a weighting based on parameters of interest, such as cost, or ease of synthesis. The collected weightings for each path are then used for comparison and evaluation of the results, to help find the optimal pathway. Another, retrosynthetic approach to the same problem is reported by Gothard (Gothard et al., 2012). Firstly, all reactions in the network that lead directly to the product are considered in terms of the previously discussed cost parameters. Once the optimum reaction has been selected for this step, the process continues backwards from the starting material of that reaction through the network until a suitable start point is identified.

The network path searching approach has also been reported as part of a potential 'early warning tool' for identifying easily available precursors to hazardous materials (Fuller et al., 2012). This utilises the same backwards search method through the network as with Gothard's sequence optimisation but, rather than looking for a single sequence, it focussing on finding simple chemical routes to regulated substances from common household materials. If a path is identified that can be readily followed by those with a basic understanding of chemistry, or access to basic chemical literature, the relevant authorities can be notified to take action.

2.8 Conclusions

In this chapter, a number of different representation styles, formats and database searching strategies have been discussed. These all have their own particular benefits in use, but for general purposes simple approaches such as SMILES and the .MOL and .RXN dominate. It is clear that cataloguing reactions in a manner that permits rapid searching and synthetic awareness requires additional treatment and curation of the data. This includes a number of methods for automated reaction centre detection and classification, such as the reaction vector approach which forms the main focus of this thesis. In the next chapter, methods for *de novo* design will be discussed, looking at the history of the technique, and the various design and evaluation methods employed.

Chapter 3:
De novo Design

3.1 Introduction

The idea of using computational technology to assist in the design of drug candidates began to take hold in the late 1980's, when it started to become possible to make comparisons between medically interesting protein structures and the three-dimensional geometry of small, drug-like molecules in order to assess the ability of the molecules to bind to key receptor sites. Having the ability to visualise the protein on screen and assess significant numbers of potential candidates without the need for expressing large amounts of the necessary protein, greatly assists in the design process.

Attempts to design worthwhile therapeutic candidates *in silico* have largely fallen into two main categories: those that evaluate large numbers of molecules already in existence (so-called high-throughput virtual screening) and those that build novel molecules in order to fit a pre-established series of parameters, working from the 'ground up' (Schüller et al., 2008). It is this *de novo* approach that this chapter will focus on.

3.2 *De novo* design tools to date

In *de novo* design, there are two key elements that a computational tool needs to implement: the method by which individual candidates are constructed, and some scoring function to evaluate the candidates.

The earliest *de novo* tools operated in three-dimensional space, producing molecules according to a set of constraints imposed by the receptor site with which they were to interact. It is these interactions that determine how effective a molecule will be therapeutically, and therefore form the basis for any scoring or comparison. However,

the fact that the three-dimensional structure of the active site needs to be known in detail drastically limits the number of potential problems that can be analysed with these methods. In addition, the computational power required to model such sites and interactions is sufficiently large as to force considerable compromises in the modelling, compromising the results (Glen, 2011). As a consequence, later work has seen a more ligand-based approach to the problem, with scoring instead being based on known active molecules.

In the ligand-based systems, the properties of the respective candidates are evaluated against the structural features of an established reference compound or indeed a series of compounds. Depending on the tool used, and the nature of the scoring function, ligand-based design methodologies can operate in either two or three dimensions. For example, the three-dimensional case may use an appropriate model of the structural features necessary for recognition (known as a pharmacophore); while the two-dimensional case performs comparisons based on topological characteristics of the reference compound(s). With ligand-based methods providing significantly increased adaptability to different targets over receptor-based approaches, while also proving to be computationally less expensive to operate in the two-dimensional case, it is no surprise that the majority of *de novo* tools produced after 1995 include some element of ligand-based scoring within their functionality (Hartenfeller and Schneider, 2011, Schneider, 2014).

Over the course of this chapter, the various scoring present in *de novo* design tools will be explored, followed by the methods by which molecules are generated *in silico*. The approaches towards searching chemical solution space will also be covered, before considering the efforts to implement synthetic feasibility checks within design tools.

3.2.1 Defining constraints in *de novo* design

In all molecular design approaches, it is necessary to ensure the generated results are relevant and useful. This can be effectively achieved by applying constraints to the solutions as they are being generated, with these taking different forms depending on what information is available about the target. For structure-based methods, the candidate molecule is positioned within the active site of the target protein with the

system attempting to maximize the quantity and strength of the favourable interactions that will lead to strong binding ('docking' the molecule). In order to obtain the geometric constraints needed to build molecules to fit a given receptor, it is first necessary to analyse the active site, looking for potentially interesting features that could support hydrogen bonding, or Van der Waals interactions. This can be done by searching for key features of the site in accordance with previous crystallographic studies (a rule-based approach) or by consideration of the energy of the system as a whole. In a rule-based approach, a basic appraisal is made that assigns atoms to bonding and non-bonding roles in accordance with previously obtained 'real world' data. Such an approach can be seen with HSITE (Danziger and Dean, 1989, Lewis and Dean, 1989) which focussed on hydrogen bonding interactions and LUDI (Böhm, 1992).

The alternative approach is to calculate energy hot spots within the virtual site, obtaining a more genuine picture of the energy of the binding site. This can be achieved via the probe atom approach seen in the work reported by Toda et al., (2010), and programs like LEGEND (Nishibata and Itai, 1991), GRID (Goodford, 1985), MCDNLG (Monte Carlo *De Novo* Ligand Generator) (Gehlhaar et al., 1995) and those in the CONCEPTS family (Pearlman and Murcko, 1993). In principle, the programs place an atom at each grid point within the active site, and calculate the binding energy. From aggregating these results, the areas of energetic activity can be determined.

For ligand-based design methods, information about the active site structure is not necessarily provided, so an alternative set of constraints are required. Many of these methods instead focus on the properties of compounds known to have activity, often screening the generated results against some form of pharmacophore model produced from the actives. In these cases, the results are evaluated according to the ability to fit the identified features of the pharmacophore, representing the sites of interaction between the ligand and the receptor. Examples of tools that use these constraints include PhDD (Huang et al., 2010), which mounts individual fragments in key sites according to the pharmacophore, which are then linked to form a molecule. Other tools enhance pharmacophores to build a more complex model for comparison, such as with the pseudoreceptor approach (Fayne, 2013). In these, the common structural features from the pharmacophore are combined with additional steric considerations, based on the three-dimensional conformations of the active compounds to create something

more akin to a traditional active site structure. Results are then scored by evaluating the goodness of fit to this model as seen with PrGen (Zbinden et al., 1998), which uses estimates of binding energy between the generated molecule and the pseudoreceptor.

In circumstances where there is insufficient data to construct a pharmacophore model, direct evaluation of the ligands based on chemical and structural properties can be used as an alternative. For structure evaluation, the most common approach is to calculate the similarity scores for the new products relative to the known active results, with structural similarity implying a similar activity profile. One of the advantages of this method is that the calculations are less complex than three-dimensional modelling, drastically reducing the computational time required. However, 'similarity' between two different molecules is subjective, and as such a number of different methods have been reported to quantify the value, like the Tanimoto coefficient (Section 2.2.2). Alternatively, design constraints can be implemented by restricting the fragments available for structure generation to a limited subset with known properties. Tools such as PRO_LIGAND (Clark et al., 1995) use these limited approaches to empirically score the molecule at each step, based on the ranks associated with the individual fragments.

3.2.2 Structure generation - atoms versus fragments

Common to both ligand-based and receptor-based approaches is the need to construct a molecular structure from scratch using some iterative process. This can be achieved via two routes, either atom-by-atom or by combining known molecular fragments together from an established library.

In the atom-based case, each step in the molecular generation process directly affects one atom of the final structure, through removal, addition or substitution performed on the structure generated previously. This is a relatively slow process, with only small changes made at each step, but provides unlimited scope for exploring the whole solution space. However, the size of the problem rapidly becomes sufficiently large to be unworkable (approximately 10^{100} possible molecules can be constructed that fit the definition of drug-like behaviour) (Walters et al., 1998, Medina-Franco, 2012), and heavy constraints have to be applied to ensure results in a reasonable time frame.

The alternative fragment-based approach, by definition, restricts the solution space to approximately 10^{13} molecules, assuming that an average drug candidate consists of a scaffold and three side chains (based on the estimate of 10,000 realistic scaffolds and the 1,000 known side chains used in drug-like molecules (Walters et al., 1998)). The process helps to reduce the computational cost by presenting fewer molecules for evaluation, but with the step size so large, the possibility of missing the optimal solution to the problem is an issue (Schüller et al., 2008). However, fragment-based systems have additional advantages to the medicinal chemist, as will be explained later.

3.2.3 Structure generation strategies

When considering the construction of the molecule, whether atom-by-atom or fragment-by-fragment, the methods used to connect individual building blocks together fall into three separate categories (Schüller et al., 2008). These are defined as:

- growing
- linking
- lattice-based structure sampling

In this section each method will be considered in turn, highlighting key programs reported, and their method of operation.

3.2.3.1 Growing approach

For the growing approach, a 'seed' atom or molecular fragment is placed within the target site, with motifs from a library added around the seed in order to optimise the interaction between the formed molecule and the target site. This is the logic behind the GenStar (Rotstein and Murcko, 1993a) and GroupBuild (Rotstein and Murcko, 1993b) programs which represent atom- and fragment-based approaches to the same problem.

In the atom-based approaches, as seen in LEGEND (Nishibata and Itai, 1991) and GrowMol (Bohacek and McMartin, 1994), an atom is positioned within the site, usually

in alignment with a three-dimensional grid. The seed atom is placed first and its type and location is chosen to ensure that the seed forms hydrogen bonds with a randomly chosen heteroatom in the target site. Thereafter each new atom is linked to a randomly selected part of the previously generated structure, with the atom type and orientation also randomised.

PRO_LIGAND (Clark et al., 1995) takes a fragment-based approach, using four different libraries with which to build the molecule at the appropriate stages. These can be ranked by the user in accordance with the type of chemistry desired, or alternatively scored empirically by the program, based on summation of individual receptor-ligand energies and bond distances. FOG (Fragment Optimised Growth) (Kutchukian et al., 2009) works in a very similar way, adding a Markov chain training approach, where reference compounds can be used to create an optimised library that is more likely to create drug-like compounds.

One of the main problems with the growing approach is that the growth method is often non-deterministic, and as such different runs will produce different results, potentially missing the optimal solution and indeed, possibly missing key interaction sites altogether due to lack of compatible geometry. In addition, the nature of the scoring and building process forces the progression of the molecule through increasingly energetically favourable areas, precluding access to any superior solutions that may have less stable intermediary stages.

3.2.3.2 Linking approach

In the linking approach (Leach and Kilvington, 1994, Leach and Lewis, 1994), the main interaction sites at the receptor are highlighted, and the molecule generation program focuses on placing structural motifs at these sites, either through pre-docking via another tool, such as with GRID and HOOK (Eisen et al., 1994), or through empirical analysis of the site at build time, as in LUDI (Böhm, 1992). By establishing points of growth at all of the important interaction sites, one of the key problems with the growth approach, that of missing key sites, is negated. In LUDI, the program attempts to connect pre-docked molecular fragments together to give a complete candidate structure that can be synthesised. The difficulty here is ensuring that the linked

molecule remains feasible, synthetically and structurally, as the linking chemistry may not be compatible with the fragment selected. In addition, due to the assumptions made in the virtual model, there is a chance that the produced molecule will have a different conformation than the modelled one, which could have a significant effect on the predicted scores, although this can also be said of growing approaches. Other methods of generating molecules via linking can be found in CONFIRM (Thompson et al., 2008), FOUNDATION (Ho and Marshall, 1993a), SPLICE (Ho and Marshall, 1993b), NEWLEAD (Tschinke and Cohen, 1993), FlexNovo (Degen and Rarey, 2006), PhDD (Huang et al., 2010) and the work of White and Wilson (2010).

SPROUT (Gillet et al., 1993), uses a combination of both the growing and linking methods to produce molecular candidates. Initially, HIPPO (Gillet et al., 1995) is run to analyse the potential receptor site. HIPPO follows rules regarding hydrogen positioning, the location of hydrogen bond acceptors and donors and any potential covalent bonding sites in order to identify key interaction sites. The rule set is based on literature results, and statistically validated by comparison with structures from the Protein Databank. SPROUT then proceeds to build initial skeletal structures using a three-dimensional subgraph methodology (Mata et al., 1995), where the edges of the graph represent the bonds and the vertices represent a generic atom type making one graph equivalent to a number of potential structures. Partial structures or fragments are positioned at the various target sites, and grown outwards until they can be joined together to form one structure - this is achieved by overlaying a template common to both on the region in question. When all fragments have been joined the atom types are also manipulated at this point to make the molecules into realistic structures that complement the binding site features.

One further approach used by some two-dimensional tools is to take known active compounds and rearrange their fragments in different manners or with different geometries to create a new lead. This approach can be found in BREED (Pierce et al., 2004) which uses ligands known to bind to a particular target site as its fragment source and in FLUX (Fechner and Schneider, 2005, Fechner and Schneider, 2007) which connects entities from a library of likely candidate fragments using randomly assigned linking groups. The BIBuilder method (Teodoro and Muegge, 2011), uses a BREED-like algorithm with a library of fragments generated via RECAP retrosynthesis of a relevant

drug library. The RECAP (Retrosynthetic Combinatorial Analysis Procedure) (Lewell et al., 1998) rules were originally designed as an attempt to heuristically retrosynthesise molecules to give stable fragments as potential starting points for new lead compounds. BIBuilder works by first fragmenting a set of molecules according to the 11 defined bond types in RECAP. The user then defines design constraints (either a known receptor for a structure-based design, or suitable ligands for a ligand-based design). New lead compounds are created by linking the generated fragments in all possible ways. Other similar approaches, such as that by Foscato et al. (2014) prevent the combinatorial explosion by filtering the produced fragments, and assessing any potential linkages for chemical compatibility prior to generating the products. Further use of the RECAP principles can be seen in the feasibility checks implemented in programs like ARChem/Route Designer (Law et al., 2009), which uses retrosynthetic rules generated automatically from reaction databases.

3.2.3.3 Lattice-based structure sampling

In lattice-based structure sampling, a lattice of carbon atoms mixed with other random atom types, or indeed molecular fragments from a selected library, is constructed in the active site. Structures are built by generating bonds along the shortest path between lattice atoms that bridge the points of interaction. This approach can be seen in BUILDER (Lewis et al., 1992). To use the program, DOCK (Kuntz et al., 1982, DesJarlais et al., 1988) must first be run. This scans a library of compounds in search of molecules or fragments that will most appropriately fit the target site, subsequently generating lattices with the selected entities placed within them. The user must then select from the generated lattices any particular regions of interest. From this, BUILDER selects a start point, and the various molecules are linked in sequence, using bridging elements again selected from a list of suitable candidates. This leads to a more interactive design tool than most programs, relying more heavily on user input. However, the automated bridging logic proves limited, both in terms of synthetic feasibility and compound diversity. These problems were largely remedied in BUILDER v.2 (Roe and Kuntz, 1995), which uses heuristic rule sets applied to the search strategies which serve to decrease needless complexity found in the molecular linkages generated by the previous tool. PRO_SELECT (Systematic Elaboration of Libraries Enhanced by Computational Techniques) (Murray et al., 1997) also uses the lattice framework

approach, adding it to the PRO_LIGAND software previously discussed, in order to enhance operation. Further information about these and similar methods can be found in the review by Cavasotto and Phatak (2011).

3.2.3.4 Scoring methods

A major issue in *de novo* design is the need to evaluate the proposed molecules after generation. One approach in structure-based *de novo* design is to first build structures using a two-dimensional method and then generate the three-dimensional structures and use docking methods to calculate the free energies of association for use as a ranking criterion. For this approach, most of the established docking methods such as AutoDock Vina (Trott and Olson, 2010) can be adapted for this, providing that suitable target information is provided for the fitting. It does not necessarily follow, however, that all functions that predict the ideal binding geometry for a molecule will be suitable for comparing different ligands. In particular, the additional computational time required to dock all potential solutions makes the process inefficient without some pre-screening approach. When these are factored in, the use of an additional bespoke scoring function to assist in selection has benefits over the usual execution time penalties associated with adding an additional process to the setup. In all cases, when looking at potential geometries for docking, it is necessary to consider the number of degrees of freedom possessed by both the site and the candidate. As more realistic treatments of conformation flexibility are added, so the complexity of the calculations increases, resulting in a trade-off between absolute accuracy and computation time (Dias and Filgueira de Azevedo Jr., 2008). As all potential candidate molecules must be screened in a number of geometries, the amount of time taken to process each molecule is a significant factor in determining the maximum number of molecules that can be evaluated.

Where molecules are generated directly in the binding site, some evaluation is required of the partial solutions to ensure that that the synthesis proceeds in the right direction, with the minimum of inactive or unsuitable results. For iterative methods, this is relatively straightforward in that the result population can be ranked at the end of each iteration, with the best performing molecules used as the parents for the next step. Any individual property of the molecule that can be expressed as a relative score can be

used in this manner, but many bespoke algorithms have been reported that combine these evaluations with other screening functions, such as TOPAS (TOPology-Assigning System) (Schneider et al., 2000). Each structure produced is scored for fitness by comparison with the target compound, using topological pharmacophores and the Tanimoto coefficient. If the results converge to an optimum (i.e. no structural changes occur over multiple generations) the process is automatically stopped; otherwise, the iterations continue using the best performing molecules as the new parents until the maximum number of iterations is reached. Approaches like TOPAS and other iterative optimisation techniques are particularly effective where large numbers of molecules could potentially be generated, as they help to cut down the potential time wasted pursuing undesirable solutions.

3.2.4 Searching strategies

One significant issue with *de novo* tools relates to the nature of the exploration of the chemical space. As previously mentioned, the number of individual molecules available to the *de novo* design tool is so large that to attempt to explore chemical space in its entirety would be completely infeasible. It is therefore necessary to find a means to sample the space to give usable results within the research timeframe. Traditionally, these search methods have included breadth- and depth-first algorithms as well as various stochastic sampling routines.

3.2.4.1 Breadth- and depth-first searching

The breadth- and depth-first searching regimes differ in the degree of storage of the search results at each individual step. When dealing with a limited search space, the two techniques have their own advantages. In breadth-first searching, all the partial solutions reported by the program are scored, with a subset taken on to pursue further, leading to a large number of simultaneous search processes. As a consequence, as the number of potential paths increases, the timeframe for the search also increases dramatically. This makes breadth-first searching more appropriate to methods that provide limited diversity within construction such as LUDI (Böhm, 1992). When using depth-first searching, on the other hand, the number of paths pursued at one time is

limited to one i.e. one partial solution at each level is retained until the end result is reached, getting to a result relatively quickly.

In both cases, the limitations placed on the searching methods may ultimately make the searching process more efficient, but the use of such techniques may result in missing the optimal candidate. The main problem is that when choosing which result to retain at a given point, the tendency is to go for the best scoring molecule at that point, which may favour an overall mediocre solution over a poor scoring intermediate that may rapidly evolve into an ideal candidate. This can be improved by utilising backtracking algorithms such as those in SPROUT to review alternative solutions should issues arise.

3.2.4.2 Stochastic sampling and searching

There are a number of different stochastic approaches to sample the space, from Monte Carlo approaches to those more closely related to genetic algorithms, and genetic algorithms themselves.

In the Monte Carlo method the solution space is sampled, with individual movements within the space occurring at random. As this unfocussed searching is computationally very expensive, the Metropolis criterion is often applied to act as a filter.

$$P = \min(1, e^{\frac{-\Delta score}{T}})$$

Equation 3.1: The Monte Carlo Metropolis Criterion.

If the movement from one molecule to the next results in an improvement in the scoring function, the result is accepted and the next modification is generated. However, if the modification results in a reduction in the scoring function, the probability that the change will be accepted is calculated in accordance with the equation above, where P is the probability of acceptance, T represents the absolute temperature of the system in Kelvin (a reference to the system entropy seen in simulated annealing processes) and Δ_{score} represents the change in the score as the result of modification. Such Monte Carlo searching has been employed in CONCEPTS/CONCERTS (Creation Of Novel Compounds By Evaluation Of

Particles/Residues at Target Sites) (Pearlman and Murcko, 1993, Pearlman and Murcko, 1996), SkelGen (Todorov and Dean, 1998) (Lloyd et al., 2003), DycoBlock (Liu et al., 1999) and SMoG (Small Molecule Growth) (DeWitte and Shakhnovich, 1996). In these cases an estimate of the free energy of the interactions is used as the scoring, either directly in the case of CONCEPTS/CONCERTS, or as part of a more knowledge-based approach as in SMoG, or the Monte Carlo *De Novo* Ligand Generator or MCDNLG (Gehlhaar et al., 1995). In all these cases, by fine tuning the 'T' parameter during the run, initial wide variations can be made to hone in on a particular molecule (referred to as 'annealing' the system). This ultimately results in one candidate being generated that scores highly, but different runs can result in differing solutions due to the random nature of the movements through chemical space.

3.2.4.3 Genetic algorithms (GAs)

Genetic algorithms work on populations of potential molecular candidates, utilising the Darwinian principles of natural selection and survival of the fittest, attempting to mimic the mutation and crossover operations present in reproduction (Back et al., 1997). The algorithms treat the individual bits within a data string, or the individual atoms in a molecule structure, as chromosomes to be manipulated in a number of different ways. Usually, one of the encoded operators is selected at random and applied to one or more chromosomes according to the rules associated with it, creating new results to be evaluated. Once the results are scored, the process repeats until the goal is reached, or the maximum number of iterations is reached. In the vast majority of cases, two distinct operators are used:

- *Mutation* – An individual component of the representation (an atom, bond or even a whole functional group) is randomly replaced with another or deleted to generate a new molecule.
- *Crossover* – Two randomly chosen molecules (the 'parents') are taken, and sections of their data exchanged at a convenient linkage point, resulting in the addition of the new combinations to the population.

The 'randomness' of these functions and operators is key to avoiding the problem of local optima, where a series of operations can find a less than ideal solution due to an

inability to adequately explore the sample space. By adding a random element, other solutions are at least considered, reducing the likelihood of this problem occurring. The degree of randomness in the system is usually influenced in some manner, such as in the 'weighted roulette wheel' selection method, where groups or molecules that score highly have a higher chance of being selected than lower scoring alternatives. This serves to guide the searching towards optimal solutions, while still retaining an element of diversity with the random element.

The Chemical Genesis program (Glen and Payne, 1995) applies these principles starting with an initial seed molecule input as a SMILES string, or other two- or three-dimensional structure. In all cases, the GA operates on a 3D conformation that has been generated from the molecule, rather than a traditional 2D chromosome. As a consequence, the standard GA operations are enhanced to include translation and rotation of either the molecular structure or an individual bond, as well as enhanced mutation operations that change atom types and add methylene or ring substituents. At each step, parent molecules are selected and a set number of the GA operations are applied to make new hybrids, with two separate lists of results for the mutation and crossover operations. These structures are then optimised using molecular mechanics and scored according to a number of criteria. These include restrictions on the volume of the manipulated structure, as well as the usual collection of molecular properties derived from the target.

MEGA (Multiobjective Evolutionary Graph Algorithm) (Nicolaou et al., 2009) uses molecular graphs for its molecular representations, with the start point defined as a particular structure input by the user or assembled at runtime from a library of suitable fragments. This start point then has the genetic operators applied to generate a population of suitable candidates that are converted into three-dimensional representations and docked into a binding site. The molecules are scored according to their interactions with the binding site and by comparisons with the Lipinski rules (Lipinski et al., 1997) to ensure they are drug-like. These scoring functions represent the multiple objectives used to guide the generation of the molecules. This same approach is seen with the SENECA program (Han and Steinbeck, 2004) and the Pareto Ligand Designer (Ekins et al., 2010), which uses a reference molecule set to identify objective values to build Pareto fronts. The Pareto efficiency approach refers to the

allocation of resources (properties and scores) in such a manner that it is impossible to improve one individual without disadvantaging another in the distribution. The Pareto front for a system is one for which the set of property allocations for a result set all meet the requirements for Pareto efficiency (these results are referred to as non-dominated results). Any solutions that meet this criteria are stored are used for the next molecule evolution step via a number of transformation rules, including those from Drug Guru (Stewart et al., 2006), until the optimisation criteria are met. Drug Guru incorporates 186 rules in the form of SMIRKS strings, as shown in Figure 2.9 in Section 2.3.2, favouring functional group transformations and a number of ring structure modifications.

CoG (Compound Generator) (Brown et al., 2004) and the Globus method (Globus et al., 1999) use a typical GA approach, with the ability to backtrack along synthetic routes. The development of each molecule is represented as an individual subgraph in a collated synthesis tree, using standard genetic operations used on the nodes, and Tanimoto similarity scoring. GANDI (Dey and Caflisch, 2008) uses a GA method for fragments, but without the use of multiple mutation operators. Instead, fragments are prearranged and docked within the active site with a number of suitable linkers chosen at random and evaluated via either two-dimensional similarity coefficients or a three-dimensional overlap function. In order to prevent a synthesis 'loop' occurring with constant re-evaluation of the same complexes a *tabu* search (as discussed by Rusu and Bulacovschi (2006), after the work by Glover) is used, where knowledge of the previous potential solutions is stored.

LeapFrog (Tripos, Kharkar et al., 2009) does not employ a genetic algorithm in the strictest sense, but it does use the standard genetic operators for its molecular constructions. Configuration of the program allows it to operate in one of three modes: suggesting improvements to a given structure (Optimise); creating a bespoke molecule (Dream); and providing a more interactive stepwise design process (Guide). EA-Inventor (Tripos) uses a more conventional evolutionary algorithm, making all operators tuneable to increase or decrease the likelihood of them being used (Feher et al., 2008).

An alternative approach to GA-like expansion from a fixed scaffold is the BOMB (Biochemical and Organic Model Builder) method (Jorgensen et al., 2006) in which a seed structure is modified extensively with different side chains in an iterative fashion to form a product. The process involves a library of over 100 cores and 600 substituents that can be overlaid to the structure to build ligands which are then optimised and scored. While not implementing all of the GA operations, the creation and optimisation methods are similar to standard mutations. LigBuilder (Wang et al., 2000, Yuan et al., 2011) works in a similar manner, but using an elitist approach to ensure that each successive generation does not result in a backward step. The program uses a GA approach, implementing the growing and linking strategies as described in Section 3.2.3, with a fixed proportion of the highest scoring results (based on the free energy of their binding with the active site) copied from one generation to the next.

In terms of exclusively two-dimensional approaches, the Nachbar method of molecule evolution (Nachbar, 2000) uses a genetic programming tree structure, as opposed to linear chains when describing molecules. The initial population is created randomly, selecting one atom and adding new bonds until its valence is complete. New atoms are then added as appropriate, with new bonds added and so forth, branching out until terminated by either the random selection of a terminal atom (hydrogen for example), or maximum depth for the system is reached. This is repeated for different start points until a population is generated. The members of the population are then scored by similarity comparisons with a reference compound. Optimisation then proceeds through the usual mutation and crossover parameters.

Other, similar methods include TOPAS (TOPology-Assigning System) (Schneider et al., 2000), as discussed in Section 3.2.3 and 3.2.4, and ADAPT (Pegg et al., 2001). The latter method combines a genetic algorithm approach with a fragment-based linking method, using DOCK 4.0 to evaluate fitness.

3.2.5 Particle swarm optimisation

Particle swarm optimisation (PSO) was first proposed by Kennedy and Eberhart (1995) as a means of simulating social behaviour *en masse*, and is best described as a subset of

evolutionary algorithms. In it, each individual candidate solution can be described as a particle within the swarm. The particles effectively move through the search space independently, but are guided by their own previous results, and the best results of the swarm as a whole. In this way, the system quickly converges to a solution.

PSO has been utilised for *de novo* design in COLIBREE (COmbinatorial Library BREEding) (Hartenfeller et al., 2008). A starting molecule is selected which will serve as the basis for each individual particle in the swarm. Each of these particles can access the library of fragments and linkers that can be combined to generate a potential new molecule for evaluation, which is stored within the structure of the particle. A score is allocated to each fragment and linker (known as a quality vector or QV) representing the suitability of the unit for the scenario in question and the likelihood of selection at a given point, in order to allow the swarm to operate. In each iteration, all the particles are informed of the highest scoring solution, which is used alongside the individual results to select the next operation. The operation is chosen via roulette wheel sampling, with the additional ability to edit the scores externally to force a particular route to be covered. A similar variant, Ant Colony Optimisation is used in tools such as MAntA (Molecular Ant Algorithm) (Reutlinger et al., 2014). In MAntA, as each fragment and linker is evaluated, a score is assigned in the form of a 'pheromone concentration', which gradually diminishes over time. Much like a real ant colony, the shorter, superior routes to a target are more likely to be covered repeatedly, which increases the concentration, creating a weighted sampling method that converges rapidly towards a solution.

3.3 Synthetic feasibility in *de novo* design

One of the main issues in *de novo* design is ensuring the synthetic accessibility of the suggested compounds. Two approaches have been developed to tackle this problem. The first is to score compounds on ease of synthesis once they have been generated (allowing the *de novo* tool to operate without restricting the sampling space) while the other is to base the structural transformation operations embedded within the *de novo* design tools on known reactions, restricting the sample space but ensuring, at least in theory, that a synthetic route remains available for any solution generated.

3.3.1 Feasibility scoring functions

Many of the *de novo* tools produced post-1995 look to implement some form of synthetic feasibility check within their scoring, imposing penalties on structures that have too many of a particular group, for example. RASSE (RAtional Space SEarching) (Luo et al., 1996) and TOPAS (Schneider et al., 2000) score molecules on the structural features present that affect synthesis (favouring esters and amides over enols and peroxides, for example, and avoiding overly large atom counts) as well as their binding affinity and other similar parameters in order to ensure suitable results. In the case of RASSE, a standard ligand-based scoring approach is used, but penalties are applied for every instance of a chemically unstable functional group or excessive use of asymmetric structural elements prior to tabulation of the results.

Another approach is to leave the scoring functionality untouched, and instead perform a more detailed, specific evaluation on the final candidates as seen in the SYLVIA structure evaluation method (Boda et al., 2007, Molecular Networks GmbH). As well as ring complexity and chirality, this method also takes into account topological and atom type features, alongside a retrosynthetic analysis of the molecule in question to identify simple, readily available precursors. Allu and Oprea (2005) take a similar approach, with the synthetic and molecular complexity (SMCM) scoring system intended for use with existing *de novo* tools. Firstly, each atom is assigned a relative electronegativity value according to empirical data, and then every bond is identified and assigned a parameter value. Next, the molecule is assessed for features that are known to be more complicated to synthesise, such as chiral centres or complex ring systems, and an additional penalty score calculated. The final score is the sum of these individual values and represents the ease of molecule generation – the higher the score, the harder the molecule is to synthesise. CAESA (Computer Assisted Estimation of Synthetic Accessibility) (Gillet et al., 1995), analyses a given structure to determine the likelihood of there being suitable starting materials. Any potential issues regarding stereochemical complexity, topological details or specific functional groups that may hinder a synthesis are highlighted, and then the candidates are ranked in order of synthetic ease. This effectively acts in the same manner as a qualified synthetic chemist would when facing the same problem, but on a larger scale.

3.3.2 Reaction-based *de novo* design

Another approach to the synthetic feasibility problem is to limit the range of transformations to known reactions. For these purposes, the Daylight SMIRKS (Daylight Chemical Information Systems) language is often used to encode the transformations (effectively structural 'difference lists'), due to its ability to characterise the reaction centre directly. Some of the earliest examples of this process in *de novo* design are the previously mentioned TOPAS (Schneider et al., 2000) and FLUX (Fechner and Schneider, 2005). In these cases, the synthetic restrictions come from a master set of 11 reaction transformations (the RECAP rules (Section 2.4.5), encoded as SMIRKS), and a fixed fragment library derived from retrosynthesis of the contents of the World Drug Index in the case of TOPAS or the COBRA library (Collection Of Bioactive Reference Analogues, a literature collection of 4,236 molecules with known structures, activities and bioavailability information) (Schneider and Schneider, 2003) in the case of FLUX. Deliberate restriction of reaction transformations to hard-coded libraries can also be useful for producing multiple candidates that follow a similar synthetic route. This is of particular benefit when attempting to build focussed arrays for high throughput screening based around particular *in silico* properties or existing results. The vProtocol method (Schürer et al., 2005) derives its rule set from a collection of synthetic literature schemes filtered for compatibility, utilised as part of a genetic algorithm based *de novo* tool. The program generates a series of products together with the reaction sequences used to create them, with the GA used to optimise both elements. One advantage of the approach is that particularly effective short sequences found by vProtocol can be added to the rule set for future use.

SYNOPSIS (SYNthesize and OPtimize System *in Silico*) (Vinkers et al., 2003) starts with a database of existing molecules, such as those available within the Available Chemicals Directory (BIOVIA) one of which is selected in accordance with a Monte Carlo function (Section 3.2.5.2). The molecule is then analysed for appropriate functional groups by query matching with the library of 70 manually coded reaction transformations. Each transformation included within the database is chosen to be suitable for a wide range of reactants, while the query elements of the SMIRKS string ensure that reactions are not applied where the structure or competing functionality would prevent their use in real life. Of those that are suitable, one is selected at random and applied to the

molecule, to generate a new molecule which is scored. The Monte Carlo function then selects another molecule, gradually annealing the system so that processor time is devoted towards improving the quality of the solutions rather than attempting to enhance the population as a whole. This simulated annealing approach was extended to multiple objectives by MOLig, which takes the Monte Carlo method to optimise around internal energy, energy of interaction, bioavailability and similarity to a reference compound.

DOGS (Design Of Genuine Structures) (Hartenfeller et al., 2012) uses a similar approach, but with the transformations encoded in a bespoke language, Reaction-MQL (Reisen et al., 2008). This depicts the individual bonds and electrons in a linear format, better suited to storage within SQL data tables or other similar storage solutions. The DOGS reaction library is designed to be sufficiently generic to be applied to a wide range of reactants. A molecular fragment library is then processed to assess the reactivity of each fragment towards each of the list reactions, and the reaction centres in each case, flagging the entries accordingly. The structure generation process is then a case of selecting a reaction from the database, and applying it to the starting molecule at the identified reaction centres. For optimisation purposes, an initial pilot study is performed with one reaction from each class in the library (the one predicted to be the most effective), with the collected results evaluated. The classes that produced the best results are then investigated in more detail; with all of the reaction they contain being used to generate compounds. If the starting material requires it, two component reactions can be processed by applying transformations to all of the suitable structural features, but only the highest scoring combinations are retained for processing due to the potential population size problems.

As an alternative to the pre-encoded systems, reactions can be directly represented as difference values between reactant and product environments as in the reaction vector approach (Broughton et al., 2003, Patel et al., 2009), (Section 2.4.1). A ‘reaction vector’ is a set of atom pair descriptors representing two and three bond distances, indicating the differences between the product and the reactant i.e. which atoms have changed. By representing reactions in this manner, transformations are reduced to the reaction centre and immediate environment, enabling any set of reactions to be encoded without reference to predefined transformation rules or atom maps, making the overall

input significantly more straightforward. It should be noted that, because the atom pair data represents only the immediate reaction environment, the same issues that affect all similar reaction centre methods are present here. For example, any functional groups that would cause reaction incompatibility in a real world synthesis (due to steric or electronic effects) will have their effect ignored if they are not directly connected to the perceived reaction centre. Careful selection of the size of the reaction centre can mitigate this effect, but can present its own problems with respect to novelty. In general, a compromise is required as the greater the number of bonds recorded as part of the reaction centre, the narrower the scope of application of the reaction to novel starting materials.

In the reaction vector *de novo* design tool, an initial database of reactions is converted to the vector representation to be stored internally and recalled as necessary. A starting molecule is input, and evaluated against the reaction vectors to determine which reactions are possible for that structure. One of these is selected at random, at which point the reaction is applied *in silico* using an algorithm that is described later in the thesis. This process is then repeated with another copy of the starting molecule to generate a population of solutions. At this point, a weighted roulette wheel sampling is used to pick a candidate for the next phase and so on for a set number of iterations. As this method is the foundation of the work presented in this thesis, a more detailed discussion of this method is included in Chapter 4.

3.4 Drug-likeness in *de novo* design

Much like the assessment of synthetic feasibility, there are two main approaches to ensuring that the results produced by *de novo* tools are useful drug candidates. These are *post hoc* evaluation of a given set of results based on a given set of criteria, or via restriction of the used transformations and structural elements to those known to be commonly used in drug design.

3.4.1 Rule-based drug-likeness evaluation

One of the most common approaches used is to filter any prospective results against a simple set of rules devised by Lipinski (Lipinski et al., 1997, Lipinski, 2004). These empirical rules are based on the principle that the majority of orally administered

drugs are lipophilic, and relatively low in molecular mass. The rules use the following criteria:

- molecular mass below 500 Da
- log P (octanol-water partition coefficient) value below 5
- no more than 5 hydrogen bond donors (N—H and O—H bonds)
- no more than 10 hydrogen bond acceptors (nitrogen and oxygen atoms)

where an orally available molecule breaks no more than two of these rules. As all of these values are multiples of five, these rules are often referred to as the rule of five, despite the original list only having four criteria. One of the key advantages to such an approach is that the properties are easily calculated from a two-dimensional structure, with software libraries such as Marvin (ChemAxon) and MOE (Chemical Computing Group Inc., 2015) able to calculate these in order to assist the evaluation of results after generation. As a result such evaluations are commonly used as a simple addition to existing *de novo* methods requiring little additional effort to implement. However, tools such as PHDD directly incorporate the rule of five into the design workflow, with non-drug-like compounds removed from the result list prior to calculation of the fitness values.

Further studies into bioavailability have led to additional rules being developed to improve the quality of the predictions (Ghose et al., 1999). These include adjustments to the molecular mass and log P ranges (180 – 500 Da and -0.4 to 5.6 respectively), as well as requiring an atom count for the molecule between 20 and 70. A related rule system for identifying good lead compounds also exists, the rule of three (Congreve et al., 2003). This suggests that leads should have log P values below 3.0, molecular masses below 300 Da and no more than three hydrogen bond donors and acceptors, as well as limiting structures to no more than three rotatable bonds.

3.4.2 Transformation-based drug-likeness evaluation

An alternative method of ensuring drug-likeness is to restrict the available transformations and fragments for *de novo* structure generation. Many of these methods have already been covered in Section 3.3.2, as the libraries and rule

definitions used in tools such as DOGS, FLUX and TOPAS to determine synthetic feasibility are derived from collections of compounds known to be bioactive with the aim of increasing the likelihood of drug-like results. The BOMB tool discussed in Section 3.2.4.3 also uses this approach, with the library of cores and sidechains preselected to ensure compatibility and bioavailability.

3.5 Conclusions

This chapter has looked at how *de novo* design tools have progressed, and the various methodologies by which they operate. Despite the initial interest being in structure-based *de novo* design, it is interesting to note that in recent years, programs have tended to focus on ligand-based design and on structure optimisation. This could be because drug discovery chemists are seeking to either improve on known entities from other companies, or their own assay screening results without necessarily wanting to design compounds from scratch. Also, two-dimensional comparisons take considerably less computation time than three-dimensional methods, making them a more efficient screening method. In any case, comparing potential candidates to a reference compound via three- and two-dimensional metrics seems to be the underlying scoring method for these programs.

It has been suggested that the lack of uptake of traditional *de novo* design tools is largely due to the failure to consider how the candidates can be synthesised. Many of the more recent *de novo* tools do consider the feasibility of synthesising the output by reference to reaction databases, either commercial or bespoke, or through some sort of review process of the final candidate molecules. One of the main advantages of the reaction vector method is that this synthetic accessibility is an integral part of the tool, due to the nature of the data collection.

Chapter 4:
Reaction Vectors

4.1 Introduction

Previous work carried out within the research group (Hristozov et al., 2011, Patel et al., 2009, Patel et al., 2008) resulted in the creation of a knowledge-based *de novo* design tool based on reaction vectors as introduced in Chapter 3. As this project enhances and extends this work, it is necessary to summarise how the tool works, and how it may be used to generate structures.

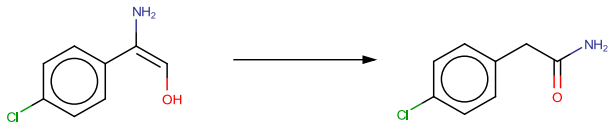
4.2 The Reaction Vector (RV) format

As previously discussed in Section 1.2 and Section 3.3.2, a reaction vector stores the structural changes that occur over the course of a chemical reaction as a difference vector. In a *de novo* design context a reaction vector can be applied to different starting materials to generate novel structures. In order to deliver the best compromise between applicability, ease of calculation and simplicity, the changes are encoded in the form of atom pair descriptors (Carhart et al., 1985). In this project, two different forms of atom pair descriptors are used, which are referred to as AP2 and AP3, with the number in the descriptor representing the length of the path involved. The AP2 descriptor refers to two atoms that are directly bonded, whereas the AP3 descriptor refers to a pair of atoms separated by two bonds. This means that AP2 descriptors encode bond information, with the AP3 descriptors providing information about the environment of the bond, specifically one bond away from the changed bond. The descriptors take the form X1(h,p,r)-S(o)-X2(h,p,r), where:

- X1 and X2 are the atoms in question
- h represents the number of non-hydrogen connections to the atom
- p is the number of Π bonds the atom contributes to
- r is the number of rings the atom is part of
- S is the path separation (i.e. whether this is an AP2 or AP3)
- o is the bond order of the connection (only relevant for AP2)

The ‘p’ parameter is calculated as follows. Initially, p is set to zero for each atom in question, before each bond for the atom is analysed in turn; if the bond is aromatic or double, p is incremented by one; if the bond is triple, p is incremented by two. As a result of this calculation, a Kekulé representation of an aromatic structure will give a different atom pair descriptor to a delocalised representation. As a Kekulé structure uses alternating single and double bonds to represent aromaticity, any atom in the ring is perceived as having at least one double bond and one single bond and therefore p will be incremented by one (assuming no other bonds), whereas in a delocalised system both ring bonds incident to an atom will be assigned as aromatic and therefore p will be incremented by two (again assuming no other bonds). In order to prevent problems with these mismatching definitions, a standardisation step is used to convert all structures to the delocalised aromatic representation prior to analysis. Additionally, any explicit hydrogen atoms are also removed from the molecule for consistency. The bond order parameter ‘o’ is assigned as follows. A single bond is 1; a double bond is 2; a triple bond is 3; and an aromatic bond is 4.

To generate a reaction vector, the atom descriptors are calculated for both sides of the reaction, with the list associated with the ‘reactant’ side subtracted from the list from the ‘product’ side to give an indication of the transformation itself. The result of this process is a set of negative atom pairs that represent atoms and bonds that are lost as part of the transformation and a set of positive atom pairs that represent atoms and bonds that are gained. An example of this process for a sample rearrangement reaction is presented in Figure 4.1, with the negative atom pairs (AP2s and AP3s) listed on the left in red, and the positive atom pairs listed on the right in green. The AP2 descriptor associated with the C—N bond that moves as part of the rearrangement does not appear in the reaction vector since there are C—N bonds in both the reactant and the product which cancel out when the difference is calculated. However, the reaction vector contains negative and positive AP3 descriptors that include the N, and these encode the changing environment of the C—N bond. It should be noted that the reaction vector code handles the reaction as shown with the reactant and product as separate entities; it has no knowledge of tautomerism or mesomerism and as such this reaction is processed as entered.



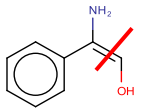
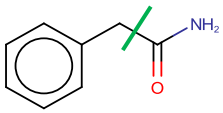
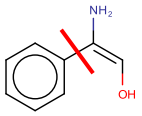
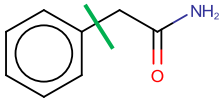
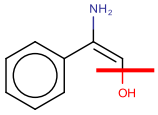
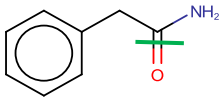
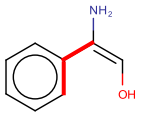
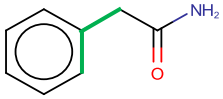
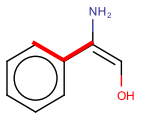
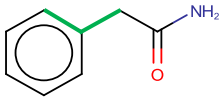
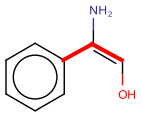
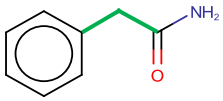
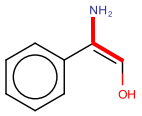
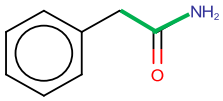
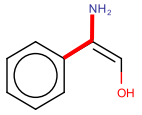
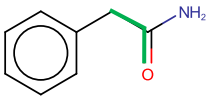
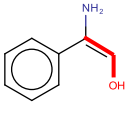
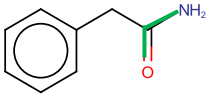
Atom Pairs removed during the reaction	Illustration of bonds removed (red line indicates removed bonds)	Atom Pairs added during the reaction	Illustration of bonds added (green line indicates new bonds)
C(3,1,0)-2(2)-C(2,1,0)		C(3,1,0)-2(1)-C(2,0,0)	
C(3,2,1)-2(1)-C(3,1,0)		C(3,2,1)-2(1)-C(2,0,0)	
O(1,0,0)-2(1)-C(2,1,0)		O(1,1,0)-2(2)-C(3,1,0)	
C(3,1,0)-3-C(2,2,1)		C(2,2,1)-3-C(2,0,0)	
C(3,1,0)-3-C(2,2,1)		C(2,2,1)-3-C(2,0,0)	
C(3,2,1)-3-C(2,1,0)		C(3,2,1)- 3-C(3,1,0)	
N(1,0,0)-3-C(2,1,0)		N(1,0,0)-3-C(2,0,0)	
N(1,0,0)-3-C(3,2,1)		O(1,1,0) -3-C(2,0,0)	
O(1,0,0)-3-C(3,1,0)		O(1,1,0)-3-N(1,0,0)	

Figure 4.1: Example of the generation of a reaction vector for a rearrangement reaction. (Wallace, 2015)

4.3 Structure generation using RVs

4.3.1 Original method

The first approach to generating and applying RVs in *de novo* design was developed by Patel (Patel et al., 2009). While the generation of RVs is relatively simple, the application of the vectors to generate new products is more complex. In the method developed by Patel this is done in an atom-by-atom and bond-by-bond approach, pursuing each possible solution in a breadth first search until all possibilities have been exhausted. First, the starting material is fragmented by removing bonds recorded in the reaction vector as being 'lost' (the negative atom pairs). Then, atom pairs are selected one at a time from the list of items to be 'gained' (the positive atom pairs), and added to the fragments from the starting material. This is done in all possible ways, starting from a seed atom (the highest numbered atom with an unsatisfied valence). All of the positive AP2s within the vector are analysed, and any that contain an atom descriptor matching that of the seed atom are used to grow the fragment in turn, with the positive AP3 descriptors used to validate the extended fragment and verify that it is consistent with the reaction vector. Should any of the extended fragments contain AP3s that are not present in the reaction vector, they are considered incorrect and eliminated. The search then moves to consider each valid extended fragment in turn. For each extended fragment, the highest numbered atom with an unsatisfied atom valence becomes the next seed atom and any remaining AP2s are used to grow the fragment as before. The process continues until the entire structure is assembled or no possible solution can be found.

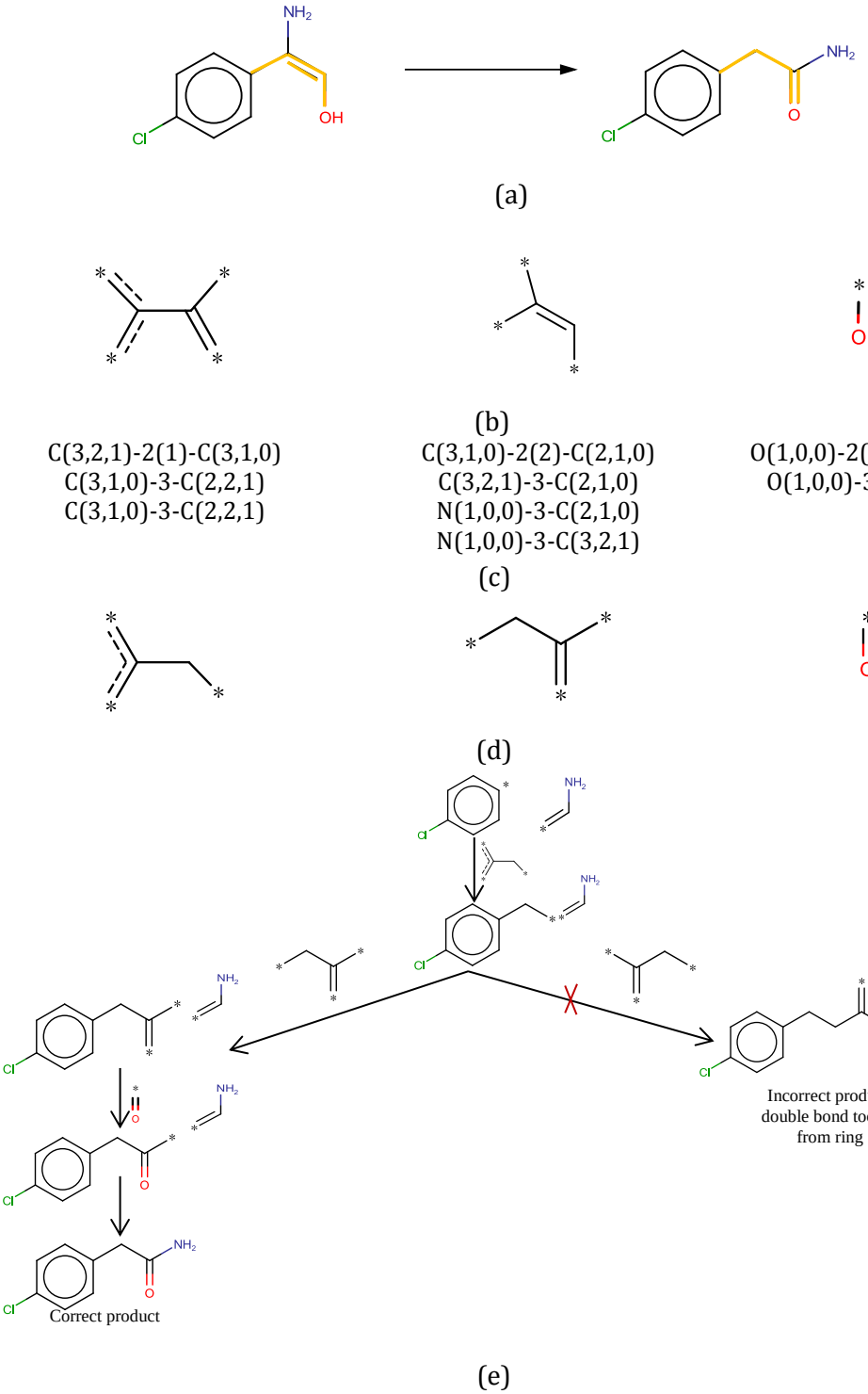


Figure 4.2: The structure generation procedure using the reaction vector method. (a) The reaction in question (affected bonds in orange). (b) The bonds to be removed from the reactant to form the fragment (* indicates attachment point). (c) The negative AP2 and AP3 descriptors (the 'negative' pairs from Figure 4.1). (d) An unordered set of bonds to be added to the fragment to form the product. (e) The structure generation procedure – fragments are assembled bond-by-bond. (Wallace, 2015)

An illustration of the structure generation process is given in Figure 4.2, whereby the RV is applied to the starting material of the reaction from which it was generated and the known product is generated. Initially, the bonds lost during the reaction are removed from the starting material to leave a collection of fragments, in this case the chlorobenzene ring, and a nitrogen atom connected to an sp2 carbon atom. The seed atom in the starting material fragments is then identified as the highest numbered atom with an unsatisified valence (the unsubstituted atom in chlorobenzene in this case). Each applicable positive AP2 is attached in turn, to build the structure bond-by-bond. However, as in general there may be multiple ways in which a given AP2 can be attached, as well as multiple AP2s, additional verification is achieved through use of the positive AP3s. In the first step, only one AP2 is applicable, as only one fragment is compatible with the attachment point on the chlorobenzene ring, and there is only one way in which it can be attached. The bond is added to the ring to extend the chlorobenzene fragment. It is not possible to attach the nitrogen atom at this stage since it is incompatible with the new unsatisfied valence which is an sp3 carbon. The search now moves to consider this extended fragment and a new seed atom is identified. The remaining AP2s are examined; only one is applicable but this can be added in two different orientations and so two extended fragments are generated. Comparing the AP3 data for the possible result structures with the positive AP3s results in the structure on the right hand branch being eliminated. Conversely, the AP3 data for the fragment on the left hand branch is consistent with the positive AP3s verifying the structure as correct. In the next step, the AP2 representing carbonyl oxygen is added and finally the two fragments are joined.

The method was tested by Patel et al., with a variety of different reaction types including epoxide reduction and formation, amide reduction and Diels-Alder reactions. The method was demonstrated to be effective for R → P, R1 + R2 → P, R → P1 + P2 and R1 + R2 → P1 + P2 reactions, although these latter two categories are more prone to failure. Overall, 85% of the RVs tested reproduced the correct product given the original starting material. This breadth-first method works well for simple cases, but due to the exhaustive approach needed to ensure the correct product is built, more complex reactions with higher numbers of atom pairs can lead to incredibly slow structure generation times. Indeed, for many large molecules, even a maximum

execution time of 60 seconds per RV is not sufficient on a 256 core HPC cluster to recreate the desired structure. As a consequence, revisions were made to permit faster application of the vectors.

4.3.2 Revised RV generation and storage (reverse fragmentation)

Hristozov et al. subsequently increased the speed and success rate of RV application for structure generation by storing additional information with the RV when it is first generated. This information is in the form of an ordered list of pre-made molecule fragments (a 'recombination path') that can be used in the structure generation process. Rather than constructing the new molecule atom-by-atom as before, these fragments are used to apply multiple atoms at once in a predetermined order, removing the need for a breadth-first search (Hristozov et al., 2011).

The recombination path data is generated during reaction vector calculation by reconstructing the product molecule from the starting material. Initially a 'reverse fragmentation' approach (Figure 4.3) is used. The name refers to the fact that, in addition to fragmenting the starting material by the atom pairs lost during the reaction, the product molecule is also fragmented using the atom pairs that are gained during the reaction. The aim of this step is to reduce both sides of the reaction to the fragments that remain unchanged by removing all changed bonds from the process. Presuming that the information encoded in the RV is sufficiently unambiguous, both sides will have identical fragments, as the structures and environment data for the fragments will be the same. However, as the atom pairs do not encode the full environment of the changed bonds, there may be multiple sites at which the fragmentation can occur. In these situations, multiple sets of fragments are generated, and must be compared systematically. If a match is found between the two sets of fragments, it is assumed that the forward synthesis for the reaction can be generated via said fragments. The necessary fragments for structure generation are then obtained by extracting the largest 'base' fragment from the product, using an MCS algorithm, and assigning the remainder of the product molecule as reagent fragments. It should be noted that, in the example in Figure 4.3, only one reagent fragment is present, but depending on the disconnections multiple fragments are possible. These fragments are represented internally as lists of atom pairs, effectively representing the substructures as complete

entities to be attached. These are then used to perform a full reconstruction in the manner previously described in Section 4.3.1, but now using fragments (sets of atom pairs), in order to determine an assembly order. The RV and ordered list of fragments are then placed into an SQL database, designed to permit rapid recall of the vectors and recombination data as necessary, further speeding up the process.

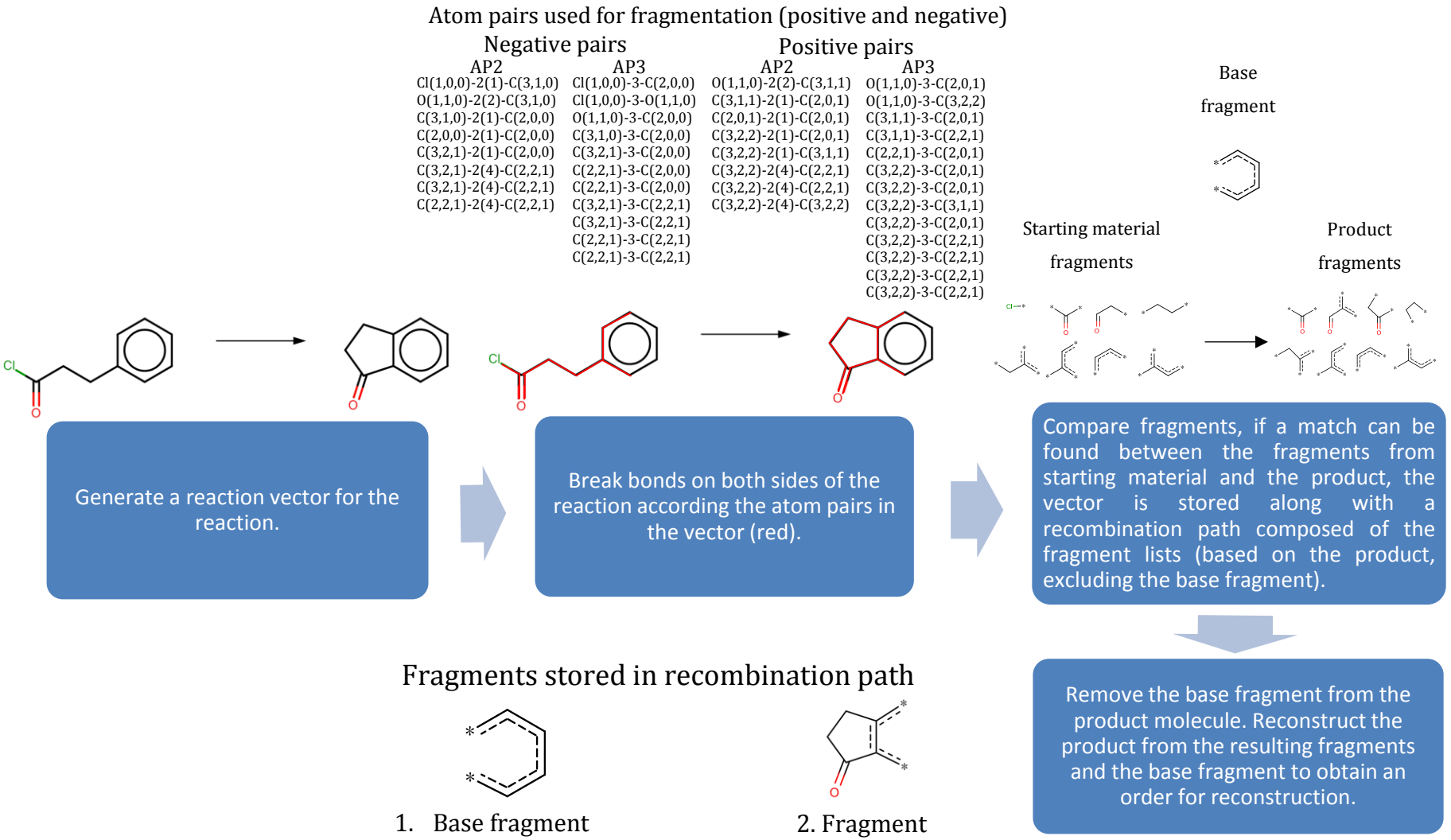


Figure 4.3: Flowchart showing the reverse fragmentation process. Note that all the acyclic atoms are 'lost' as they become part of a ring in the product. On recombination, the ring fragment (labelled 2 above) is stored as the recombination path (Wallace, 2015)

While the reverse fragmentation method is effective in the majority of cases, when this process cannot produce a result, the original method by Patel (Section 4.3.1) is used instead, attempting to construct the product molecule via the breadth-first search process as previously described. If this method is successful, then a path to the solution is stored as an ordered list of atom pairs. This recombination path then enables an ordered step-by-step reconstruction of the product that, while slower than the reverse fragmentation approach, is considerably faster than the breadth-first approach described by Patel et al. If this also fails to produce the correct structure, then the reaction vector is not stored in the database. An analysis of the reverse fragmentation approach carried out by Hristozov (Hristozov et al., 2011) demonstrated that this approach has a higher success rate than the previous method, with 89.8% of the 5,695 reactions tested successfully reproduced, compared with 85%. The analysis also claims an average run time of 0.015 seconds per reaction for the new method, with a maximum execution time of 30 seconds, compared to an equivalent maximum run time of up to 5 minutes in the case of the original method, as reported in the original PhD thesis (Patel, 2009). However, not all reaction types can be reproduced to the same degree of success. For example, Fischer indole synthesis reactions were successfully reproduced in only 41% of the 230 cases tested for the new method. However, no equivalent analysis by reaction type was made for the original method to enable a comparison.

4.3.2.1 RV-based structure generation

Once the reaction vector database has been created, it can be utilised to generate novel structures by applying the vectors to different starting materials. The atom pairs of the starting material are compared to the set of negative atom pairs belonging to a particular reaction vector, and if all of the necessary features are present (or a suitable subset is present that can be combined with an external reagent) then the vector can be applied. The structure generation process is outlined in Figure 4.4. Once an RV has been selected, the negative atom pairs are removed from the starting material, and the fragments from the recombination path are added according to the order previously recorded. This will lead to a new structure being created. However, in order to ensure that the molecules produced are chemically sensible, each produced molecule is checked before being reported as a result: the molecule is loaded into the RDKit library (Landrum) and subjected to a full molecule sanitisation process which includes cleaning up non-standard valence states, verifying the aromatic states for rings are

correct and valid, and calculating hybridisation states. Any structures that are not considered chemically stable and sensible (such as those with atoms in higher than allowed variance states, or incompatible aromatic systems) are rejected at this stage. The structure generation process can be repeated for all of the vectors in the database, and all possible functional sites on the molecule, until every possible molecule is generated.

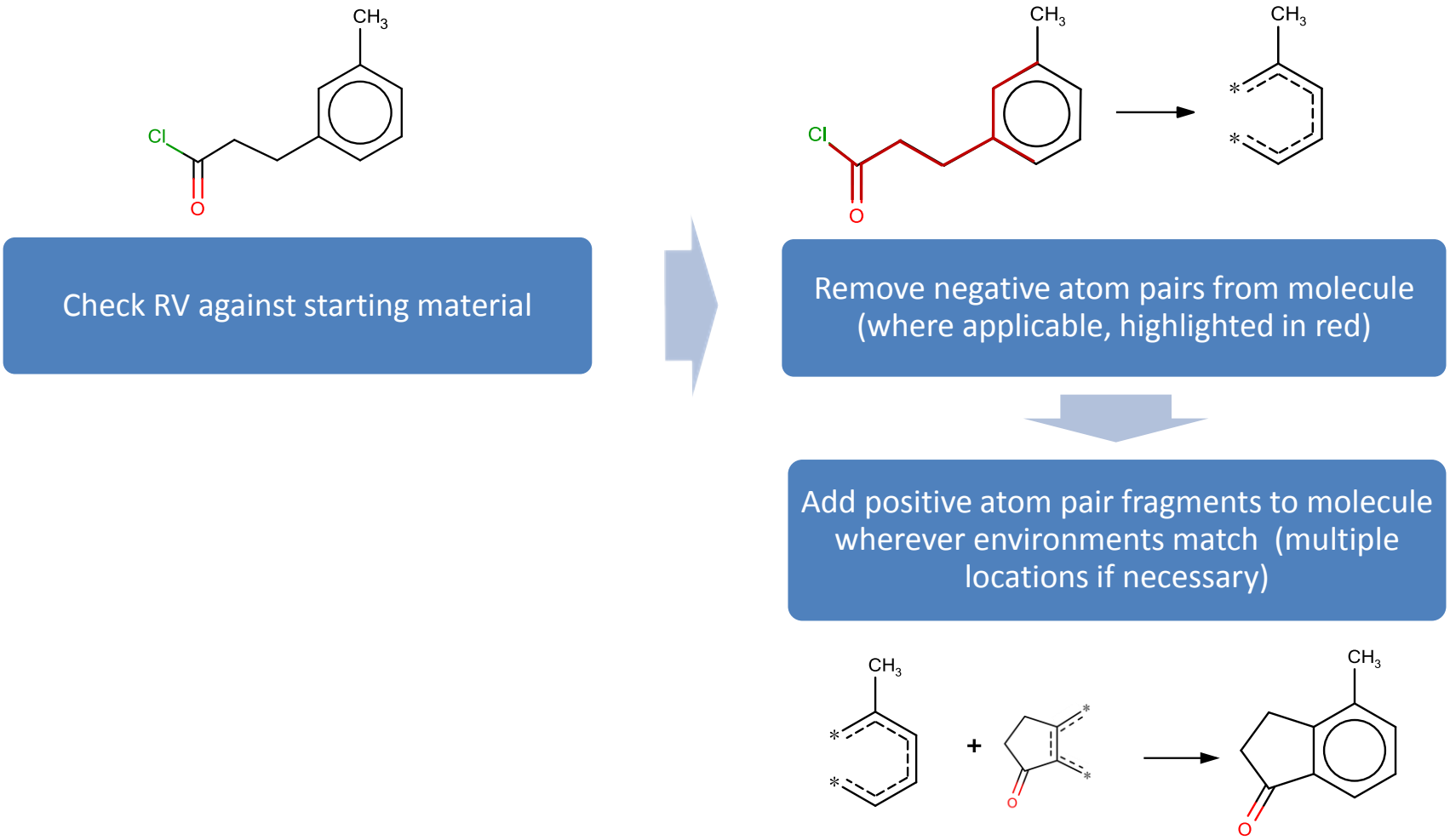


Figure 4.4: Simple example of the structure generation process, using the RV from Figure 4.3. (Wallace, 2015)

4.3.3 Additional features

4.3.3.1 Handling multiple reactants

The reaction examples discussed so far have consisted of molecular rearrangements and other similar simple reactions. As indicated previously the RV method is also capable of supporting reactions where two reagents are combined to produce one or more products, as illustrated in Figure 4.5.

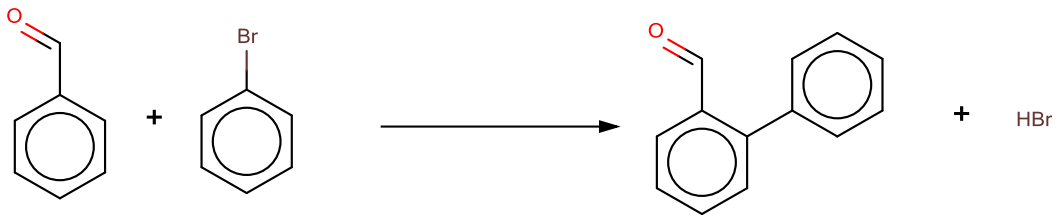


Figure 4.5: Example of a two component reaction taken from J. Med. Chem. (Wallace, 2015)

In these cases, the reaction vector encodes the negative atom pair descriptors of both starting materials. If the starting material used for the structure generation step does not contain all of the negative atom pairs it is possible to search in a database of reagents for a molecule that contains the missing atom pairs in order to use the reaction vector. The revised RV method developed by Hristozov et al. encodes reagent information directly alongside the generated recombination path, so that an external database is not required. However, to increase the number of products alternative reagents can also be used via a database, with any appropriate molecule replacing the stored reagent. The use of the external reagent generation method is illustrated in Figure 4.6. In this case, any reagent in the pool that contains the same atom pair environments as the original reagent is used to generate structures in the same way, resulting in additional products being generated.

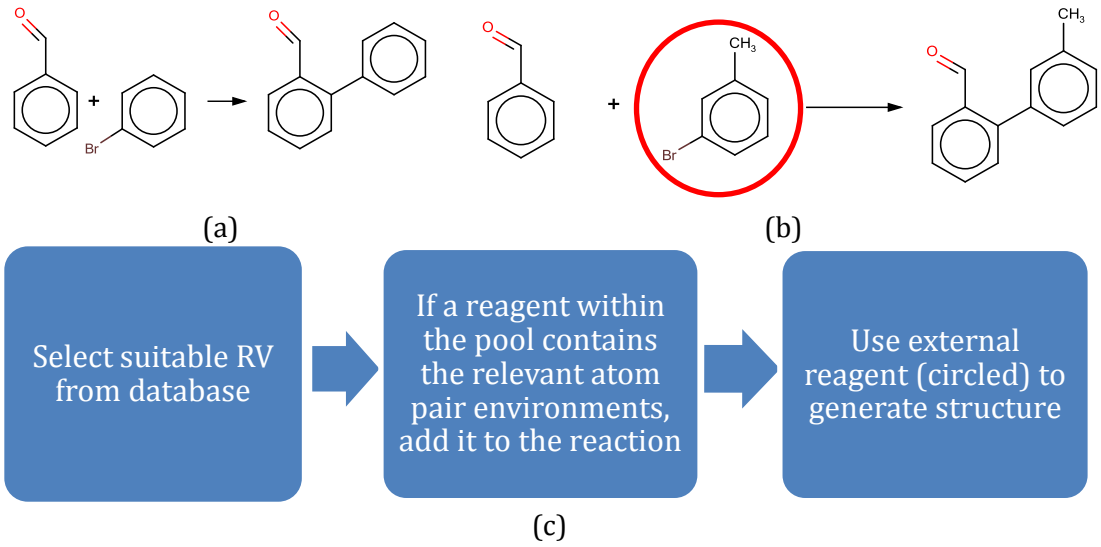


Figure 4.6: Example of the use of the external reagent generation. (a) The reaction from which the RV is derived. (b) An alternative reagent (circled) that contains atom pairs needed to apply the RV. (c) Flowchart describing the external reagent process. (Wallace, 2015)

4.3.3.2 Reaction balancing

As the RV is based on the differences between the two sides of the reaction, any mismatches in the number and type of atoms between the sides may result in problems when applying the RV. As part of the original Patel method for generating RVs, a reaction cleaning tool was designed to reduce these problems. This tool seeks to correct imbalances in reactions between the carbon atom counts on the reactant and product sides. The first step is to determine the number of reactants and products in the reaction. In situations where more than one product is listed, the reaction is split further into separate, one product reactions (so R1 + R2 → P1 + P2 is split into two reactions, R1 + R2 → P1 and R1 + R2 → P2), as illustrated in Figure 4.7.

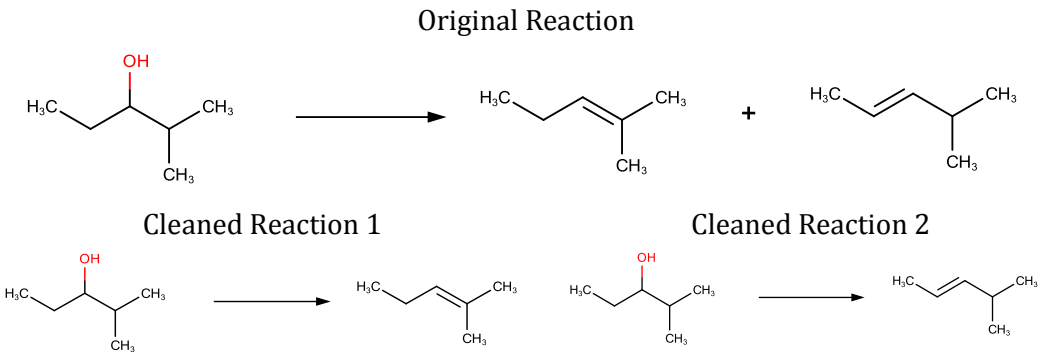


Figure 4.7: Example of a dehydration reaction that is cleaned by separating into two distinct reactions. Only carbon containing molecules are shown.(Wallace, 2015)

The carbon atoms in these reactions are then counted again, to see if the reaction has now become balanced. Should there still be a mismatch, atom mapping information from the reaction is used to identify any missing fragments. In this process any atoms on the reactant side that do not have mapped counterparts in the product side are combined into a new, stable product molecule. This process is repeated for the product side, creating new reactants out of unmapped product atoms. If there are still atom imbalances at this point, additional copies of each reactant and product are added to balance the stoichiometry, with the carbon count repeated at each addition. If none of these approaches work, reactions with more than one reactant are analysed with each reactant removed in turn to identify any reagents that are not involved. A flow chart illustrating the whole process is shown in Figure 4.8. It should be noted, however, that this tool is not essential for use with the revised fragment based approach, as this does not require a perfect atom balance to operate.

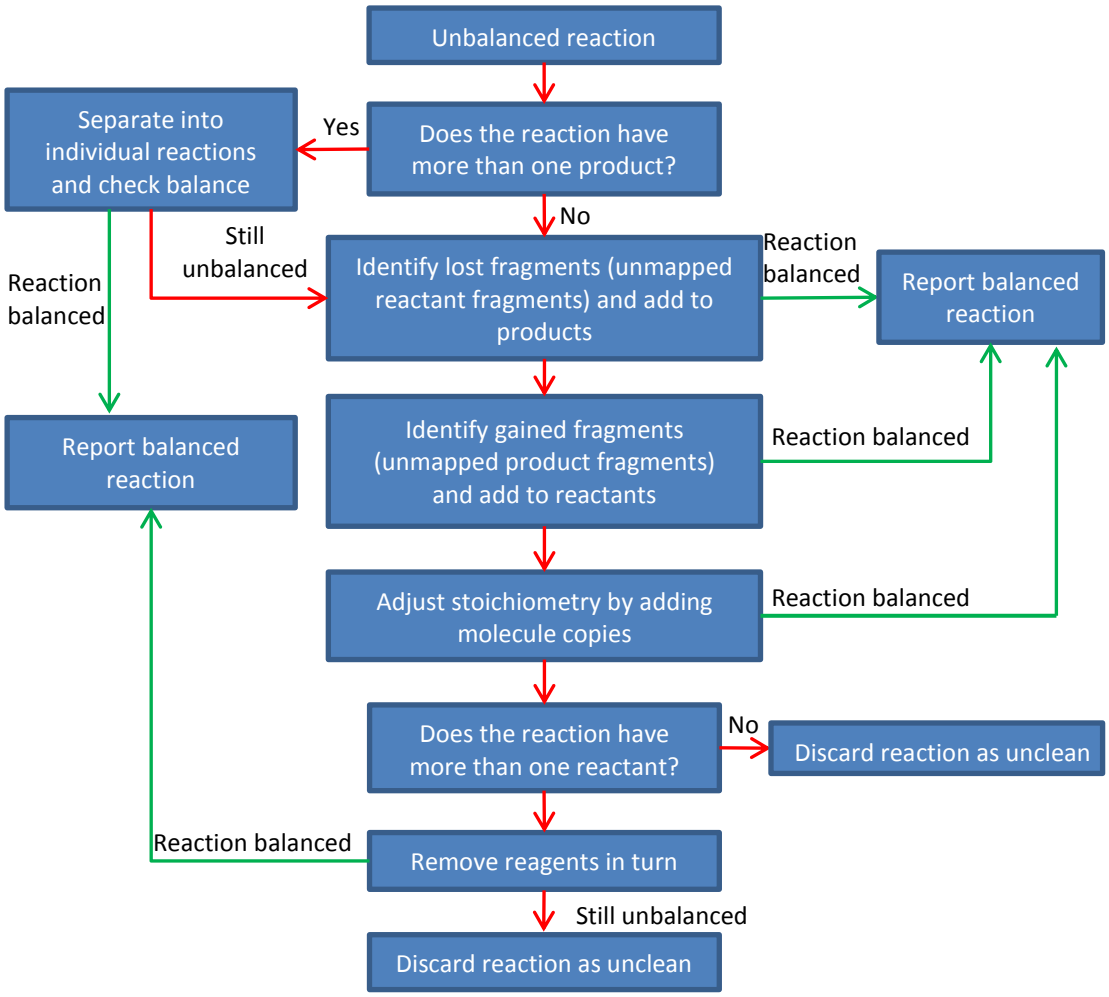


Figure 4.8: Summary of the reaction cleaning algorithm. At each step, the carbon count is rechecked, with the algorithm stopped if the reaction is balanced (green route). If not, the algorithm continues (red route).

4.4 Conclusions

In this chapter, the reaction vector approach developed by Patel et al. and then refined by Hristozov has been summarised, including how the reaction vector is calculated and applied to generate product molecules. Using atom pairs as the descriptor retains sufficient data from a reaction to ensure that it will not be applied inappropriately, while also permitting novel molecules to be made. While RVs have been shown to be useful for *de novo* design (Gillet et al., 2012, Gillet et al., 2014), they have limitations when considering multi-step reactions since intermediates may not score well thus preventing potentially useful molecules from being found. Furthermore, the application of RVs can lead to very long execution times. The next chapter describes an approach to generating reaction sequences and reaction sequence vectors with a view to overcoming these limitations.

Chapter 5:
Reaction networking

5.1 Introduction

One of the main problems in *de novo* design is that the exploration of all possible structures within a given solution space is impossible, due to the combinatorial explosion. Rather than pursue every possible compound, it is necessary to find some way of scoring and evaluating the population of candidates at each generation, focussing on the routes most likely to give usable products. However, in a reaction sequence, such scoring methods become problematic. In these circumstances, the intermediates in the sequence may be given significantly worse scores than the starting material, for example, due to the structural contribution of protecting groups or similar features. As a consequence, potentially useful routes can be rejected.

As finding a scoring method that can account for the disparity between intermediate and final structures is a very complex problem, an alternative approach can be considered for use during the *de novo* process itself, which is to skip past the intermediates and execute entire sequences within one execution step. This chapter explores preliminary work aimed at developing networks of reactions with a view to using this approach to perform these multi-step processes.

Section 5.2 describes the methods used in database preparation to transcribe and store reaction sequences. In Section 5.3, the KNIME nodes and workflows developed for testing and cleaning the input data are discussed and demonstrated. Section 5.4 shows how this data can be expressed in the form of a network, linking molecules via known reactions. This concept is extended in Section 5.5, where an external knowledge base consisting of single step reactions is processed and sequence data is generated by linking reactions according to common reaction components.

5.2 Collation of a set of reaction sequences

While many reaction databases exist that can be mined for reaction information (Table 2.2, Section 2.3.2), the majority of these store the reactions as individual entities, with only limited sequence data available such as the synthesis information in the Reaxys AutoPlan synthesis planner (Elsevier). In order to develop a *de novo* method based on sequences it is therefore necessary to develop a method for creating reaction sequences. To test if this would be feasible, a preliminary experiment was conducted whereby a small set of sequences was collected manually. This set was then split into its component reactions and methods were developed to reconstruct the sequence data algorithmically. Should this process succeed, this implies that it will be possible to create sequence information for any collection of reaction data. This section describes the preparation of this test set, and the ways in which the set was used for tool development and evaluation.

5.2.1 Literature abstraction

Following the procedure outlined by Roughley and Jordan (2011), the SciFinder database (Chemical Abstract Services, 2011) of journal articles was searched for structure activity relationship (SAR) papers that contain suitable reaction schemes. The search was restricted to those papers published in 2008 in three significant medicinal chemistry journals (Bioorganic and Medicinal Chemistry, Bioorganic and Medicinal Chemistry Letters and the Journal of Medicinal Chemistry) with further restriction to three drug companies (AstraZeneca, GlaxoSmithKline and Pfizer). The resulting papers were analysed by hand for reaction schemes (here defined as any collection of reactions that leads to a defined product), with each step redrawn and saved as a BIOVIA .RXN file. In order to ensure both sides of the reactions were balanced in terms of atom counts (see Section 5.3.1), the reagent information listed in the quoted method was used, excepting in cases where the only information was from a reference to a previously published paper, in which case data from that method was taken. If the reaction could not be balanced using this information, or the reaction sequence proved too complex to transcribe (more than eight molecules on one side), the scheme was rejected and not added to the database. In total, 102 reaction schemes were collated; examples are shown in Table 5.1 in the form of the individual reaction steps that combine to form the sequences. The reaction ID field in this table is made up of the paper number, the number associated with the product in the paper and finally the reaction number. For example, 19_2701 represents the 19th paper to be reviewed and

the first reaction in the sequence to make product **27** from that paper. In these particular examples, many of the sequences share the same first reaction, only differing at the second. For product **30**, however, the sequence is entirely different.

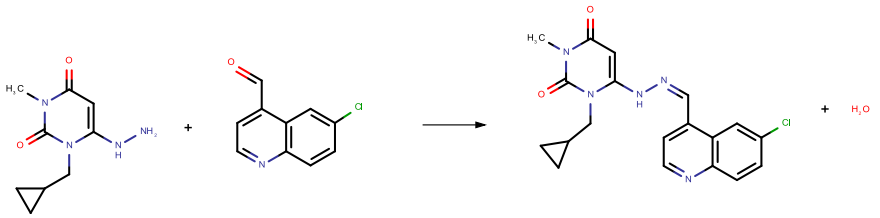
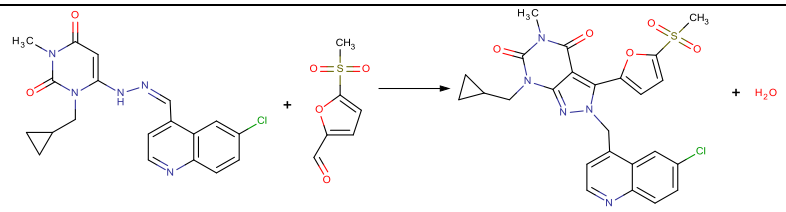
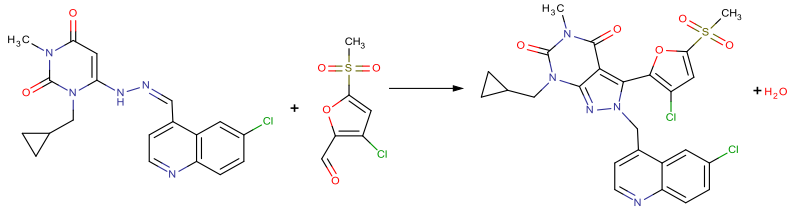
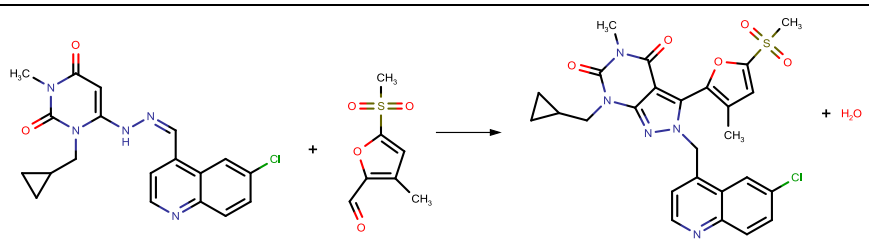
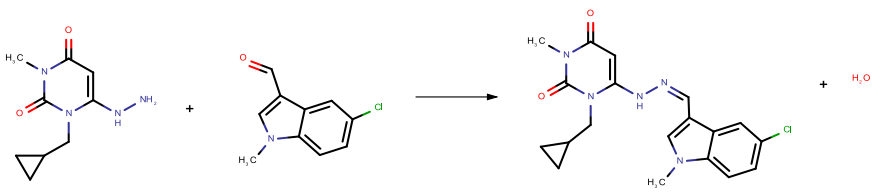
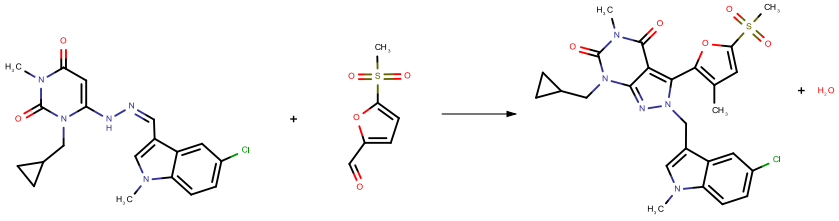
Reaction ID	Image
19_2701	
19_2702	
19_2802	
19_2902	
19_3001	
19_3002	

Table 5.1: Example reactions from the reaction database. All examples taken from Basarab, G. S., Hill, P. J., Rastagar, A. & Webborn, P. J. H., 2008. Design of Helicobacter Pylori Glutamate Racemase Inhibitors as Selective Antibacterial Agents: A Novel Pro-drug Approach to Increase Exposure. *Bioorganic & Medicinal Chemistry Letters*, 18, 4716-4722, where they were represented in the Kekulé form. (Wallace, 2015)

Due to the analytical focus of the papers, many of the reaction schemes used contain information about the preparation of series of analogous compounds, usually made by following the same or similar reaction sequence with appropriately different reactants. One such scheme is shown in Figure 5.1, and illustrated in Table 5.2.

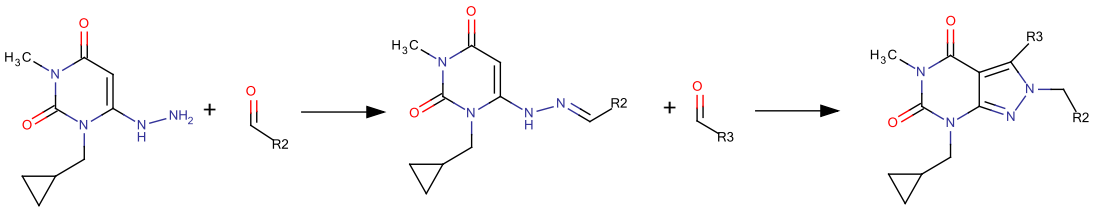


Figure 5.1: Generic representation of the reaction scheme associated with paper '19' from Table 5.1, represented in the Kekulé form.(Basarab et al., 2008, Wallace, 2015)

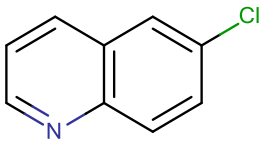
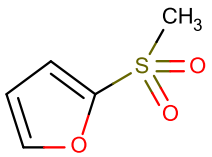
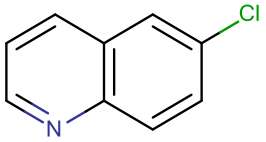
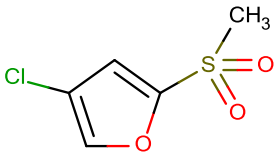
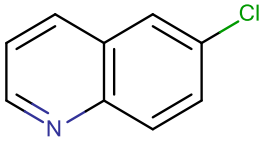
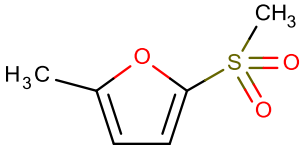
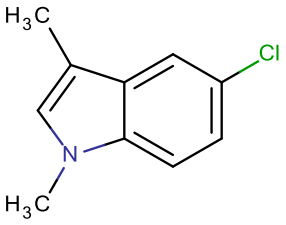
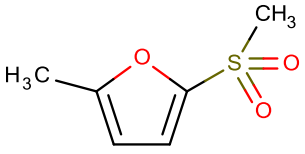
Product ID	R2	R3
27		
28		
29		
30		

Table 5.2: Breakdown of reactants used in scheme '19', using the product ID from Basarab et al, represented in the Kekulé form.(Basarab et al., 2008, Wallace, 2015)

In these cases, all of the individual sequences were enumerated leading to 424 reaction sequences being collated in total, representing 1544 individual reaction steps. This includes a number of duplicates, either due to the implementation of the sequence branching, or coincidental duplicates (where the same reaction or an equivalent form is present in more than one unrelated sequence, due to being used in schemes in different papers). After removing duplicate reactions, there are 974 unique reaction steps, which were used to make the test set.

5.2.2 File format creation and data set processing

The sequences were represented as a '.SCMX' file with CML validated connection tables for the reactions in the sequence, and a reference to the paper for each step of the sequence was included. This XML form was used for both storage and processing of the initial database, due to its efficient storage capabilities, and the ability to use ChemAxon's Marvin libraries (ChemAxon) to read and manipulate the data. To produce the .SCMX file, each individual RXN file representing a reaction step was processed via a Java program to generate individual CML strings. These strings are converted via a second program to build one single file. Once encoded, the individual reaction steps can be processed within KNIME (Berthold et al., 2008) using the XPath query engine. A breakdown of the sequences by size is presented in Figure 5.2.

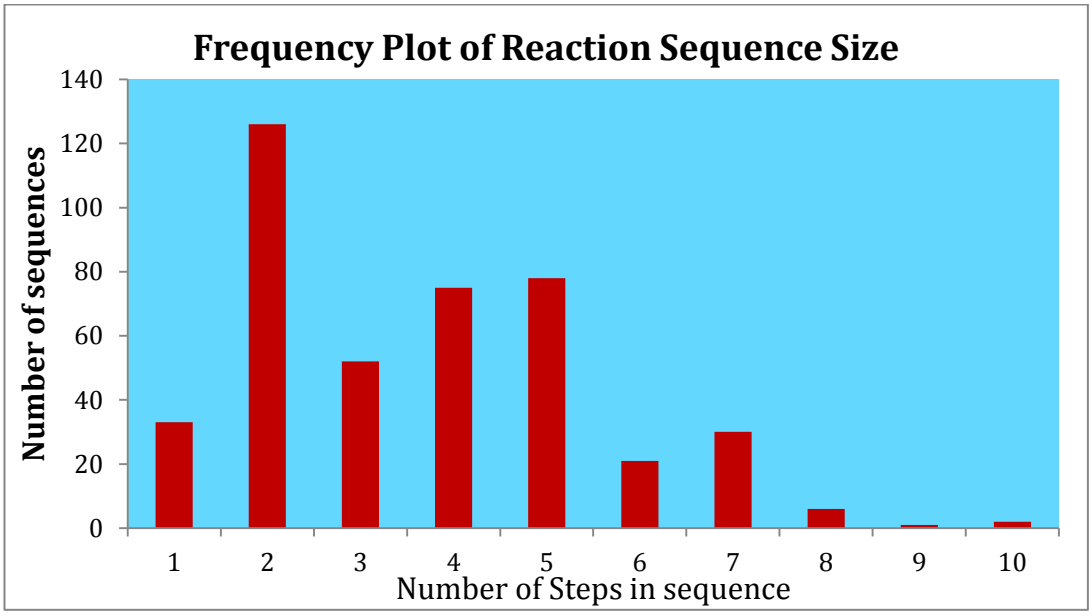


Figure 5.2: Frequency plot of reaction sequence size for the test set.

This limited data set seems to contradict the assumptions made by Carey et al. (2006), namely that the average synthetic scheme used in drug preparation contains 8.1 steps,

with heavy reliance on protecting group chemistry. However, this average appears to include all the steps involved in preparing the starting material. This would lengthen the sequences relative to the ones shown here that begin with the starting materials already prepared. In the above example, the large number of two-step sequences in this set skews the average towards three and four steps, although from a set this small it is difficult to draw meaningful conclusions as to trends.

The bulk of the data processing within this project was carried out using the KNIME data mining system, and a workflow was produced to process the .SCMX files into two SQL data tables. As before, XPath queries were used to read the sequence identifier, the reference and the reaction information. For reaction handling, the CML data was imported, canonicalised and converted to Reaction SMILES via the ChemAxon Marvin library incorporated within KNIME, as the alphanumeric nature of the format makes the data easier to process within Java using standard text processing methods. Samples of the output from these processes are shown in Table 5.3 and Table 5.4.

Scheme ID	Number of Steps	Step ID	Reaction ID	Reference
19_27	2	0	ID	
19_27	2	1	19_2701	
19_27	2	2	19_2702	

Table 5.3: Sample table of a reaction sequence as seen in Table 5.1.

Scheme ID	Reaction ID	Reaction SMILES String
19_27	19_2701	CN1C(=O)C=C(NN)N(CC2CC2)C1=O.ClC1=CC2=C(C=C1)N=CC=C2C=O>> CN1C(=O)C=C(N\N=C/C2=C3C=C(Cl)C=CC3=NC=C2)N(CC2CC2)C1=O.O
19_27	19_2702	CN1C(=O)C=C(N\N=C/C2=C3C=C(Cl)C=CC3=NC=C2)N(CC2CC2)C1=O.CS(=O)(=O)C1=CC=C(O1)C=O>> CN1C(=O)N(CC2CC2)C2=NN(CC3=C4C=C(Cl)C=CC4=NC=C3)C(C3=CC=C(O3)S(C)(=O)=O)=C2C1=O.O

Table 5.4: Table of reaction data for processing from Table 5.3.(Reaction SMILES split at product portion for increased legibility) (Basarab et al., 2008, Wallace, 2015)

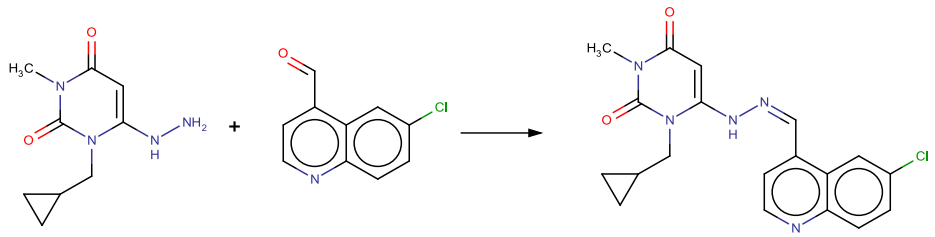
5.3 Curation of the reaction data

In any collection of data, whether gathered through automated processing or manual transcription, the likelihood of duplicate entries and other minor errors being present in a set is high. Having duplicates or imbalances in the reaction set can cause significant problems when trying to establish connections between reactions or when generating structures. In this section, the creation and usage of tools made to detect and eliminate duplicates and errors is discussed.

5.3.1 Reaction atom balancing

By using the Reaction SMILES format to store the data within KNIME, the individual molecular representations are automatically 'cleaned' by the loading algorithm. This generates a new molecule object from the original atom data, ensuring all charges and valences are correct, with explicit hydrogens removed where specified. This ensures that all mesomeric structures (such as nitro groups) are represented in a standard manner, and all aromatic structures are handled consistently, with Kekulé rings converted to the aromatic form. However, there is no guarantee that the reaction data

input is necessarily valid or balanced in terms of atoms. If any reagents or side products are omitted, problems can result with structure generation due to atoms required to construct the product molecule being missing. Consequently, identifying these issues at an early stage is essential, and so an atom balance checker was written. This counts each atom present in the molecules on the reactant side of the reaction and lists them by type, before doing the same for the product side. The counts for both sides are then compared to check for equality. Any mismatch between reactant and product is logged, listing the reaction in question and the nature of the imbalance (both the element symbol and the atom counts) to allow for easier correction by hand (see Figure 5.3).



Atom counts for individual components and sides

Reactant 1	Reactant 2	Product
H14 C9 N4 O2	H6 C10 N1 O1 Cl1	H18 C19 N5 O2 Cl1
H20 C19 N5 O3 Cl1		H18 C19 N5 O2 Cl1

Figure 5.3: Example of the atom count process for an unbalanced reaction from scheme '19'. (Basarab et al., 2008, Wallace, 2015)

In the example above the imbalance was fixed by adding H_2O to the generated products. After running the tool, a number of unbalanced reactions were discovered throughout the data set, usually involving the need to add the small molecules that were omitted in the original papers to the product side of the reaction, or the need to add equivalents of particular molecules to parts of the reaction to ensure stoichiometry.

5.3.2 Detection of duplicate reactions

Duplicate reactions were removed by processing with a set of KNIME nodes. The individual components of a reaction were sorted in alphabetical order of their SMILES strings to prevent issues where B + A → C would not be detected as a duplicate of A + B → C. Each individual reaction was then compared to the remainder of the database and any duplicates were removed, amending the scheme data to point to the first matching reaction in the set.

5.4 Reaction network generation

This section describes how the individual reaction steps are connected to form a reaction network of the form shown in Figure 5.4.

This approach is similar to the reaction graph creation method described in Section 2.3.1, with molecules on the nodes, and linking reactions on the edges. By forming a directed network, where the sense of each edge is to move from the reactant of the reaction to the product, a path through the network is then representative of a reaction sequence.

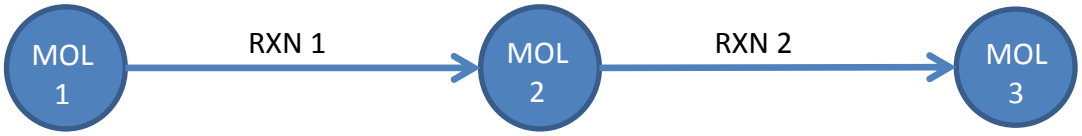


Figure 5.4: An illustration of the reaction network approach. Nodes (circles) representing molecules are linked by reactions (edges, arrowed).

The reaction network is generated using a KNIME workflow as seen in Figure 5.5. This contains three bespoke KNIME nodes, highlighted in green. The first KNIME node processes cleaned reactions in turn, outputting the reactant and product molecule strings in separate columns of a temporary storage table. A second KNIME node sorts the table by the individual molecules, assigning a unique hash value to each, and listing which reactions the molecule participates in, either as a reactant or a product. Finally, a comparator KNIME node takes the sorted data and produces a network compatible with KNIME. As previously discussed in Section 5.3.1, all reactions used in the network are pre-processed so that all structures are represented in a consistent manner. This greatly facilitates the network process by ensuring all representations of a given molecule are identical, and so errors are reduced. To form the network, the first reaction is inserted as two connected nodes. Next, the second reaction is taken and compared to all molecules present in the network. If neither the reactant nor the product is present in the network, the reaction is added as a new set of two nodes, with the reactant and product connected via a single directed edge, reactant to product. However, if one of the molecules matches, a new connection is made, depending on which molecule matches, with a new edge representing the reaction and new molecule(s) being added as nodes respectively. This process is demonstrated for a three reaction system in Figure 5.6, with an image shown in Figure 5.7.

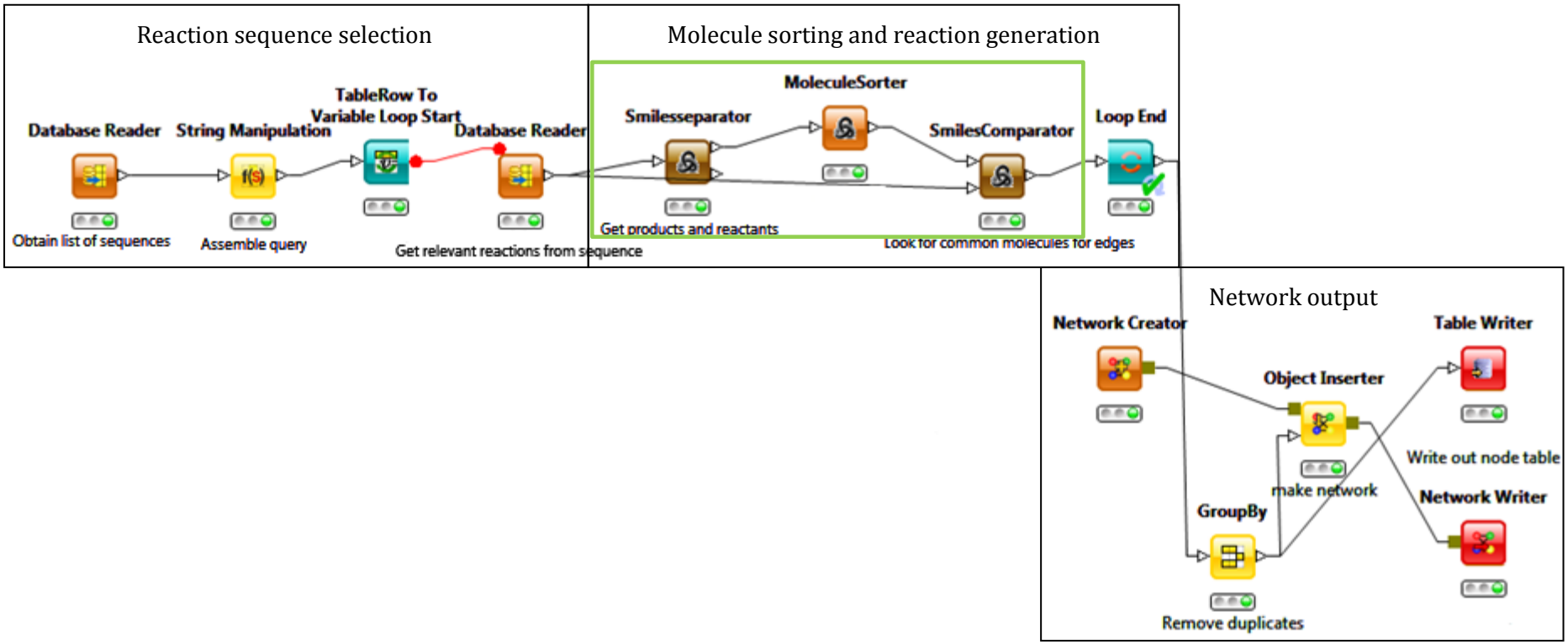


Figure 5.5: KNIME workflow showing the generation of the reaction network.

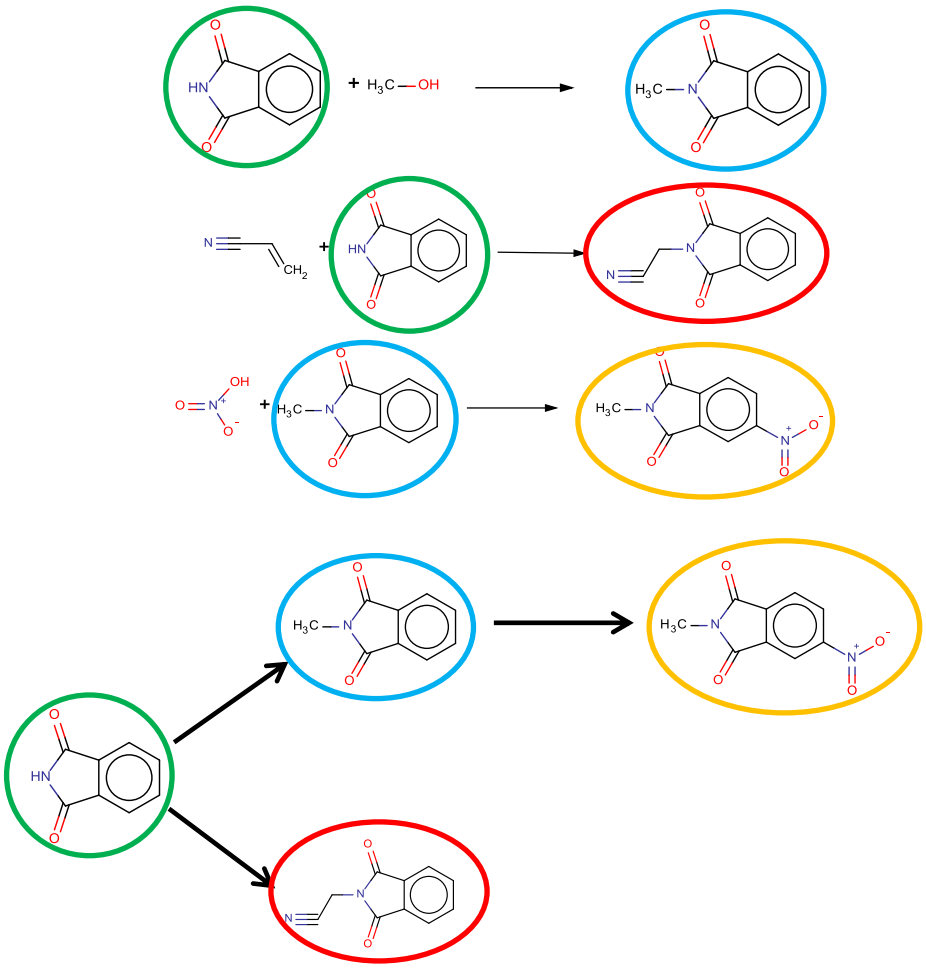


Figure 5.6: Example of network construction for three reactions from the database. Only one reactant and one product are considered for each reaction (circled, top) to be collated into the network based on their relative roles (bottom). (Wallace, 2015)

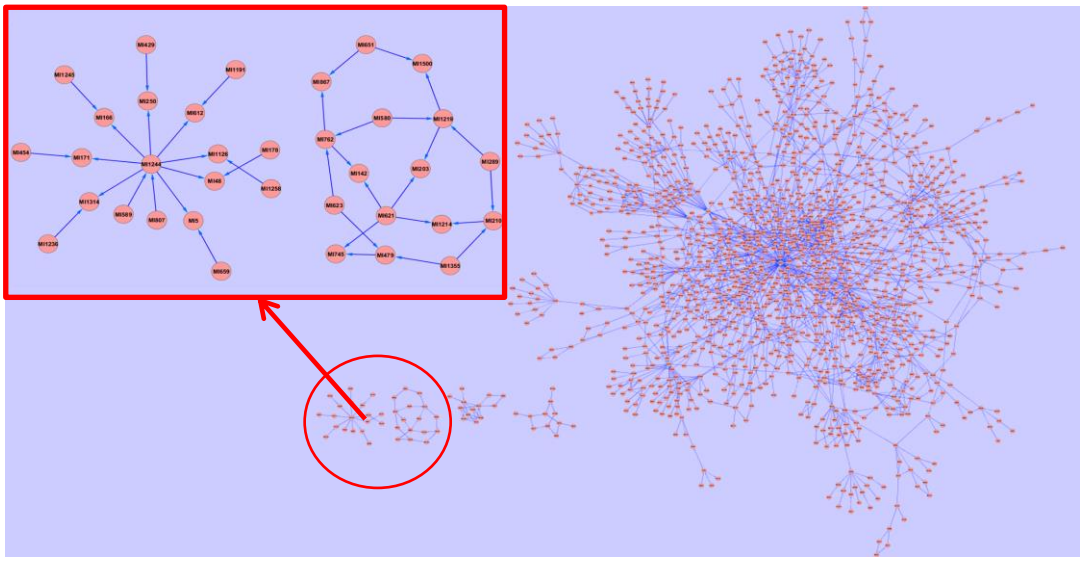


Figure 5.7: Images of the original small database expressed in terms of a molecule transformation network.

In Figure 5.6, one reactant and one product from each reaction is used to build the sample network. In the first version of network building each component in a reaction was considered independently, however, this led to many otherwise unrelated reactions being connected due to common small molecules such as water and methanol. This effect explains the dense region to the right of Figure 5.7. If these connections are allowed to remain in the network, the sequences that are produced will be overly long, and make little sense from a synthetic perspective, as the main products and starting materials will have little connection to one another. It was therefore necessary to introduce some rules to define which components should be considered when forming the network. These rules are applied to the individual reactions. Firstly, any molecule with fewer than three heavy atoms was removed from the reaction, before selecting the molecule with the largest atom count on the reactant and product sides. If either of these molecules had a weight above 500g mol^{-1}, the next largest molecule on that side was selected in its place where available, to avoid incorrect connections due to heavy reagents or catalysts. An example is shown in Figure 5.8, where one reagent and hydrogen bromide are removed from the network, while retaining the intent of the reaction.

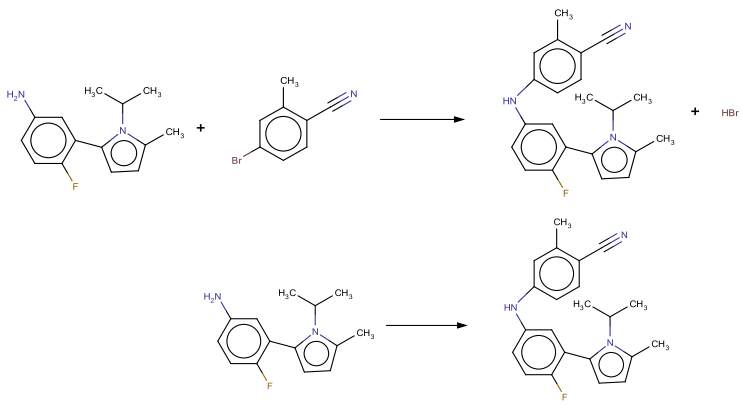


Figure 5.8: Result of filtration step for a sample reaction from the database. Top: Original reaction. Bottom: Filtered result.(Jones et al., 2008) (Wallace, 2015)

This selection process results in each reaction being reduced to one reactant and one product. This is important, as it means that the process of generating the reaction network is inherently lossy in nature. In order to retain a source of the full reaction data for later use and recall, the full reaction string is also stored within the relevant parts of the network as a feature, in the manner reported in Section 5.4.1. This balanced

reaction can be used to generate RVs as mentioned in Chapter 4 where needed for structure generation. An image of the filtered network is shown in Figure 5.9.

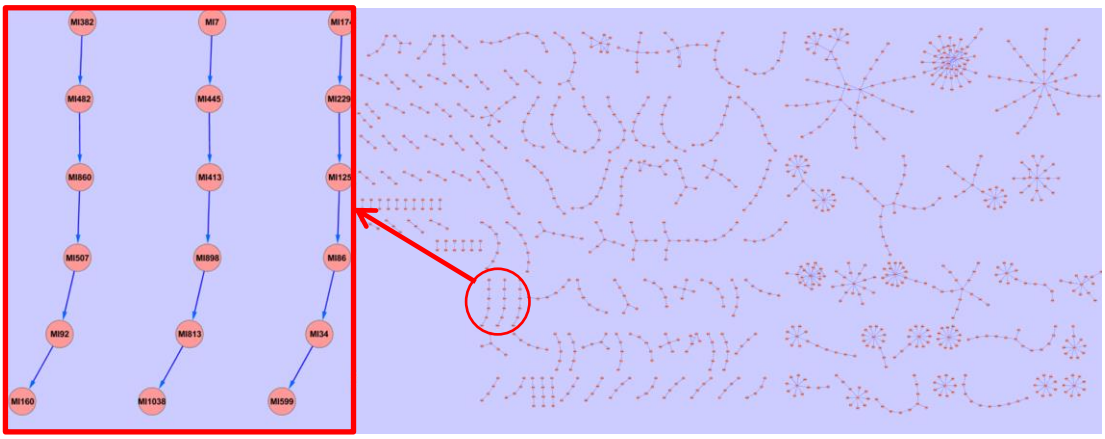


Figure 5.9: Images of the reaction network generated from the test set, with small molecules removed (Expansion of network portion highlighted).

In general, the reaction network will consist of a number of disconnected graphs, each of which represents a reaction sequence with the branching indicating sequences that have some steps in common but which diverge in later steps. Each node in the network represents a molecule, with the connected paths leading from it representing synthetic routes to potential product molecules. In addition to the simple linear paths that represent reproductions of the original sequences, there are also a number of interconnected 'wheels' towards the top right of the figure. These represent collections of reaction sequences with one common molecule at the centre, with the spokes indicating a series of analogues that could be made, based on the use of different reactions. The network also contains some new sequences found as a result of new connections linking the original sequences. Examples of sequences that show single and multiple routes to a product are shown in Figure 5.10.

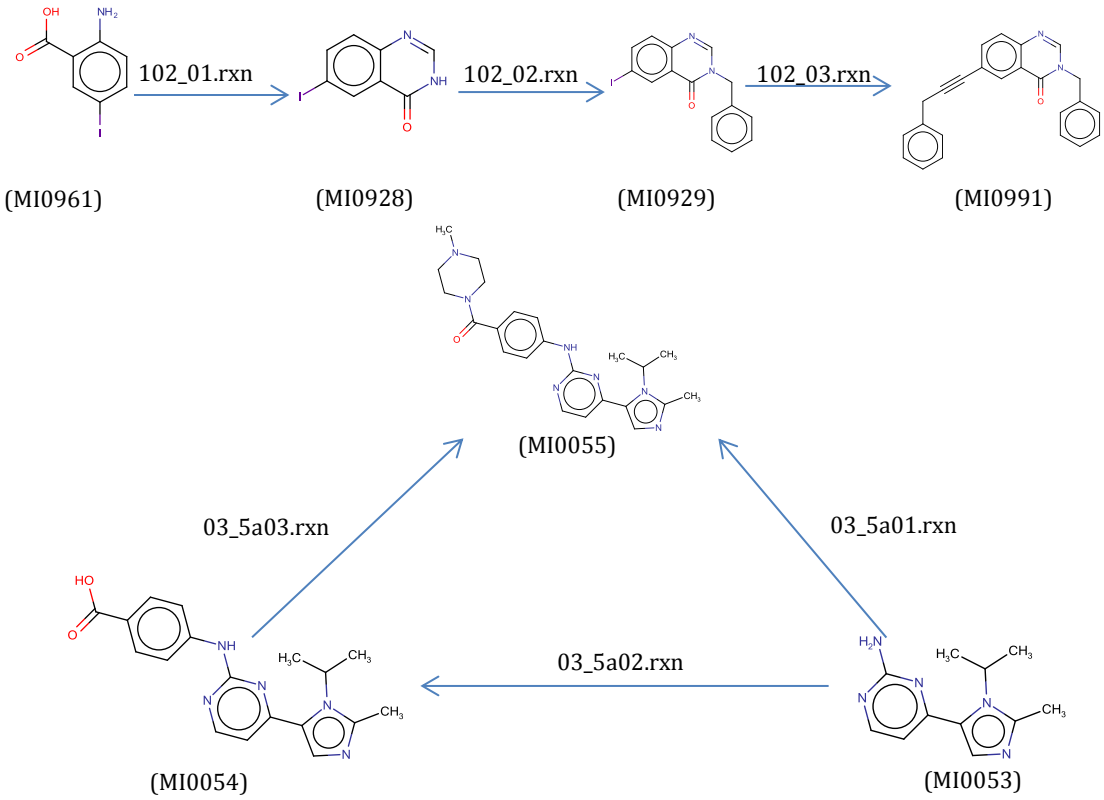


Figure 5.10: Demonstration of the molecule transformation network.
Top: Single route reported based on sequence 102. Bottom: multiple routes to one product. (Li et al., 2008, Wallace, 2015)

As Figure 5.10 shows, any path through the network represents a reaction sequence, and therefore, by iterating through all the available paths, the reaction sequence data can be regenerated. Where multiple reactions lead to the same end points, cycles can form that have to be considered when processing the network. However, in these circumstances, the only cycles of note are reversible reactions and rearrangements that occur over multiple reaction steps. No further action is needed in these respects, as the directionality of the network edges is sufficient to prevent recursion occurring. In order to collect reaction sequence information from a reaction network in an efficient manner, the network is first split into its subgraphs, so that each discrete portion of the network is analysed separately. A subgraph is then loaded into a KNIME node which interrogates all the possible paths between nodes using a variant of the Dijkstra algorithm (Dijkstra, 1959), with those paths between 'terminal' nodes recorded. For these purposes, a terminal node is one that represents a natural start or end in the network, i.e., it has only outbound edges (start) or inbound edges (end). In the case of the cycle shown in Figure 5.10, MI0053 is the only node that is considered as a start

point, and MI0055 is the sole end point, with both routes between the two recorded. After aggregating all of the paths and assigning a sequence ID to each, a data table is output to an SQL database as previously described, using the workflow in Figure 5.11. This form is then used as the new reaction sequence database, with the network retained only for visualisation purposes.

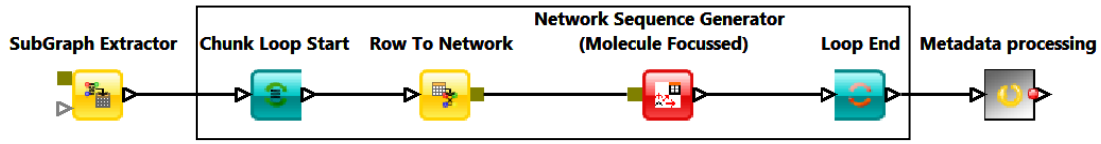


Figure 5.11: KNIME workflow showing the network sequence generator.(Generation portion highlighted).

Once all the subgraphs have been processed, it is possible to generate a breakdown of the sequences by size including any new sequences that have been discovered, to compare with the original distribution, as shown in Figure 5.12.

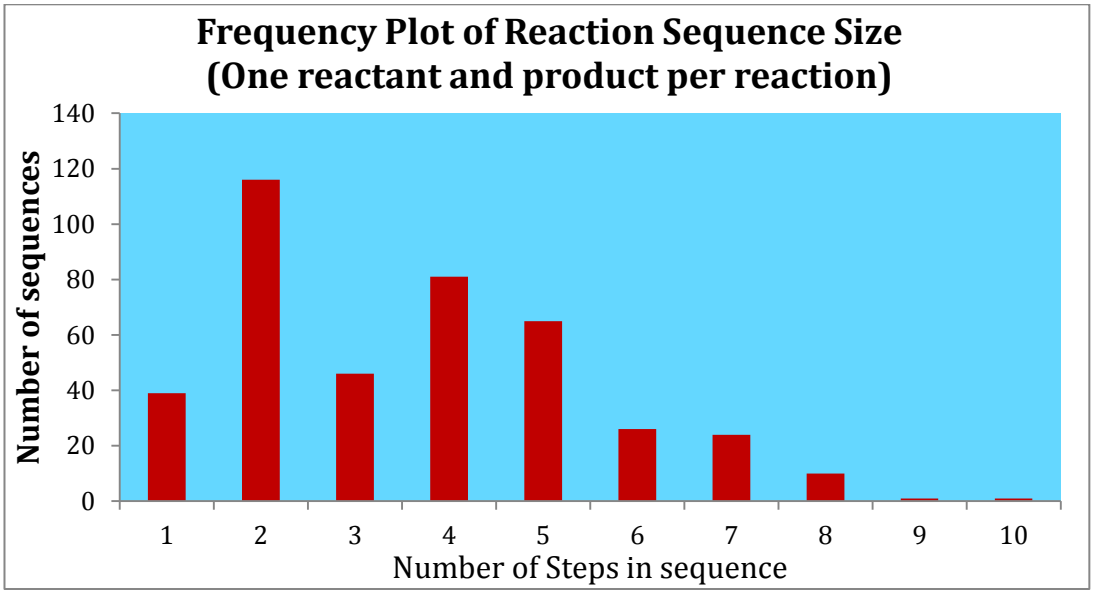


Figure 5.12: Frequency plot of reaction sequence size for the network, when only one reactant and one product are used. This includes any newly created sequences.

When the reactions are pre-processed to consist of only one reactant and one product prior to generating the network, some of the side connections that were part of the original network form, but do not represent the true intent of the reaction, are no longer present, hence the difference between Figure 5.12 and Figure 5.2. An analysis of the extracted sequences was performed, comparing each extracted sequence with the list of those originally recorded. The evaluation confirmed that all the known reaction sequences within the database are successfully reproduced in this form, with some

existing within longer network paths. Additionally, a number of new sequences were obtained from the network, as listed in Table 5.5. An example of an interconnection between sequences is shown in Figure 5.13.

Number of steps	Number of additional sequences
1	0
2	10
3	5
4	0
5	3
6	8
7	6
8	0
9	0
10	1

Table 5.5: New sequences found from the test set.

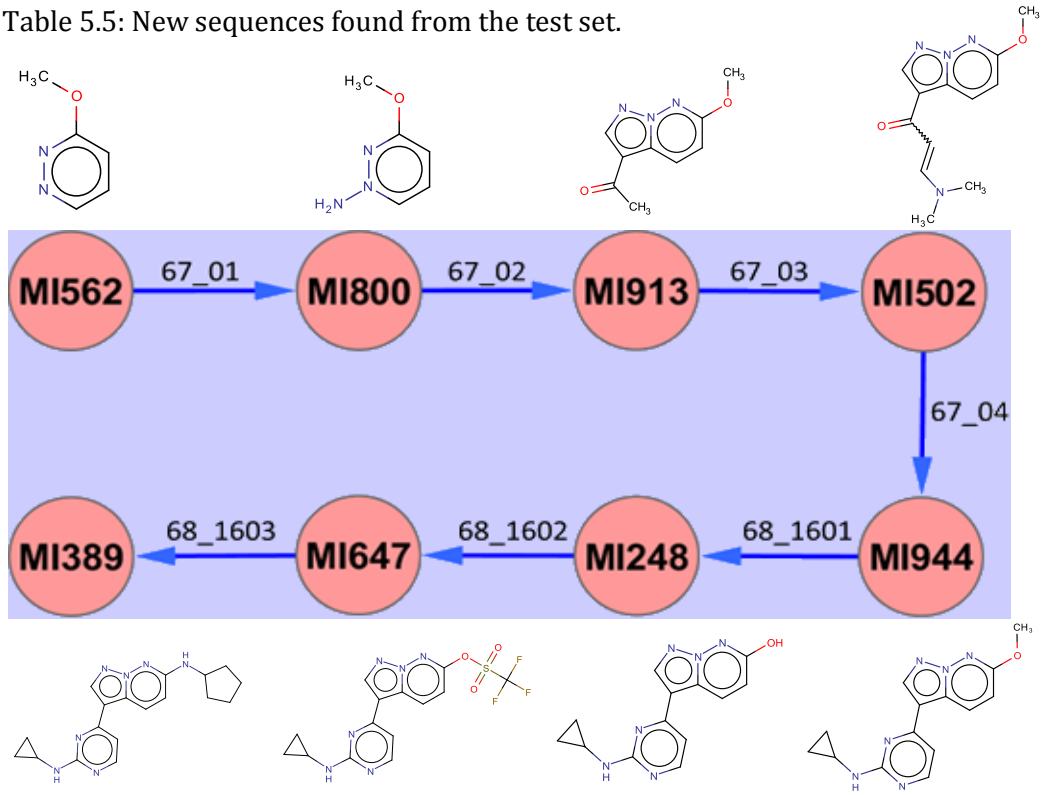


Figure 5.13: Chart showing an example of a new connection within the network.(Cheung et al., 2008, Stevens et al., 2008, Wallace, 2015)

In Figure 5.13, the compound identified as MI502 is the end product of sequence **67**, as well as the starting material for sequence **68**. Through connections like these, new routes can be uncovered in sufficiently large data sets.

5.4.1 Sequence database property addition

When sequences are extracted from the network and entered into the sequence database, additional information is also stored including the original references from which the sequence is derived. An example of the output as viewed through the Cytoscape network visualisation tool is shown in Figure 5.14.

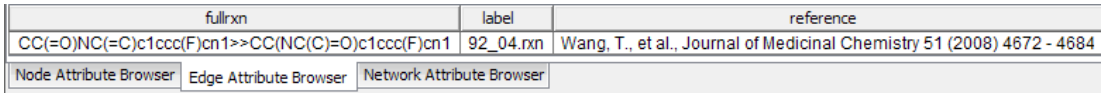


Figure 5.14: Example output from selection of an edge in the reaction network. (Wang et al., 2008)

By embedding the molecule data as SMILES strings within the attributes, it becomes possible to visualise the molecules within Cytoscape. Using the ChemViz software (UCSF), on selection of a network feature, this data can be passed through the Chemistry Development Kit (Steinbeck et al., 2003) and displayed within the network window, as seen in Figure 5.15.

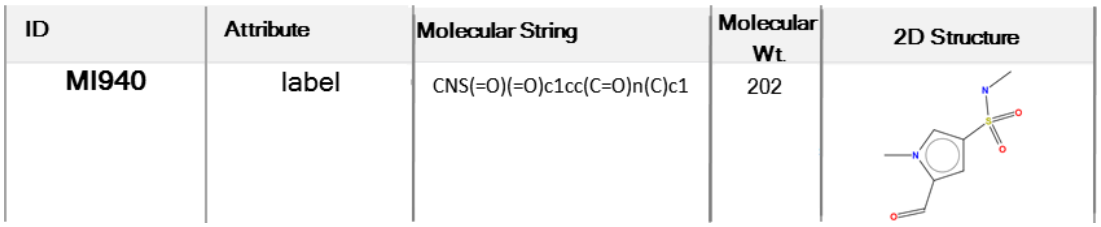


Figure 5.15: Example output from ChemViz on a given node of the reaction network.(Schnute et al., 2008, Wallace, 2015)

5.5 Using external databases and knowledge bases

5.5.1 Data processing and input

After confirming that the network tools can reliably reproduce the sequences in the manually created test set, the same process was used to generate larger networks, and subsequently lists of reaction sequences for other databases. This was achieved via utilising a knowledge-base of reactions previously collated by members of the CADD research group at Lilly UK (Hristozov et al., 2011). The database consists of 24,489 reactions abstracted from a number of papers from the Journal of Medicinal Chemistry (Patel et al., 2008), packaged as a BIOVIA RDFile. By using the RDFile parser built for

the KNIME system by the Lilly UK group as part of the Erl Wood Chemoinformatics tools, these reactions were converted into the existing database format. After importing, it was necessary to expand some of the wildcards found in the data. There are two different wildcards, one that represents generic halogen atoms as an 'X', and one representing any given atom as '*'. For compatibility with the reaction tools, the any atom wildcard was replaced with a carbon atom, and the X symbol was replaced with F, Cl, Br and I, in turn. Once these wildcards were fully enumerated a total of 25,610 reactions were made available for use.

5.5.2 Analysis of enlarged reaction network

After using the tools from Section 5.3 to clean the reaction data, a reaction network was created as before (Section 5.4.1) with the rules described above applied to limit each reaction to one reactant and one product. When the network is run through the sequence generator, 45,308 individual sequences of two or more steps in length are detected, with an average sequence length of 1.57 steps if single step reactions are included (shown in Figure 5.16, with the full data recorded in Table 5.6). If single step reactions are excluded, the average sequence length for the remainder is 6.72 steps. The sequences extracted from this network were collated into a data set referred to in the rest of this work as 'JMC1'.

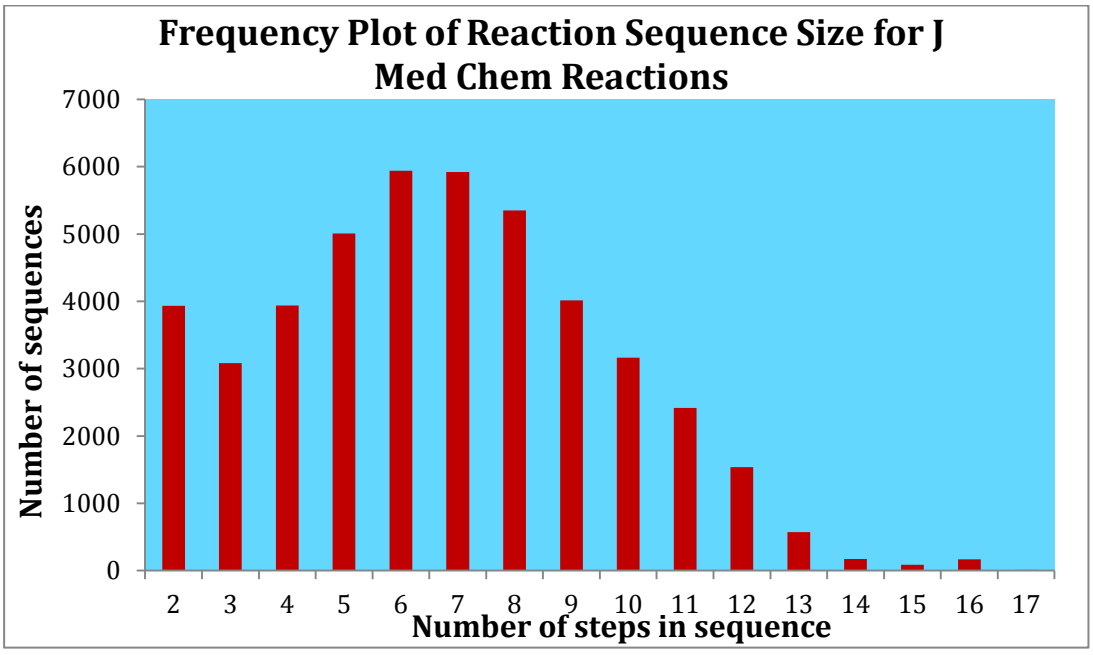
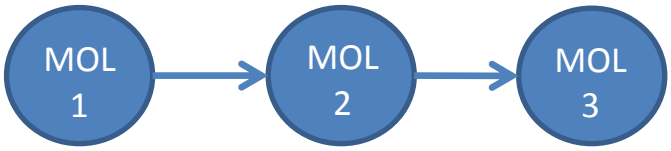


Figure 5.16: Frequency plot of reaction sequence size for the population

Number of steps	Number of sequences
1	25610
2	3932
3	3083
4	3937
5	5008
6	5939
7	5922
8	5352
9	4015
10	3162
11	2415
12	1539
13	573
14	170
15	87
16	165
17	12

Table 5.6: Table of the full sequence summary.

Another approach to generating sequences is to include all of the partial paths within the network (i.e. sequences in the middle of existing paths), as opposed to just using the longest paths. An illustration of this approach for a collection of three molecules is shown in Figure 5.17. Using this method, the distribution of sequences is more even, with an average sequence length of 5.32 steps per sequence (shown in Figure 5.18, with the full data recorded in Table 5.7), and 6.41 steps when single step reactions are removed.



Sequence Length	Molecules involved in sequence
1	MOL 1 → MOL 2 MOL 2 → MOL 3
2	MOL 1 → MOL 2 → MOL 3

Figure 5.17: Illustration of the additional sequences found within an existing path. In the original case, only the final sequence would be reported.

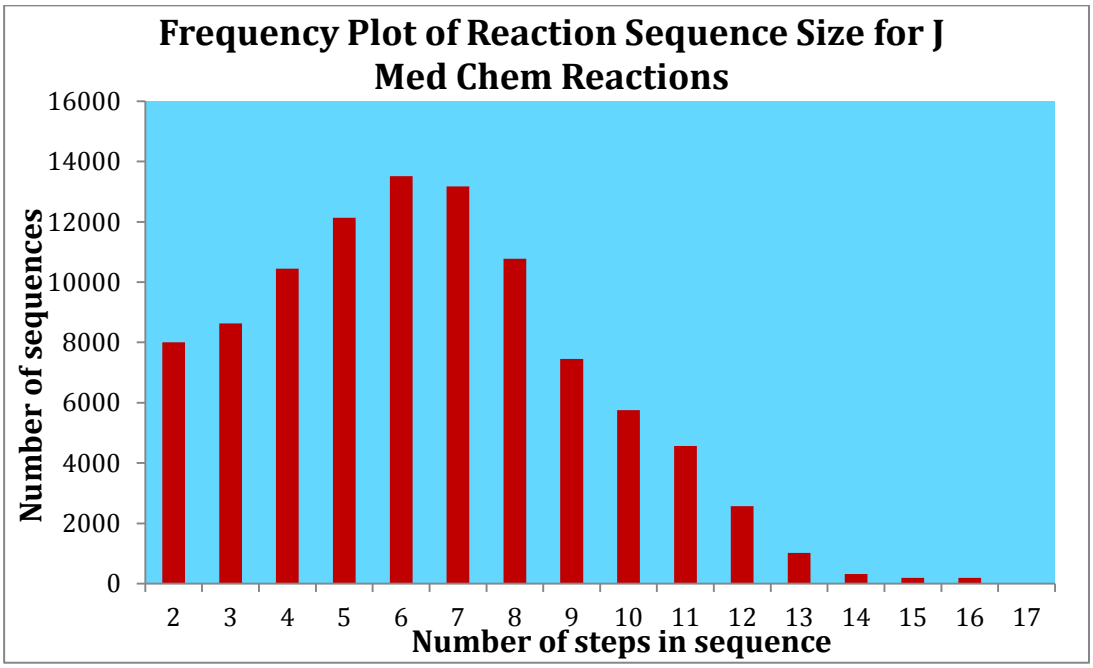


Figure 5.18: Frequency plot of reaction sequence size for the full population.

Number of steps	Number of sequences
1	25610
2	8000
3	8632
4	10452
5	12137
6	13519
7	13176
8	10771
9	7458
10	5758
11	4563
12	2569
13	1014
14	318
15	187
16	187
17	12

Table 5.7: Table of the sequence summary for the full population.

This collection of sequences (hereafter referred to as 'JMC2') is potentially more useful, in that it contains a larger number of sequences, with a more even profile.

5.5.3 Database analysis by atom pair content

For *de novo* design use, it is desirable for a given reaction collection to represent as diverse a range of transformations. One method of analysing a collection for diversity is to study the reaction centres for each reaction stored. Since the RV contains a representation of the reaction centre for a given reaction, it is possible to generate RVs for the entire collection and group reactions on the basis of identical negative atom pairs.

The negative atom pairs are relevant for this analysis since they represent the reaction features that must be present in a molecule in order for the RV to be applied. An analysis of the J. Med. Chem. reactions used to make the JMC1 and JMC2 databases was performed in this manner, grouping the RVs according to negative atom pairs via KNIME (Berthold et al., 2008). A frequency distribution for the groups was produced and is shown in Figure 5.19, with an expansion of the early portion shown in Figure 5.20.

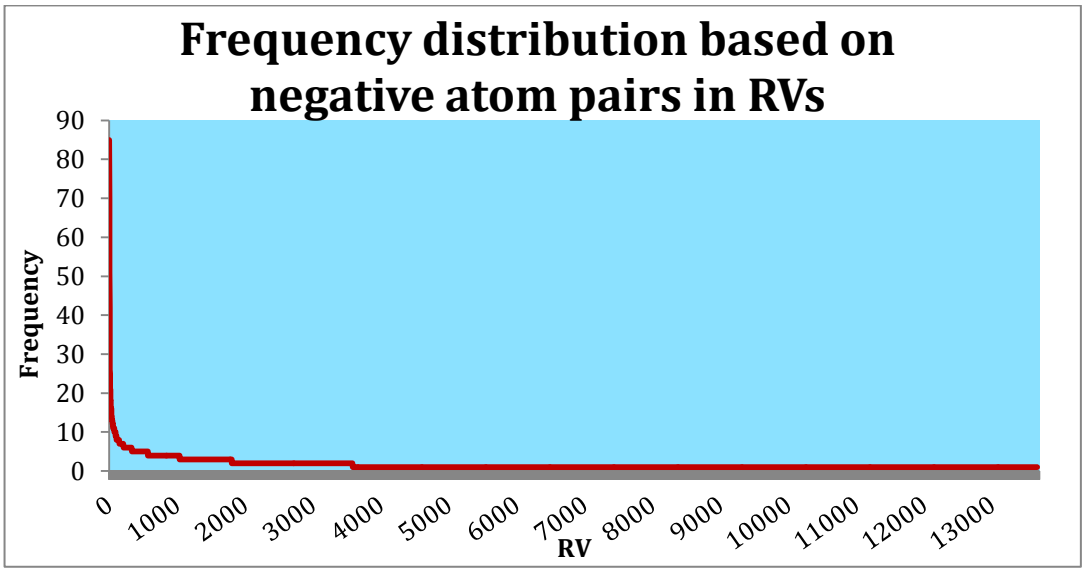


Figure 5.19: Frequency distribution curve based on the negative atom pairs in the JMC1 reaction data set.

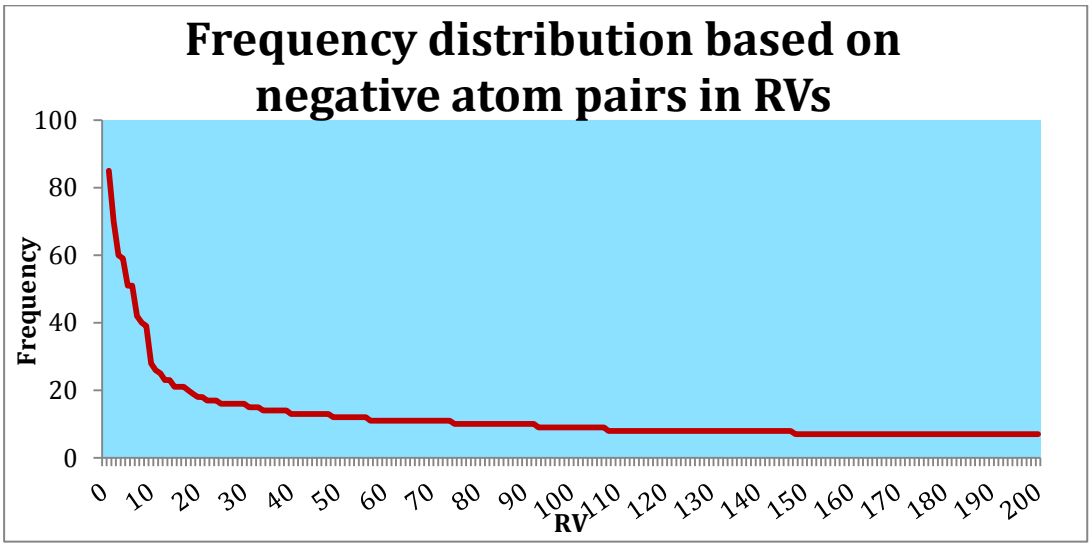


Figure 5.20: Expansion of the first 200 entries in Figure 5.19.

It is clear from the steep drop off and long tail in the distribution that there is a significant bias towards particular reaction types, with some having particularly high levels of representation. This property is well known, as discussed by Garagnani and Bart (1977). The distribution appears to follow Zipf's law (Adamic, 2011), where the frequency of a given entry is inversely proportional to its rank in the frequency table. If this is the case, a plot of the frequency value and relative rank of each entry on a log-log graph will be linear. Such a plot for this distribution is shown in Figure 5.21.

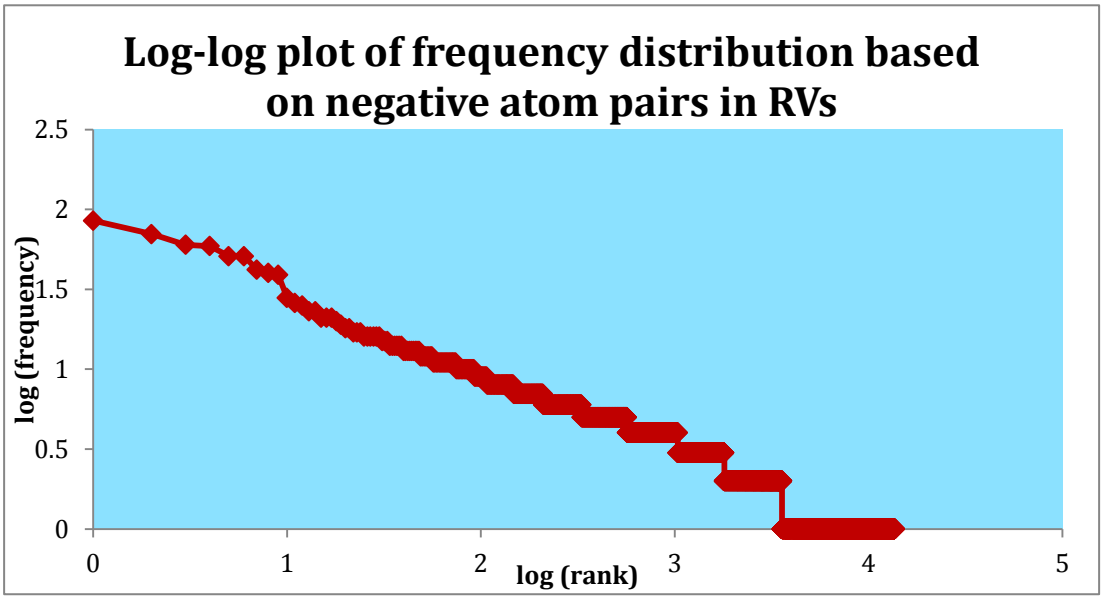


Figure 5.21: Log-log plot of the frequency distribution of negative atom pairs in the JMC1 reaction data set.

It can be seen that there are deviations from true linearity, suggesting that the distribution is not perfectly Zipfian. However, the inverse relation between rank and frequency is clear. In terms of the chemistry represented, the five sets of groups of negative atom pairs that are most frequent are shown in Table 5.8.

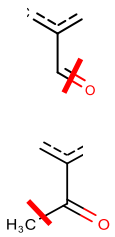
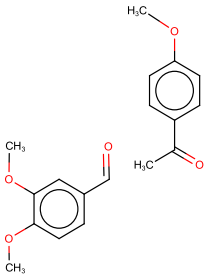
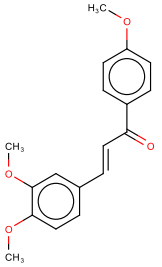
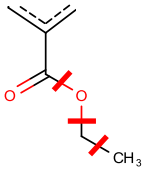
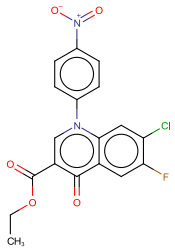
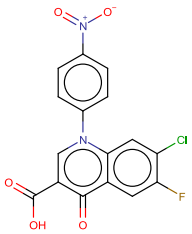
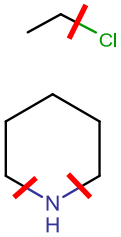
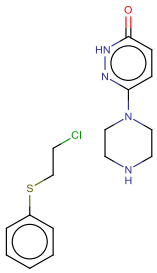
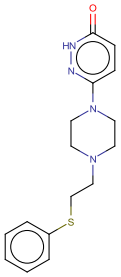
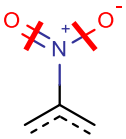
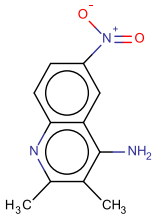
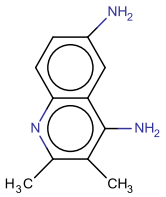
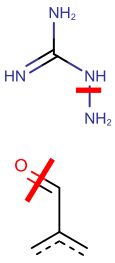
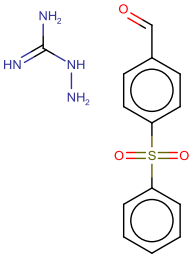
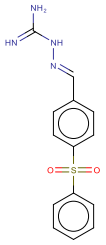
Negative atom pairs (duplicates indicate multiple entries)	Number of reactions represented	Reaction centre structure	Sample reactant(s)	Sample product
C(3,1,0)-2(1)-C(1,0,0) O(1,1,0)-2(2)-C(2,1,0) C(3,2,1)-3-C(1,0,0) O(1,1,0)-3-C(1,0,0) O(1,1,0)-3-C(3,2,1)	85			
C(2,0,0)-2(1)-C(1,0,0) O(2,0,0)-2(1)-C(2,0,0) O(2,0,0)-2(1)-C(3,1,0) C(3,1,0)-3-C(2,0,0) O(2,0,0)-3-C(1,0,0) O(2,0,0)-3-C(3,2,1) O(2,0,0)-3-O(1,1,0)	70			
Cl(1,0,0)-2(1)-C(2,0,0) N(2,0,1)-2(1)-C(2,0,1) N(2,0,1)-2(1)-C(2,0,1) Cl(1,0,0)-3-C(2,0,0) N(2,0,1)-3-C(2,0,1) N(2,0,1)-3-C(2,0,1)	60			
N(3,1,0)-2(1)-C(3,2,1) O(1,0,0)-2(1)-N(3,1,0) O(1,1,0)-2(2)-N(3,1,0) N(3,1,0)-3-C(2,2,1) N(3,1,0)-3-C(2,2,1) O(1,0,0)-3-C(3,2,1) O(1,1,0)-3-C(3,2,1) O(1,1,0)-3-O(1,0,0)	59			
N(2,0,0)-2(1)-N(1,0,0) O(1,1,0)-2(2)-C(2,1,0) N(1,0,0)-3-C(3,1,0) O(1,1,0)-3-C(3,2,1)	51			

Table 5.8: Representation of the five largest groups of partial RVs. The red lines indicate bonds broken in the reaction centre structure. (Wallace, 2015)

Unsurprisingly, given the fact that the reactions were collated from SAR explorations in the literature, the majority of the most common reaction centres in the set have some form of carbonyl content or aromatic character. As it is these features that are used as the selection criteria for deciding whether an RV is applicable, starting materials with these features will be more likely to give good results.

This method of grouping compares RVs on the negative AP2 and AP3 data, making the groups particularly sensitive to minor changes in environment. This will affect the nature of the grouping, as very similar reaction centres that differ in their immediate environment will be treated as separate entities, rather than being considered together. Making the comparison using just the AP2 content gives a distribution with the same skew, but with fewer groups overall (10,344 versus 13,669). A frequency distribution for the JMC1 reactions using the AP2 content is shown in Figure 5.22, with an expansion in Figure 5.23 and a log-log plot in Figure 5.24. As before, the distribution is not perfectly Zipfian, but shows a definite inverse relation between rank and frequency.

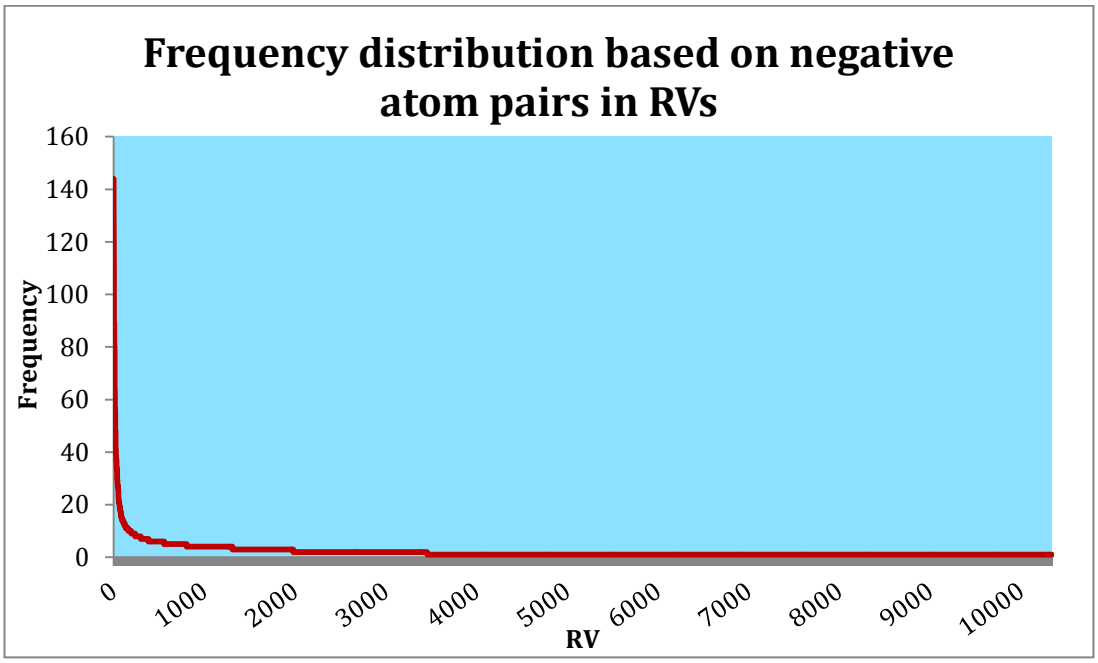


Figure 5.22: Frequency distribution curve based on the negative AP2 content in the JMC1 data set.

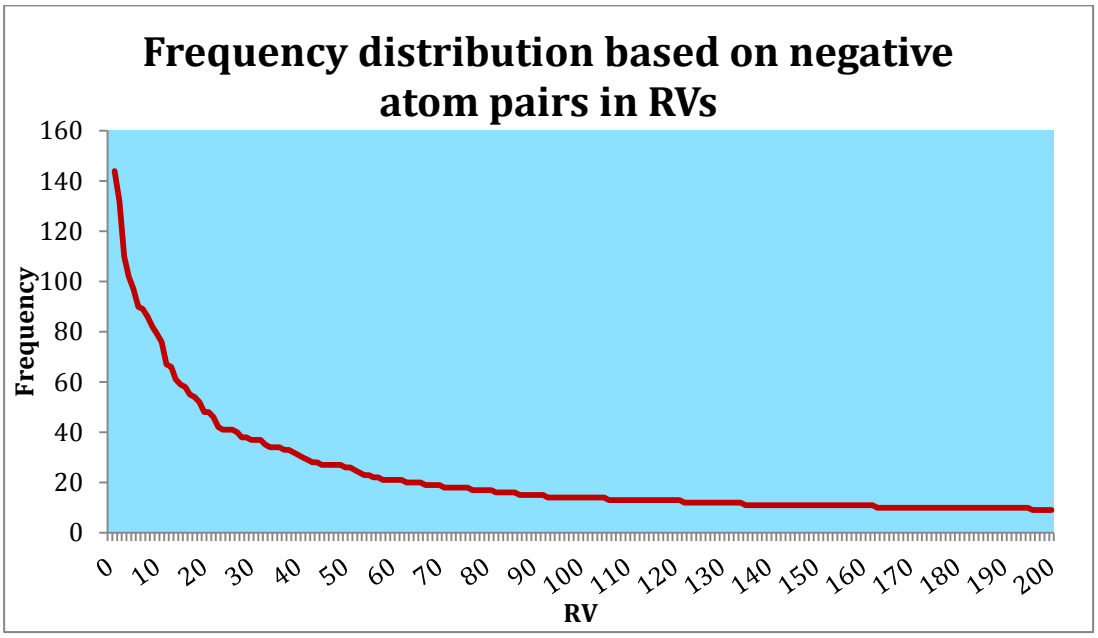


Figure 5.23: Expansion of the first 200 entries in Figure 5.22.

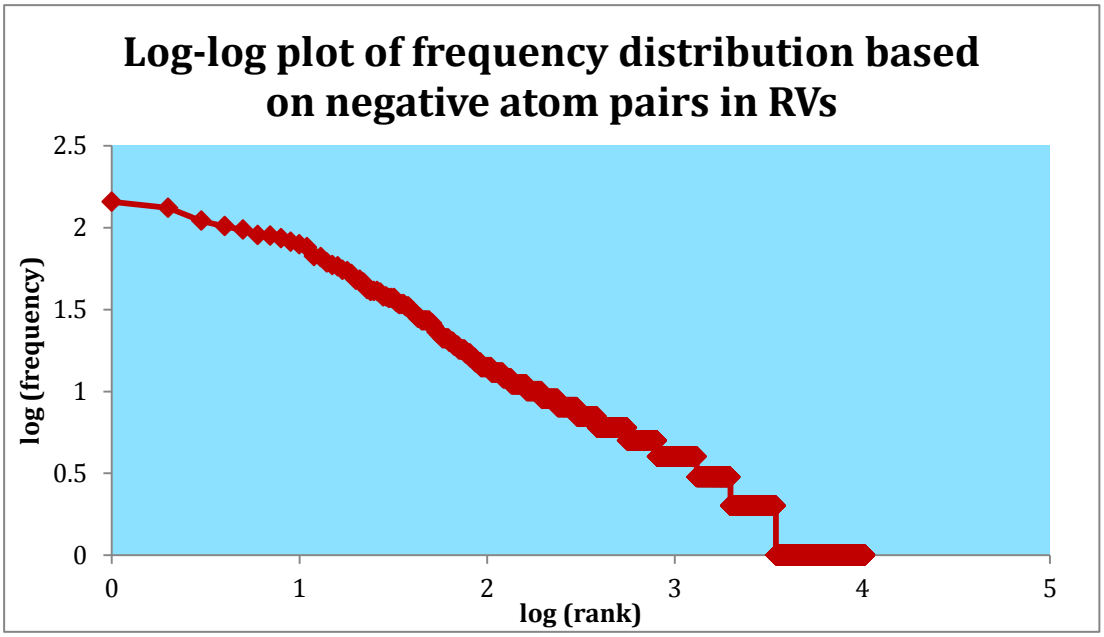


Figure 5.24: Log-log plot of the frequency distribution of negative AP2 content in JMC1.

Analysing the most popular reaction centres (Table 5.9) makes it clear that ignoring the AP3 content has reduced the number of unique groups of atom pairs. The most frequent atom pair groupings show a tendency towards nitro, amine and ether groups, as seen in Table 5.9. These functionalities are very common in SAR chemistry, as part of lead optimisation processes, and as a result tend to be heavily represented.

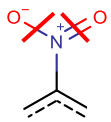
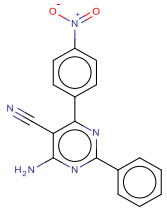
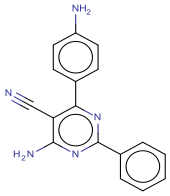
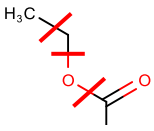
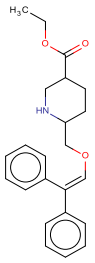
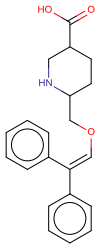
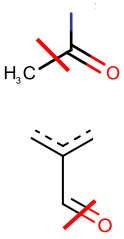
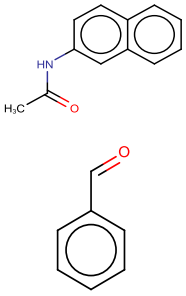
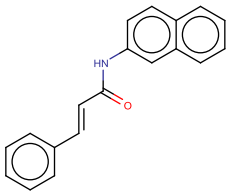
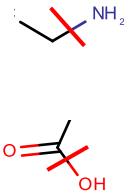
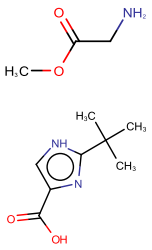
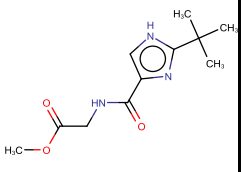
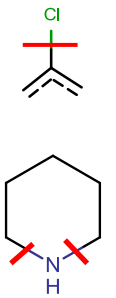
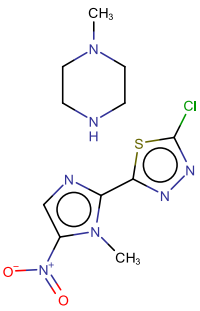
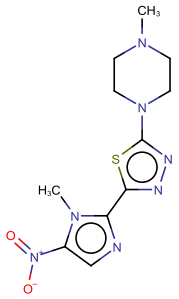
Negative atom pairs (duplicates indicate multiple entries)	Number of reactions represented	Reaction centre structure	Sample reactant(s)	Sample product
N(3,1,0)-2(1)-C(3,2,1) O(1,0,0)-2(1)-N(3,1,0) O(1,1,0)-2(2)-N(3,1,0)	144			
C(2,0,0)-2(1)-C(1,0,0) O(2,0,0)-2(1)-C(2,0,0) O(2,0,0)-2(1)-C(3,1,0)	132			
C(3,1,0)-2(1)-C(1,0,0) O(1,1,0)-2(2)-C(2,1,0)	110			
N(1,0,0)-2(1)-C(2,0,0) O(1,0,0)-2(1)-C(3,1,0)	97			
Cl(1,0,0)-2(1)-C(3,2,1) N(2,0,1)-2(1)-C(2,0,1) N(2,0,1)-2(1)-C(2,0,1)	90			

Table 5.9: Representation of the five largest groups of partial AP2 RVs. The red lines indicate bonds broken in the reaction centre structure. (Wallace, 2015)

In both frequency distributions, it is clear that there are a considerable number of reaction types that are underrepresented. A significant proportion of these occur in a single reaction in the database only, as shown by the long tail in Figure 5.24. Some examples of such reaction centres are illustrated in Figure 5.25. The low occurrence of these functional groups in the database suggests that they may be of limited use for *de novo* design, assuming that the underlying database is typical of the reactions carried out in medicinal chemistry.

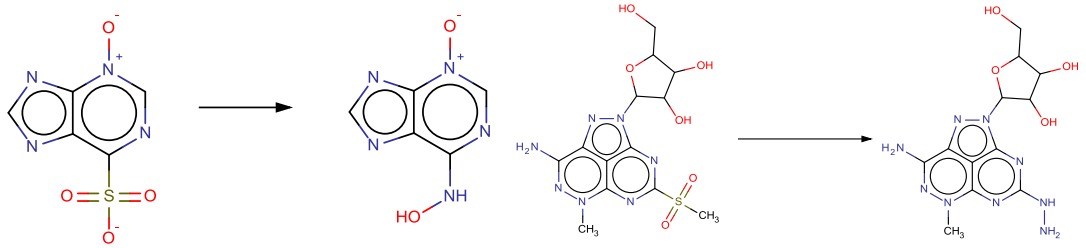


Figure 5.25: Examples of reaction centres for which only single partial RVs exist in the JMC database. (Wallace, 2015)

5.5.3.1 US Patent Reaction set

While the JMC1 and JMC2 databases are directly derived from medicinal chemistry research data, both of the frequency distributions retain a long tail of reactions that are not widely applicable. As a result, it is worth analysing other collections of reactions in order to see if this effect is common. One readily available source of reactions is the collection published by NextMove Software (Lowe and Sayle, 2014). This consists of over a million reactions from the US Patent database which are considerably more complex than the J. Med. Chem. reactions, with multiple reactants and products encoded, alongside catalysts and other agent molecules. Some of these reactions are incomplete or otherwise invalid, causing problems with loading and processing. To produce manageable databases comparable with the previous experiments, a random number generator was used to select two sets of 22,500 reactions from the pool. While these sets are similar in size to the original data set used to produce the JMC1 and JMC2 sequence databases, the fact that these reactions were randomly selected from a large pool rather than a series of related papers means that the likelihood of connections between the data will be smaller. Additionally, the presence of other agent molecules can cause problems with the reaction network creation, as there is an increased chance of selecting the wrong molecule as the intended reactant and product of the reaction. However, the existing network method can be used to categorise the data set. The first

random sample has an average sequence length of 1.26 which is considerably lower than that seen for the J. Med. Chem. data. This would make sense, given that the syntheses represented in patents are likely to involve more specialised starting materials, without the optimisation steps seen in medicinal chemistry studies. There is also a significant reduction in the maximum length of the sequences, as can be seen in Figure 5.26 and Table 5.10.

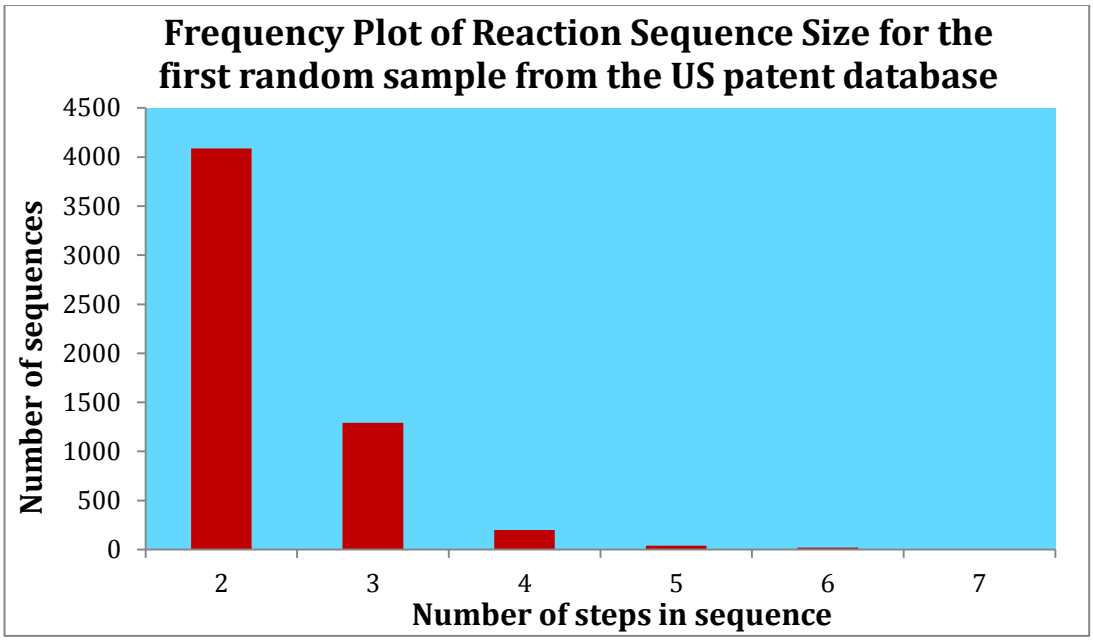


Figure 5.26: Frequency plot of reaction sequence size for the first random sample extracted from the US patent database.

Number of steps	Number of sequences
1	22500
2	4087
3	1293
4	198
5	40
6	19
7	2

Table 5.10: Table of the sequence summary for the first random sample extracted from the US patent database.

When comparing the number and size of the groups produced from the negative AP2 content with that from the previous datasets, the first collection of patent data closely resembles the J. Med. Chem. reaction sets. The frequency distribution is illustrated in Appendix A, Section A-1, showing a similar distribution to the J. Med. Chem. set. However, the precise nature of the most common RV groups shows some significant differences, as seen in Table 5.11. Overall, 5,485 unique groups of negative atom pairs were recorded, with the biggest groups representing reaction centres containing the same ether abstraction and nitro and amine processes seen with the JMC1 reaction set, but with even higher frequencies. In addition, the boronic acid and bromine reaction centre associated with the common Suzuki coupling process is significantly more common here than with the JMC1 set, highlighting its heavy usage in the kind of process chemistry represented in the patents.

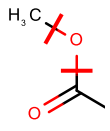
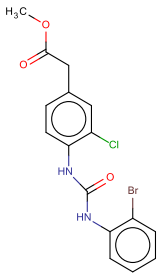
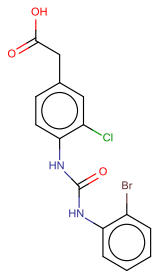
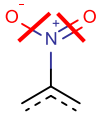
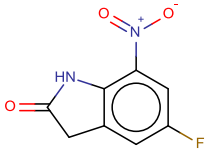
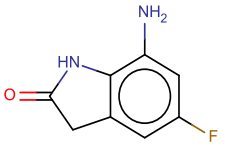
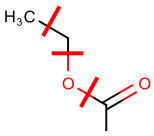
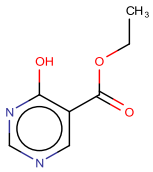
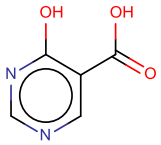
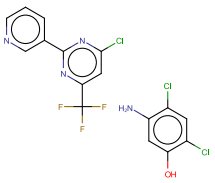
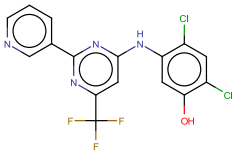
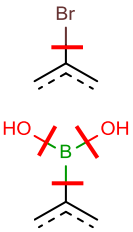
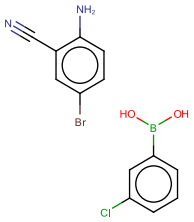
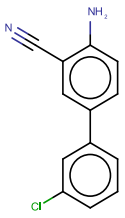
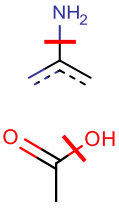
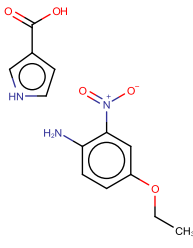
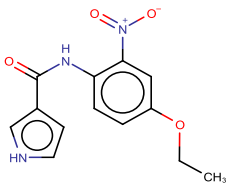
Negative atom pairs (duplicates indicate multiple entries)	Number of reactions represented	Reaction centre structure	Sample reactant(s)	Sample product
O(2,0,0)-2(1)-C(1,0,0) O(2,0,0)-2(1)-C(3,1,0)	397			
N(3,1,0)-2(1)-C(3,2,1) O(1,0,0)-2(1)-N(3,1,0) O(1,1,0)-2(2)-N(3,1,0)	329			
C(2,0,0)-2(1)-C(1,0,0) O(2,0,0)-2(1)-C(2,0,0) O(2,0,0)-2(1)-C(3,1,0)	326			
Cl(1,0,0)-2(1)-C(3,2,1) N(1,0,0)-2(1)-C(3,2,1)	171			
Br(1,0,0)-2(1)-C(3,2,1) C(3,2,1)-2(1)-B(3,0,0) O(1,0,0)-2(1)-B(3,0,0) O(1,0,0)-2(1)-B(3,0,0)	129			
N(1,0,0)-2(1)-C(3,2,1) O(1,0,0)-2(1)-C(3,1,0)	121			

Table 5.11: Representation of the five largest groups of partial AP2 RVs. The red lines indicate bonds broken in the reaction centre structure. (Wallace, 2015)

The heavy skew in the frequency distribution shows that the degree of underrepresentation of certain groups remains high. Some examples of the reaction centres for which only one example exist in the collection are shown in Figure 5.28. These represent straightforward reactions which are not that common in drug design, particularly with the complexity of some of the starting materials used, such as in the bottom example. However, the top example represents a fairly straightforward reduction using Diisobutylaluminium hydride (DIBAL, a bulky reducing agent), which would be relatively commonplace. In fact, there are 26 examples of this kind of reaction in the database, but all have differing partial RVs. Part of the problem in this case is the unusual method of reporting the use of DIBAL within the database, not using the traditional 'bridged' layout around the central aluminium atom. This can also occur with other reagents and leaving groups, where more obscure structures lead to a number of singletons in the data set.

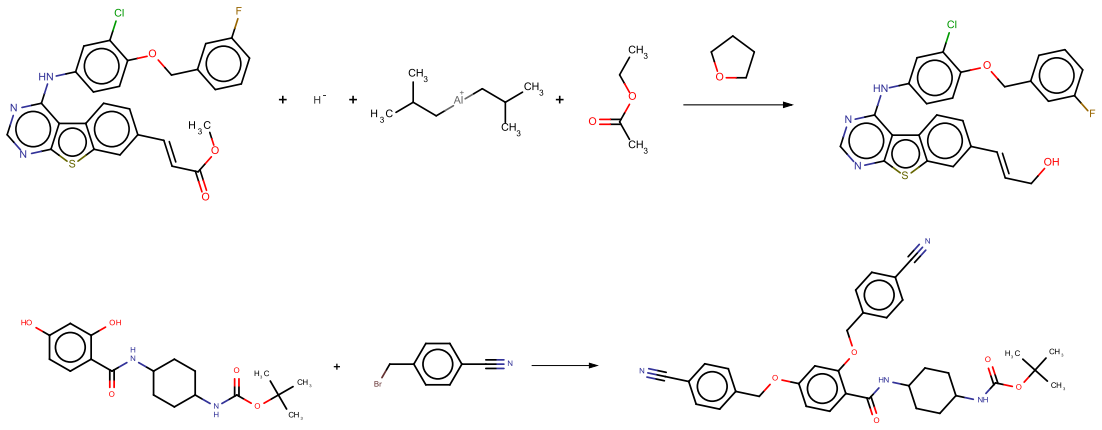


Figure 5.27: Examples of reaction centres for which only single examples exist in the first US patent database. (Wallace, 2015)

If the random sampling is truly representative of the data set, the distribution and content of the partial RVs of the second randomly selected set should be similar to the first set, with heavy representation of ether abstraction and nitro group conversions. This data set has an even more pronounced bias towards shorter sequences, with an average sequence length of 1.20 steps, as seen in Figure 5.29 and Table 5.12.

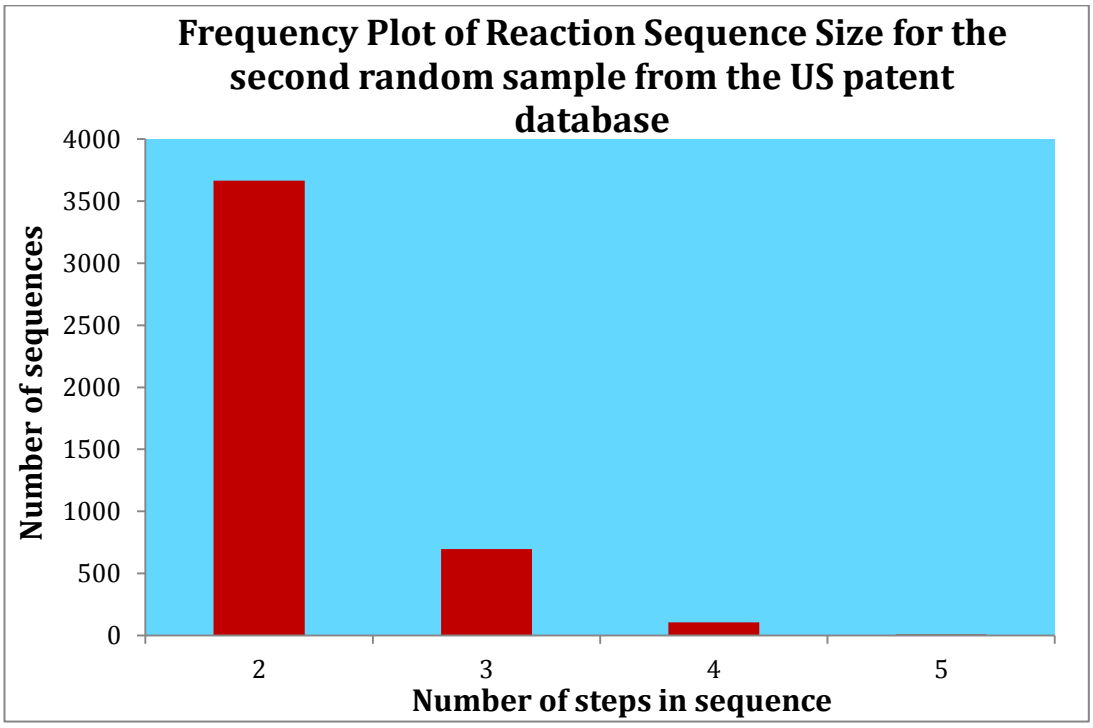


Figure 5.28: Frequency plot of reaction sequence size for the second random sample from the US patent database.

Number of steps	Number of sequences
1	22500
2	3667
3	697
4	105
5	7

Table 5.12: Table of the sequence summary for the second random sample from the US patent database.

In this set, 5,511 groups were recorded, but overall there is very little difference between this set and the previous example from the patent database. The frequency distribution of the atom pair groupings in this set is shown in Appendix A, Section A-2, along with information regarding the most common reaction centres. The fact that the log-log plots for both samples are nearly identical would indicate that the random sampling is indeed indicative of the data collection as a whole. A comparison between the two sets is shown in Appendix A, Section A-3.

5.6 Conclusions

In this chapter, the creation of a network of reactions was described by linking individual reactions according to common reaction components. This network approach was then used to generate reaction sequences for different collections of medicinal chemistry reactions. The reaction collections were also grouped by first generating RVs and grouping them according to identical negative atom pairs. The collections were shown to have significant biases towards particular functional groups. As might be expected for reactions used in medicinal chemistry there was a bias towards reactions that act on aromatic rings and on amine and carbonyl functionality. The approach was then extended to collections of reactions from the US Patent database, demonstrating the ability to assemble usable material from any reaction collection. In the next chapter, the reaction sequences will be used to construct a variant of the reaction vector method capable of representing the entire sequence as a single transformation for *de novo* design purposes.

Chapter 6:
Reaction sequence encoding

6.1 Introduction

The work in Chapter 5 established a way of creating reaction sequences from a collection of single step reactions via the formation of a network. However, it is not yet possible to use these in *de novo* design, as the existing RV format is limited to encoding a single step at a time. In this chapter a revision of the RV algorithm will be reported that permits the encoding of whole sequences as a single transformation. This should eliminate any issues caused by the application of multiple RVs in molecule optimisation methods, while also being significantly faster for enumeration. This new method will then be compared with the existing RV tool in terms of the total number of molecules generated in a *de novo* context, as well as their novelty. Additionally, the various sequences produced from the reaction collections will be analysed via their atom pair content as in Chapter 5, to see if the same skew in functionality is present as in the RV case.

6.2 Handling of reaction sequences

In Chapter 5, it was demonstrated that reaction sequences can be extracted from a reaction network by tracing paths through the network and recording an ordered list of the nodes and edges from start to finish. With these sequences recorded, attempts can be made to develop methods to represent them in a manner that makes them effective within the existing *de novo* framework. Ideally, these approaches would permit structure generation via application of all relevant stored sequences regardless of length in a manner that is faster than applying the individual reaction vectors from the sequence in turn, while remaining relatively easy and quick to implement. Thus, the aim is to store all of the information required to generate the product of a sequence in a single transformation step, in a comparable manner to reaction vectors. If the format of the data storage could be made to be compatible with that of the original method, this would be of additional benefit, as this would permit the original tools and workflows to

be appropriated, with only minimal changes. In the next section, three such methods are described and validated.

6.2.1 Reaction sequence vectors (RSVs)

6.2.1.1 Direct Method

The simplest method of representing reaction sequences that is compatible with the *de novo* toolset is to create an artificial chemical reaction in which the start and end points of the sequence are directly linked (see Figure 6.1). The resulting 'compressed sequence' can then be converted into a single difference vector and stored as before, giving an advantage over the original setup which would require as many individual vectors (and therefore structure generation iterations) as there are steps in the sequence, as seen in Figure 6.2. Using the same format as the original vectors is a considerable benefit, as it permits the reuse of the existing structure generation code, greatly simplifying the further development of the tool set. For clarity, and to avoid confusion with the earlier forms of reaction vector (RV) reported this reaction vector form will be referred to as a reaction sequence vector (RSV).

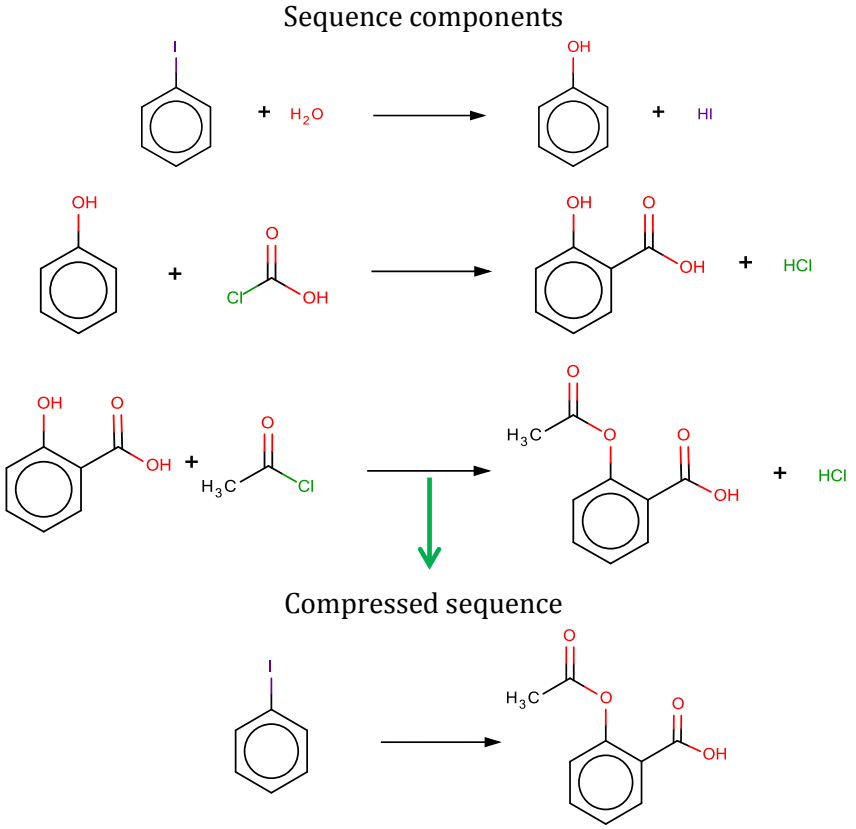


Figure 6.1: Illustration of the sequence compression process using a sequence from JMC2. (Wallace, 2015)

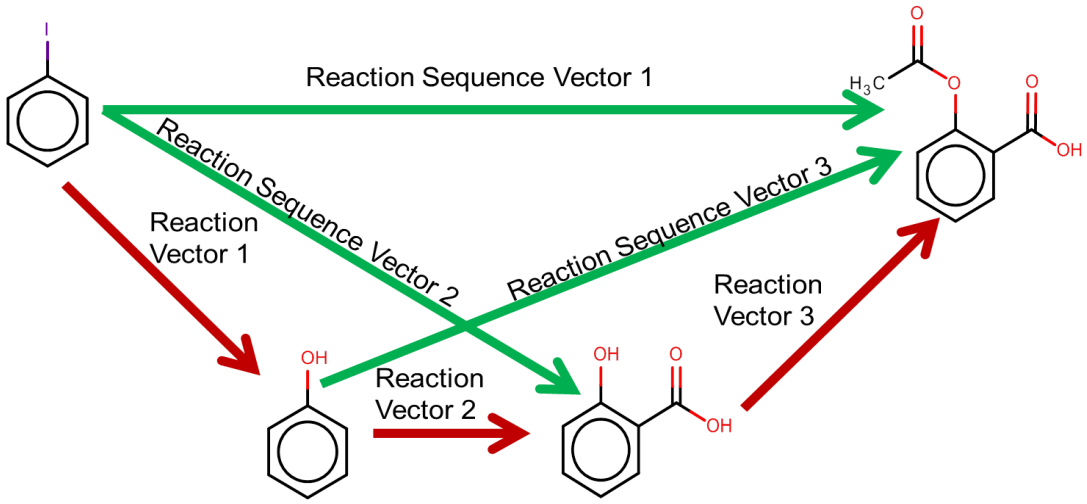


Figure 6.2: Comparison of the reaction vector (RV) and reaction sequence vector (RSV) based approaches to structure generation. The RSV method aims to enable direct transformation from start to end point of the sequence, without having to generate each individual reaction in turn. (Wallace, 2015)

As with the original RV method, the creation of the RSV is straightforward as it is a simple list of the differences in the atom pair descriptors between the two sides of the compressed reaction sequence. The RSV resembles an RV in structure, albeit with a larger list of descriptors stored, due to the increased number of changes encoded. As with the original reaction vector method, some additional information is required in the form of a recombination path to ensure it can be quickly and effectively applied to other molecules to generate the new products. To create this, the reverse fragmentation approach as described in Section 4.3.2 is applied to the start and end molecules in the sequence and an ordered list of bonds or fragments (the recombination path) is created. The path is then stored with the RSV in the database as before, to enable the final product to be generated from the starting material in an efficient manner.

The main issue with using the reverse fragmentation approach to generate the recombination path data is that as the molecules in the reaction get larger, so does the number of possible sets of fragments that can be generated, as a result of multiple matches to the atom pairs. Particularly complex molecules, or those with a large dissimilarity between reactant and product, need considerably larger amounts of memory as the list of fragments stored increases exponentially, eventually reaching a

point where it is impossible for the system to store all of the combinations. Should this occur, or if the reverse fragmentation method fails for any other reason, the breadth first search method (Section 4.3.1) is used instead to generate recombination data by constructing the necessary ordered bonds atom-by-atom in a brute-force approach. This process is considerably slower than the reverse fragmentation approach, and prone to failure, especially when there is a large difference between the reactant and product. Even if the RSV is confirmed as suitable for reproducing the sequence on which it is based, there are still potential problems. In the earlier work it was assumed that in order for an RV to be effective in generating new molecules, there is a need for all non-hydrogen atoms to be balanced on both sides of the reaction to ensure that the correct product can be generated from the constituent molecules (Patel et al., 2009). This would imply that using an RSV created through direct connection of the molecules at the start and end of a multi-step reaction is likely to fail when used for novel structure generation due to missing atom pairs as a result of the absence of reagent information. While the reverse fragmentation approach does not necessarily require this atom balance, no assessment was performed of the effectiveness of the tool for unbalanced reactions. Therefore, two other methods of vector preparation were also investigated.

6.2.1.2 Additive and subtractive methods

In order to add reagent information, two different methods can be used. One is based on the addition of molecules to create reagents, while the other is based on subtraction of the starting material from the product. These are illustrated in Figure 6.3, for a simple, two step sequence, alongside the direct method previously discussed. In both cases, these approaches utilise the full reaction data encoded as part of the reaction network (Section 5.4), and the reactant and product assignments already obtained during the network processing.

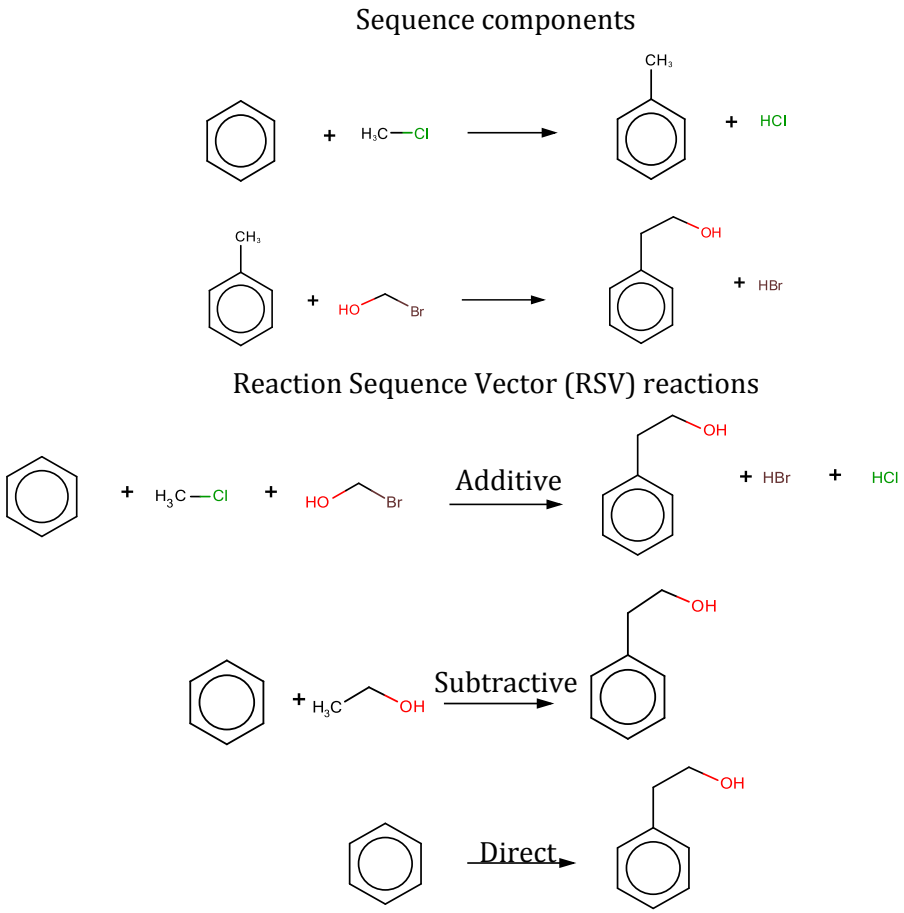


Figure 6.3: Examples of the different methods of generating reaction sequence vectors from a typical two step reaction sequence. (Wallace, 2015)

In the additive approach, the individual reagents in each step of the sequence are added to the start molecule with the RSV being the difference between the atom pairs in the product and the sum of the atom pairs in the starting molecule and all of the reagents in the sequence. The reagents for each reaction step are identified by removing the designated reactant, and then collecting the molecules that remain on the reactant side. While including all of the reagent molecules does not strictly balance the reaction, it ensures that all of the relevant atom types needed for the transformation are present in the correct numbers.

In the subtractive approach, a "super reagent" is created by subtracting the starting molecule from the product of the sequence, and added to the left hand side of the reaction. To perform the subtraction, a maximum common subgraph comparison is made between the two molecules using the Indigo library (EPAM Life Sciences) (Figure 6.4). The super reagent is formed by subtracting the maximum common subgraph (highlighted in red) from the product to generate a substructure. The RSV is then the difference between the product and the combined starting molecule and super reagent.

In this case, the reactant and product designations come directly from the reaction network, with only the first and last reactions of the sequence being considered.

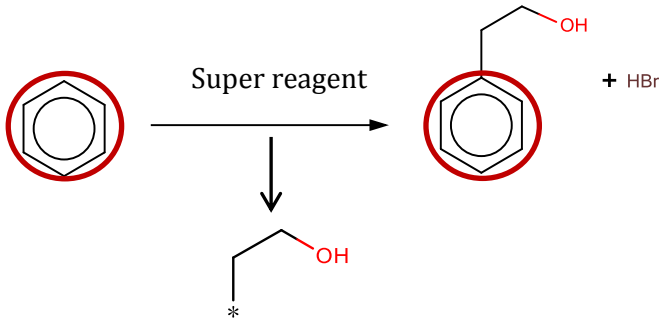


Figure 6.4: Illustration of the subtractive method (maximum common subgraph highlighted). Note that side products are not included. The * indicates the attachment point. (Wallace, 2015)

6.2.1.3 Comparison of Methods

In order to determine which of these three RSV generation methods is the most effective for sequence encoding and reproduction, all three were validated using 6,500 two step sequences found within the smaller reaction network produced from over 25,000 reactions from the Journal of Medicinal Chemistry (JMC1, Section 5.5). RSVs were prepared from the sequences according to each of the three methods.

For each method and each sequence, the RSV was generated and then applied to the known starting molecule using the *de novo* tool and the resulting products were assessed against the expected product. If the correct product molecule was found, the sequence was said to have been successfully reproduced. The RSV approaches were also compared with the original reaction vector method, in which RVs generated from the individual reaction steps were applied in turn. For consistency, these single reactions were processed in the same manner as the sequences i.e. only one reactant and one product molecule were included, and reagents were removed. The results of these experiments are presented in Table 6.1 and Table 6.2. To aid comparison, two versions of the RVs were also used: one where reagent data was included when generating the RV and one where the RV was generated without reagent data, i.e., where the reactions were reduced to a single reactant and product prior to generating the RV.

	Additive method	Subtractive method	Direct method	Executing individual reactions in turn (original approach, with reagent data)	Executing individual reactions in turn (without reagent data)
Number of sequences successfully reproduced	5080	5123	5305	5720	4509

Table 6.1: Table comparing methods of reaction sequence vector generation for 6,500 two step sequences.

	Additional sequences reproduced compared to method		
Initial Method used	Additive	Subtractive	Direct
Additive		510	380
Subtractive	495		348
Direct	587	570	

Table 6.2: Table demonstrating where some sequences are reproduced in one method, but not another. The rows and columns represent the unique sequences reproduced in one method compared to the other, (e.g. 510 sequences were reproduced in the subtractive method that were not produced in the additive method).

Table 6.1 shows that the original RV method with reagents reproduces more sequences than any of the new approaches. However, considering the RSVs, the direct approach is the most effective for reproducing the reaction sequences, despite the heavy atom imbalance that results from excluding any additional reactant data. This would imply that additional reactant data is not essential to the structure generation process in this case, although the RV case shows better results where reagent data is present. The difference results from the fact that, for certain sequences, the reagent data produced exceeds the amount of material that can be represented within the vector framework, which is designed to support a maximum of three separate molecules on the reactant

side. In cases where this is exceeded, the vector cannot be reproduced and therefore the sequence is recorded as failing. It is interesting to note that Table 6.2 shows that each approach is able to accomplish the reproduction of some sequences that are not possible by the others. In the cases where the direct method is outperformed, this is due to the presence of the additional reagent data.

6.3 Reaction sequence validation

6.3.1 Sequence reproduction tests

In order to further determine the effectiveness of the direct method for *de novo* design, the sequence reproduction experiment was repeated and extended to a randomly selected subset of reaction sequences from the original JMC1 data set (Chapter 5). By splitting the experiment into separate groups according to sequence length, it should be possible to determine if this has any effect on the ability of the RSV method to encode and reproduce the contained chemistry. It is expected that the degree of success in reproducing sequences will be inversely proportional to the sequence length, as the more steps there are, the greater the difference between the start and end points, and thus the greater likelihood that the necessary material will be missing due to ambiguity in assigning atom pairs. A breakdown of the results by sequence is listed in Table 6.3.

Number of steps	Sequences successfully reproduced	Percentage reproduction
2	5240/6500	80.6%
3	1471/2731	53.9%
4	559/1172	47.7%
5	187/462	40.5%
6	71/189	37.6%
7	23/66	34.8%
8	10/19	52.6%
9	5/10	50%
10	2/7	28.6%
11	2/4	50%
12	1/3	33.3%

Table 6.3: Table showing the success rate for reaction sequence reproduction.

While the overall figures are not particularly impressive (an overall success rate of 55.5% for the database as a whole), the experiment shows that there are a number of issues with memory allocation that can be worked around to improve the quality of the results. At this point, given the general failure rate, it is difficult to determine whether the apparent relation between the reproduction success and sequence length is significant.

6.3.2 Improvements to the algorithm

On further analysis of the RSV failures, it became clear that the sequences that failed to reproduce successfully were due to the corresponding RSV not being generated. This is due to the procedure for creating the recombination path following generation of the RSV (Section 6.2.1) failing. The vast majority of these failed sequences triggered error messages associated with being 'too large' for the reverse fragmentation to handle, with the fragment combinations exceeding the memory threshold. As processing power has increased considerably since the vector code was originally designed, it is possible to simply increase the amount of memory available to store the combinations, and thus

permit these complex sequences to be handled by more powerful computers. This slows down the generation of the vector database due to the greater number of combinations that can be tried, but does not have significant impact on the time taken for the structure generation process itself. With the code amended to take this into account, a new experiment was performed using the sequence database to see how much of an improvement has been made.

By running the same sequences through the code with the expanded memory allocation, the overall success rate increased from 55.5% to 74.7%, which is a significant improvement. However, there remain a number of errors that cannot be resolved via memory related fixes alone. These are largely related to the recombination method producing the wrong molecule (or no molecule at all) due to errors in the fragmentation processes. By adding data logging features to the vector generation code, it was possible to trap errors during the recombination path generation step without going through the structure generation process, and thus determine the cause of these issues. The overall results of the experiment are listed in Table 6.4, categorised by the reason for the failure.

Type of failure	Number of reported failures
Fragments generated (forward and reverse), but no path.	1542
Forward fragments invalid/empty.	633
Reverse fragmentation fails due to memory issues, forward cannot find path.	283
Path finding times out.	419
No valid fragments generated.	240
Total	3117

Table 6.4: Report of failures in the sequence vector system.

There are five different categories of reproduction failures. The main causes of failure are where the recombination path code simply times out, without any results being found, or the algorithm fails quickly without any fragments being generated. Usually this is due to the differences between the sides of the reaction being too great to result in meaningful fragmentation via any of the existing methods, resulting in an attempt to

build the recombination path atom-by-atom. While increasing the memory allocation for fragments can improve matters in some cases, in order to permit all of these examples to be encoded will require far more memory than is available with standard computer hardware, and as such, these problems remain unfixable.

The same issue can also manifest itself in a slightly different way, where one side of the reaction fragments correctly, while the other side does not. This seems to be more of a problem when all examples of a particular atom environment change bond order and type (due to cycle formation or condensation, for example) over the course of the reaction. If this situation cannot be reversed through simple fragmentation of a particular bond, there will be insufficient material to permit a correct reconstruction based on the stored descriptors. An example of such a reaction is presented in Figure 6.5.

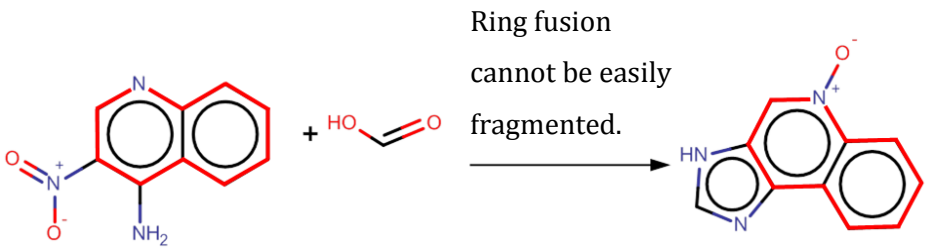


Figure 6.5: Example of a failing reaction in the data set, where ring fusion confuses the fragmentation code (MCS highlighted). (Wallace, 2015)

In order to increase the number of sequences that can be processed, an attempt was made to add further reagent data to enable processing of the sequence using the subtractive method (Figure 6.6). Reactions that fail processing with the direct method are passed to the subtractive method and processed again with the super reagent added. This leads to successful reproduction in the majority of cases. Note that in the example in Figure 6.6, the super reagent does not have a fully satisfied valence as no atoms are added to the fragments after calculation in order to keep the reaction balanced.

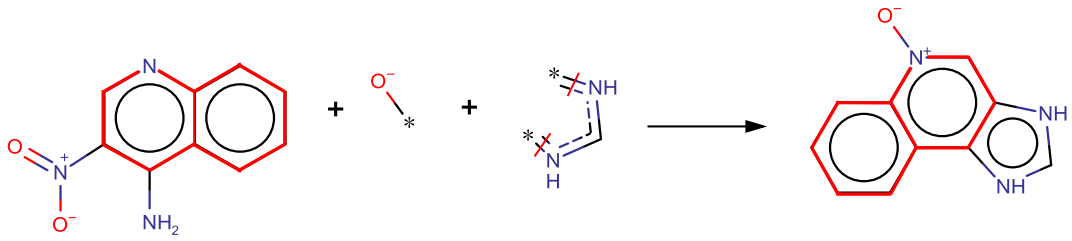


Figure 6.6: Revised version of the reaction from Figure 6.5, using the super reagent data to generate the ring fusion fragment (MCS highlighted).The red lines in the super reagent indicate bonds broken, the asterisks represent points of attachment(Wallace, 2015)

After increasing the memory allocation to the algorithm, fixing an apparent bug with the database handling and permitting the subtractive RSV method to generate additional reagent material where necessary, another attempt was made to reproduce all of the reaction sequences contained within the J. Med. Chem. subset. This is summarised in Figure 6.7, with the full details in Table 6.5. The second run was far more successful, with 8,582 sequences successfully reproduced, giving a 76.3% success rate overall. Over the whole of the sample set, it appears that there is no significant relationship between the number of steps in the sequence and the rate of success, although the number of sequences of five steps or above is so small, it is difficult to draw strong conclusions. For those sequences with over 400 examples (the solid line), a slight downward trend can be observed. Where there are fewer sequences (the dashed line) it is difficult to observe any trend.

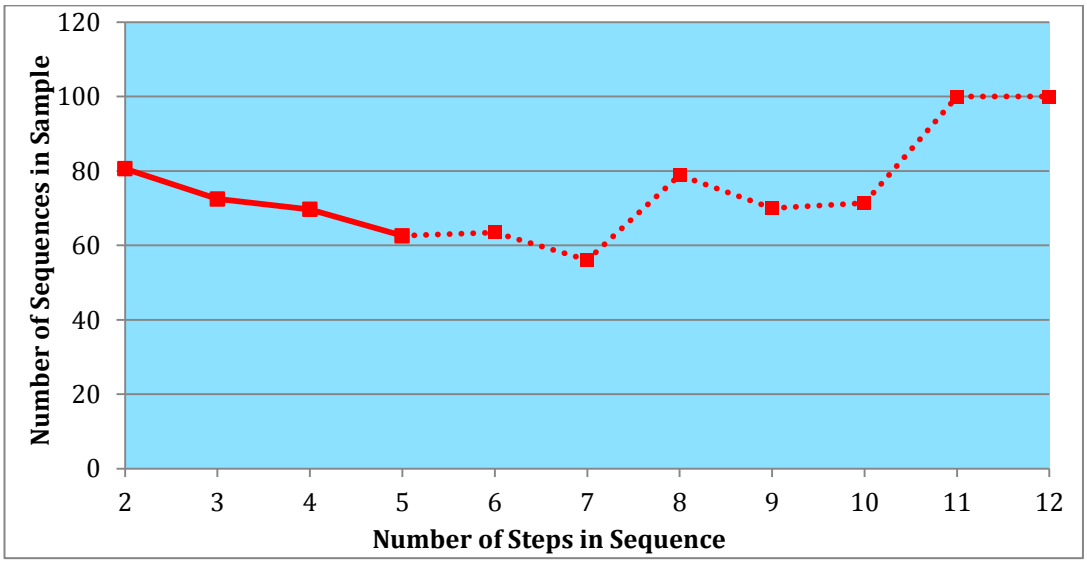


Figure 6.7: Graph showing the relationship between sequence length and percentage success. The solid line represents the sequence lengths for which there are sufficient numbers to draw conclusions over trends, while the dashed lines have too few to be useful.

Number of steps	Sequences successfully reproduced	Percentage reproduction
2	5240/6500	80.6%
3	1980/2731	72.5%
4	817/1172	69.7%
5	289/462	62.6%
6	120/189	63.5%
7	37/66	56.1%
8	15/19	78.9%
9	7/10	70%
10	5/7	71.4%
11	4/4	100%
12	3/3	100%

Table 6.5: Table showing the success rate for reaction sequence reproduction with the revised method.

Because of the lack of long sequences in the original database, the full content of the expanded reaction network JMC2 (Section 5.5) was used to carry out the same

experiment. This network includes all sequences that start or finish partway through a longer path, and as such includes the synthesis of all possible intermediate molecules. As a result, the network contains 124,354 reaction sequences generated from the 22,694 J. Med. Chem. reactions previously curated, as shown in Figure 6.8 and Table 6.6. With larger numbers of sequences available, it is possible to observe a general downward trend as the sequences get longer and more complex, as can be seen for sequence lengths 2 to 13 which have over 400 sequences stored (the solid line). Of the reaction sequences stored (including single step sequences), 93,557 give unique vectors, and 92,767 can be successfully reproduced, giving an overall reproduction rate of 99.2%, which is considerably better than the subset of JMC1 previously studied.

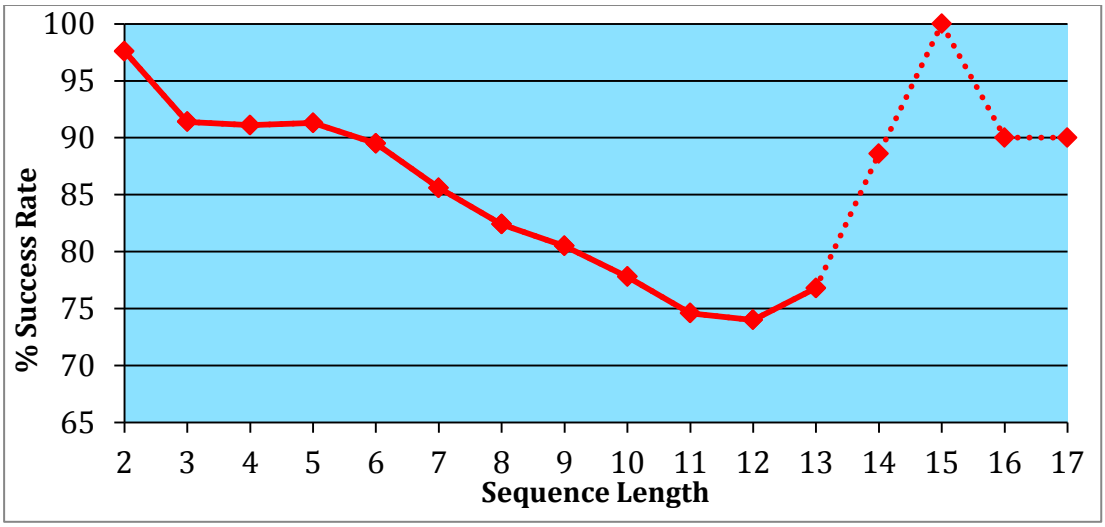


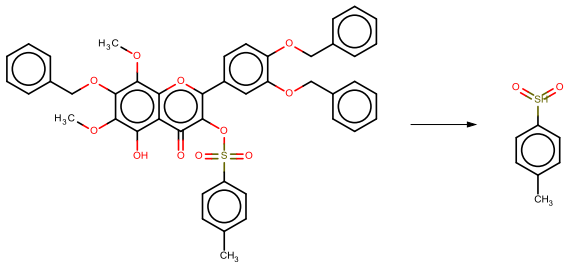
Figure 6.8: Graph showing the relationship between sequence length and percentage success for the expanded network. As with Figure 6.7, the solid line represents those sequences for which there are sufficient numbers to draw conclusions over trends, while the data points connected by the dashed lines have too few examples to be able to generalise.

Number of steps	Sequences successfully reproduced	Percentage reproduction
2	7808/8000	97.6%
3	7890/8632	91.4%
4	9522/10452	91.1%
5	11081/12137	91.3%
6	12100/13519	89.5%
7	11279/13176	85.6%
8	8875/10771	82.4%
9	6004/7458	80.5%
10	4480/5758	77.8%
11	3404/4563	74.6%
12	1901/2569	74.0%
13	779/1014	76.8%
14	282/318	88.6%
15	187/187	100%
16	168/187	90%
17	11/12	90%

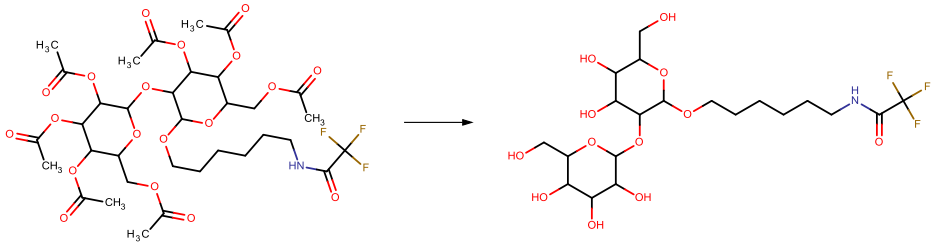
Table 6.6: Table showing the success rate for reaction sequence reproduction with the revised method, using the expanded network.

Further efforts to improve the performance of the method were considered, but it was felt that the potential gains that would result would be outweighed by the complexity involved in modifying the algorithm at this stage. In particular, the types of reactions that the current code struggles with are those which are less appropriate from a drug design standpoint, due to the components having overly high molecular weights, or being highly complex in terms of the number and types of bonds, making fragmentation difficult. Some examples of such reactions are listed in Figure 6.9. It should be noted that all of these reactions do not participate in sequences, and have no connection to others within the reaction network.

Requires fragmentation of large molecule in multiple places (decomposition)



Requires handling of large fragments



Overly complex end product, cannot fragment correctly

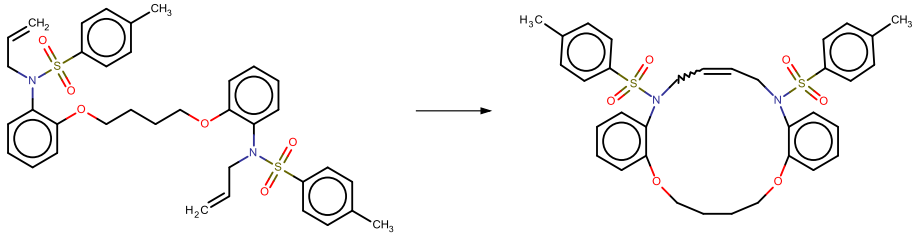


Figure 6.9: Examples of reactions that fail using the *de novo* algorithm, with the reasons for failure. (Wallace, 2015)

These highly specialised reactions are unlikely to offer many opportunities for *de novo* use, and as such their omission should not adversely affect the ability of the tool to generate useful novel molecules.

6.3.3 Comparison of RV and RSV for *de novo* design

With two different approaches to performing structure generation available (the individual RVs and the corresponding RSV), the number of molecules generated by each approach is likely to vary in size considerably. As the RSV approach misses out molecules generated by the intermediate steps, and the reagent data is explicitly encoded within the RSV, the amount and novelty of the produced molecules is likely to be much smaller. This could potentially lead to results that are insufficiently diverse to be worthwhile in a *de novo* context. To compare the numbers of novel molecules generated, the 6,379 three step sequences from the JMC2 data set that produce unique RSVs were extracted from the database. All of these unique RSVs were applied to the

1,043 unique starting materials from the initial reactions in the sequence, using the direct RSV method. The number of unique, novel (i.e. not present in the network) molecules was recorded. This process took 10 minutes to execute on an i7 workstation. This experiment was then repeated for the RVs of each reaction step in each sequence in turn. For each of the sequences, the individual RVs (including the original full reagent data) in each step were extracted and duplicates were removed. Each of the 760 unique RVs in the first set of reaction steps was applied to each starting material to give a set of 1,820 single step products. The 376 unique RVs from the second steps of the sequences were then applied to each of the unique products to give a set of 12,128 two step products, before repeating this again for the 819 RVs in the third set (this process is illustrated in Figure 6.10). The number of unique products following the final step was then recorded, and is summarised in Table 6.7. Because of the increased numbers of intermediates involved in this process, the overall execution time was considerably longer, taking approximately 90 minutes to complete on the same i7 workstation as the RSV experiment.

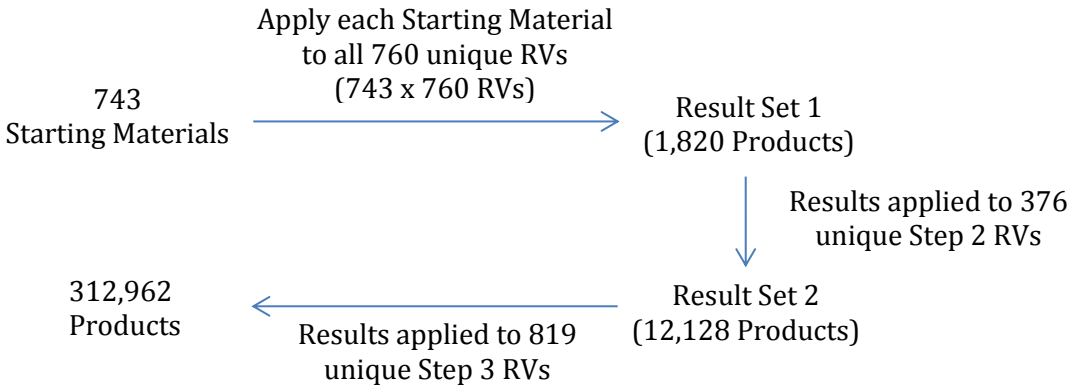


Figure 6.10: Illustration of the stepwise RV experiment.

Method of structure generation	Unique, novel molecules generated
RSV approach	33,976
RV approach	312,962

Table 6.7: Comparison of the result populations generated by the different structure generation approaches.

Because of the way this experiment is conducted, not all of the product molecules from the RV approach are generated from applying three consecutive reactions to the

original starting materials (for example, some may be the product of only one or two reactions, starting from one of the intermediate points). However, there are a considerable number of the generated products that can be tracked through an entire three step sequence, as shown in Figure 6.11.

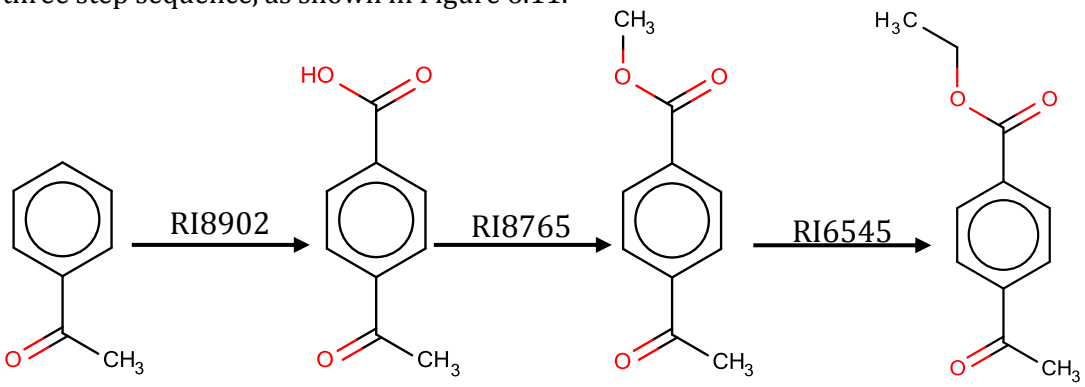


Figure 6.11: A sample route seen in the stepwise RV. (Wallace, 2015)

While the individual reaction based approach produces considerably more molecules, the population from the sequence based method still appears to be of sufficient size and scope to be useful from a *de novo* standpoint. However, for *de novo* design purposes, longer sequences needed to be analysed using the RSV method, with greater focus placed on reviewing the diversity of the molecules produced.

6.3.4 Molecule novelty assessment

In order to assess the applicability of the RSV method to *de novo* design as a whole, the full JMC2 reaction network of 93,557 unique sequence vectors (as mentioned in Section 6.3.2) was tested for its ability to generate novel molecules.

The goal of the experiment was to determine the amount of product novelty that can be produced on application of the reaction sequence vectors, and to compare this with what can be achieved with the original reaction vectors. To assess this, the structure generation tool was used to apply the JMC2 database of RSVs to a set of 500 starting material molecules selected at random from the reaction database. This process took an average execution time of 25 minutes per starting material on the i7 workstation mentioned previously, with the overall time being proportional depending on the number of vectors that were applicable. The number of novel unique molecules produced (i.e. those absent from the reaction network) was then recorded. A summary of the distribution of the molecular weights of the 500 starting molecules is presented in Figure 6.12, with further information regarding the hydrogen bond donors and

acceptors present in Figure 6.13 and 6.14. By selecting starting materials at random, the intention was to cover as much of the whole database in terms of functional group properties and general characteristics as possible. As generated, the sample favours hydrogen bond donors over acceptors, and is skewed slightly towards molecules that have molecular weights below 200g mol^{-1}. These properties should result in a series of starting molecules very similar to traditional drug precursors.

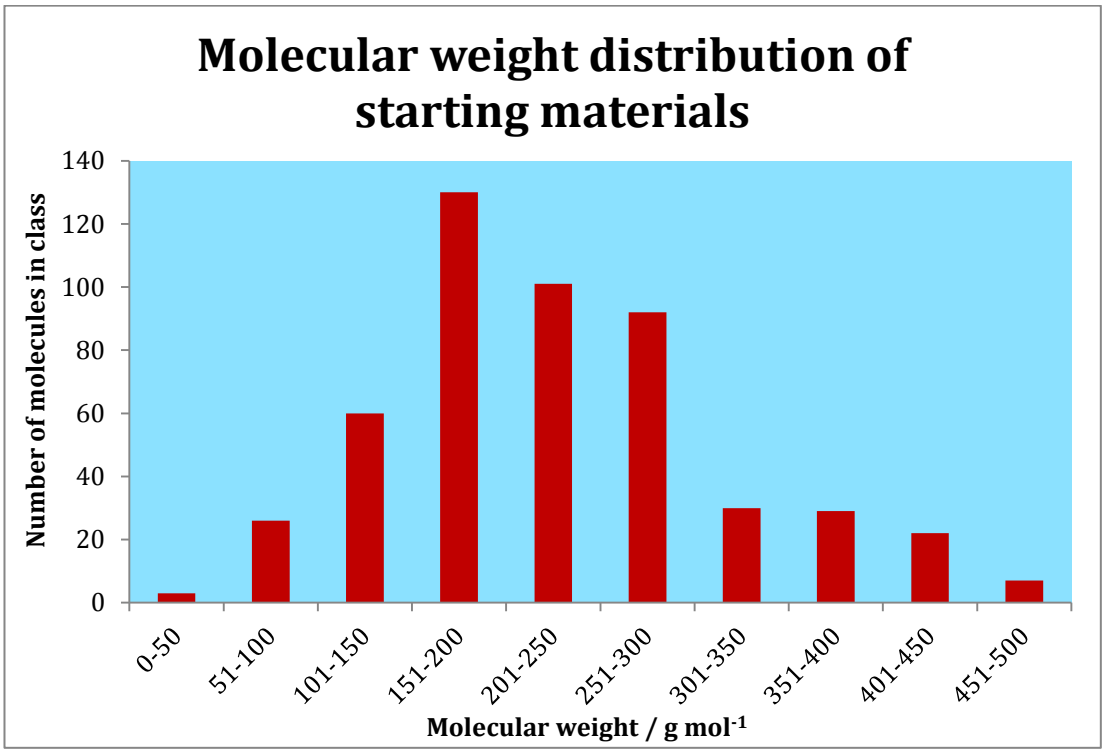


Figure 6.12: Molecular weight distribution for the 500 starting material molecules.

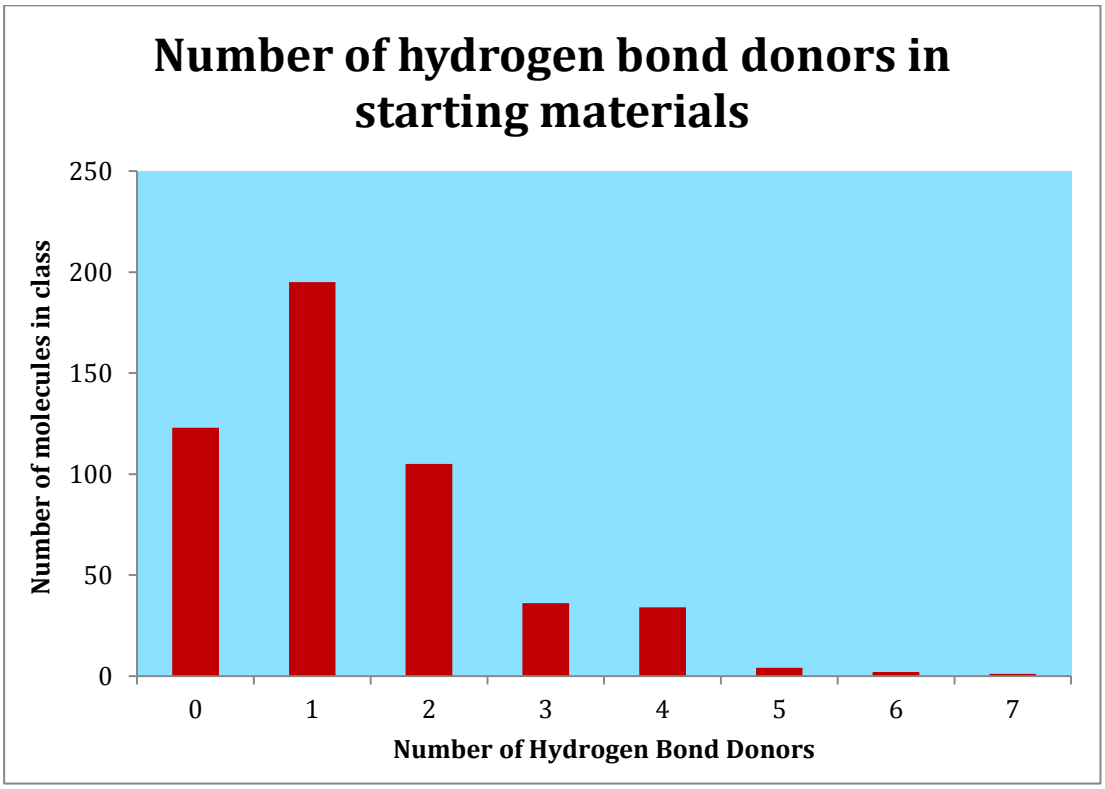


Figure 6.13: Hydrogen bond donor distribution for the 500 starting material molecules.

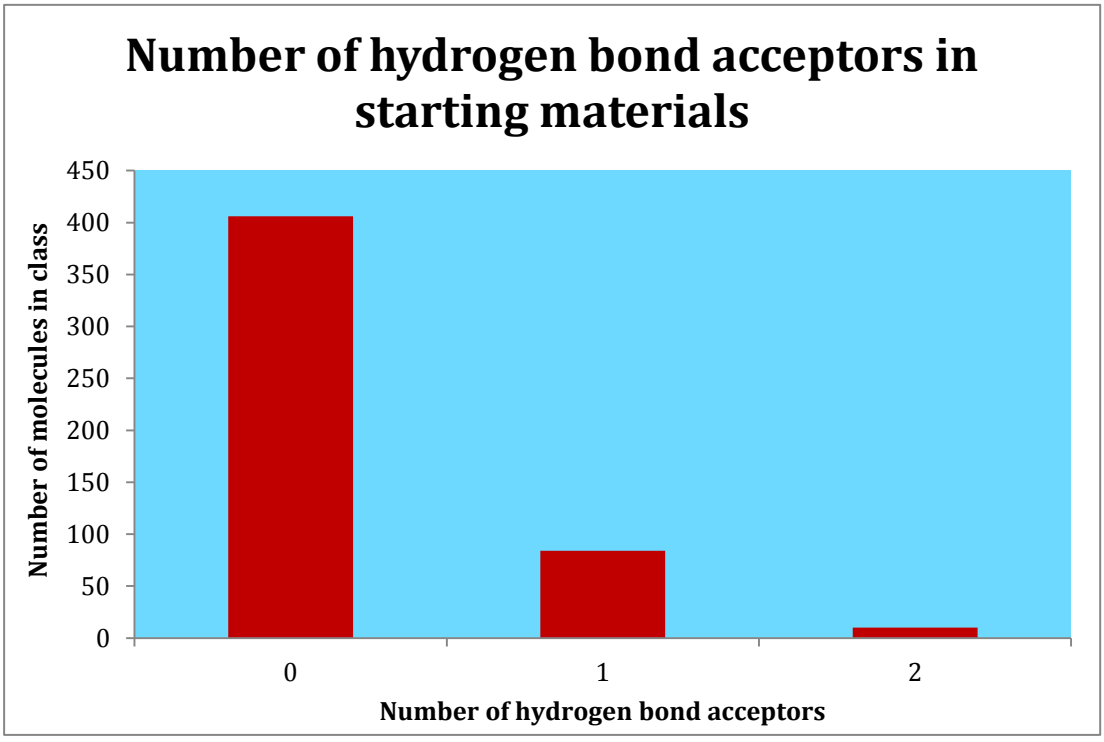


Figure 6.14: Hydrogen bond acceptor distribution for the 500 starting material molecules.

A summary of the products from this experiment is given in Table 6.8, along with a frequency plot ordered by number of product molecules per starting material in Figure 6.15. Overall, an average of 137 molecules is produced per starting material, with 68,703 unique products generated in total. The largest number of unique products generated was from a simple alkene. The large number of products generated from this particular compound is due to the presence of RSVs that can act on both the saturated atoms in the chain (of which there are several) as well as the unsaturated atoms. Figure 6.16 shows a histogram of the average number of products sorted by the number of steps in the sequence used to generate them, while Figure 6.17 shows a frequency plot of the number of products generated per RSV.

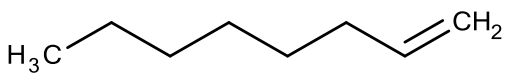
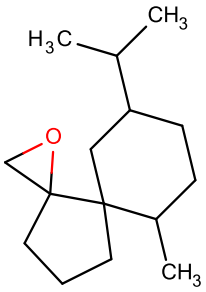
Most molecules generated	1,742 
An example molecule which results in no product molecules being generated.	0 
Total number of molecules generated	68,703
Average	137.4 molecules per starting material

Table 6.8: Summary of the results, applying the RSVs in JMC2 to 500 randomly selected starting materials. (Wallace, 2015)

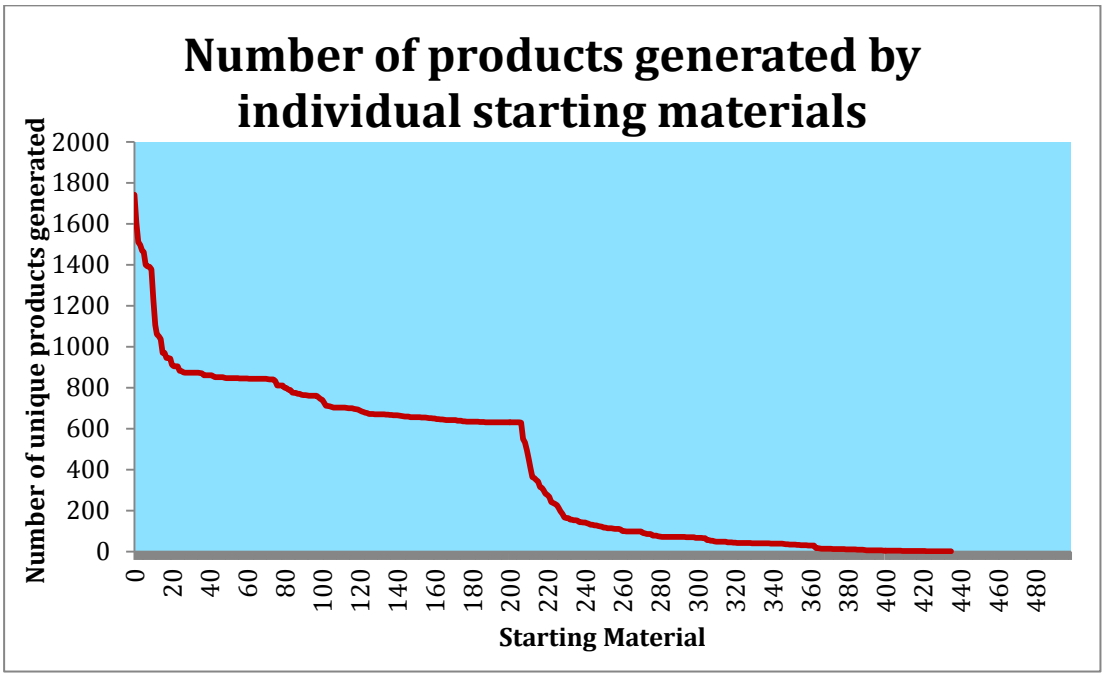


Figure 6.15: Plot showing the number of unique products generated from each starting material from the JMC2 RSVs.

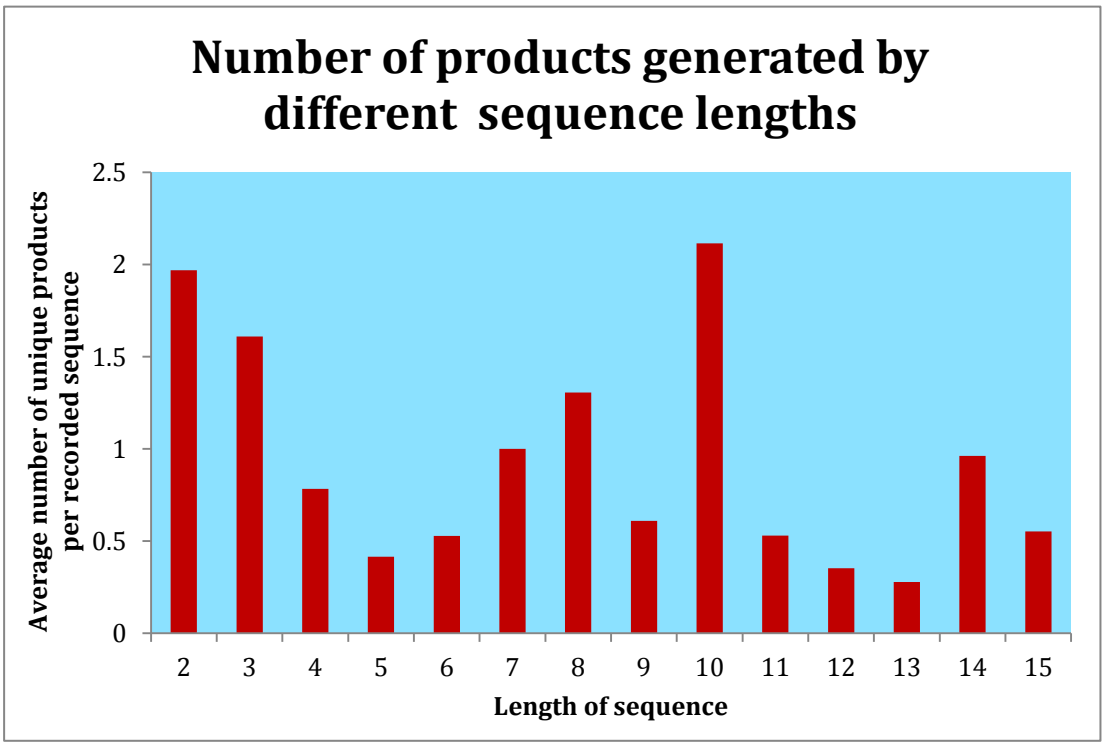


Figure 6.16: A breakdown of the products, arranged by sequence length from the JMC2 RSVs.

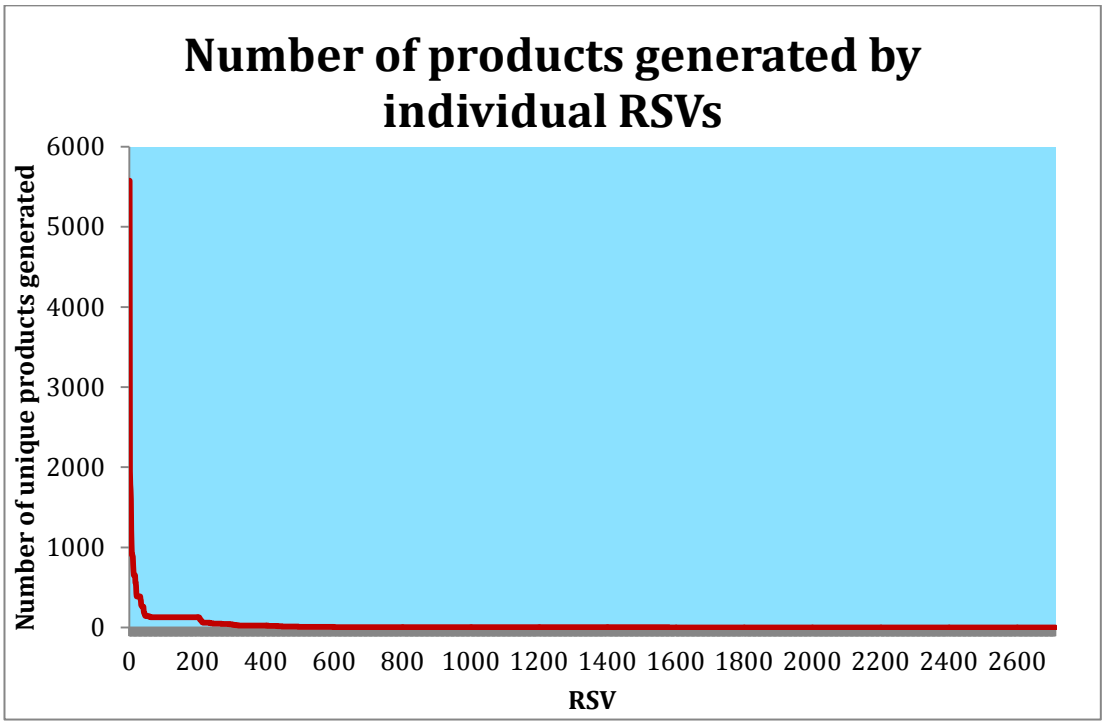


Figure 6.17: Frequency plot showing the number of RSVs applicable to each starting material from the JMC2 RSVs.

Figure 6.15 shows a great variation in the number of novel products generated with 210 starting materials producing in excess of 500 products. There is then a gradual decline, with the next 20 starting materials producing an average of 300 products each, before a further decline to below 100 products, and for the last 64 molecules, no unique products generated at all. Reviewing the molecules in question, it is the simpler, more drug-like starting materials with common functionality that lead to greater number of products, as shown by the examples in Table 6.9. For example, the most products in this case were generated from molecules with some carbonyl or double bond functionality.

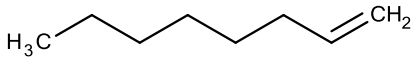
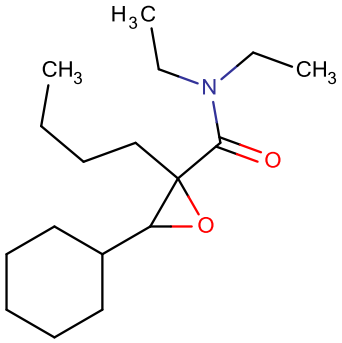
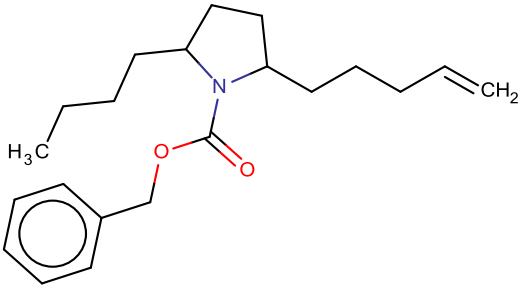
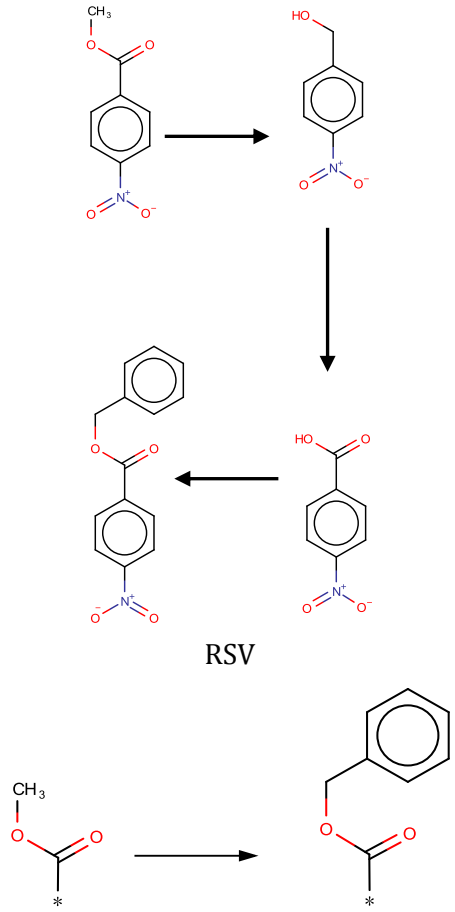
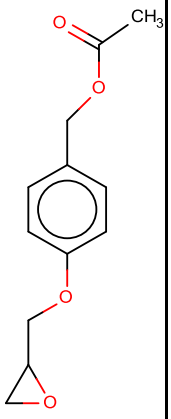
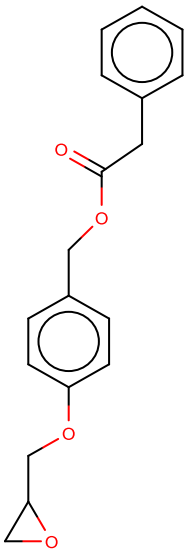
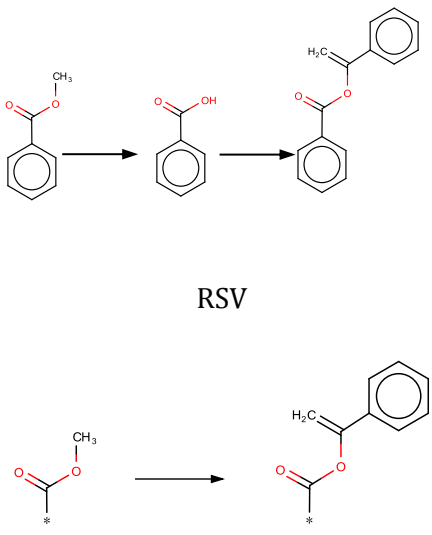
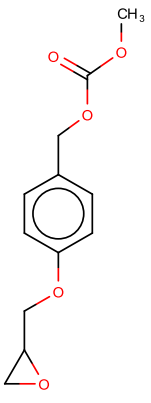
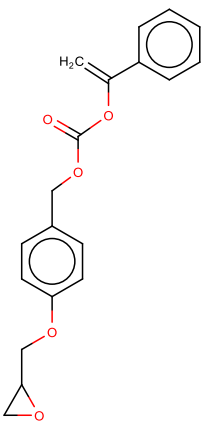
Structure	Number of unique, novel products generated
	1,742
	1,604
	1,511

Table 6.9: The three starting materials that generated the most unique products in the sampling experiment from the JMC2 RSVs. (Wallace, 2015)

A breakdown by RSVs (Figure 6.17) offers a number of surprising results, considering the nature of the network and the sample of starting materials used. Firstly, it is clear that, despite the number of RSVs, relatively few sequences are actually used to generate products, with only 2,700 of the 93,557 RSVs being used. Of these, only a quarter of that figure are used to generate more than ten unique products, and 41 of these generate in excess of 250 products. This is an interesting result, as it suggests that there is one key portion of the network that is used for structure generation, with the vast majority of the network proving irrelevant to these simple starting materials.

Table 6.10 shows examples of: RSVs that generated the most products; the original reaction sequences from which they were generated; together with a starting material and the product generated from it. With a few noted exceptions, these operate on aromatic species, further confirming the apparent dominance of the more generic addition chemistry in the test set.

Sequence ID	Sequence information	Sample reactant	Sample product
SCM_17413	Sequence reactions  RSV		
SCM_54554	Sequence reactions  RSV		

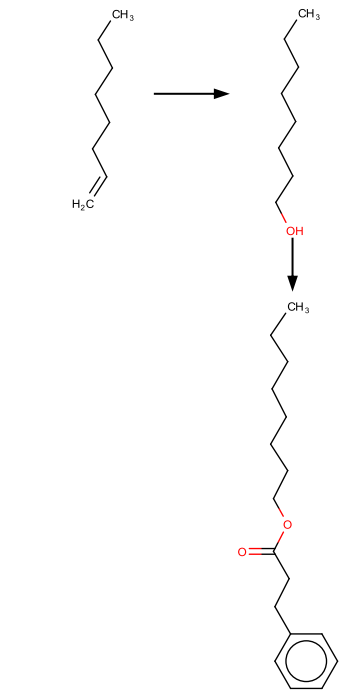
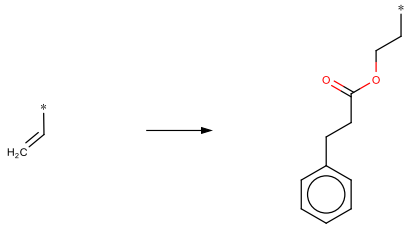
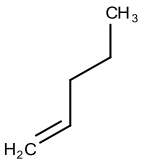
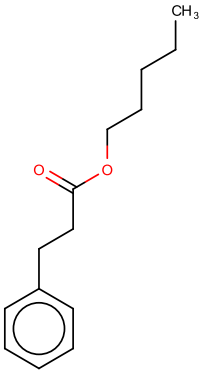
Sequence ID	Sequence information	Sample reactant	Sample product
SCM_62626	Sequence reactions  RSV 		

Table 6.10: Illustration of the three most frequently used JMC2 RSVs. (Wallace, 2015)

Studying the overall frequency distribution of RSVs, it appears that approximately 600 RSVs (0.6% of the total) are responsible for the majority of the products. Again, this is down to the relative applicability or otherwise of the sequences represented, with simpler processes being more applicable than the more convoluted sequences. When this consideration is extended to sorting by sequence length, it can be seen that once sequences get particularly long the likelihood of them being applicable is reduced, with more products generated through RSVs of shorter sequences.

In order to test the wider application of the reaction sequence data, the same experiment was performed with a second series of 500 molecules extracted from the

reagent pool. This pool was generated as part of a previous project, where reagent molecules detected within the reaction database were removed from the records as part of a database cleaning operation, and stored in a separate file. These reagent molecules are not directly related to the stored sequences, but should provide sufficient functionality to resemble the typical small molecule pool used in *de novo* design. The completion time for this experiment was approximately the same as with the other starting materials, averaging 25 minutes for each input molecule. A summary of the distribution of the molecular weights of these molecules is presented in Figure 6.18, along with the hydrogen bond donor and acceptor profiles in Figure 6.19 and 6.20. For comparison, the distributions of these properties for the original collection of starting materials are also included.

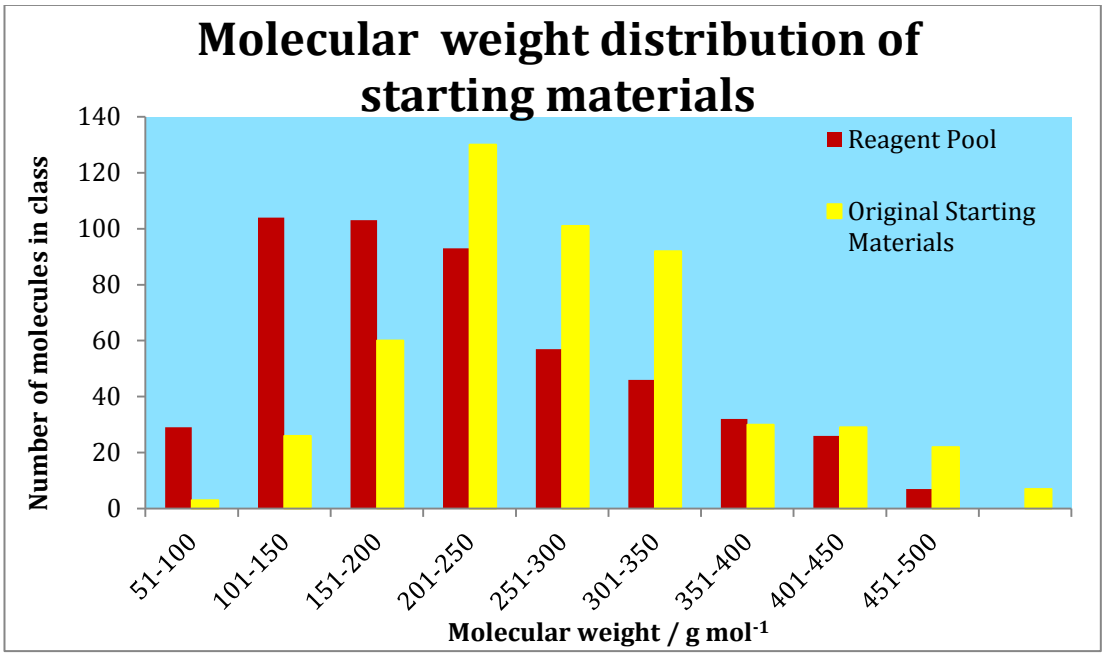


Figure 6.18: Molecular weight distribution of the starting material collections.

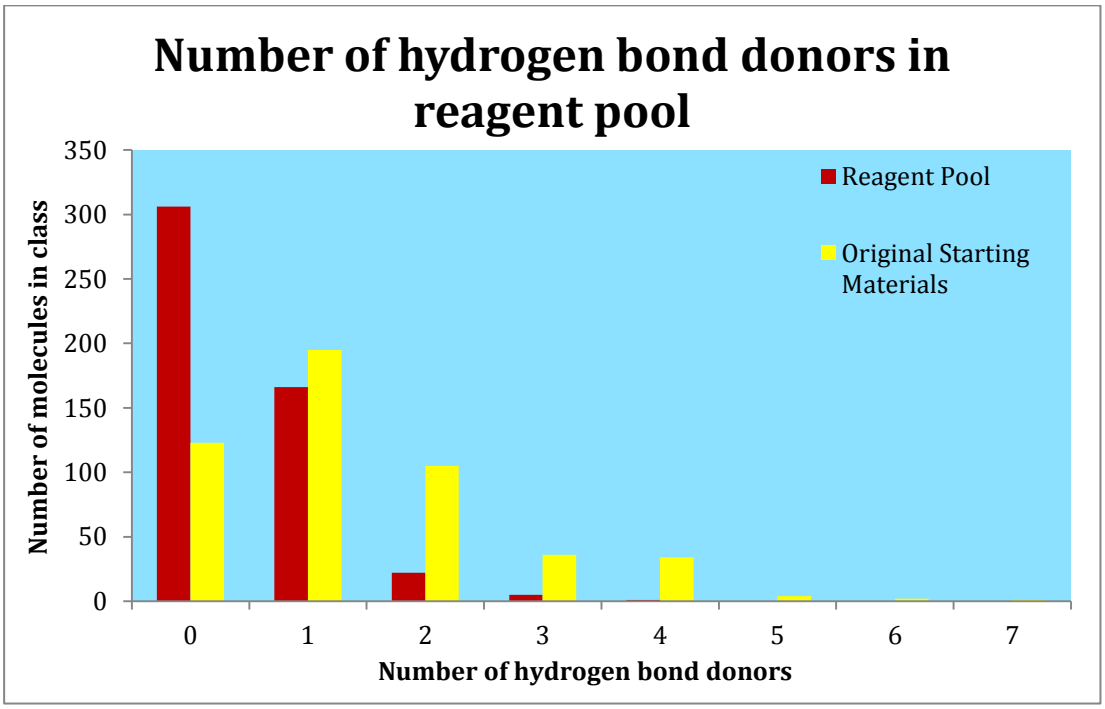


Figure 6.19: Hydrogen bond donor distribution for the starting material collections.

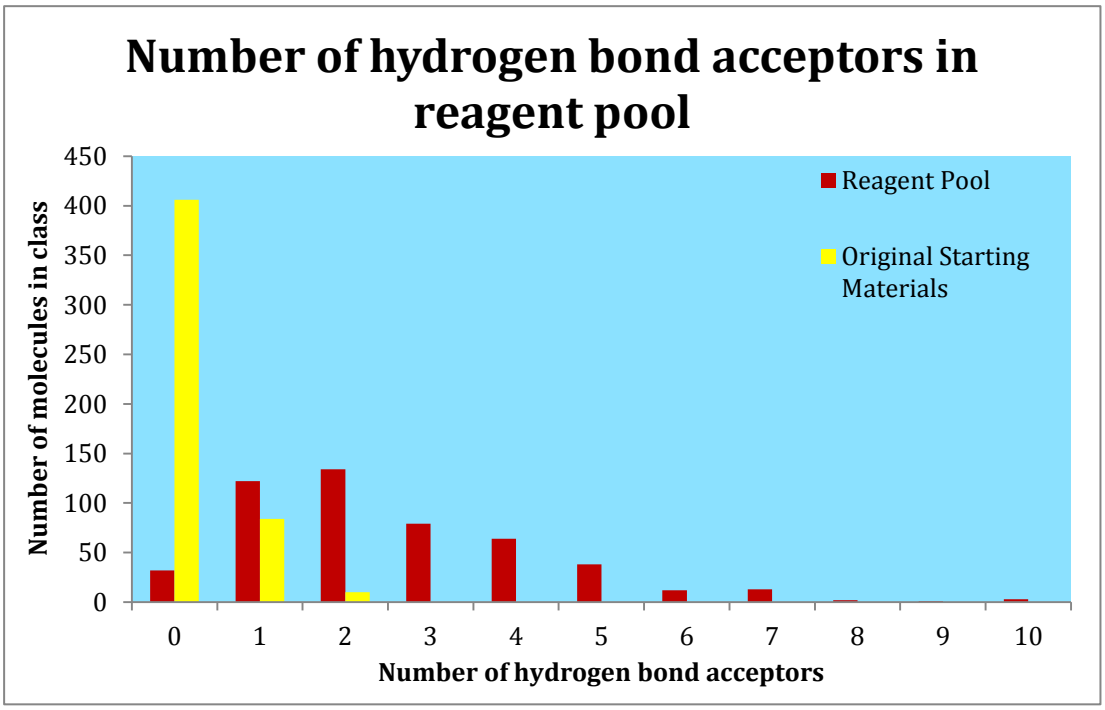


Figure 6.20: Hydrogen bond acceptor distribution for the starting material collections.

It should be noted that this distribution of molecular weights favours a lower range than the original as shown in Figure 6.12. In terms of functionality, however, the sample is now biased in favour of hydrogen bond acceptors. This gives a different

activity profile to the previous set that may favour a different portion of the reaction network.

A summary of the data produced from this experiment is given in Table 6.11, along with frequency plots ordered by product molecule, the sequence lengths and the particular RSVs involved (Figure 6.21, 6.22, 6.23). Overall, an average of 179 molecules is produced per starting material, with 89,881 generated in total. This is higher than with the previous pool of starting materials. However, in this case the most products come from a molecule containing a functionalised benzene ring, while the least products come from molecules like carbon tetrabromide, a molecule with very little functionality to exploit. Looking at the most commonly applied sequences, two of these are shared with the previous sample (SCM_17413, and SCM_62626), so only the sequence that is not shared is shown in Table 6.12. It should be noted that this sequence (SCM_6297) is not a realistic synthetic route, and is produced in this form in the database as a result of the automated connection of related reactions. A more realistic approach to the same goal would be via a Grignard reagent operating on the ester. Such a transformation is present in the database, but is removed as a duplicate RSV. This highlights an issue with the sequences that are stored with the RSVs. When multiple sequences lead to duplicate RSVs, it would make more sense to retain the sequence having few steps, rather than making an arbitrary choice as occurs currently. Table 6.13 shows the three most frequently used starting molecules.

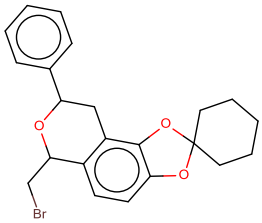
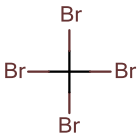
Most molecules generated	1,416 
An example molecule which results in no product molecules being generated	0 
Total number of molecules generated	89,881
Average	179.8 molecules per starting material

Table 6.11: Summary of the results of the molecule novelty experiment from the reagent pool, using the JMC2 RSVs. (Wallace, 2015)

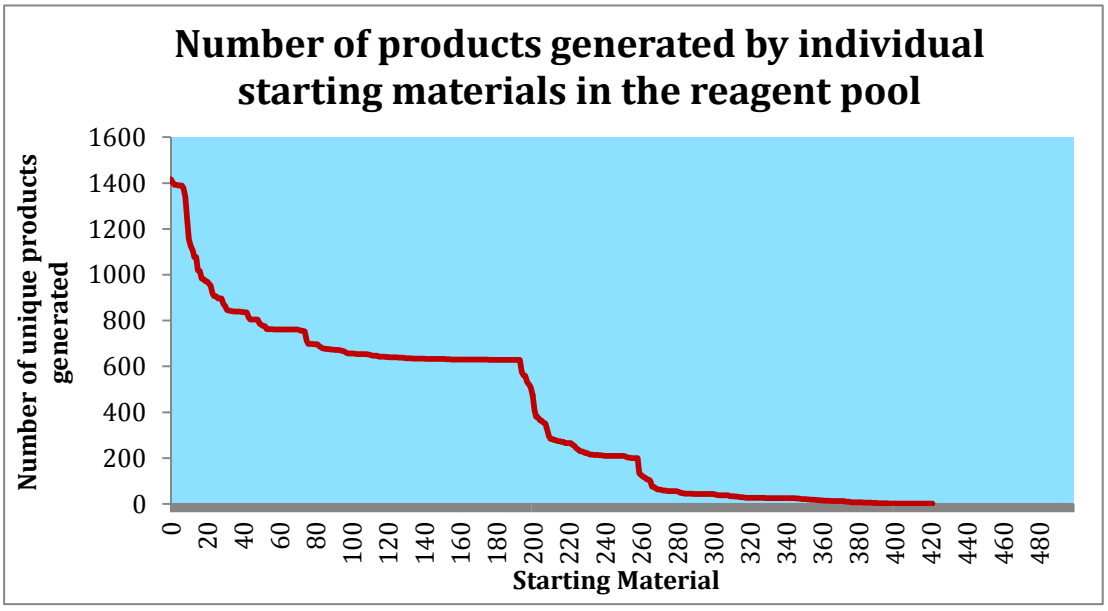


Figure 6.21: Frequency plot showing the number of unique products generated from each starting material in the reagent pool, using the JMC2 RSVs.

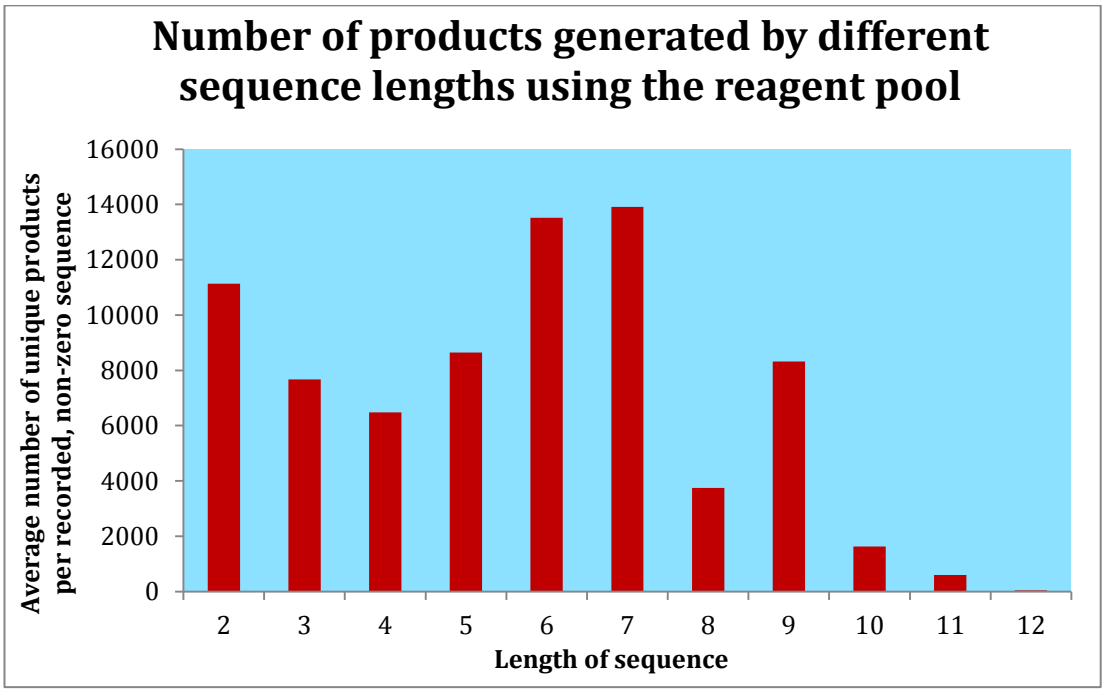


Figure 6.22: A breakdown of the products, arranged by sequence length, using the reagent pool and the JMC2 RSVs.

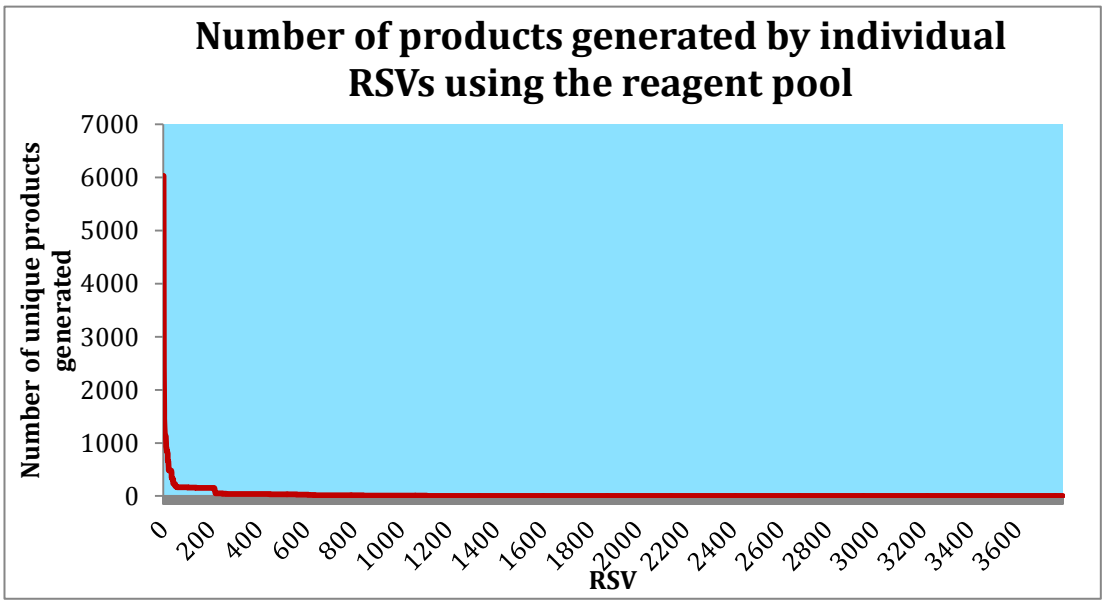


Figure 6.23: Frequency plot showing the number of RSVs applicable to each starting material in the reagent pool, using the JMC2 RSVs. Only the non-zero values are shown, the vast majority of the RSVs remain unused.

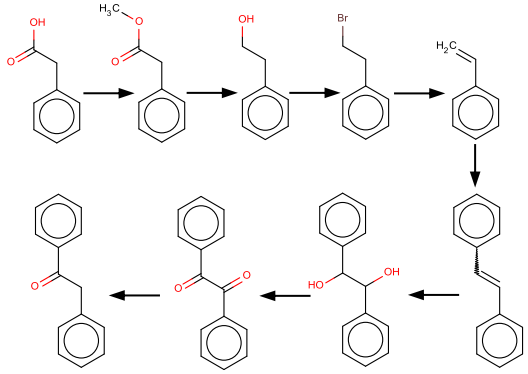
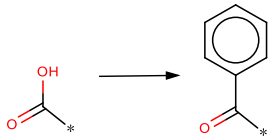
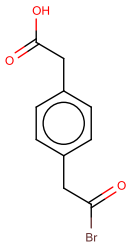
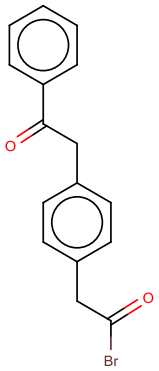
Sequence ID	Sequence information	Sample reactant	Sample product
SCM_6297	Sequence Reactions  RSV 		

Table 6.12: Example of some of the most frequently used RSVs with the reagent pool and the JMC2 RSVs. Two of the three most frequently used RSVs are shared with the other experiment. (Wallace, 2015)

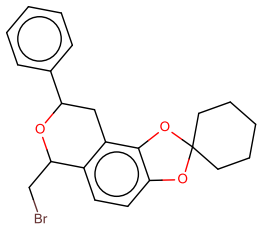
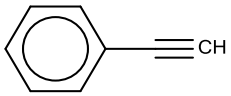
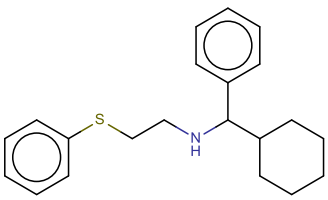
Structure	Number of unique, novel products generated
	1,416
	1,404
	1,393

Table 6.13: The three starting materials that generated the most unique products in the sampling experiment using the reagent pool and the JMC2 RSVs. (Wallace, 2015)

The same general trends in generation of products according to starting materials and RSVs can be seen as before, with many of the most frequently used sequences in this case being the same as those seen with the previous set. Comparing the three most frequently used RSVs with the previous experiment shows significant overlap, with only one RSV being different (shown in Table 6.12). Figure 6.21 shows that, as before, there is a heavily skewed distribution of products generated from particular starting materials, but with a lower product total overall. In total, 3,792 RSVs are applied to generate structures in this case. The same steep descents and long plateaux are present in this distribution as in the one for the previous experiment (Figure 6.17), indicating that a limited subset of the sequences are used to generate the products. However, in this case the number of molecules generated at each step is slightly larger than the previous case, with a steeper tail off towards the end of the distribution. Overall, 78 molecules lead to no products being generated due to a lack of applicability.

Considering the breakdown by sequence length in Figure 6.22, it appears that the profile of products generated relative to sequence length is more evenly distributed. However, the most commonly used transformations seem to be consistent over the two runs. Reviewing the breakdown by RSV (Figure 6.23) again reinforces the suggestion that the issues with apparently only accessing limited portions of the network in these experiments are genuine effects rather than limitations of the sampling. In both cases the majority of the product generation is carried out by a small portion of the total network, with around 95% remaining unused. The fact that so little of the recorded network appears to be used for both samples suggests that there are issues that require further investigation.

Looking at the overlap between the two sets of results, there are 2,821 RSVs used to generate structures in both experiments. This represents the majority of the useful sequences in both cases, suggesting that only a very narrow range of reaction centres are relevant for these kinds of starting materials. However, in terms of the structures generated, the overlap is relatively small, with 2,909 molecules common to both experiments. Examples of some of these are illustrated in Figure 6.24, with these common molecules being based around ring structures, and in some cases carbonyl functionality.

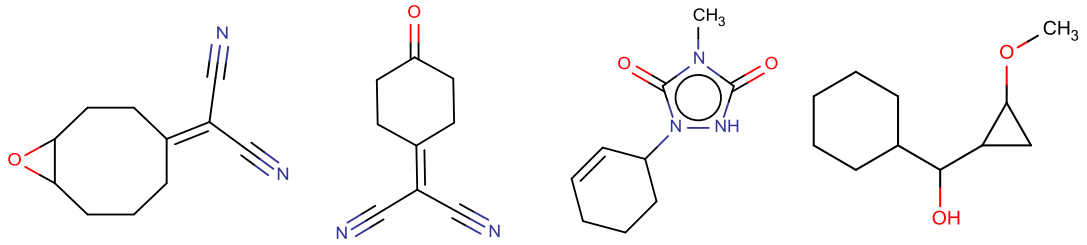


Figure 6.24: Examples of molecules produced from both sets of starting materials. (Wallace, 2015)

In order to ascertain the sensitivity of the output of structure generation to sets of RSVs derived from different sources to see if the issues regarding the relatively small numbers of applicable sequences occur with other collections, the two sets of 26,000 reactions abstracted from the NextMove (Lowe and Sayle, 2014) collection of patent data in Section 5.5.3.1 were used to perform similar experiments. As mentioned previously, the analysis by Schneider et al. (2014) indicates that these reactions are more complicated than those considered in the JMC1 and JMC2 collections, meaning that fewer of these are likely to generate RSVs. However, there should be sufficient data for a meaningful experiment. It should be noted that the patent data sets contain atom mappings that can theoretically be used to identify the intent of the reactions. However, for ease of comparison it was decided to disregard this and use the existing methods of processing the reactions.

The reaction networks produced for each data set (as detailed in Section 5.5.3.1) were used to produce RSVs using the direct method. The two RSV collections were then used to generate products using the original set of 500 starting materials selected for the first experiment. As before, the number of novel unique molecules produced from the RSVs in this database was recorded. A summary of the number of products generated is given in Table 6.14, along with a frequency plot ordered by number of product molecules in Figure 6.25. Overall, an average of 2,421 molecules is produced per starting material, with 1,210,733 generated in total. As before, the most products are generated for a starting material with multiple functional groups and attachment points, in this case featuring two benzene rings that can be functionalised in a number of ways, but those with more complicated groups only provide results that fail the stability check. It should be noted that more of the starting materials from the experiment generated products in this case as opposed to the JMC2 experiment, with only 26 recording no products at all. Figure 6.26 shows a histogram of the average number of products sorted by the number of steps in the sequence used to generate

them, while Figure 6.27 shows a frequency plot of the number of products generated per RSV.

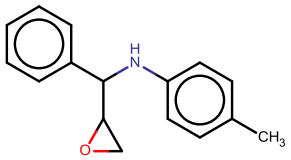
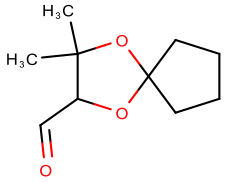
Most molecules generated	9,829 
An example molecule which results in no product molecules being generated	0 
Total number of molecules generated	1,210,393
Average	2,420.79 molecules per starting material

Table 6.14: Summary of the results of the molecule novelty experiment for the first patent data collection. (Wallace, 2015)

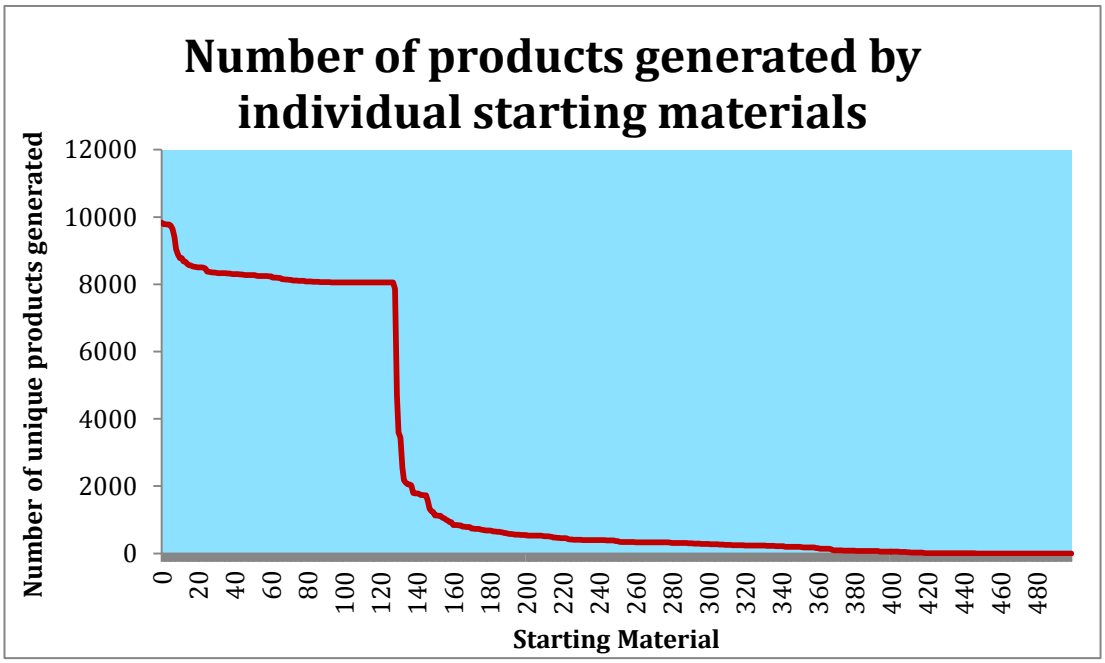


Figure 6.25: Plot showing the number of unique products generated from each starting material for the first patent data collection.

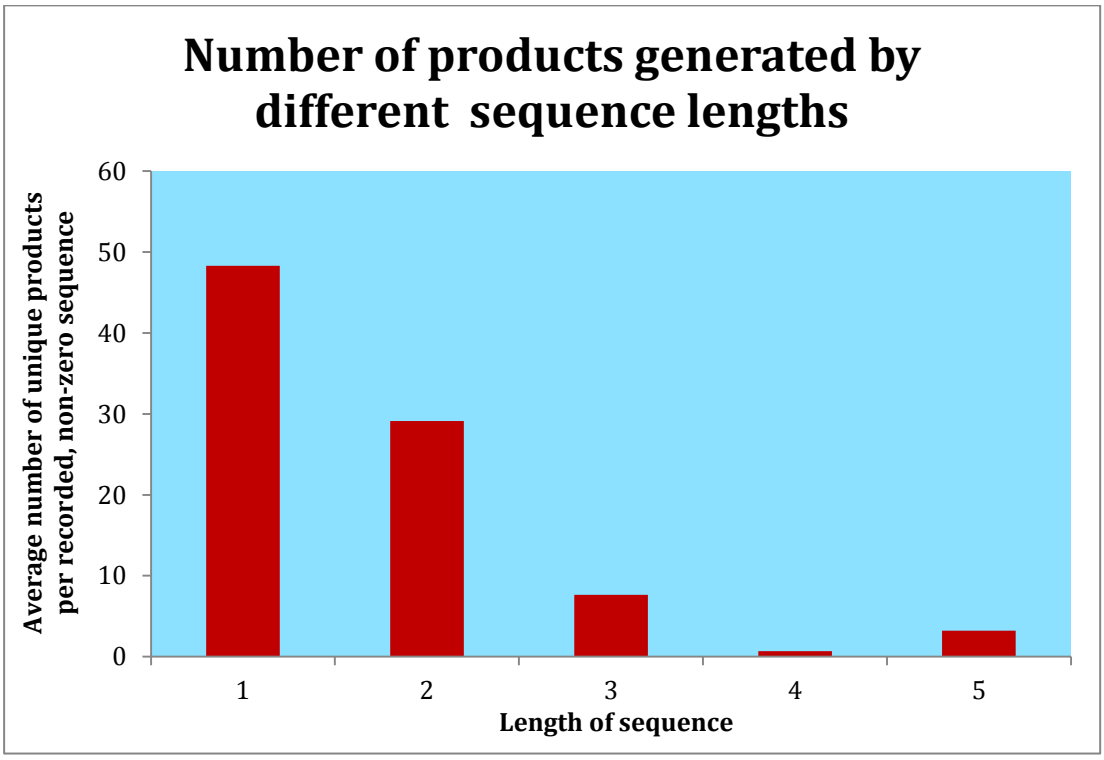


Figure 6.26: A breakdown of the products, arranged by sequence length, for the first patent data collection.

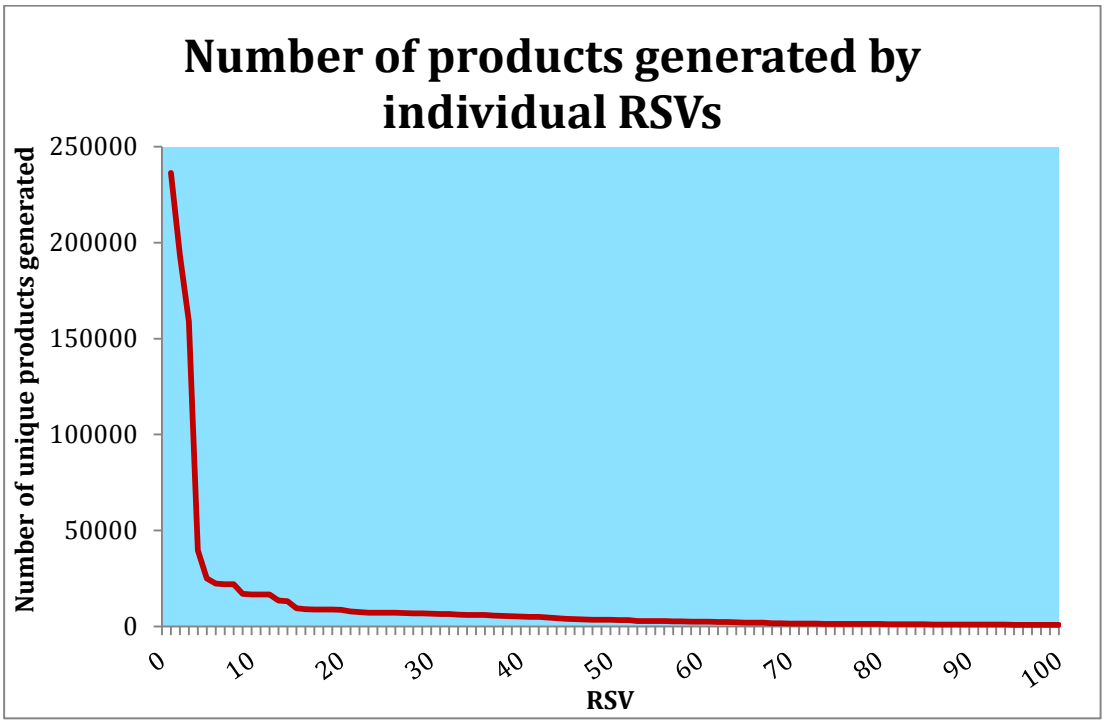


Figure 6.27: Partial frequency plot showing the number of RSVs applicable to each starting material for the first patent data collection.

As with the JMC data sets, the vast majority of the products are generated by a relatively small portion of the RSVs. The larger number of products seen with this set (1,210,393) would suggest that there is a greater likelihood of applying the RSVs in this set to our starting materials, despite the relatively short sequence length.

The second data set taken from the patent information is very similar in characteristics to the first, with slightly fewer sequences present (26,981 in total). Despite this lower number of sequences, more unique products were made from this set than the previous example, with 1,470,740 generated, averaging at 2,941 per molecule. Once again, not all of the 500 starting materials generate products, with 35 producing nothing in this case. A summary of the number of products generated from the sampling experiment is given in Table 6.15, along with a frequency plot ordered by the number of the generated product molecules in Figure 6.28. Figure 6.29 shows a histogram of the average number of products sorted by the number of steps in the sequence used to generate them, while Figure 6.30 shows a frequency plot of the number of products generated per RSV.

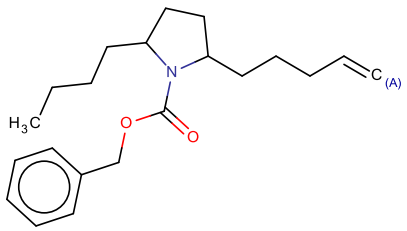
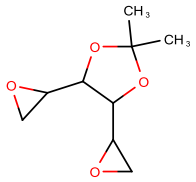
Most molecules generated	12,260 
An example molecule which results in no product molecules being generated	0 
Total number of molecules generated	1,470,740
Average	2,956.74 molecules per starting material

Table 6.15: Summary of the results of the molecule novelty experiment for the second patent data collection. (Wallace, 2015)

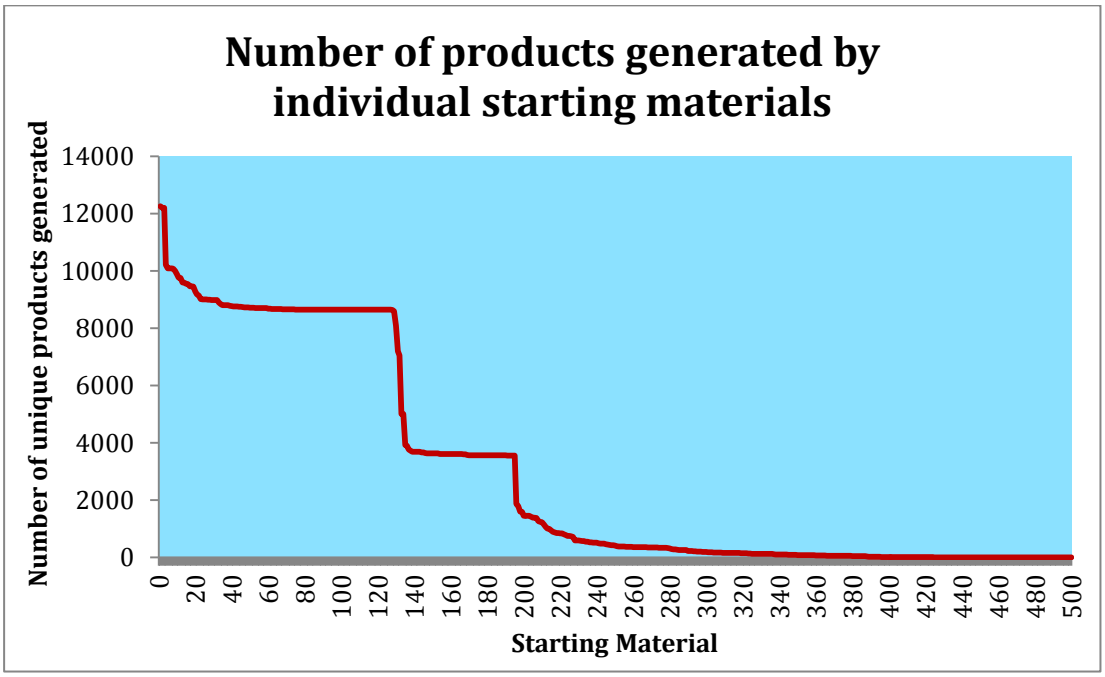


Figure 6.28: Plot showing the number of unique products generated from each starting material for the second patent data collection.

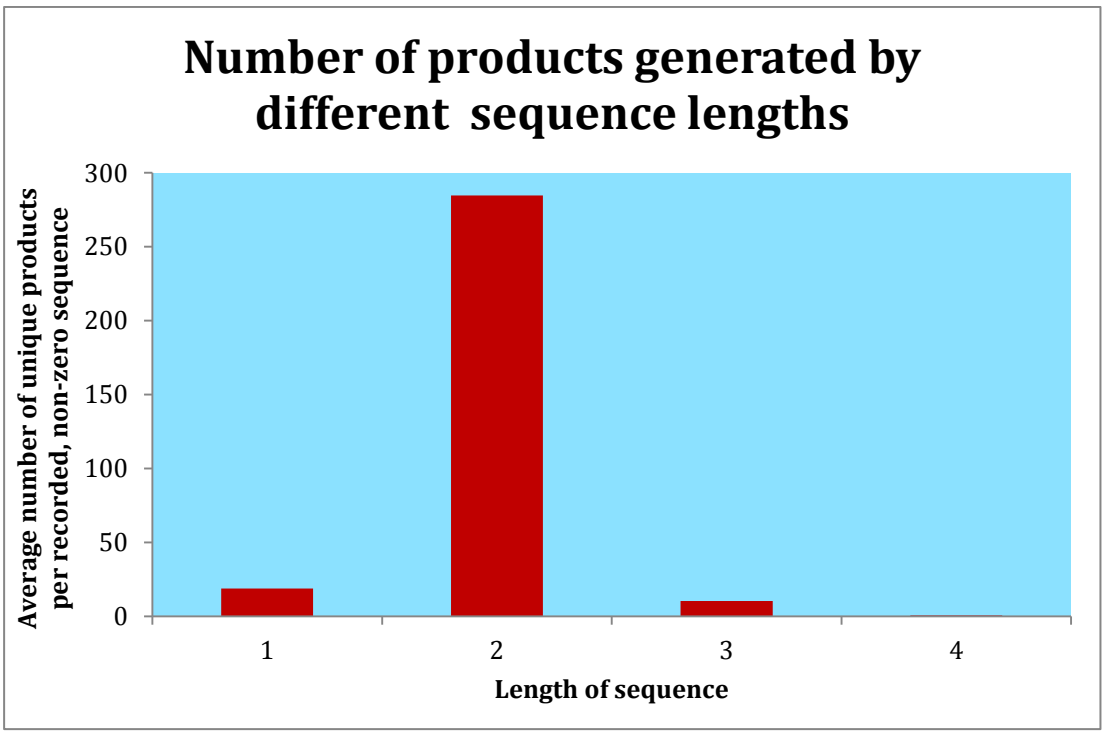


Figure 6.29: A breakdown of the products, arranged by sequence length for the second patent data collection.

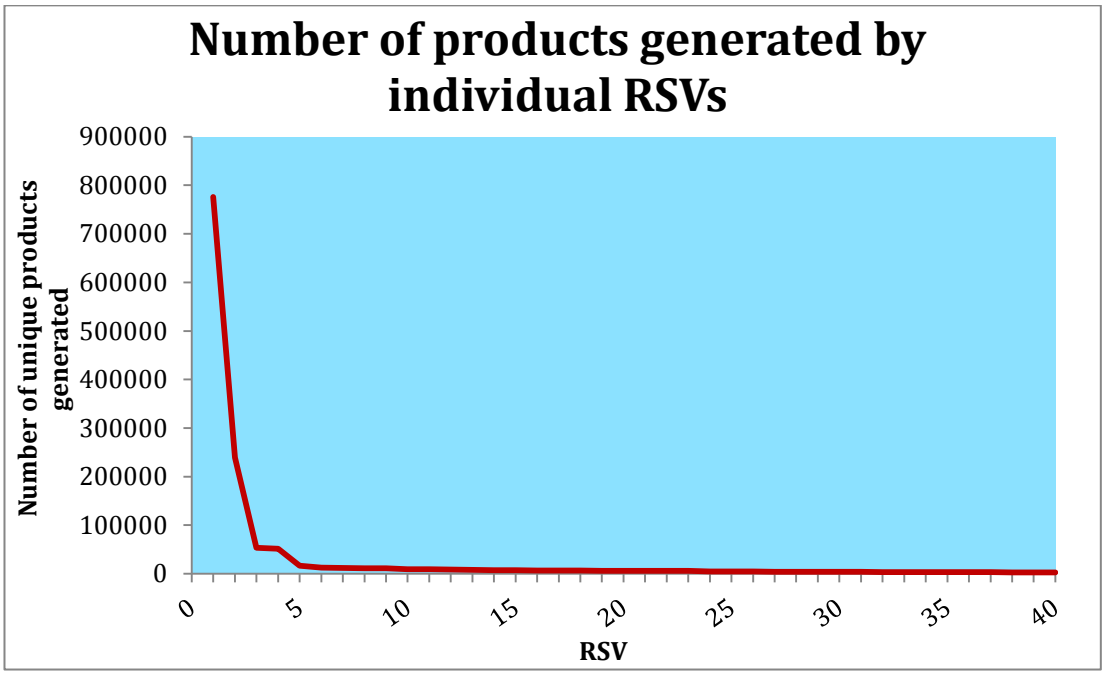


Figure 6.30: Partial frequency plot showing the number of RSVs applicable to each starting material for the second patent data collection.

Again, the vast majority of the unique products are made from only a few of the stored RSVs in the database, and the general trends are similar to the first patent derived set.

6.3.5 Network analysis by RSV content

The skewed distribution of the numbers of products generated for the RSVs and starting materials has potential implications for a *de novo* tool, in that populations could either become too large to realistically handle, or too small to generate interesting molecules, depending on the nature of the input. This has importance when considering the composition of the ideal reaction network and subsequent database of RSVs generated from it. To be useful in a *de novo* context a reaction network would ideally cover as much of the potential solution space as is possible. This means that it would need to contain sufficiently diverse transformations to ensure that it is applicable to all starting materials of interest. At the same time, sequences that are irrelevant to the used starting materials should not be present, as the time taken to search the RSV collection to generate structures increases with the collection size. In the sampling experiments large sections of the reaction network proved to be irrelevant to the drug-like starting materials used, so in this case a much smaller database could be used to achieve the same results with greater efficiency. One of the

simplest methods to analyse the collected RSVs for applicability to a given set of starting materials is to study the reaction centre, in a similar manner to the database analysis in Section 5.5.3.

As discussed previously, the simplest grouping approach for RVs is by negative atom pair content, since these determine the characteristics required in a starting material for that transformation to be applicable. RSVs store data in the same manner, albeit in larger amounts, so this same approach can be used to study transformations over whole sequences. The stored RSVs from the JMC2 network (92,767 unique RSVs) were grouped based on identical negative atom pairs. The number of groups indicates the number of different types of functionality that the network can be applied to, whereas the relative numbers of RSVs in each group indicates if the network favours one type of reaction centre over any others. An illustration of the distribution of RSVs in groups is shown in Figure 6.31, with an expansion in Figure 6.32 and a log-log plot for the distribution in Figure 6.33 which approximates Zipf's law similarly to the RV distributions analysed in Section 5.5.3.

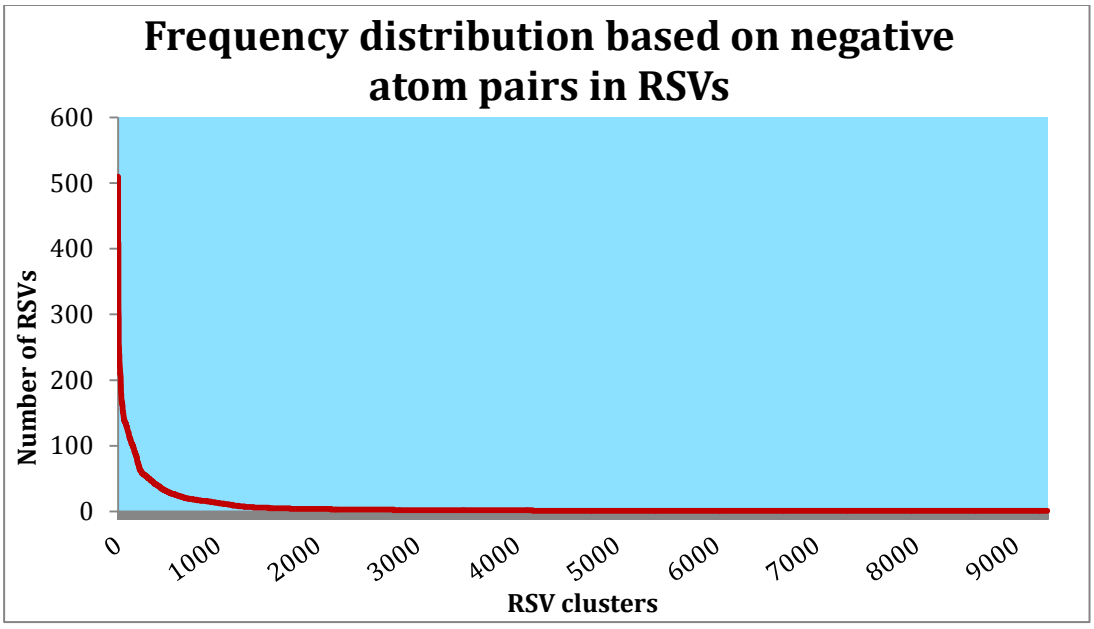


Figure 6.31: Frequency distribution curve based on the negative atom pairs in the JMC2 data set.

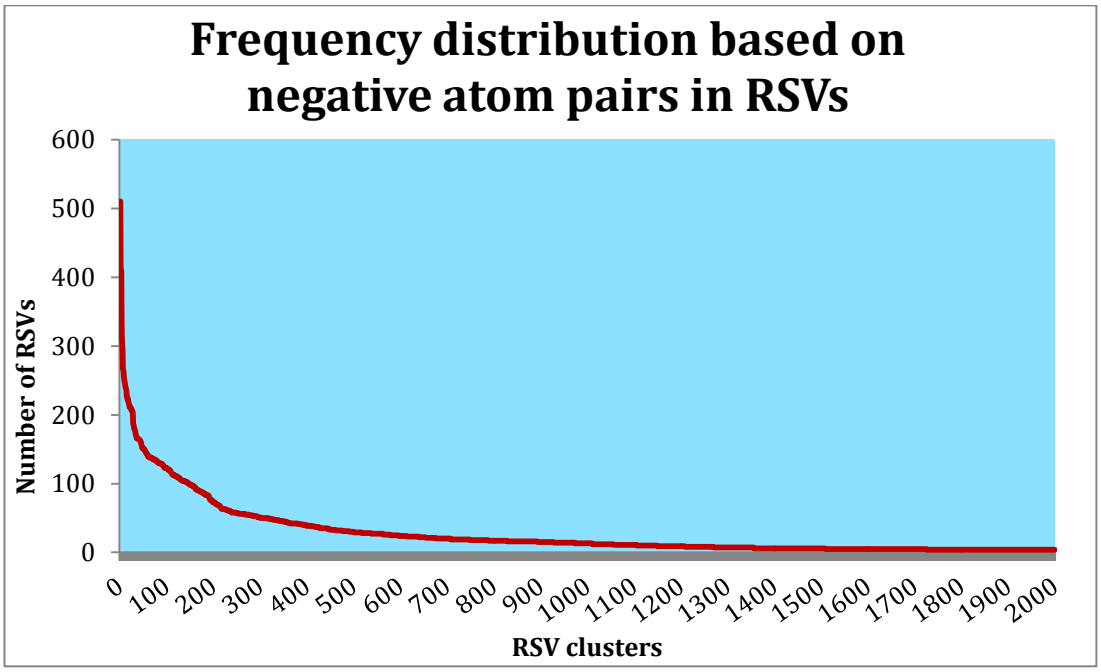


Figure 6.32: Expansion of the first 2000 entries in Figure 6.31.

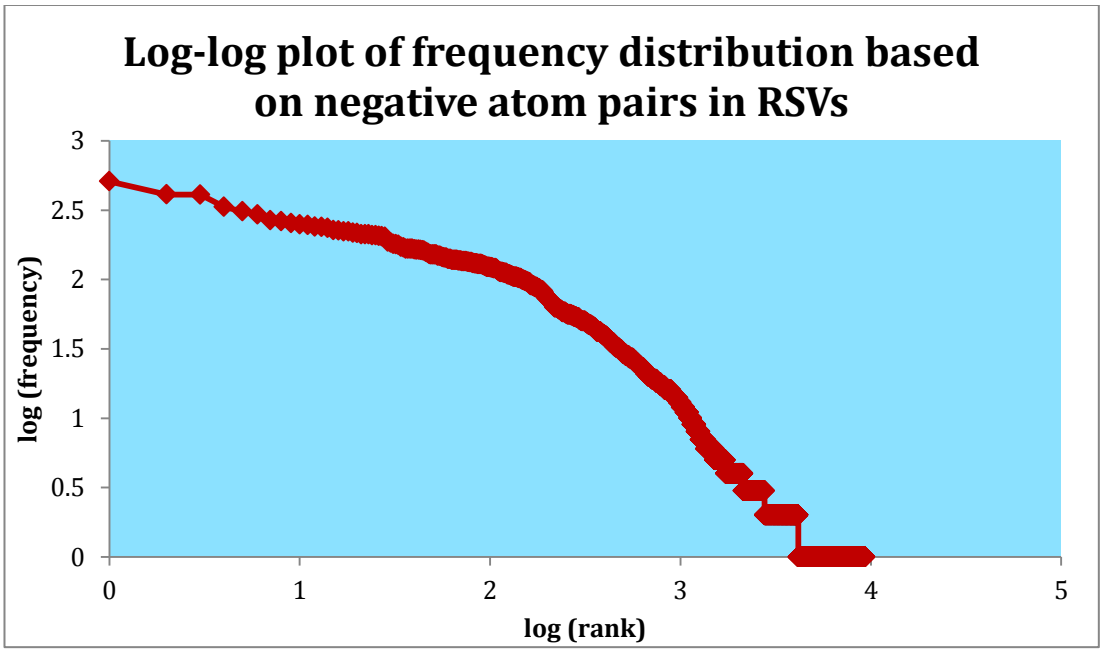


Figure 6.33: Log-log plot of the frequency distribution based on the negative atom pairs in the JMC2 data set.

The frequency distributions show that there is indeed a significant amount of skew in the nature of the reactant functionality, with only 9,316 groups generated, and only 1,105 of these representing more than ten sequences. The five most frequent partial

RSVs were tabulated, as shown in Table 6.16. These are very simple in nature, and represent structural features that would be expected to be present in many starting materials. There is a significant bias towards aromatic structural species which are common to many drug precursors and indicates a potential lack of diversity within the RSV database.

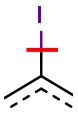
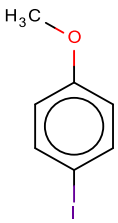
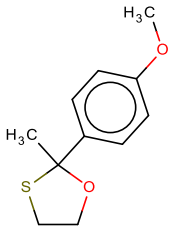
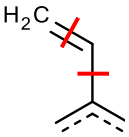
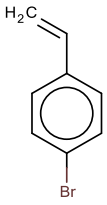
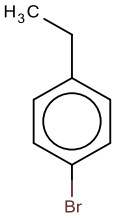
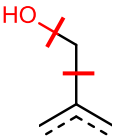
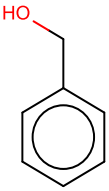
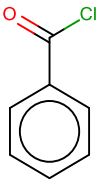
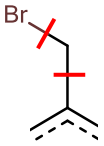
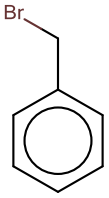
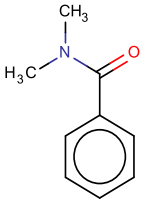
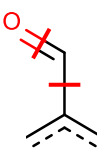
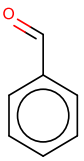
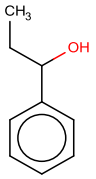
Negative atom pairs (duplicates indicate multiple entries)	Number of sequences represented	Reactant functionality structure	Example reactant	Example product
I(1,0,0)-2(1)-C(3,2,1) I(1,0,0)-3-C(2,2,1) I(1,0,0)-3-C(2,2,1)	510			
C(2,1,0)-2(2)-C(1,1,0) C(3,2,1)-2(1)-C(2,1,0) C(2,2,1)-3-C(2,1,0) C(2,2,1)-3-C(2,1,0) C(3,2,1)-3-C(1,1,0)	410			
C(3,2,1)-2(1)-C(2,0,0) O(1,0,0)-2(1)-C(2,0,0) C(2,2,1)-3-C(2,0,0) C(2,2,1)-3-C(2,0,0) O(1,0,0)-3-C(3,2,1)	409			
Br(1,0,0)-2(1)-C(2,0,0) C(3,2,1)-2(1)-C(2,0,0) Br(1,0,0)-3-C(3,2,1) C(2,2,1)-3-C(2,0,0) C(2,2,1)-3-C(2,0,0)	334			
C(3,2,1)-2(1)-C(2,1,0) O(1,1,0)-2(2)-C(2,1,0) C(2,2,1)-3-C(2,1,0) C(2,2,1)-3-C(2,1,0) O(1,1,0)-3-C(3,2,1)	309			

Table 6.16: Representation of the five largest groups of partial RSVs for the JMC2 data.The red lines indicate bonds broken in the reaction centre structure.(Wallace, 2015)

There are a number of groups that have very few examples. In total, 4,355 reaction centres exist that have only one listed example, all of which represent chemistry that is unlikely to be included in a general purpose *de novo* experiment. These centres tend to contain multiple functional groups involved in the transformation, or contain metal ions in the key leaving groups. Some examples of these reaction centres are shown in Figure 6.34. It should be noted that only 20 of the partial RSVs representing ten or more sequences have reaction centres that contain obscure metal ions or heavily specialised structures.

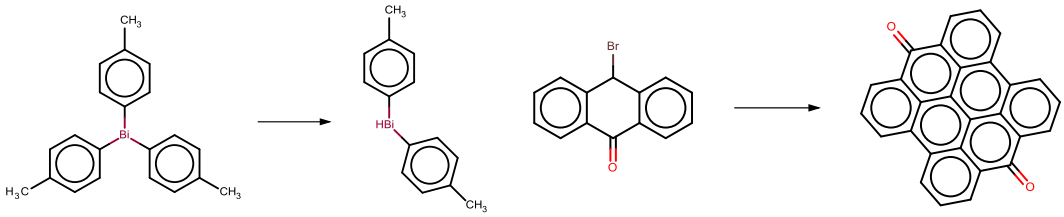


Figure 6.34: Examples of reaction centres in JMC2 for which only single partial RSVs exist. (Wallace, 2015)

The RSVs in JMC2 include all intermediate sequences. For example, a sequence of three steps in length (R1→R2→R3→ R4) will produce six RSVs (R1→R2, R1→R3, R1→R4, R2→R3, R2→R4, R3→R4). This potentially biases the RSVs towards reactant functionality that may not be useful for sequence-based *de novo* design. For example, the functionality could include protecting and deprotecting chemistry. To determine if this is an issue, the smaller JMC1 set was used for the same experiment, as this does not contain the intermediates. Looking at the frequency distribution, the same pattern emerges of a heavy skew towards particular reactant functionalities as shown in Figure 6.35. However, a different set of partial RSVs dominates, as can be seen in Table 6.17. The log-log plot for this distribution (Figure 6.36) is also similar to the Zipf's law plots seen previously, suggesting this more complete collection is similar to the reaction databases in terms of range of functionality.

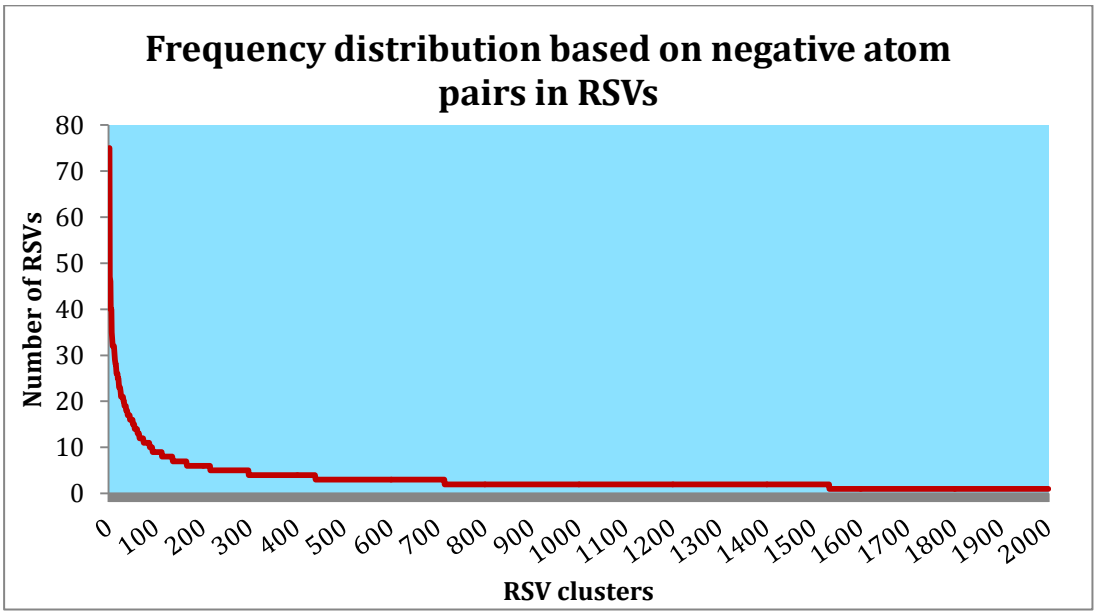


Figure 6.35: Partial frequency distribution curve based on the 'lost' atom pairs in the JMC1 data set.

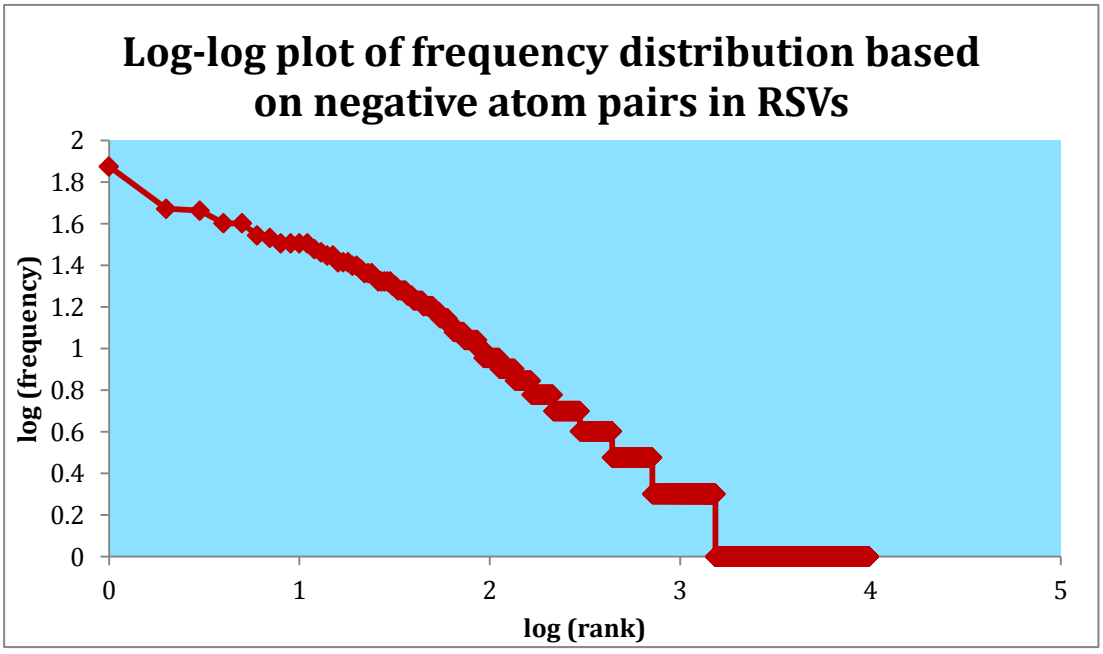


Figure 6.36: Log-log plot of the frequency distribution based on the 'lost' atom pairs in the JMC1 data set.

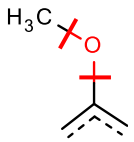
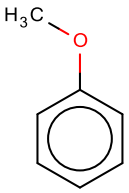
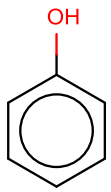
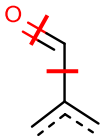
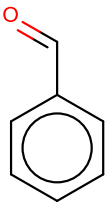
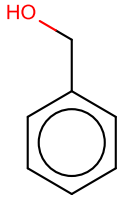
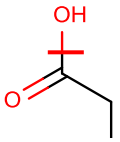
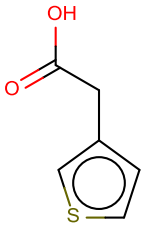
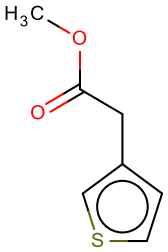
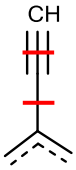
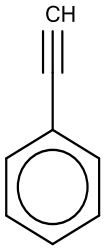
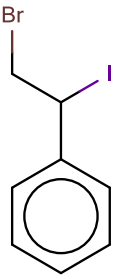
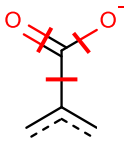
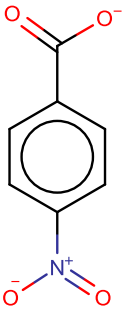
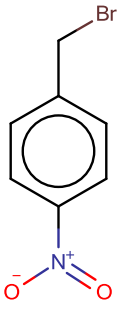
Negative atom pairs (duplicates indicate multiple entries)	Number of sequences represented	Reactant functionality structure	Example reactant	Example product
O(2,0,0)-2(1)-C(1,0,0) O(2,0,0)-2(1)-C(3,2,1) C(3,2,1)-3-C(1,0,0) O(2,0,0)-3-C(2,2,1) O(2,0,0)-3-C(2,2,1)	75			
C(3,2,1)-2(1)-C(2,1,0) O(1,1,0)-2(2)-C(2,1,0) C(2,2,1)-3-C(2,1,0) C(2,2,1)-3-C(2,1,0) O(1,1,0)-3-C(3,2,1)	46			
O(1,0,0)-2(1)-C(3,1,0) O(1,0,0)-3-C(2,0,0) O(1,1,0)-3-O(1,0,0)	46			
C(2,2,0)-2(3)-C(1,2,0) C(3,2,1)-2(1)-C(2,2,0) C(2,2,1)-3-C(2,2,0) C(2,2,1)-3-C(2,2,0) C(3,2,1)-3-C(1,2,0)	40			
C(3,2,1)-2(1)-C(3,1,0) O(1,0,0)-2(1)-C(3,1,0) O(1,1,0)-2(2)-C(3,1,0) C(3,1,0)-3-C(2,2,1) C(3,1,0)-3-C(2,2,1) O(1,0,0)-3-C(3,2,1) O(1,1,0)-3-C(3,2,1) O(1,1,0)-3-O(1,0,0)	35			

Table 6.17: Representation of the five largest groups of partial RSVs for the JMC1 data. The red lines indicate bonds broken in the reaction centre.(Wallace, 2015)

While four of the five most common reaction centres are different from the JMC2 case, small, simple structural features still dominate for both the JMC1 and JMC2 databases. In fact, only 11 of the groups containing more than ten sequences encode anything other than aromatic species with carbonyl groups attached. It should be noted that in the JMC1 list, while aromatic functionalities dominate, the presence of non-aromatic functionality within the five largest groups, suggests different characteristics between the two network types. However, the JMC1 distribution is still highly skewed, with 8,246 of the 9,781 partial RSVs representing only one example.

Looking at the first random sample extracted from the US patent database, grouping by negative atom pairs gives 7,767 groups in total, with 20 of these representing over 100 sequences, and 6,559 groups only containing one sequence. The frequency distribution for this data set is shown in Figure 6.37, with a table of the most popular groups in Table 6.18.

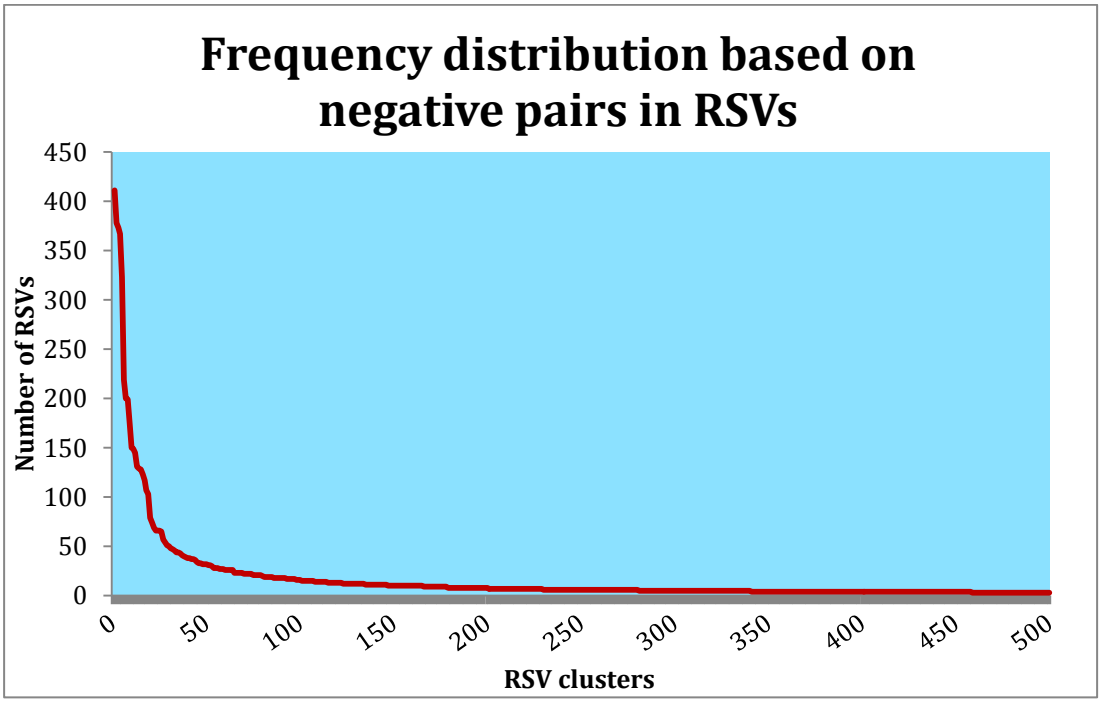


Figure 6.37: Partial frequency distribution curve based on the negative atom pairs in the reaction sequence database for the first random sample extracted from the US patent database.

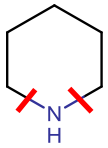
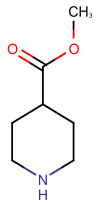
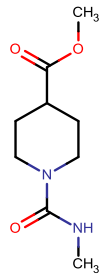
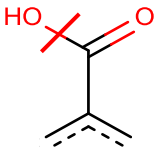
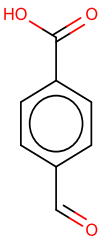
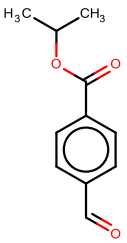
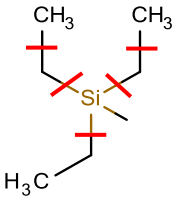
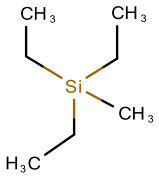
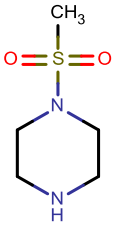
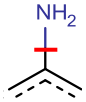
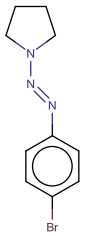
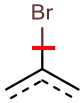
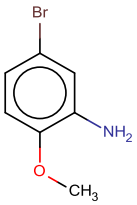
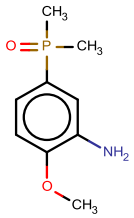
Negative atom pairs (duplicates indicate multiple entries)	Number of sequences represented	Reactant functionality structure	Example reactant	Example product
N(2,0,1)-2(1)-C(2,0,1) N(2,0,1)-2(1)-C(2,0,1) N(2,0,1)-3-C(2,0,1) N(2,0,1)-3-C(2,0,1)	157			
O(1,0,0)-2(1)-C(3,1,0) O(1,0,0)-3-C(3,2,1) O(1,1,0)-3-O(1,0,0)	135			
C(2,0,0)-2(1)-C(1,0,0) C(2,0,0)-2(1)-C(1,0,0) C(2,0,0)-2(1)-C(1,0,0) Si(3,0,0)-2(1)-C(2,0,0) Si(3,0,0)-2(1)-C(2,0,0) Si(3,0,0)-2(1)-C(2,0,0) C(2,0,0)-3-C(2,0,0) C(2,0,0)-3-C(2,0,0) C(2,0,0)-3-C(2,0,0) Si(3,0,0)-3-C(1,0,0) Si(3,0,0)-3-C(1,0,0) Si(3,0,0)-3-C(1,0,0)	134			
N(1,0,0)-2(1)-C(3,2,1) N(1,0,0)-3-C(2,2,1) N(1,0,0)-3-C(2,2,1)	133			
Br(1,0,0)-2(1)-C(3,2,1) Br(1,0,0)-3-C(2,2,1) Br(1,0,0)-3-C(2,2,1)	104			

Table 6.18: Representation of the five largest groups of partial RSVs for the first random sample extracted from the US patent database. The red lines indicate bonds broken in the reaction centre structure.(Wallace, 2015)

The most popular reactant functionalities in this set are very similar to those seen in JMC2, with small aromatic molecules, and simple structural features favoured. It should be noted that the silyl molecule in the third example in Table 6.18 is an example of one of the issues with the network construction. On occasion, the identification of the wrong molecule as the reactant or product leads to unusual partial RSVs being associated with the sequences, and as such, unrelated sequences are grouped together. The preference for small aromatic molecules is seen with the second patent data collection, although there are significant differences in the frequency distribution of the negative pair groupings. As shown in Figure 6.38, the largest group from the second set is only one quarter of the size of the equivalent group in the first set, with 6,368 of the 7,541 reaction centre groupings only containing one stored example. Additionally, looking at the most popular groups for this set (as seen in Table 6.19) shows that all of the five most popular reaction centres are different from the first collection, with the Suzuki coupling reagents seen when analysing the individual reactions becoming prominent once again. In both patent sets, however, significant portions of the collection of sequences remain unused.

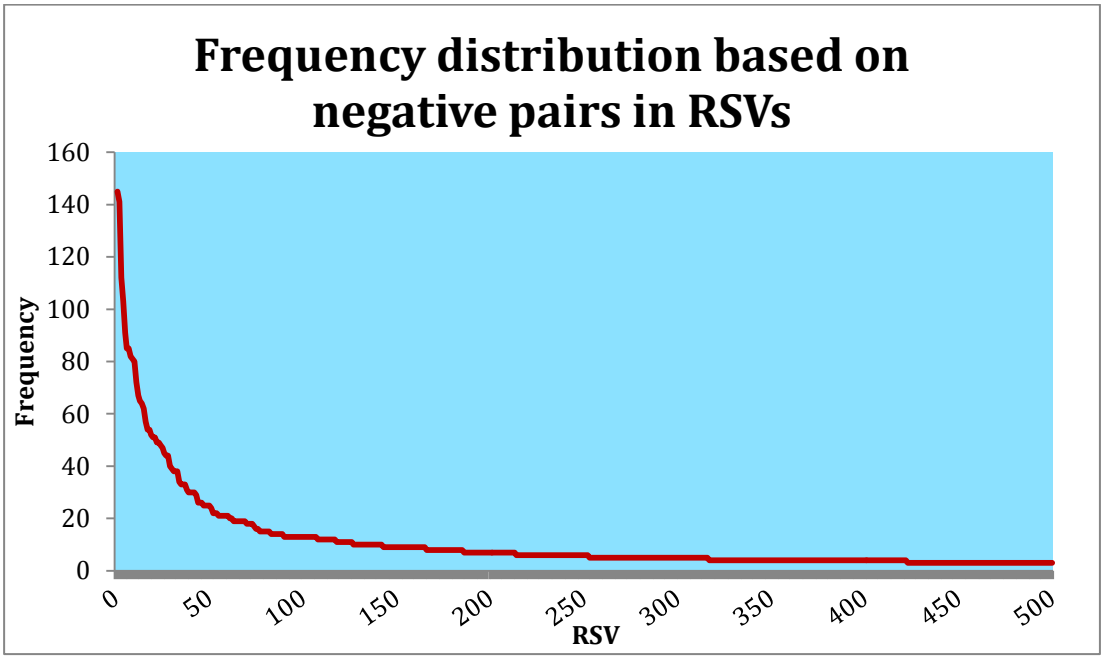


Figure 6.38: Partial frequency distribution curve based on the negative atom pairs in the reaction sequence database for the second random sample extracted from the US patent database.

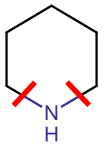
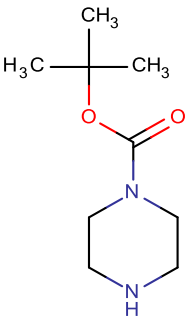
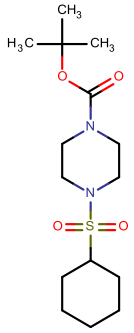
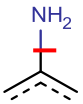
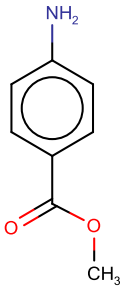
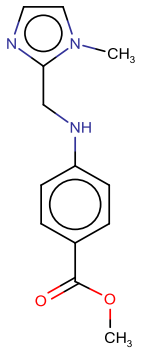
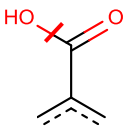
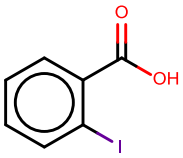
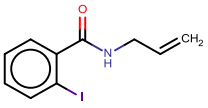
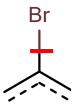
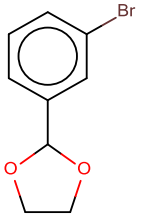
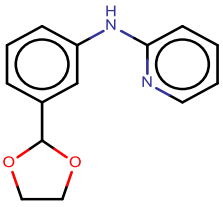
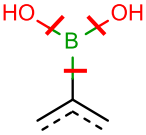
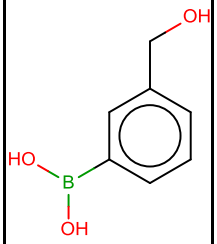
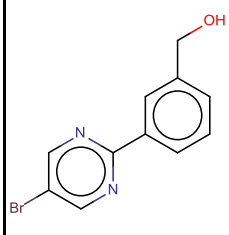
Negative atom pairs (duplicates indicate multiple entries)	Number of sequences represented	Reactant functionality structure	Example reactant	Example product
N(2,0,1)-2(1)-C(2,0,1) N(2,0,1)-2(1)-C(2,0,1) N(2,0,1)-3-C(2,0,1) N(2,0,1)-3-C(2,0,1)	145			
N(1,0,0)-2(1)-C(3,2,1) N(1,0,0)-3-C(2,2,1) N(1,0,0)-3-C(2,2,1)	141			
O(1,0,0)-2(1)-C(3,1,0) O(1,0,0)-3-C(3,2,1) O(1,1,0)-3-O(1,0,0)	112			
Br(1,0,0)-2(1)-C(3,2,1) Br(1,0,0)-3-C(2,2,1) Br(1,0,0)-3-C(2,2,1)	103			
C(3,2,1)-2(1)-B(3,0,0) O(1,0,0)-2(1)-B(3,0,0) O(1,0,0)-2(1)-B(3,0,0) C(2,2,1)-3-B(3,0,0) C(2,2,1)-3-B(3,0,0) O(1,0,0)-3-C(3,2,1) O(1,0,0)-3-C(3,2,1) O(1,0,0)-3-O(1,0,0)	104			

Table 6.19: Representation of the five largest groups of partial RSVs for the second random sample extracted from the US patent database. The red lines indicate bonds broken in the reaction centre structure.(Wallace, 2015)

6.4 Conclusions

In this chapter the reaction vector approach was extended to consider reaction sequences from which RSVs were generated. The RSVs were then applied in a *de novo* design context. While there are fewer novel molecules produced using RSVs compared to applying RVs iteratively, the advantages of speed and simplicity of RSVs make these potentially useful for *de novo* design. However, the disadvantage is the reduced number and diversity of molecules produced. This reduction in diversity is also demonstrated by the skewed distribution of the frequency of application of the RSVs and the limited number of RSVs that are applicable to a given set of starting materials. This was shown to be the case for different sets of starting materials and different sources of RSVs. This analysis also suggests that it may be possible to do some pre-analysis of RV and RSV collections to eliminate those that are unlikely to be useful and to avoid the over-representation of particular types of reaction. However, this was not explored further in this thesis. The next chapter compares the use of RSVs and RVs in a number of drug design scenarios.

Chapter 7:
SAR Exploration with Reaction Sequence Vectors

7.1 Introduction

This chapter explores the application of the RSV methods to various drug design scenarios, including exploration of Structure Activity Relationship (SAR) information and *de novo* design in general. These methods include the identification of multiple routes to the same product, and learning more about the interrelation between reactions in the database for synthesis planning.

From a molecular design perspective, there are a number of different features that are required to make the tool more effective to a medicinal chemist than simply providing a report of the generated structures. Given that the tool is intended for use in compound and synthetic route suggestion, the ability to visualise the alternate routes to generate compounds and near analogues would be of significant benefit.

7.2 Alternate route identification

In Chapter 6, individual reactions were used to generate a reaction network (Chapter 5) from which RSVs are extracted to use as part of a reaction transformation library. The network may contain more than one route between a given start and end molecule, however, given that the RSV generation ignores intermediate compounds, these will collapse to a single RSV. As a consequence, such sequences will only be recorded in the library once, albeit with both sets of sequence identifier information. KNIME (Berthold et al., 2008) code was written to retrieve the original reactions used to make the RSVs, permitting a comparison of the different routes available. Figure 7.1 illustrates one such example, in which two routes from the same starting material (green) to the same product (red) can be compared.

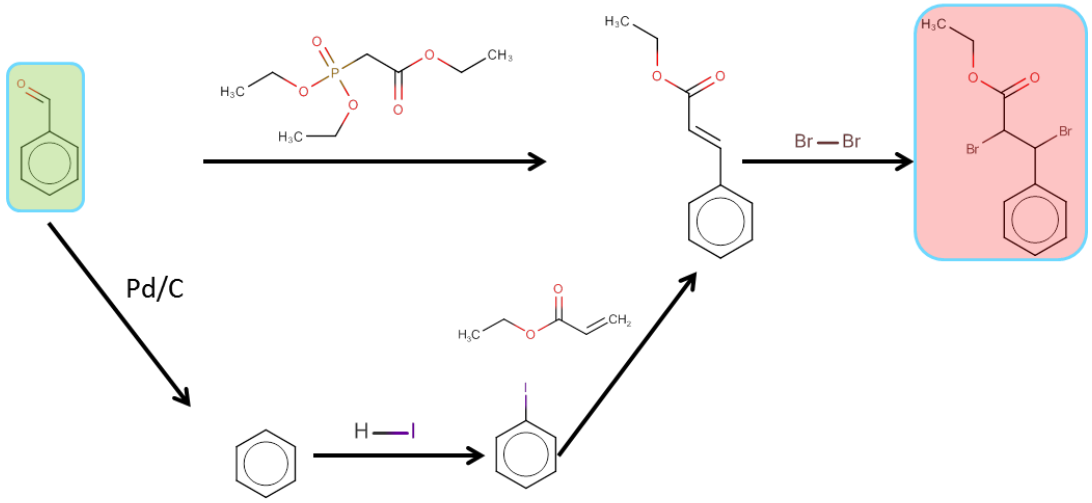


Figure 7.1: Illustration of multiple routes to the same product. (Wallace, 2015)

For a more complete SAR exploration, additional processes are required to expand upon the stored sequences to fully exploit the stored knowledge. In the next section, the separate processes used to analyse the data will be considered in turn, followed by a case study of an SAR evaluation.

7.3 SAR proof of concept

When a lead compound is identified as part of a screening programme, a period of fine-tuning and optimisation is required to determine how to improve its efficacy and physico-chemical property profile. The usual method for achieving this is to generate as many analogues as possible, with different functionality in key areas, and screen each of these in turn (known as a structure-activity relationship evaluation, or SAR evaluation). An illustration of such an evaluation for a relatively simple molecule is given in Figure 7.2, which shows how many different points of interest may be utilised in any given case.

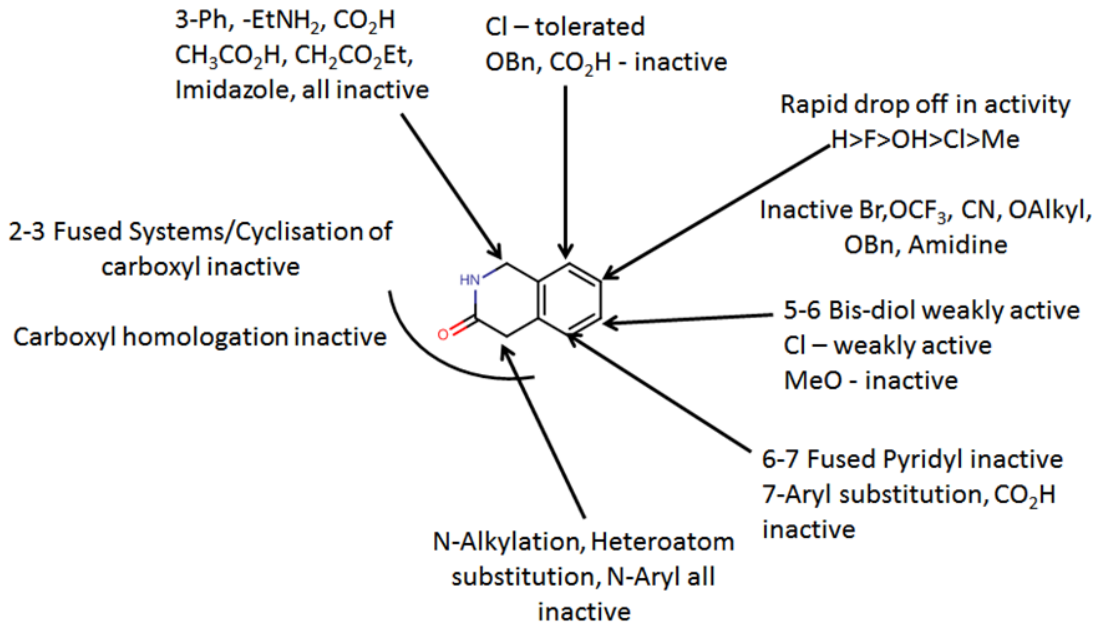


Figure 7.2: A systematic SAR evaluation for a simple drug-like molecule. (Wallace, 2015)

The structure generation program developed in this project is ideally suited to such an exploration, as it can be configured to systematically transform a starting molecule into all possible products, based on a knowledge base of reactions and reaction sequences. Consequently, the generated products will represent compounds that can be generated in one or more synthetic steps and will therefore provide a much more exhaustive SAR exploration than traditional approaches, which consider a single reaction step only. The RSVs encode multi-step reactions that can be used to extend the potential range of chemistry covered relative to RVs. Depending on the desired usage, a simple filtration of the products can then be used to highlight those molecules of interest, such as those that are structurally similar to a known active compound, at which point further investigations of their syntheses can be carried out.

7.3.1 SAR exploration example 1 – cilomilast synthesis

In order to determine the usefulness of the RSV algorithm for SAR evaluation, the reactions taken from the JMC2 database were augmented with additional SAR data. The SAR data consists of reactions collated from the SAR papers described in Section 5.2.1, using the method reported by Roughley and Jordan (2011). By combining the two sets of reactions together and forming a network as in Section 5.4, a data set was produced that focusses on drug discovery chemistry (referred to as 'JMCRoughley'). When

merged there are 26,235 reactions in total. The network generated from this database produced 125,787 unique RSVs including all intermediate pathways, with an average length of 5.06 steps (shown in Figure 7.3, with the full data recorded in Table 7.1). As all intermediate sequences found in the network are included in this collection, the number of RSVs greatly exceeds the number of reactions (as explained in Chapter 5, a sequence of three steps in length (R1→R2→R3→R4) will produce six RSVs (R1→R2, R1→R3, R1→R4, R2→R3, R2→R4, R3→R4)). As with other collections of reactions from literature, relatively short sequences are more common, as the preparation steps are excised.

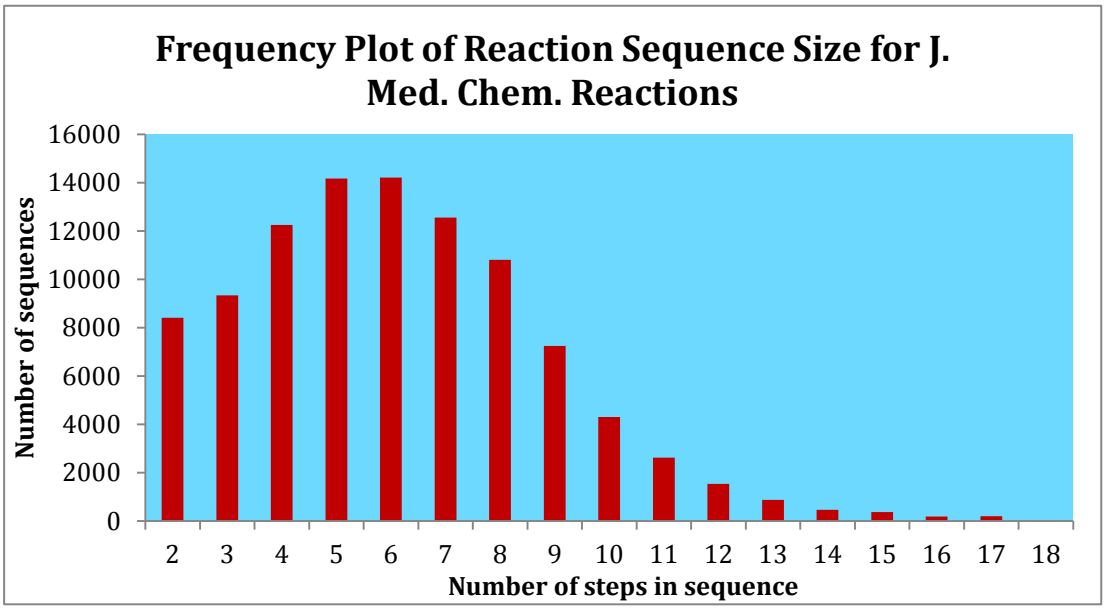


Figure 7.3: Frequency plot of reaction sequence size for the population.

Number of steps	Number of sequences
1	26235
2	8413
3	9333
4	12249
5	14168
6	14208
7	12549
8	10803
9	7241
10	4300
11	2628
12	1542
13	880
14	464
15	372
16	190
17	197
18	15

Table 7.1: Table of the full sequence summary.

A simple drug design sequence that forms part of the collection of reactions in the JMCRoughley database was used as a proof of concept of SAR exploration. The aim was to use the library of RSVs associated with the database to explore structures that can be generated from the known starting material used to synthesise the anti-asthma compound cilomilast, as published in Lednicer's collection of organic synthesis methods (Lednicer, 2007) and illustrated in Figure 7.4.

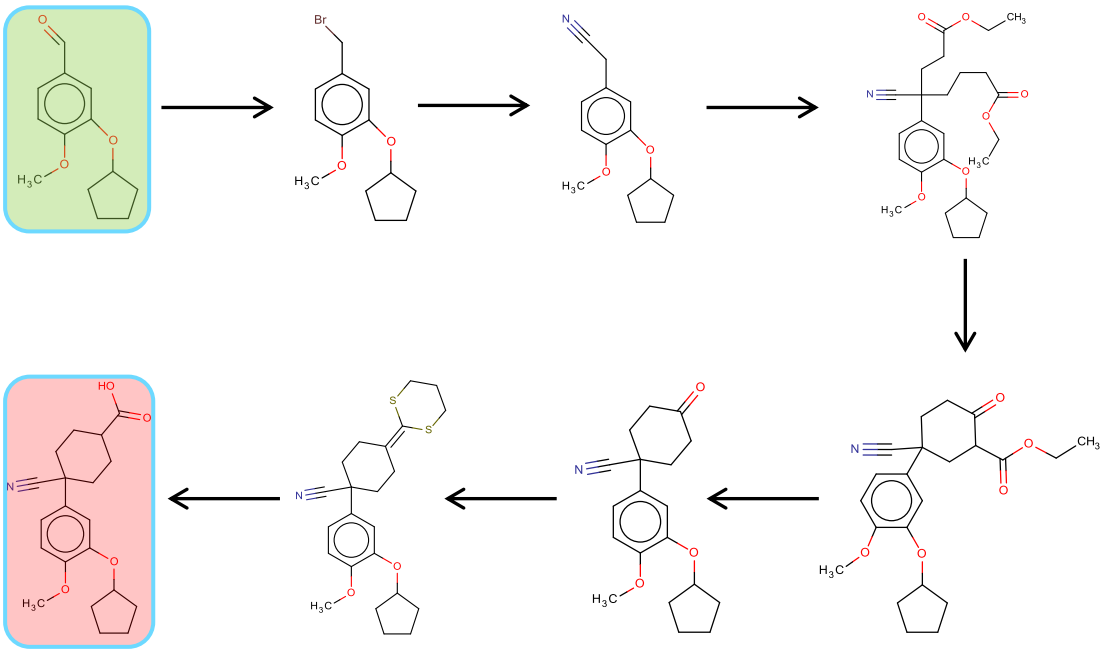


Figure 7.4: Literature synthesis route to cilomilast. (Lednicer, 2007)

The original starting material (highlighted in green in Figure 7.4) was used with the JMCRoughley database of RSVs, using the structure generation tool. In total, 4,030 products were generated, in addition to cilomilast itself, due to the application of several RSVs in the database. The products were then filtered on 0.8 Tanimoto 2D structural similarity to cilomilast using the Indigo structural fingerprints (EPAM Life Sciences). After filtering seven molecules remained, four of which were not found within the original literature sequences. These near neighbour molecules (shown in Figure 7.5) are intended to provide structural novelty, while remaining sufficiently similar to the original product to be considered worthwhile to study. The molecule shaded in red is the original product, while those shaded in blue are intermediates in the literature route to cilomilast. The unshaded molecules represent novel products.

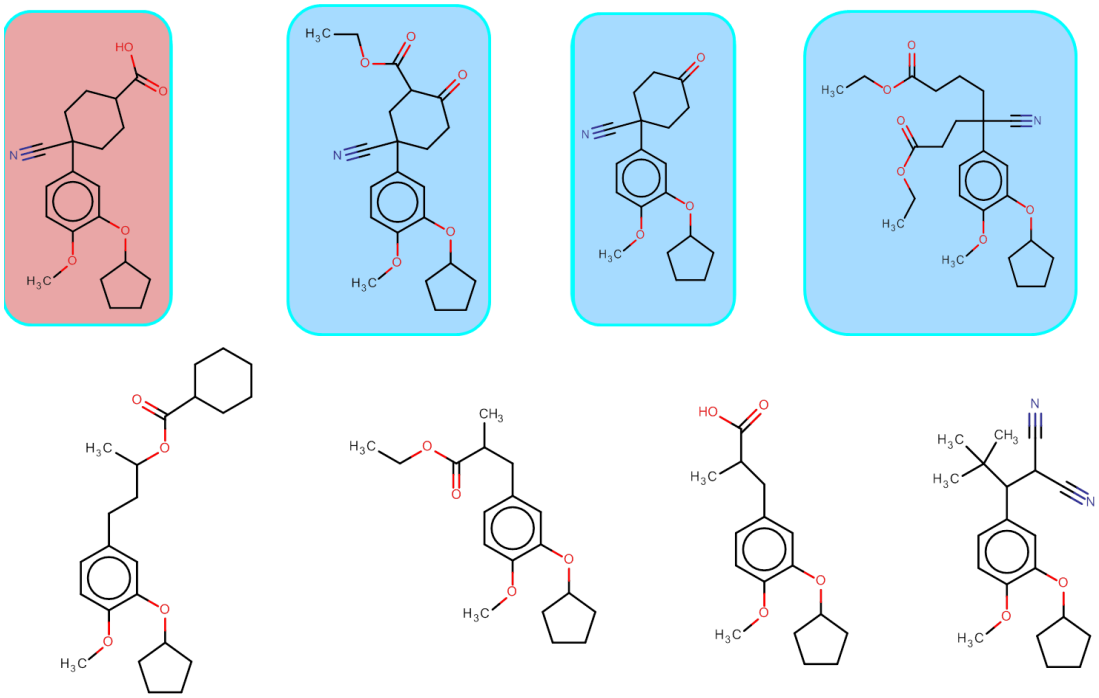


Figure 7.5: The 'Near Neighbour' products produced by the structure generation tool, including some molecules from the literature route (blue). (Wallace, 2015)

By referring to the identifiers associated with the sequences used to generate the near neighbours, potential synthetic routes can be retrieved. The various synthetic routes used are illustrated in a spider diagram in Figure 7.6, taking the original starting material as a start point. The literature route to cilomilast (highlighted in red) is shown with black arrows, with the other routes indicated by colour coded arrows originating from the starting material (green).

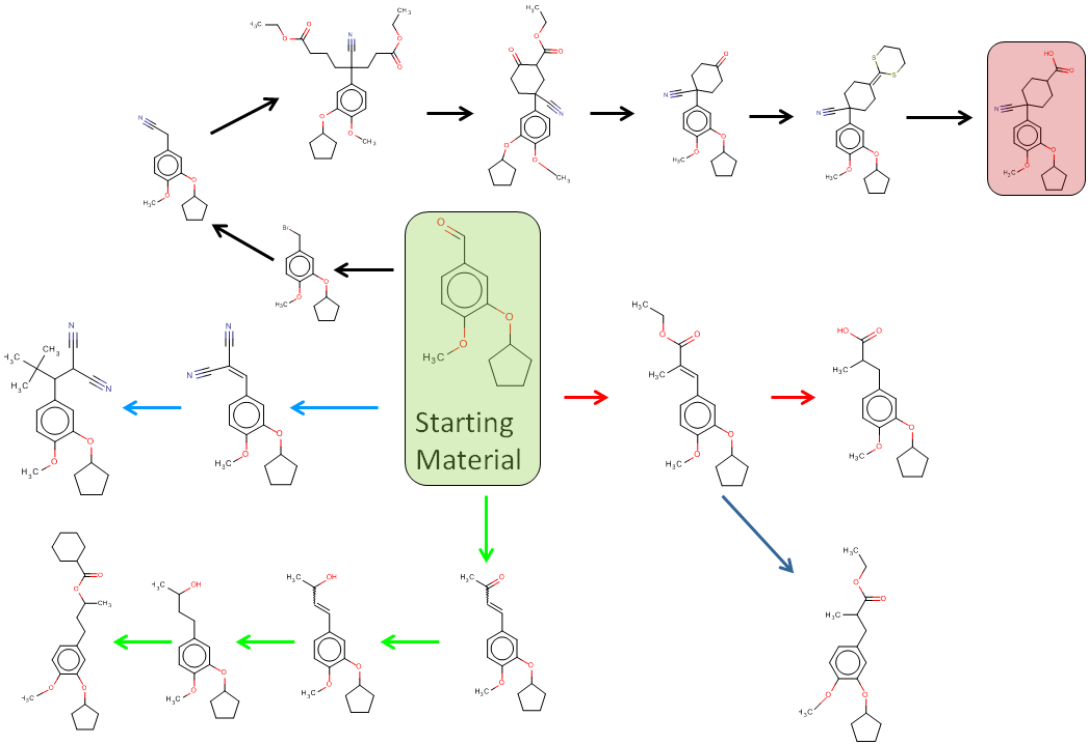


Figure 7.6: Spider diagram showing the routes to the near neighbours of cilomilast. (Wallace, 2015)

While the original literature route is indeed detected and represented (black arrows), it is interesting to note that the various new analogues are produced from much shorter sequences, with the novel products proving to be very accessible from a synthetic standpoint.

7.3.2 SAR exploration example 2 – hydroxamates

Another example is based on a literature study involving the generation of 15 hydroxamates (Bailey et al., 2008) through R group modification. These compounds are of particular interest for topical applications to treat fibroplasia, a condition in which excessive fibrous tissue forms near wounds or infection sites. The mechanism of action is to intervene with the mechanism of collagen production and deposition by binding to procollagen C-proteinase (PCP), preventing these excessive tissues from forming.

7.3.2.1 Structure generation

The original literature route is shown in Figure 7.7, with the starting material in green, and the 15 literature products shown in Figure 7.8. This starting material was used as input to the structure generation tool, using the JMCRoughley RSVs as before.

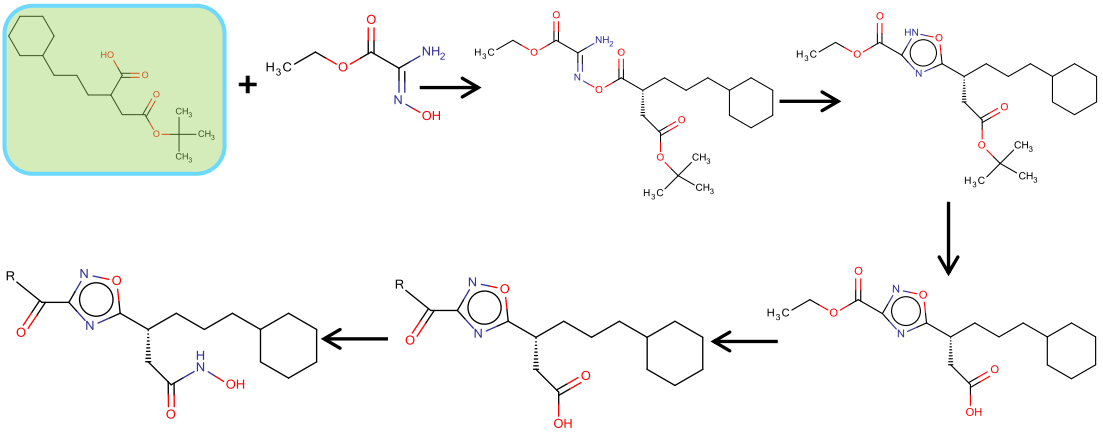
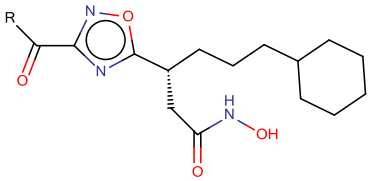


Figure 7.7: Generic literature route to hydroxamates based on the published starting material (green). (Bailey et al., 2008)



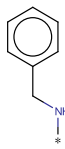
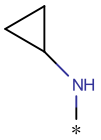
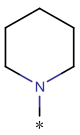
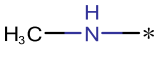
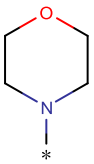
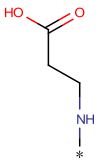
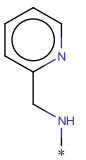
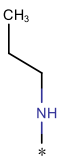
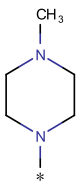
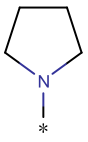
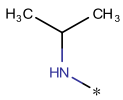
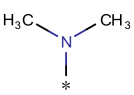
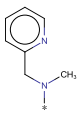
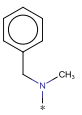
R groups				
				
				
				

Figure 7.8: The hydroxamate products generated in the original literature. (* indicates attachment point) (Bailey et al., 2008)

In total, 1,514 unique products were generated. The results were filtered by similarity to the originally reported product molecules as before. In total 23 near neighbour molecules (having a Tanimoto coefficient value between 0.99 and 0.8 relative to at least one of the products) were generated, with 14 unique molecules left after filtering for duplicates. These near neighbour products are summarised in Figure 7.9. It should be noted that the main structural difference between these compounds and the literature products is a change to the structural scaffold (highlighted), with all but one of the R substituents being identical to those represented in Figure 7.8.

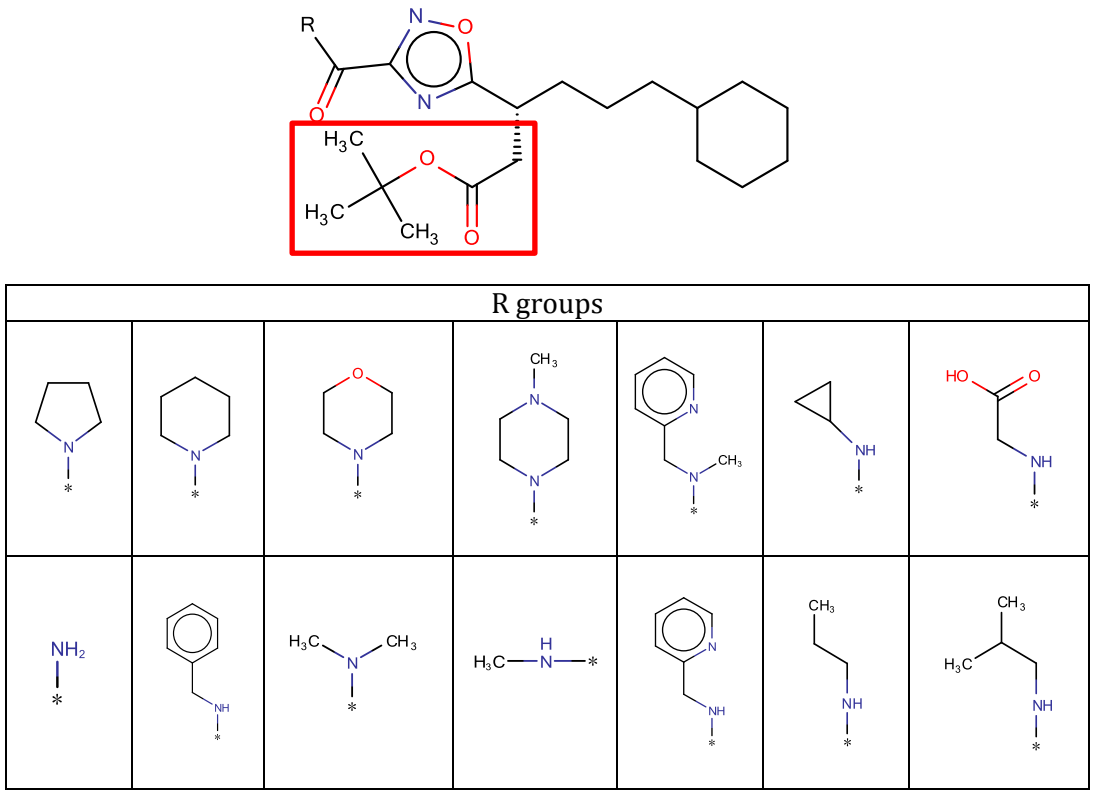


Figure 7.9: New 'Near Neighbour' products produced by the structure generation tool.(* indicates attachment point) (Wallace, 2015)

The 14 near neighbours, along with the reaction pathways used to generate them, could be of significant interest to the medicinal chemist. This is because they represent potentially interesting areas of SAR space near the published analogues that have not been fully explored. A literature search for the near neighbours in SciFinder (Chemical Abstract Services) reveals that eight of these compounds have literature references to a patent granted to the authors of the original study related to this activity class, and can therefore be considered as potential treatments for fibroplasia. However, no activity information is recorded for these compounds specifically, as these lack the zinc

chelating hydroxamic acid group needed to be effective. The remaining six compounds are not in SciFinder, nor are they referenced as part of the generic hydroxamate structure in the original paper and so there is no evidence that the compounds have been tested for therapeutic activity. Thus, the *de novo* tool has identified analogues that are relatively similar in structure to the compounds with known activity, but without such activity themselves. Lowering the similarity threshold to 0.3 shows that one compound containing a hydroxamate derivative group is generated by the *de novo* tool, (shown in Figure 7.10) but this otherwise bears no relationship to any of the original literature examples, lacking any aromatic character. In addition, a large number of compounds are present in the expanded data set with carboxylic acid groups that can weakly bind with the zinc ion, but the expected activity of these is considerably lower than those already reported. It should also be noted that there may be issues with the synthesis of these molecules, in that the ester group in the structural scaffold would need to be protected first to permit functionalisation of the ester bearing the R group.

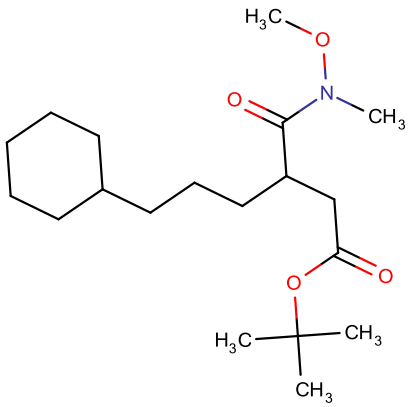


Figure 7.10: Structure produced that contains a hydroxamic acid group for zinc chelation. Similarity coefficient to the literature hydroxamates is 0.352 (Wallace, 2015)

In this context, the more analogues of a particular compound that are produced, the more likely it is that an SAR relationship can be found that enables the key properties of the drug to be optimised. As a result, any synthesisable molecule that is similar, but not identical, to a given start point is worth studying. However, since there may be very large numbers of these depending on the structures involved, additional filtration of these may be required, such as looking for the presence of key substructures, or consideration of their likely interactions with the site of interest. In the next section, the ways in which these types of result molecules can be studied will be explored.

7.3.2.2 Molecule PCA analysis

In order to determine how the near neighbours are related to one another and to the literature compounds in property space, a Principal Component Analysis (PCA) (Abdi and Williams, 2010) was performed based on parameters generated by the Chemistry Development Kit (CDK) (Steinbeck et al., 2003). As these parameters include 3D structural features, conformations were generated using the CDK tools, which include a force field based model builder. The PCA method aims to simplify the description of a complex data set by visualising the similarities and differences between data points graphically. This is achieved through the conversion of a set of descriptors into a smaller set of principal components that are linear combinations of the original descriptors. The full set of topological parameters available through CDK were used (such as the distances between carbon, nitrogen and oxygen atoms in the molecule, the nature of carbon hybridisation states, and measures of charge and polarisability). In addition, in order to assess shape similarity (for receptor binding, or other similar activity models), a number of geometric parameters were also included, such as the moments of inertia for the molecules in free space. A complete list of the descriptors used is given in Appendix B. The combination of descriptors is performed in such a way that each principal component is independent and orthogonal to the others. The molecules are plotted in a 3D space that is constructed using the first three principal components. The molecules are then displayed as 3D conformers using the CheS-Mapper (Gutlein et al., 2012) PCA tool available in KNIME, which automates the whole PCA process. This plot can then be navigated as a real time 3D visualisation, with the molecule representations given colour coding based on the 2D similarity values.

Figure 7.11 shows a 3D PCA plot of the literature hydroxamate molecules and the near neighbours generated using RSVs. The plot is coloured according to the structure similarity values, with yellow molecules representing the literature products, red molecules representing those generated molecules within 0.8 Tanimoto similarity of at least one literature product (based on the same Indigo structural fingerprints as before), and blue molecules representing the remainder of the products generated.

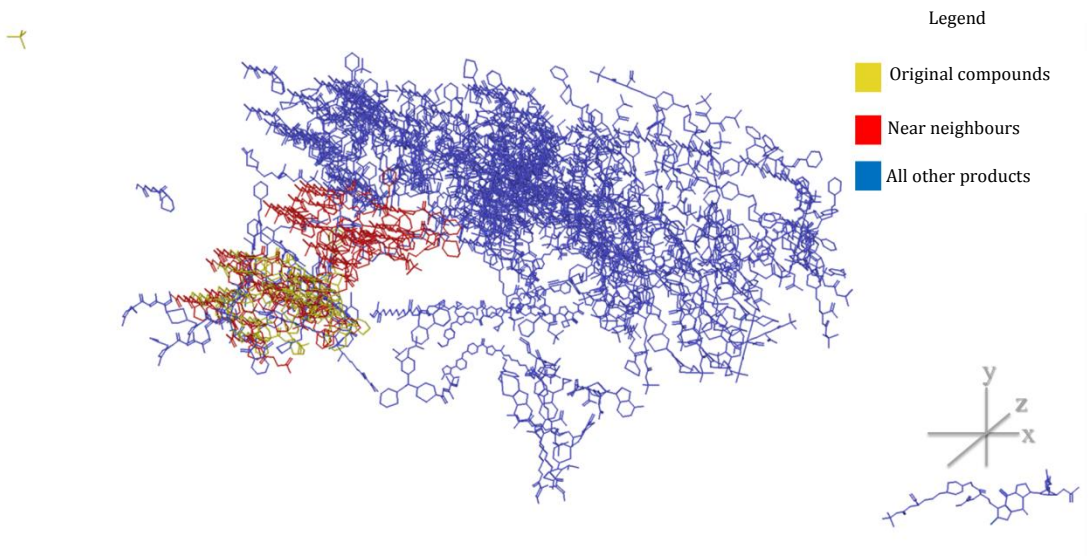


Figure 7.11: 3D PCA plot of the generated hydroxamates and associated products. (Bailey et al., 2008, Wallace, 2015)

The figure shows that a very large number of molecules has been generated covering a wide area of topological space. Figure 7.12 shows only those molecules that are within the 0.8 Tanimoto similarity range calculated previously based on the literature products, while Figure 7.13 shows an expansion of the area around the known products, showing all generated molecules in the immediate vicinity.

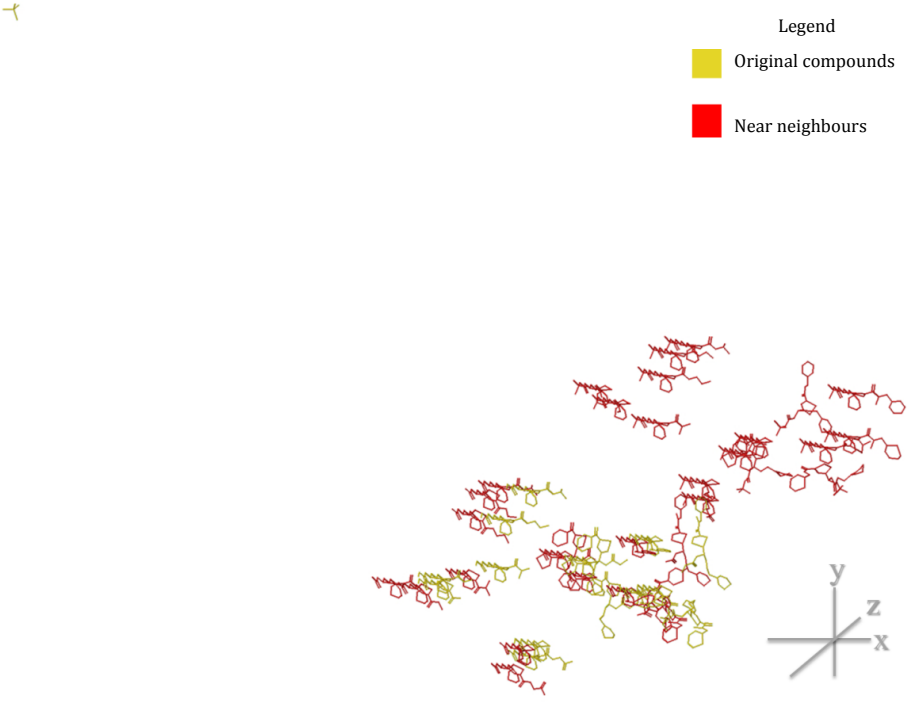


Figure 7.12: 3D PCA plot of the generated hydroxamates and near neighbours. (Bailey et al., 2008, Wallace, 2015)

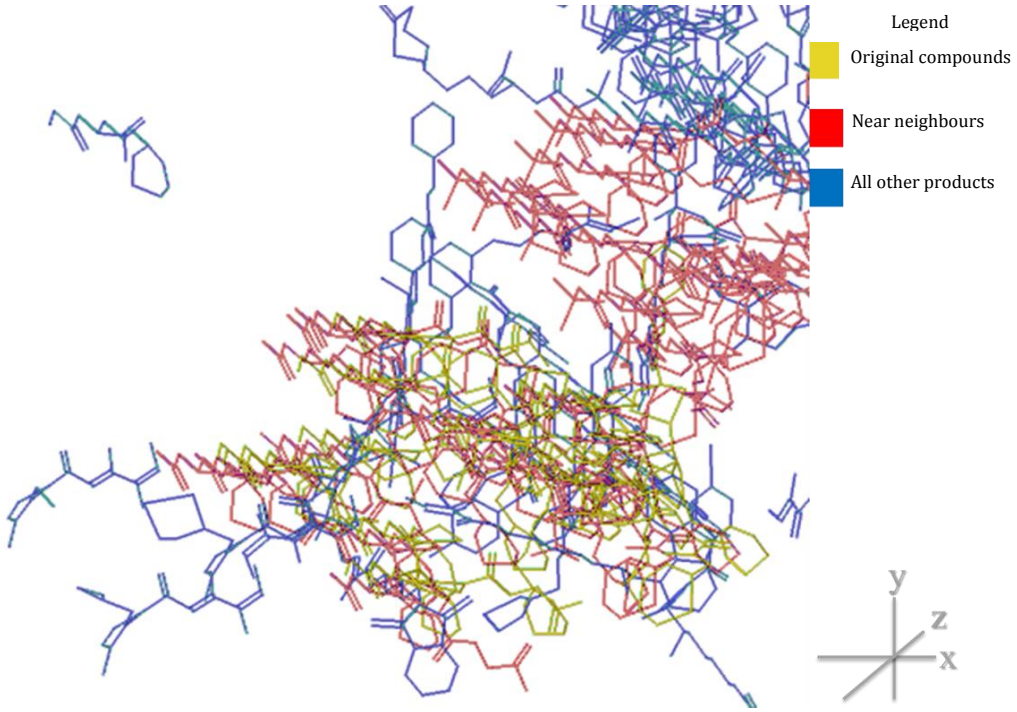


Figure 7.13: Expanded 3D PCA plot showing the relationship between the generated hydroxamates, near neighbours and other products. (Bailey et al., 2008, Wallace, 2015)

Looking at the yellow molecules alone, it can be seen that large areas of the PCA space are underrepresented. Adding in the near neighbours helps to fill in these gaps, and offers some new ideas that the medicinal chemists may want to evaluate. The blue molecules on the other hand represent molecules that may be close in descriptor space, but more distant in terms of Tanimoto similarity, as illustrated in Figure 7.12. These examples could also be worth investigating, in terms of potential scaffold hopping.

7.3.2.3 Expanding the reaction network

The molecules generated thus far are the result of applying RSVs to a known starting material. While this can lead to some exploration of SAR space, as shown in Chapter 6, RSVs limit the number of molecules that can be generated compared to application of the individual RVs that represent each step in a sequence. Further exploration of the SAR space and additional synthetic routes can be identified by extracting the individual RVs expanding out from a selection of the routes identified by the RSVs. The result is a more generalised set of reaction steps, which can be used to provide more information on the intermediate reactions that could occur between different pathways.

It is first necessary to retrieve the reaction sequences for the relevant RSVs used to generate the products. From these sequences, it is possible to obtain the individual reaction steps, and generate RVs from these. As a complete enumeration of products from each step is required, all of the components of each relevant reaction (including reactants, reagents and products) are placed into a communal molecule pool. These are then used as starting materials for the structure generation process, increasing the chemistry space represented by the network.

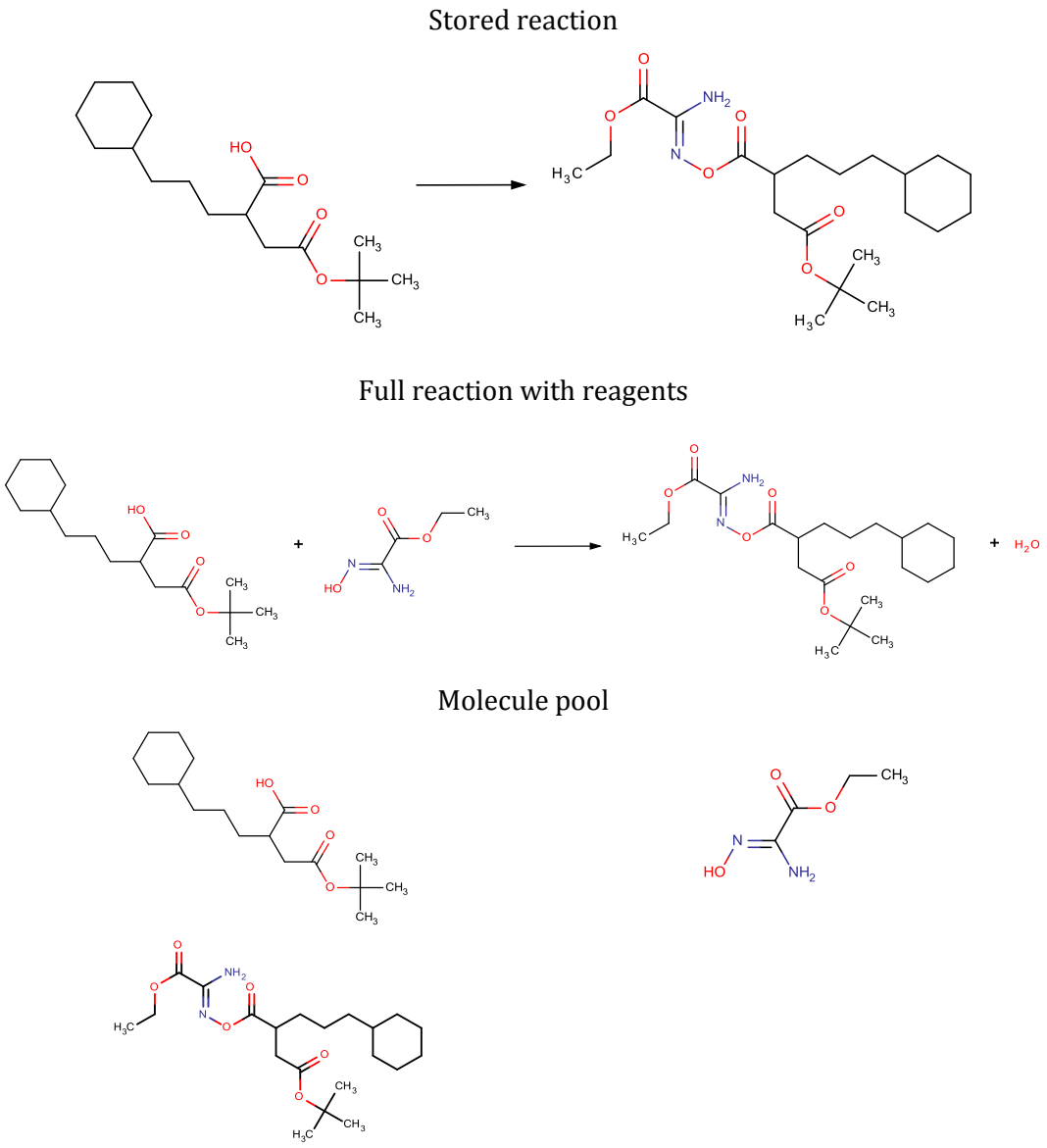


Figure 7.14: Examples of the different reaction and reagent types collected. (Bailey et al., 2008, Wallace, 2015)

The collected RVs are then applied to the starting materials with the generated products forming the starting materials for the next reaction step and so on, until the

sequence is completed. This process is then repeated for all sequences used to generate products of interest, until every molecule (starting material or product) has been used with every RV. The four key steps of the process are shown in the flowchart, given in Figure 7.14. As in Section 6.3.3, by using the individual steps rather than just the sequence as a whole, more information about side products and intermediates is collected. This effectively creates an expansion of the network in the region of these sequences, as seen in Figure 7.15.

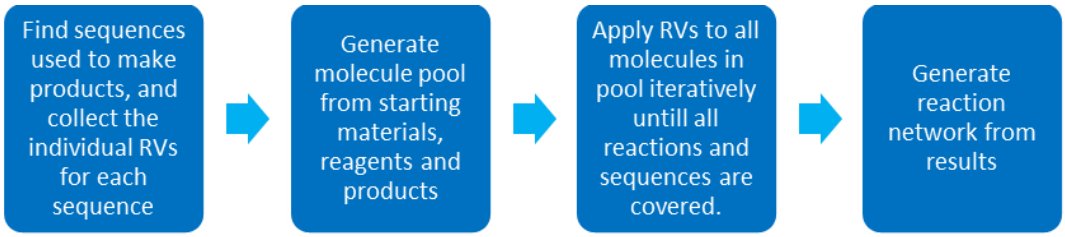


Figure 7.15: Flowchart showing the expanded network generation process.

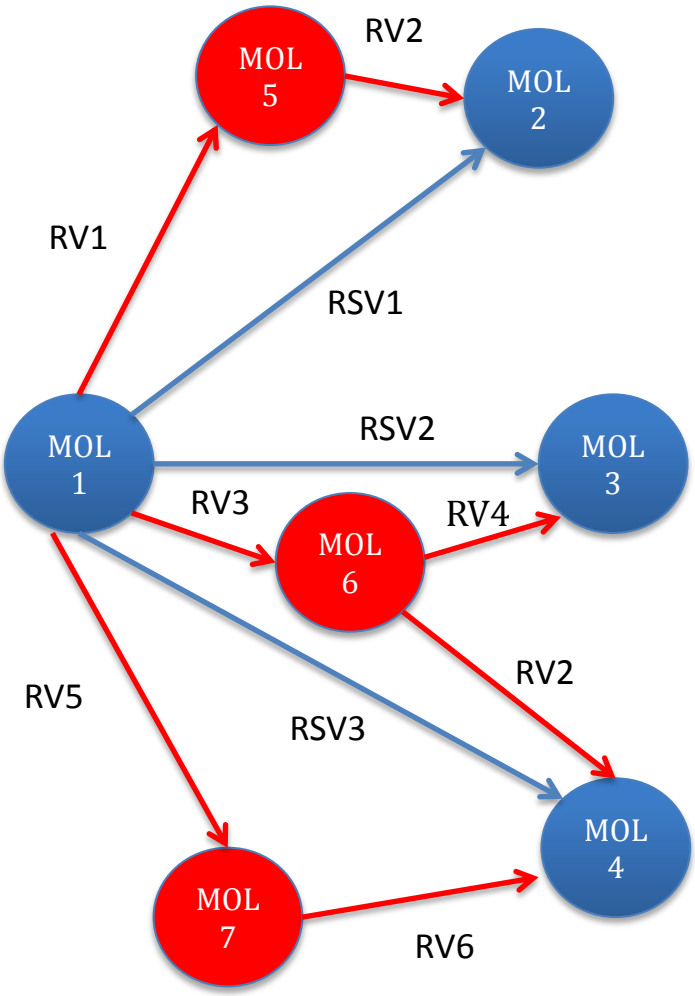


Figure 7.16: Illustration of an expanded reaction network. All routes leading to MOL2, 3 and 4 are shown, although in reality other endpoints may exist.

Figure 7.15 shows the expansion from a single start point (MOL1) via three separate sequences. The blue edges and nodes represent the molecules produced from the stored RSVs, labelled as MOL2, MOL3 and MOL4. However, if the relevant RVs are used instead of the RSVs, additional molecules are generated (represented by the red nodes and edges). In this limited example, the addition of intermediates enables the discovery of two routes between MOL1 and MOL4 for review. As well as the expected route derived from RSV3 (RV5 and RV6), a combination of RV3 and RV2 also produces the same result. This latter route is only discovered when including RVs and intermediates in the network.

Figure 7.16 shows an expanded network based on the hydroxamate synthesis which can be used to identify different synthetic routes and other interesting compounds. In this case, the individual highlight boxes represent starting materials which can be used to generate the desired products, the green box representing the literature starting material, and the yellow box representing an alternative starting material. These materials are then shown with the routes used to generate the desired products (highlighted in red).

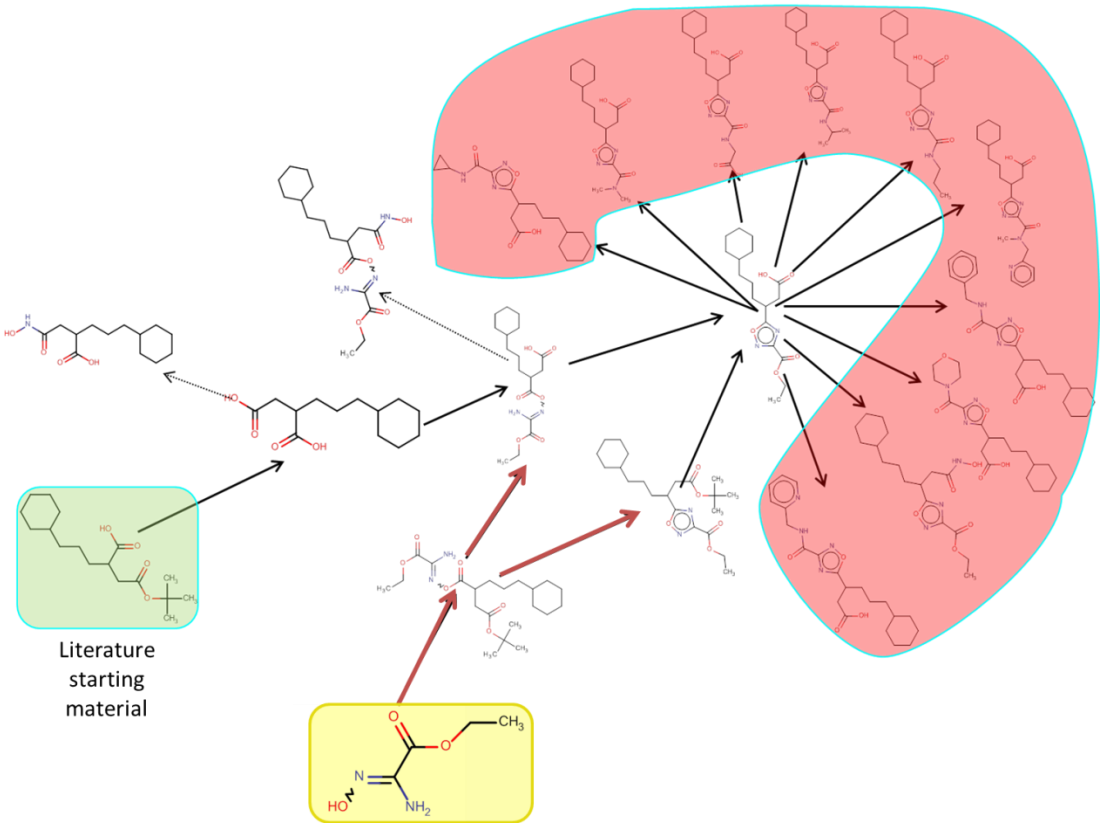


Figure 7.17: Illustration of interconnected routes found by expanding the RSV network to include RVs. The red arrows show the start of an alternative route via different starting material. (Wallace, 2015)

As can be seen in Figure 7.16, the individual paths through the network constitute sequences. The small size of this network enables the synthetic paths to be visualised by embedding the 2D structures for the molecules on the nodes. This is implemented within KNIME using the Prefuse graph library (Heer et al., 2005), and can be used with any KNIME network data. The method permits the user to select individual sequences or groups of sequences for expansion. Once the image is created, it is possible to zoom in on key areas of the network, and focus on regions a certain number of reaction steps away from a given molecule, giving a greater ability to interact with the presented data. This makes the approach more useful from a medicinal chemistry standpoint, as these smaller, more detailed network views can be used as part of a synthesis planning step, showing all of the possible products formed from a given start point, as well as any alternative pathways. In the examples in Figure 7.16, the two identified 'starting materials' are actually the starting material and reagent from the literature route. However, the expansion approach shows that the reagent in yellow can be used as part

of a route proceeding through different intermediates and using different reactions, making it interesting in its own right.

This expansion method also highlights a further use case for the approach, where the different vector methods can be used to highlight the best routes to particular products, effectively using RSVs for initial sampling, before reviewing the relevant RVs in more detail. This would also highlight any potential competing side reactions that can cause problems with synthesis.

7.3.3 SAR exploration example 3 – biaryl carboxamides

A third application was investigated, extracted from a paper outlining the synthesis of biaryl carboxamides as agonists for motillin receptors (Westaway et al., 2008). Regulating motillin binding is integral to the treatment of gastroparesis, where digestion of food is delayed due to partial paralysis of the stomach. The literature synthetic route for these compounds is outlined in Figure 7.17, and the reported literature products shown in Figure 7.18. The starting material (highlighted in green in Figure 7.17) and the JMCRoughley set of RSVs produces a set of 48 molecules after removal of the literature products, which are shown in Figure 7.19. In this case, the new molecules are not dramatically different from the originally reported products, with some representing new combinations of existing R1 and R2 groups.

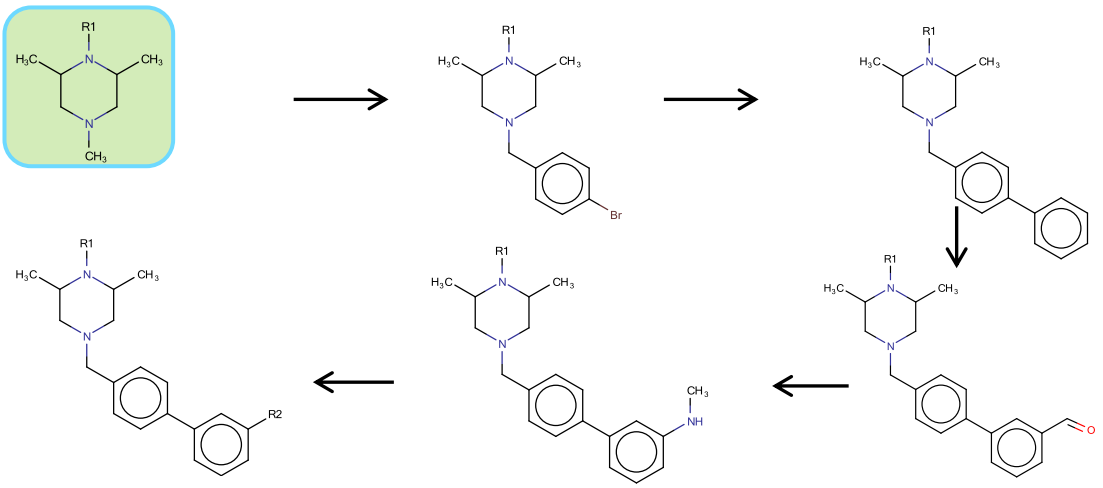
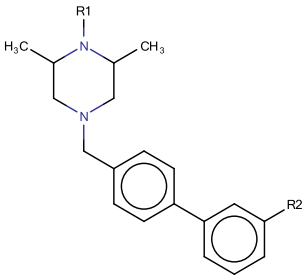
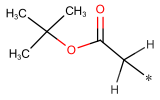
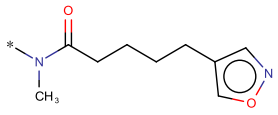
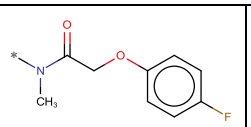
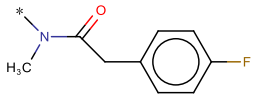
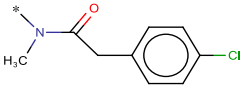
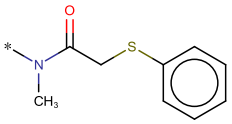
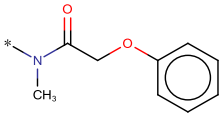
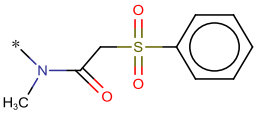
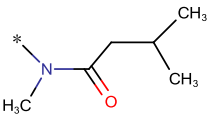


Figure 7.18: General scheme for the synthesis of carboxamides. (Westaway et al., 2008)



R1 = 
R2 groups
   
   

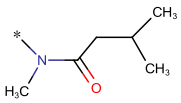
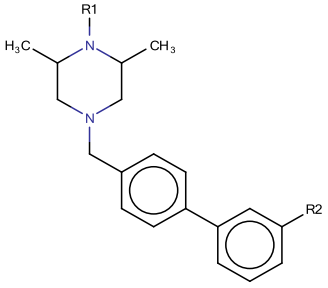
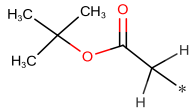
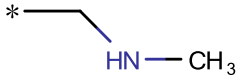
R1 = 
R2 groups


Figure 7.19: Reported literature products generated from the carboxamide literature route. (* indicates attachment point) (Westaway et al., 2008)



R1 = 	
R2 groups	
	

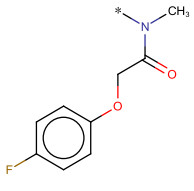
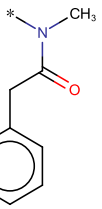
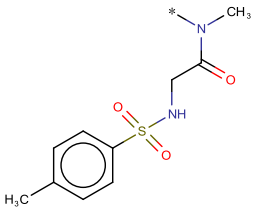
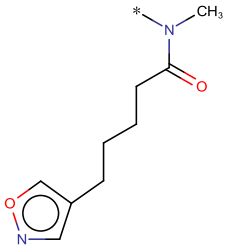
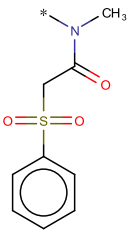
R1 = 				
R2 groups				
				

Figure 7.20: Examples of 'near neighbour' products produced by the structure generation tool from the carboxamide route. (* indicates attachment point) (Wallace, 2015)

In this case, as the number of generated products is relatively small (48 in total), it is not necessary to filter the structures by similarity before generating a PCA plot as seen in Figure 7.20.

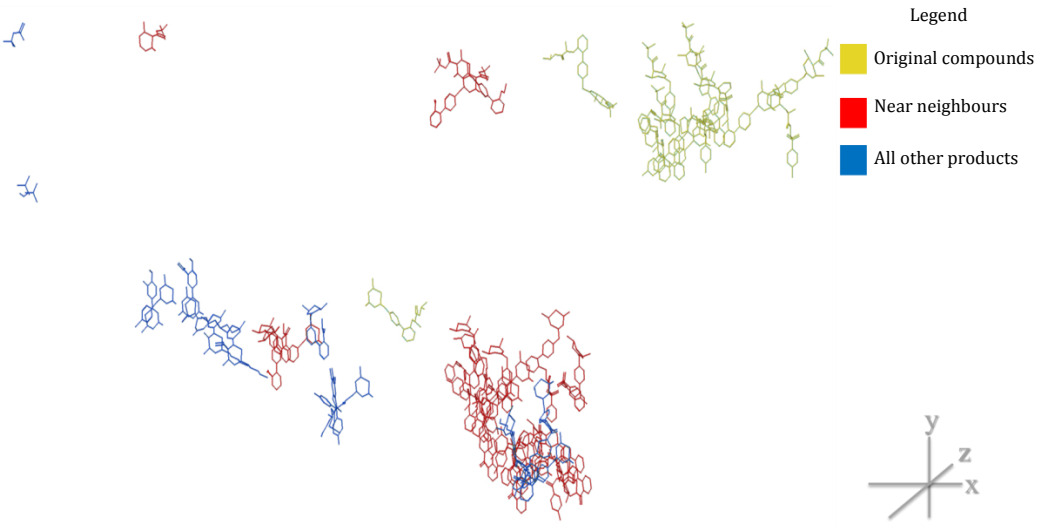
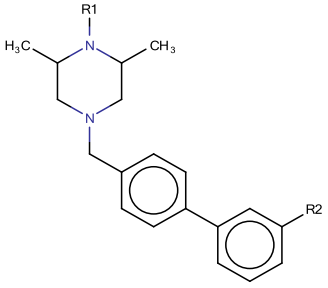


Figure 7.21: PCA analysis of the products of the carboxamide sequences. (Westaway et al., 2008, Wallace, 2015)

Using the same topological and geometric parameters as in the previous PCA plots shows the similarity in properties between the near neighbours and the original products, with the near neighbours overlapping the original compounds in property space far less than in the previous example. While the set of molecules generated is relatively small compared to the hydroxamate example, those that are close to the reported carboxamides still provide interest for further study. As before, the blue molecules in Figure 7.20 indicate compounds that may not be obvious to researchers, but are sufficiently similar in property and three-dimensional space to be valuable potential targets. Some examples of these are shown in Figure 7.21.



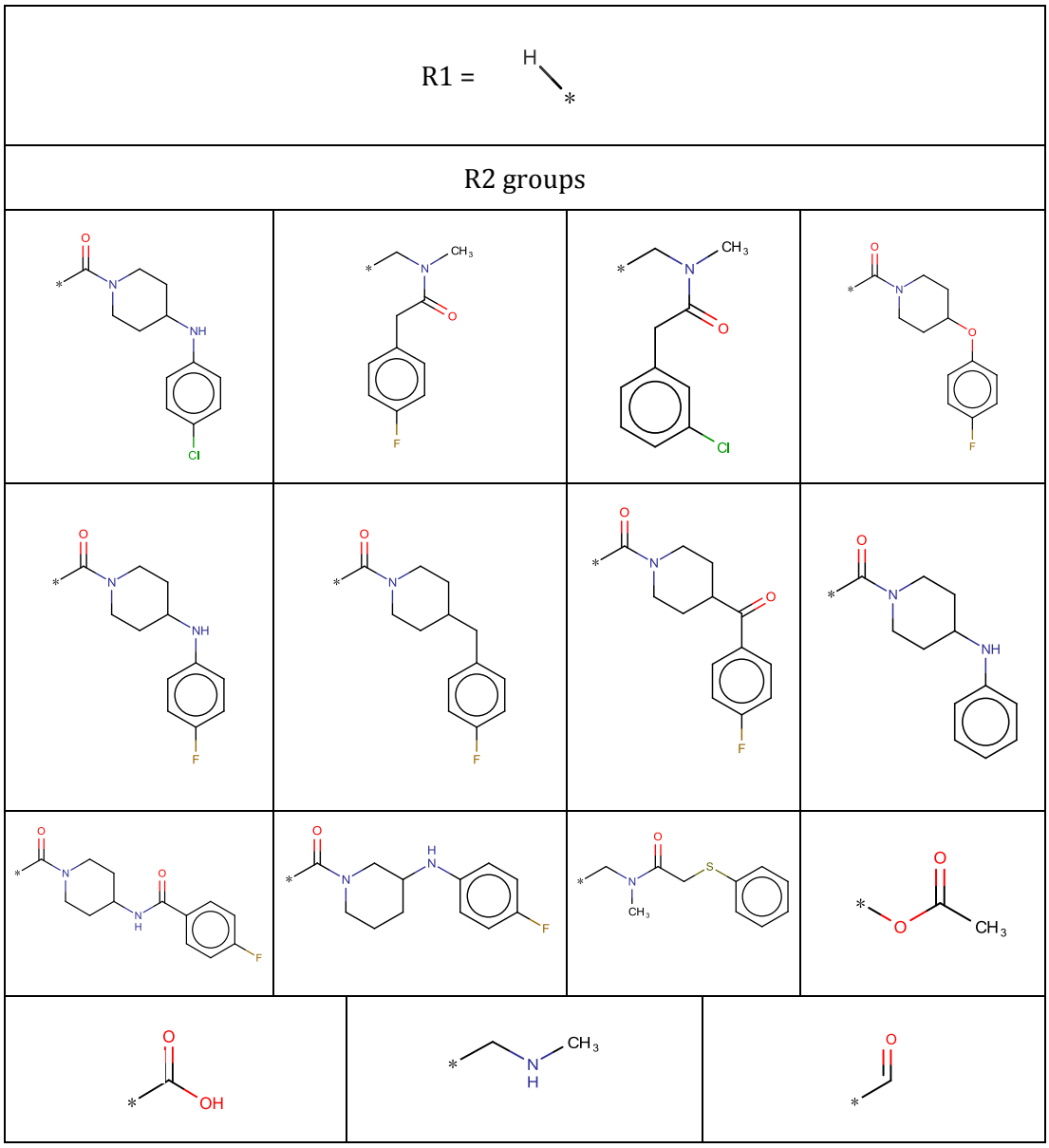


Figure 7.22: Examples of products from the carboxamide route that are outside of the similarity threshold but are of interest. (* indicates attachment point) (Wallace, 2015)

A similarity search for the near neighbour molecules within the SciFinder database returns no references outside of the original paper, suggesting that these compounds are yet to be evaluated for activity.

7.3.4 SAR exploration example 4 – substituted alkynes

In order to further test the network approach, a fourth test set was produced, from a reaction sequence used to provide functionalised alkynes as feedstock for 4-sulfamoyl pyrroles used in statin synthesis (Park et al., 2008). The generic scheme for the synthesis is shown in Figure 7.22, with the literature compounds summarised in Figure 7.23.

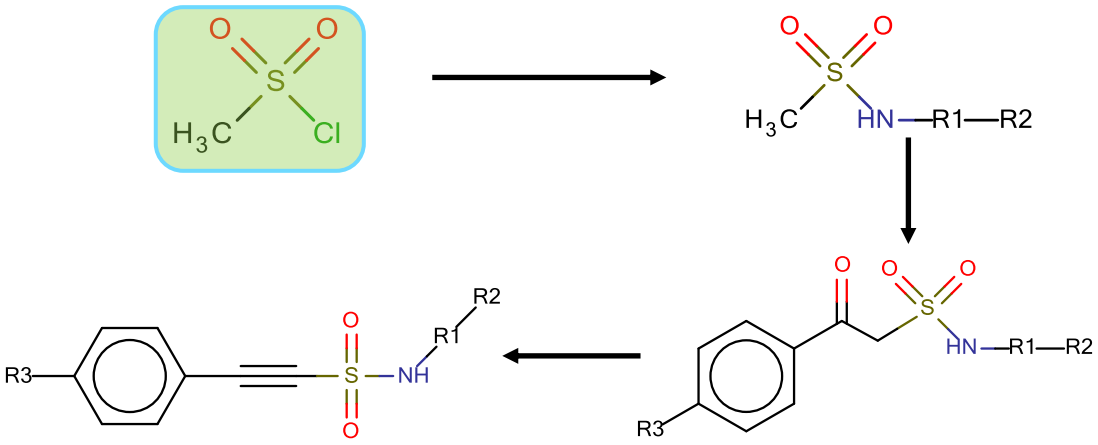
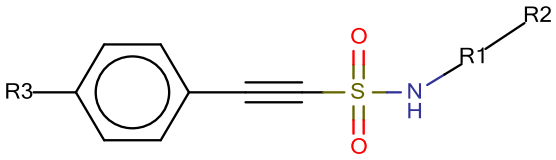
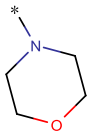
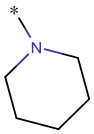
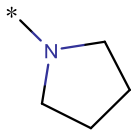
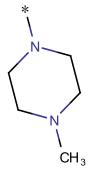
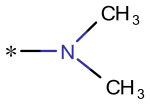
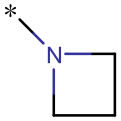
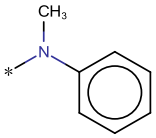
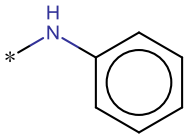
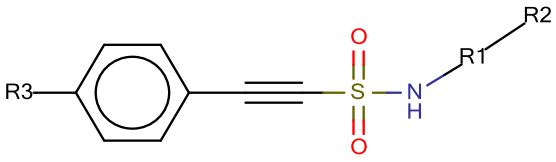


Figure 7.23: Generic route to 4-sulfamoyl alkynes. (Park et al., 2008)



R3=H			
N R1 R2			
			
			



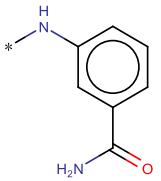
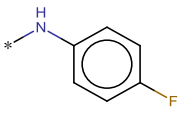
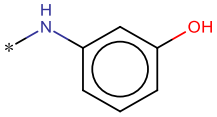
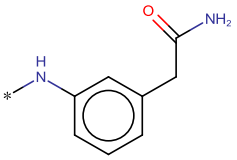
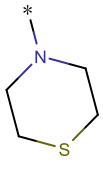
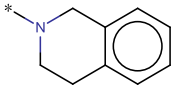
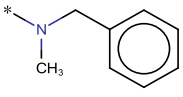
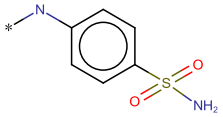
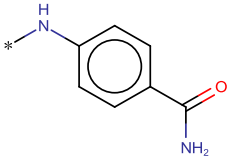
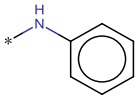
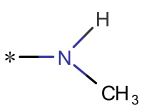
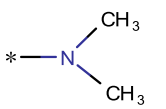
R3=F				
N R1 R2				
				
				
				

Figure 7.24: Reported alkyne products generated from the literature route. (* indicates attachment point) (Park et al., 2008)

If the original starting material is used with the JMCRoughley database as before, 1,707 products are generated, but only three are near neighbour molecules using the Tanimoto score of 0.8 as a threshold, which are highlighted in yellow in Figure 7.24. Altering the similarity threshold to show near neighbour molecules with a Tanimoto score of 0.6 relative to the known products leads to 16 additional molecules of interest, as summarised in Figure 7.24. However, with two exceptions, none of these contain the alkyne functionality required. A substructure search for this functionality through the entire result set returns no additional results.

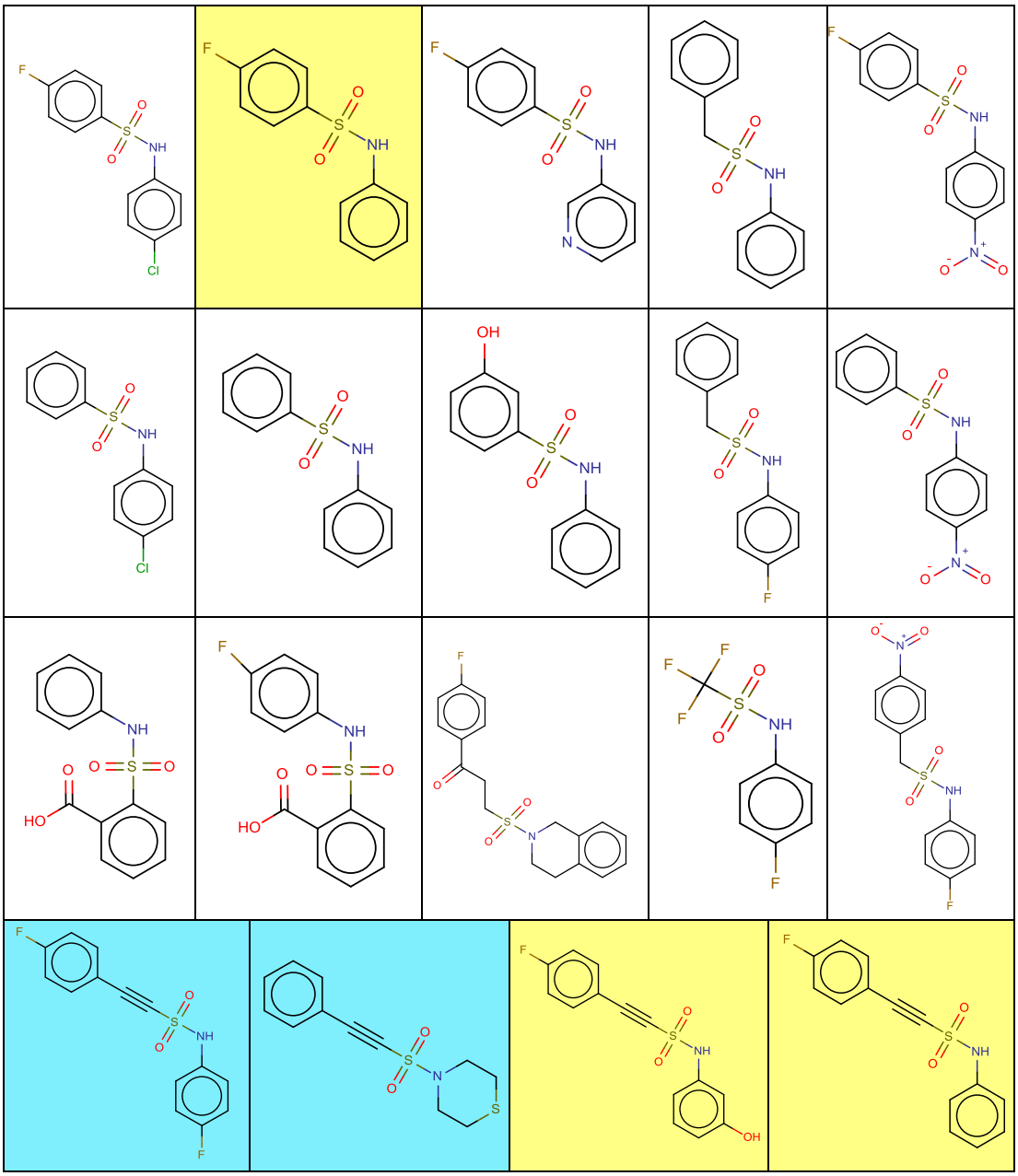


Figure 7.25: Near neighbour alkyne products, including extended similarity threshold. (Wallace, 2015)

A SciFinder search for the near neighbours with alkyne functionality indicates that the two compounds highlighted in blue in Figure 7.24 were identified as part of the patent covering the initial study, but with no reported activity. The other two molecules are not found. The data set can be expressed in a PCA plot, as shown in Figure 7.25. However, when focussing on the near neighbours, it can be seen that there is more of an overlap in property space, with all of the compounds existing in the same narrow region of the plot, as illustrated in Figure 7.26.

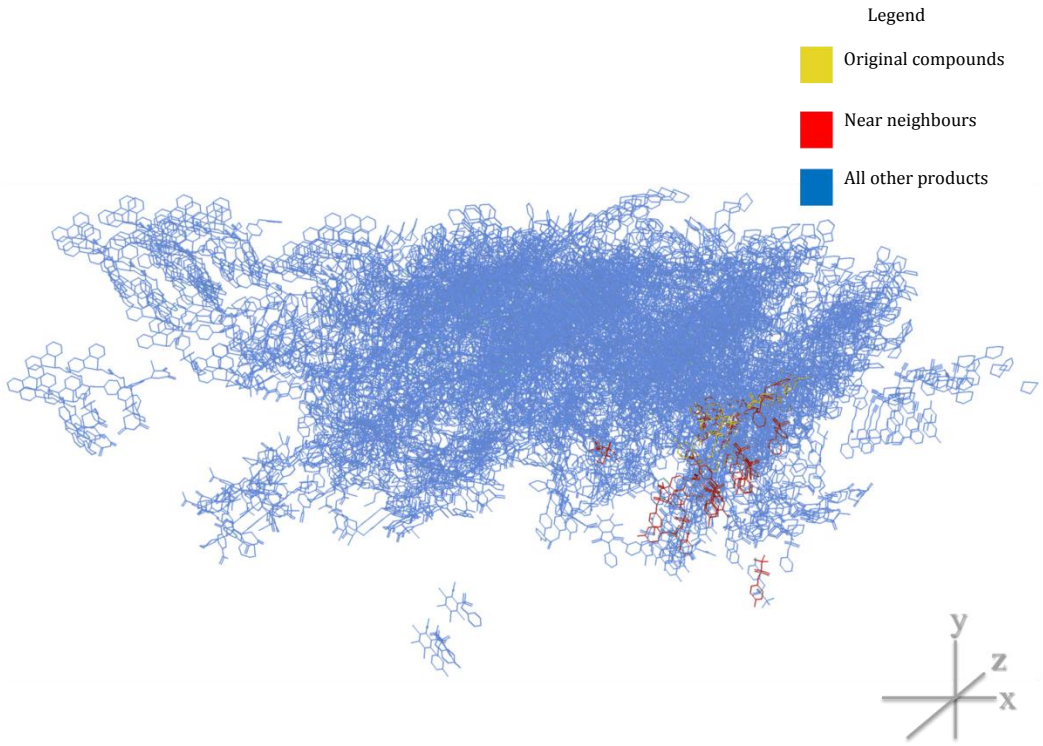


Figure 7.26: PCA analysis of the alkyne products, including selected other products. (Park et al., 2008, Wallace, 2015)

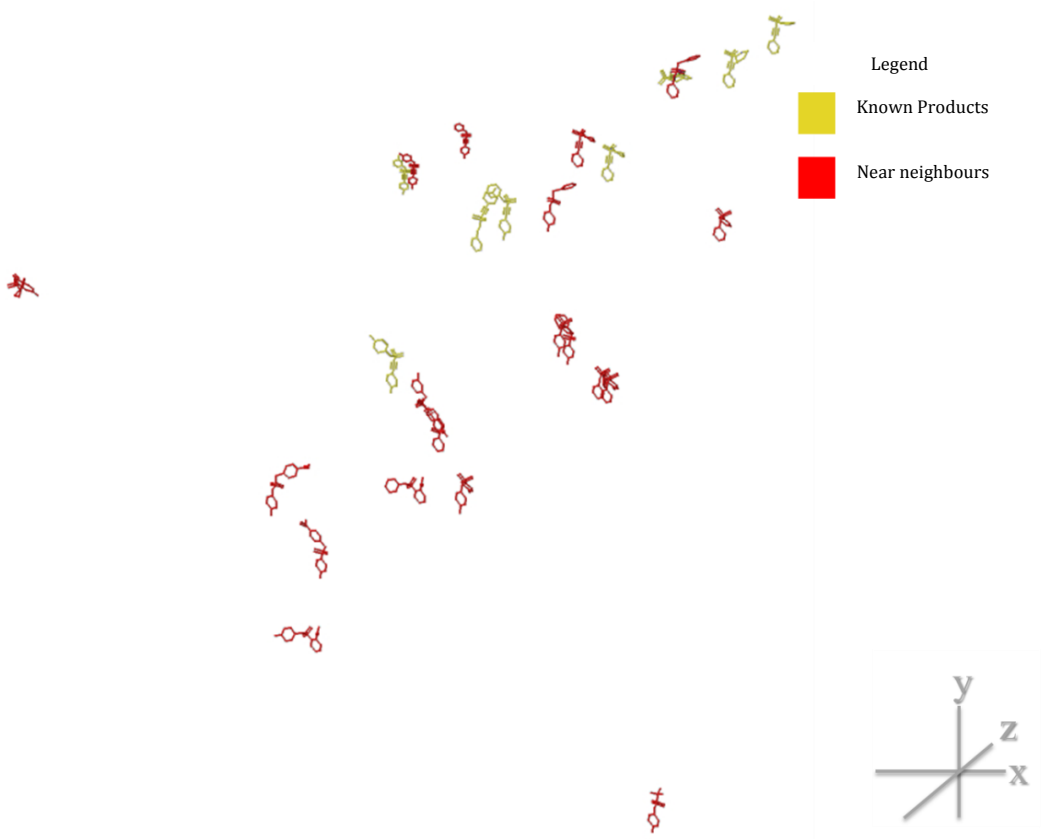


Figure 7.27: PCA analysis of the alkyne products. (Park et al., 2008, Wallace, 2015)

The reaction network associated with the production of the near neighbour molecules shows that the key to producing new analogues in this scheme is via modification of the functionalised benzene ring. The relevant portion of the reaction network is shown in Figure 7.27, expanding the network using the sequence RVs as before, with some of the more similar near neighbour molecules (using the Tanimoto threshold of 0.8) highlighted in blue. The network shows that there is some degree of branching between the synthetic routes, suggesting that there may be worthwhile alternative routes to be explored.

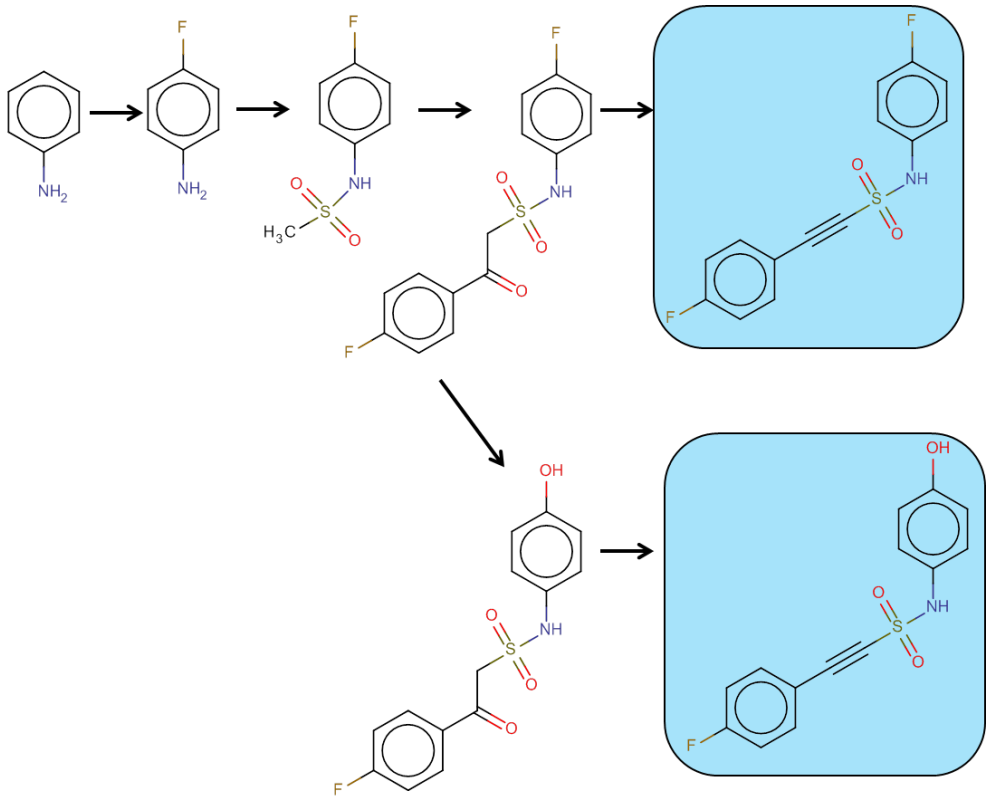


Figure 7.28: Portion of the reaction network showing the generation of near neighbour molecules. (Park et al., 2008, Wallace, 2015)

7.4 Conclusions

As part of the validation method, it was shown that alternative routes to compounds within the reaction network can be found in a number of ways, either through direct interrogation of the stored vectors, or via the assembly of SAR-focussed 'mini networks'. This latter approach is also useful when combined as part of a PCA approach, in which all of the molecules of interest from a result set can be compared based on their properties, with the necessary synthetic pathways highlighted for use. The method has been demonstrated on a number of SAR examples from literature, showing full synthetic routes to the desired products, as well as near neighbour molecules that may be of interest to the medicinal chemist. By combining the RSV approach with some RVs, the area of solution space around a molecule can be expanded, possibly suggesting alternative routes and new compounds of interest.

Chapter 8:
Structure generation with reaction sequence vectors

8.1 Introduction

The case studies outlined in Chapter 7 investigated lead optimisation scenarios by exploring molecules that could be made from a given starting material. In the first part of this chapter, the aim is to investigate the ability of the RSVs to generate molecules that are predicted to be active using a set of starting materials and an SAR model. As one of the main aims of the RSV approach is to generate molecules that are synthetically accessible, the results are also assessed using a computer model for predicting synthetically accessible as well as manually by synthetic chemists. The chapter then describes a comparison of the RSVs and RVs for de novo design based on the same data sets.

8.2 Prediction of activity and structural feasibility

As previously discussed in Section 7.3.2.3, RSVs can be used to quickly generate structures and highlight alternative routes to a given product. These generated structures can be quickly evaluated, with the routes to products of interest analysed further via the relevant RVs. To assess the feasibility of this, a study in activity and structural feasibility prediction was performed. Six data sets were extracted from a paper by Sutherland et al (2004). In Sutherland's work these were used to assess the relative accuracy of different approaches to build QSAR models, and as such each data set represents a well curated collection of molecules covering a wide activity range. The QSAR models used in the experiments described here are support vector machine (SVM) regression models (Cortes and Vapnik, 1995) based on Sutherland's work and were trained and provided by Lilly.

An SVM regression model can be used to predict the value of a property. For the models used here, each molecule in the training data was represented using a Lilly in-house structural fingerprint and has an associated pIC_{50} value (representing the negative log

of the standard IC_{50} measure) measured by Lilly. In SVM regression, the descriptor space is transformed into a higher dimensional space with linear regression performed in this space (Xue et al., 2004). An unknown molecule is then fitted to the regression function to estimate the unknown quality, the pIC_{50} value in this case.

All of the molecules within each data set were used as starting materials with the structure generation tool to produce all possible structures, using the JMC Roughley RSVs. The fingerprints for each generated molecule were calculated and the pIC_{50} value of each generated molecule was predicted using the appropriate SVM model. These result molecules were compared with existing compounds within the class on structure and the number of unique molecules recorded, as well as the range of predicted activity values. The results for each of the inhibitor classes are summarised in Table 8.1, with histograms indicating the pIC_{50} range shown in Figures 8.1 to 8.6. It should be noted that some of the larger, more complex molecules were too structurally dissimilar to the existing inhibitors, and as such, the SVM model was unable to predict pIC_{50} values in these cases. In total, 5,521 of the generated products across the compound classes could not be predicted by the model, and have been excluded from the activity ranges reported in Table 8.1. For the purposes of comparison, suggested threshold pIC_{50} values provided by Lilly are included (such as the value of 7.0 for the Ace data set). These represent the minimum pIC_{50} value a molecule has to have to be considered as worthy of further research interest. Compounds at or above the threshold are referred to as 'considered' or 'predicted' active. Looking at the collected data, there are relatively few unique molecules produced in the Cox-2, Dhfr and Gpb cases, considering the high number of starting materials. In these data sets, there are relatively few differences between the molecules in the sets in terms of their structure. As a consequence, the limited diversity of the transformations in the RSV database result in a large number of identical structures, with the various functional groups on the scaffolds being interconverted between one another.

Inhibitor class	Number of molecules in class	Number of unique molecules generated	Number of compounds without predicted activity	Suggested threshold pIC_{50} value to be considered active	pIC_{50} range for starting materials, (higher is better)	Predicted pIC_{50} range for new products, (higher is better)
Angiotensin converting enzyme (Ace)	114	17789	1585	7.0	2.14-9.90	2.94-9.20
Benzodiazepine (Bzr)	147	13900	1085	7.52	5.52-8.92	6.60-8.63
Cyclooxygenase-2 (Cox-2)	282	8534	589	6.0	4.03-9.00	4.73-8.73
Dihydrofolate reductase (Dhfr)	361	12276	1018	6.52	3.30-9.80	4.05-9.03
Glycogen phosphorylase B (Gpb)	66	2052	88	6.0	1.30-6.80	1.80-4.52
Thermolysin (Therm)	76	13818	1156	6.0	0.52-10.17	1.75-8.02

Table 8.1: Summary of the results of the RSV structure generation experiment with the Sutherland inhibitors.

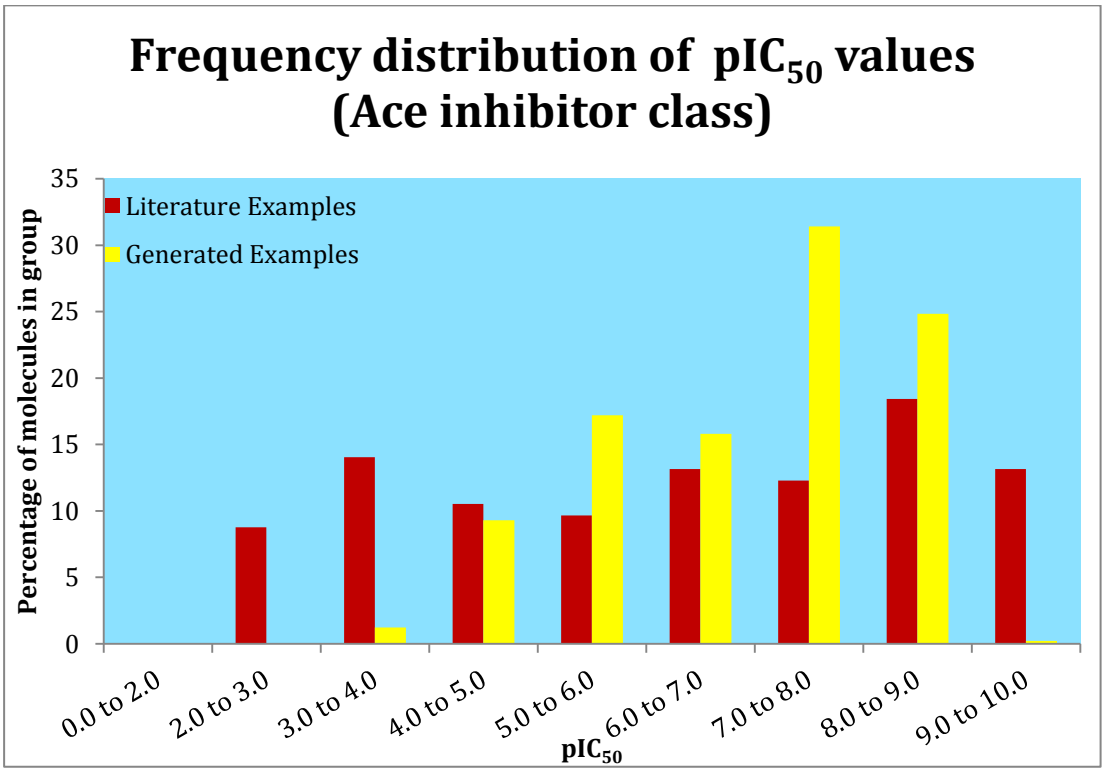


Figure 8.1: Frequency distribution of pIC_{50} values for the literature and generated examples in the Ace inhibitor class.

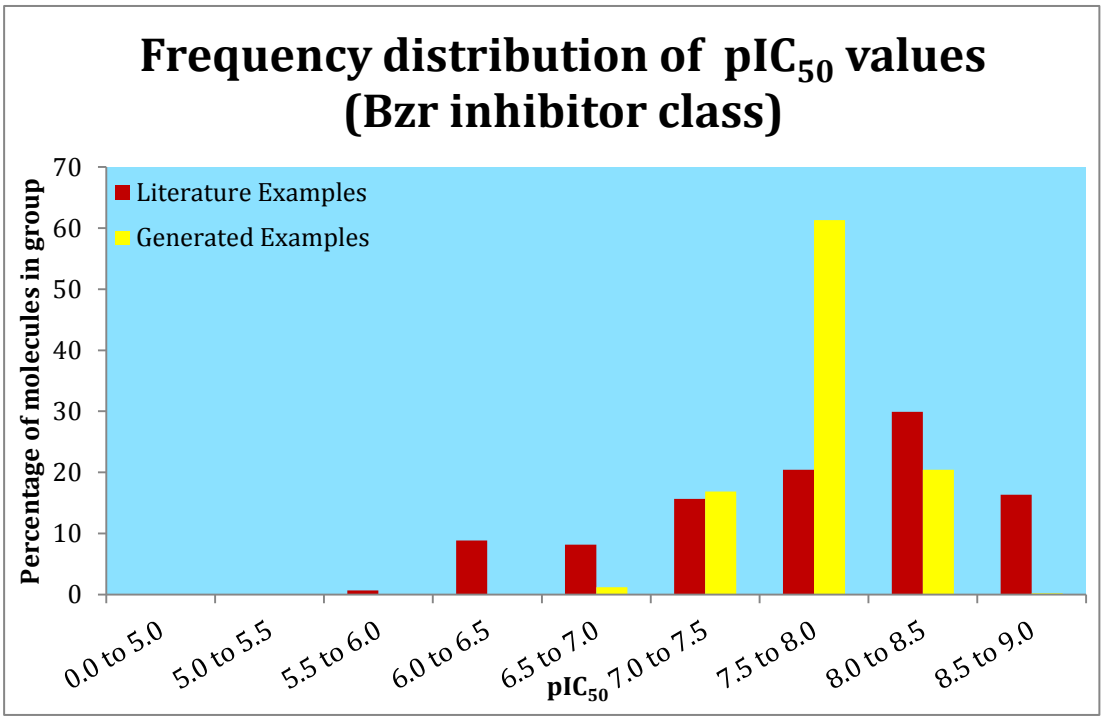


Figure 8.2: Frequency distribution of pIC_{50} values for the literature and generated examples in the Bzr inhibitor class.

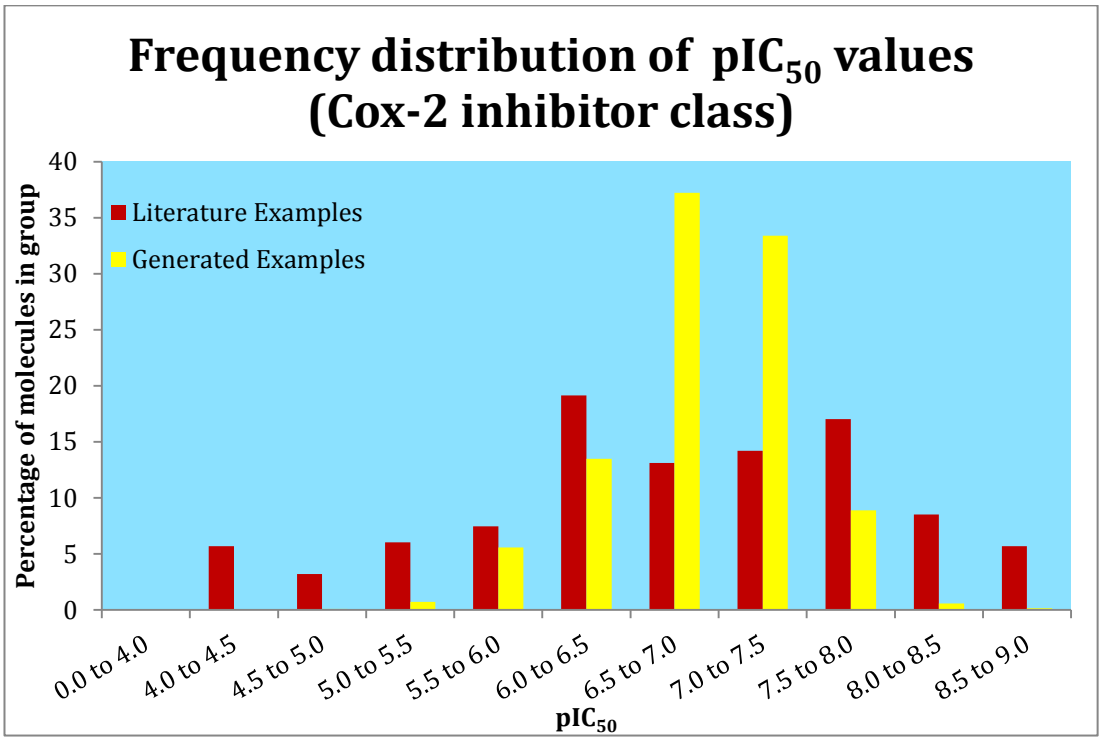


Figure 8.3: Frequency distribution of pIC_{50} values for the literature and generated examples in the Cox-2 inhibitor class.

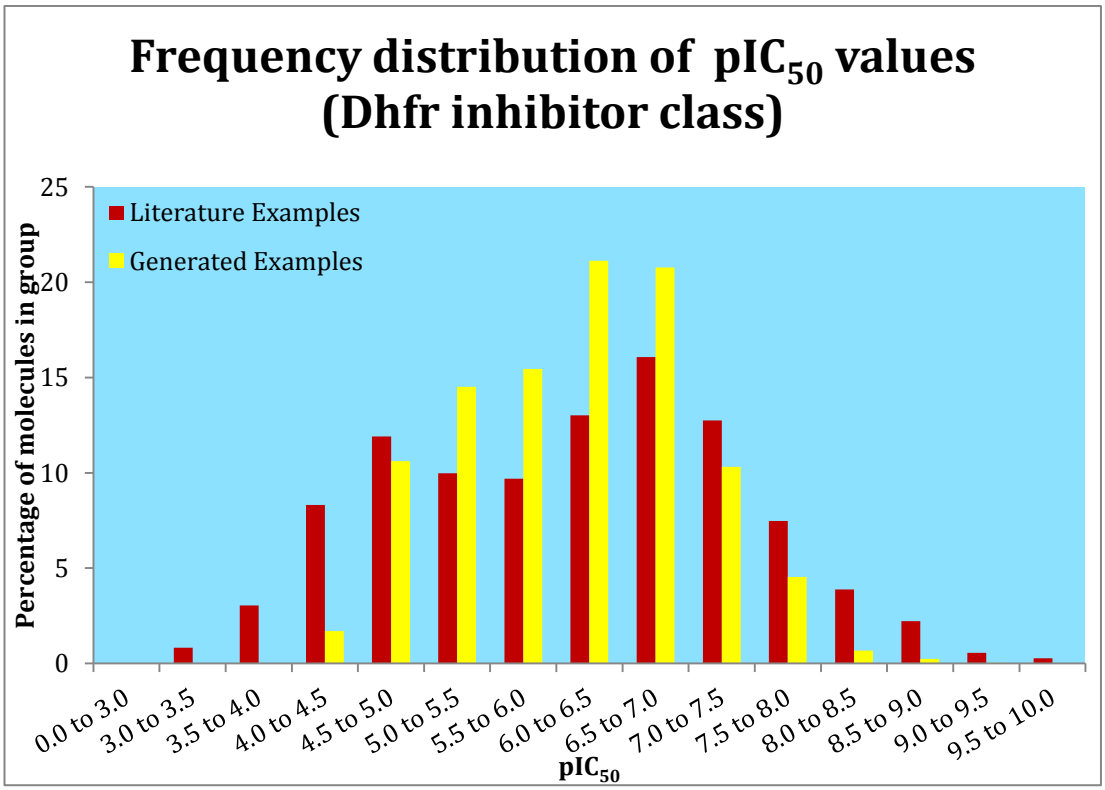


Figure 8.4: Frequency distribution of pIC_{50} values for the literature and generated examples in the Dhfr inhibitor class.

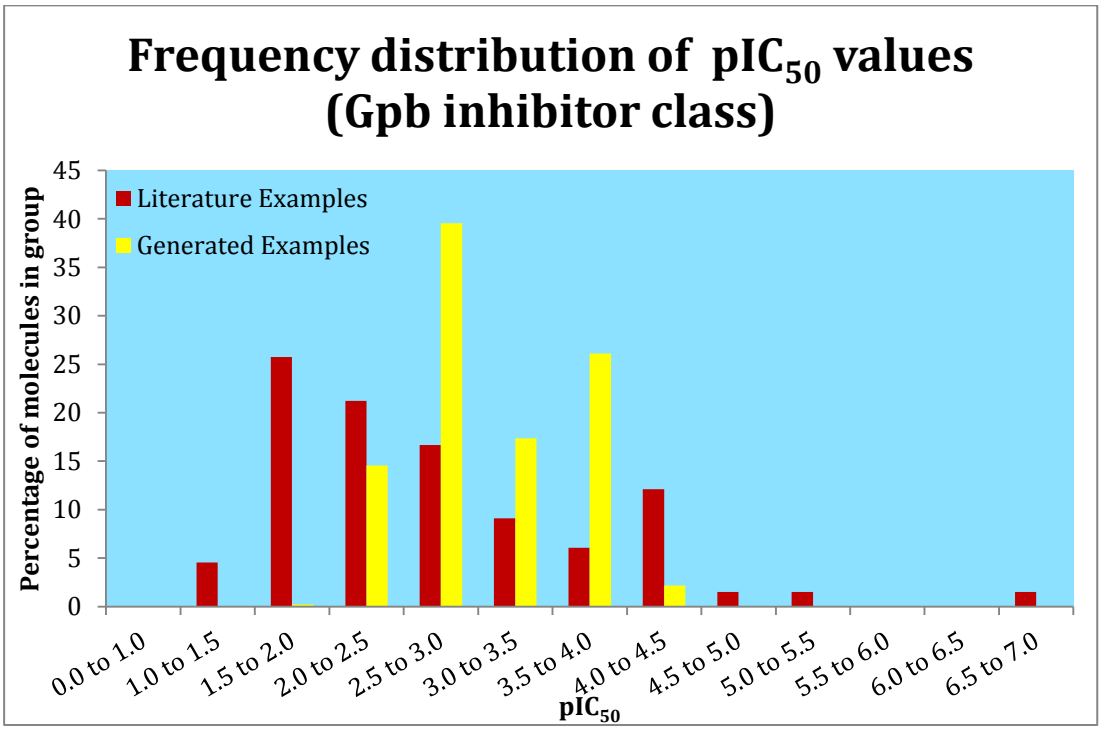


Figure 8.5: Frequency distribution of pIC_{50} values for the literature and generated examples in the Gpb inhibitor class.

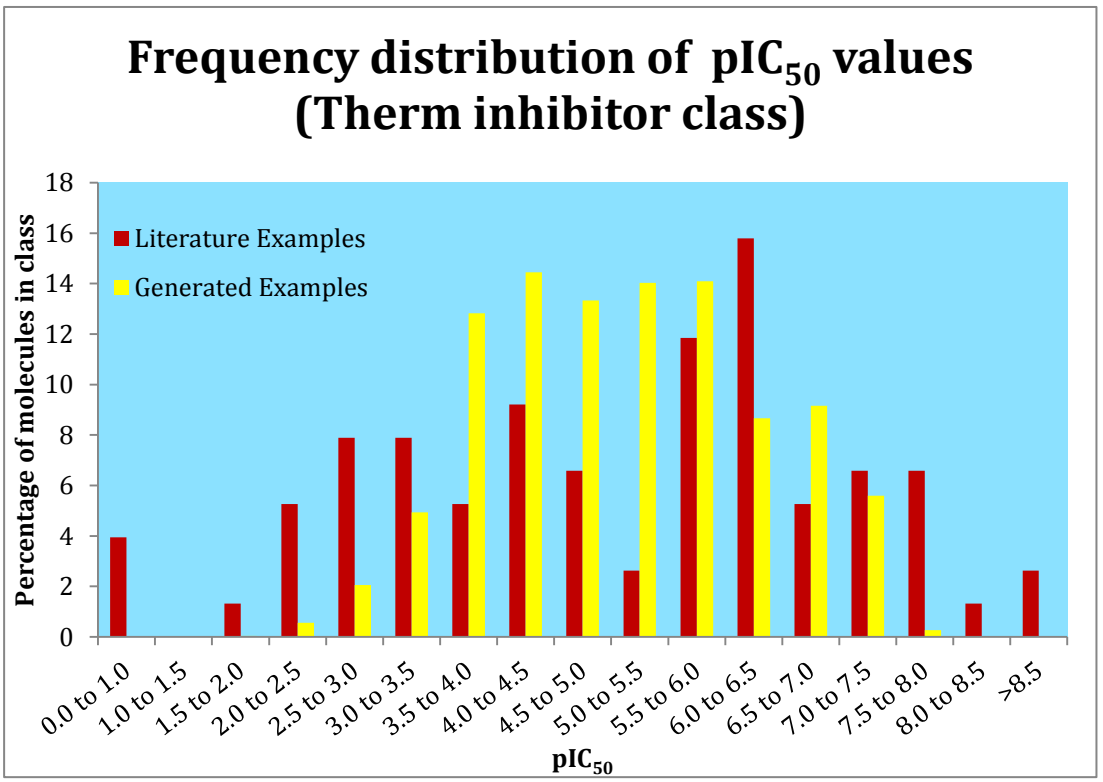
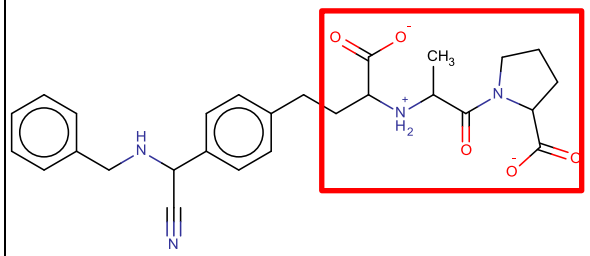
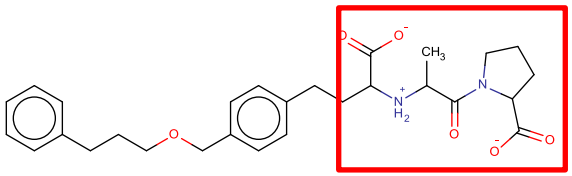
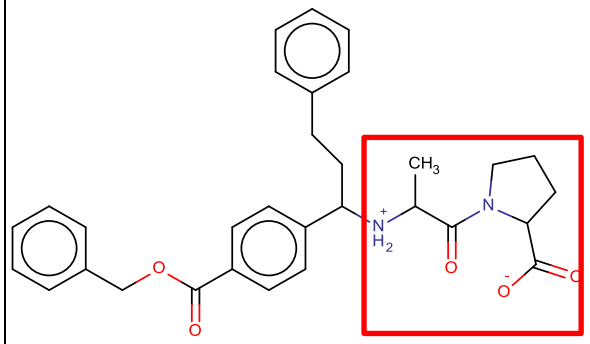


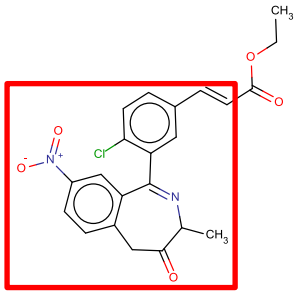
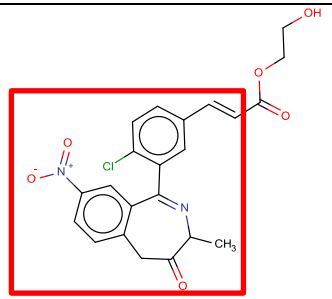
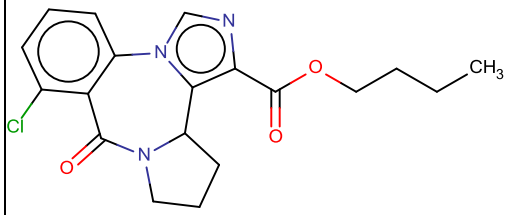
Figure 8.6: Frequency distribution of pIC_{50} values for the literature and generated examples in the Therm inhibitor class.

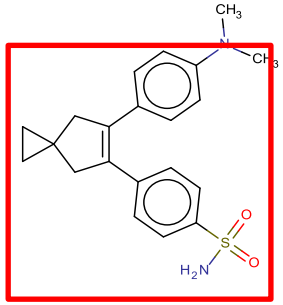
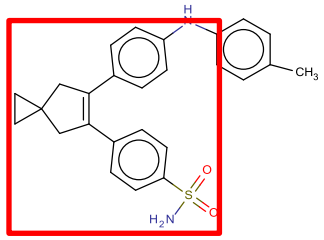
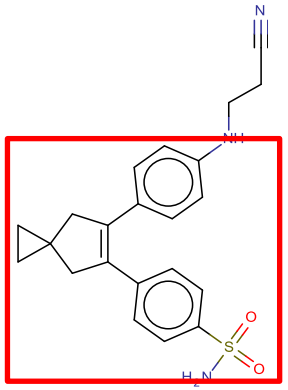
Considering the data as a whole, the RSVs generated from the JMC Roughley sequences appear to be able to construct a significant number of inhibitors of interest. For the new products the SVM models predict inhibition activities that fit well within the ranges of the existing classes. A comparison of the frequency distributions of the original inhibitors with the generated products in each class shows that the proportion of compounds in the 'active' range of pIC_{50} concentrations (above the suggested activity threshold) is greater for the generated compounds than for the original literature set. The one exception is the Gpb class, where the proportion of active compounds is particularly low for both the original data set and the generated molecules. With so few active compounds in the literature set, the range of structural features associated with activity by the SVM model will be restricted in diversity. Since the other starting materials in the set lack these structural features, the likelihood of producing active molecules is greatly reduced. In this case, the *de novo* tool generates no molecules above the threshold.

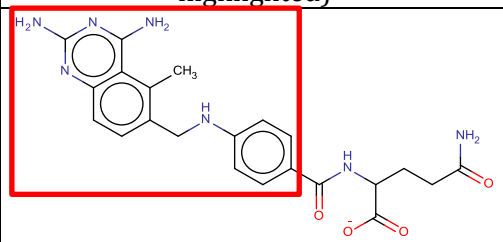
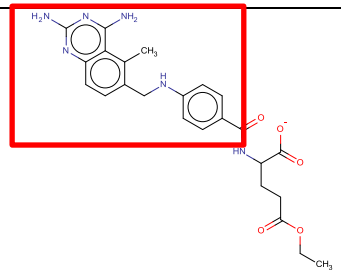
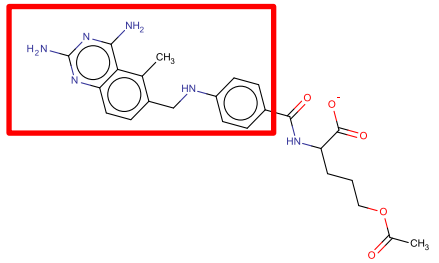
Looking at the histogram plots for each class in more detail shows that the newly generated compounds have pIC_{50} values that cluster around the threshold values of the activity ranges. The original sets on the other hand are biased towards the most active part of the ranges, and contain compounds with higher activity values. This is to be expected, as the datasets contain genuine drug candidates. The RSV generation approach appears to be useful as a means of generating compounds of interest, covering a wide range of activities, if not necessarily the most active compounds. It should also be noted that some of these new products have molecular weight values that are in excess of those seen in drug-like compounds, and as such would be less suitable for therapeutic use. Selections of the highest predicted active molecules that have molecular weights below 500g mol^{-1} are shown in Table 8.2, with key scaffold features that match those in typical inhibitors of each class, highlighted.

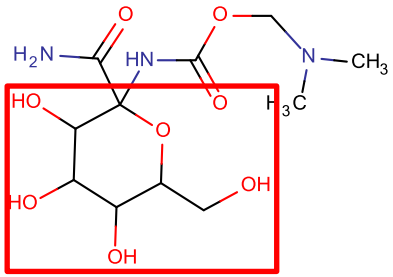
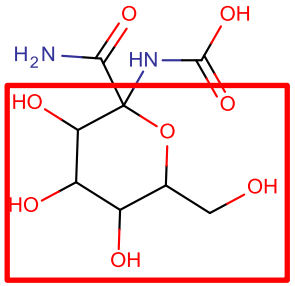
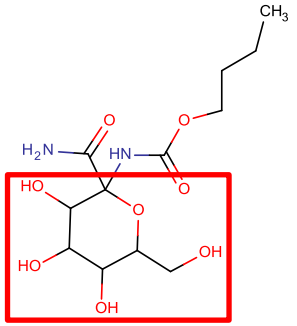
The 20 molecules with highest predicted activities in each compound class were searched for in SciFinder, to determine if any of them had previously been investigated for inhibitor activity. No reported activity values could be found for any of these compounds, but there are results of interest in each inhibitor class. Taking the example of Ace (representing inhibition of the angiotensin converting enzyme), two of the molecules were found in SciFinder, albeit without any assessment of their inhibitor activity in any recorded context.

Inhibitor class	Structure (inhibitor scaffold highlighted)	Predicted pIC_{50}
Ace		8.84
Ace		8.83
Ace		8.81

Inhibitor class	Structure (inhibitor scaffold highlighted)	Predicted pIC_{50}
Bzr		8.63
Bzr		8.62
Bzr		8.55

Inhibitor class	Structure (inhibitor scaffold highlighted)	Predicted pIC_{50}
Cox-2		8.73
Cox-2		8.62
Cox-2		8.58

Inhibitor class	Structure (inhibitor scaffold highlighted)	Predicted pIC_{50}
Dhfr		9.02
Dhfr		8.95
Dhfr		8.94

Inhibitor class	Structure (inhibitor scaffold highlighted)	Predicted pIC_{50}
Gpb		4.51
Gpb		4.34
Gpb		4.32

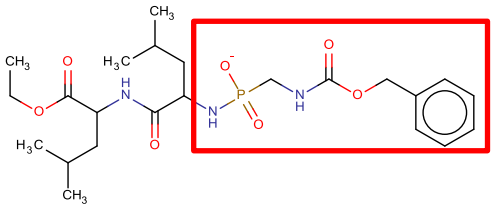
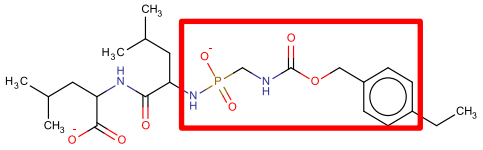
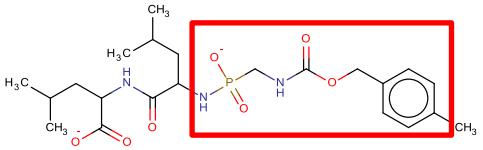
Inhibitor class	Structure (inhibitor scaffold highlighted)	Predicted pIC_{50}
Therm		7.38
Therm		7.34
Therm		7.32

Table 8.2: Selection of the most active compounds from each inhibitor class. (Sutherland et al., 2004, Wallace, 2015)

All of the 20 most active examples in the Ace set were analogues of L-Proline and L-Leucine, both of which are known as Ace inhibitors (Cushman and Ondetti, 1991) (Vrielink et al., 1996). These amino acid derivatives feature as part of the starting material scaffold. The inhibition of angiotensin converting enzyme (Ace) is key to the treatment of blood pressure and hypertension. This works by preventing the production of Angiotensin II, a peptide that causes blood vessels to constrict. The L-Proline and L-Leucine analogues both have an affinity for the zinc containing active site in the angiotensin I enzyme, selectively inhibiting it while leaving angiotensin II unaffected and thus reducing side effects.

In a similar vein, the highest predicted active compounds from the Thermolysin (Therm) class are complex derivatives of L-Leucine, as this enzyme binds hydrophobic amino acids for degradation (Khan et al., 2009). None of the compounds searched for in SciFinder returned records, but a structure similarity search confirmed their relation to Leucine, highlighting the similarity in the scaffolds. Thermolysin is a protein secreted by many infecting bacteria, such as Staphylococcus and Legionella. It specifically catalyses peptide bond hydrolysis, and is key to the reproduction of the bacteria within the host. Inhibiting the operation of this enzyme is therefore of interest in the development of new antibiotics that are effective on drug resistant bacteria strains.

In the Bzr class, (representing drugs acting on the benzodiazepine receptor site), none of the most active compounds could be found in Scifinder. However, the most active compounds found by the structure generation method have structures that incorporate the benzodiazepine functional group in a manner that does not resemble existing drugs. Drugs containing this group are known to be of therapeutic benefit for conditions concerning the central nervous system, increasing the effect of neurotransmitters (Sieghart, 1994). However, the originally discovered benzodiazepine compounds are known to have serious side effects, including issues with patients developing chemical dependency. Consequently, much research effort is going into modifications to the scaffold or derivatives with a different inhibition mechanism.

The results for cyclooxygenase-2 (Cox-2) are interesting, in that they are all relatively simple benzenesulfonamides, many of which have structures similar to those in the Markush structure expressed in a research patent filed by Talley et al. (2002), albeit without any of the specific molecules having their own entries in SciFinder. Cyclooxygenase-2 is an enzyme that regulates inflammation and the transmission of pain (Green, 2001). By specifically targeting this enzyme, anti-inflammatory drugs can

be made that keep side effects such as the formation of ulcers to a minimum. However, other effects such as cardiovascular problems or strokes were discovered with some Cox-2 inhibitors, leading to efforts to retain the positive effects while making the compounds safer to use. Benzenesulfonamides such as those generated are of particular interest as a form of inhibitor that appears to offer a compromise between high efficacy, with a significant reduction in side effects.

Analysis of the dihydrofolate reductase (Dhfr) compound class indicates that most of the highest predicted active compounds are glutamic acid derivatives, often containing diaminopyridine functionality. In this case one of these molecules was found in SciFinder, namely 1-[4-[[(2,4-diaminopyrido[3,2-d]pyrimidin-6-yl)methyl]amino] phenyl]-ethanone. Dihydrofolate reductase is a key enzyme used in the synthesis of purine and thymidine. Selective inhibition of this enzyme is effective as part of a treatment program for various cancers, by reducing the rate of tumour growth to permit other therapies to take effect. Literature searches show that glutamic acids have been investigated as anti-tumour agents acting via preventing production of purine, while also inhibiting the action of the thymidylate synthase enzymes (Gangjee et al., 2003). By inhibiting both components of the Dhfr mechanism, the overall therapeutic effect is greatly enhanced.

The generated inhibitors of glycogen phosphorylase B (the Gpb class) have similar structural features to those in the Dhfr case, in that they both use acid derivatives with high binding affinity to the active site. While all of the compounds searched for in SciFinder fit this general structure, none of them had specific entries in the database. In this case, the active compounds are mainly derivatives of carbamic acids, containing simple sugar structures that resemble glucose. Glycogen phosphorylase B is an enzyme that releases glucose from the stores of glycogen in the liver. By preventing this conversion process, selective inhibition of this enzyme helps to regulate blood sugar levels in Type-2 diabetes patients, as liver glucose production is known to increase for sufferers of this condition (Baker et al., 2005). As the inhibitor molecules resemble the conversion products for the enzyme, they effectively block the receptor, preventing further glycogen adsorption and conversion.

It should be noted that in all cases the activity of the starting materials did not directly correlate with the activity of the final products generated. In general, a wide range of pIC_{50} values were derived from each start point, with some regression to lower values

from the most active starting materials. This regression is due to the increase in molecular weight and distortion of the existing structural motifs caused by taking existing inhibitors and forcing further modification. Using late stage intermediates instead of actual inhibitors as the start point in these experiments would permit molecular growth to occur naturally without such distortion. However, since the necessary structural and modelling data is no longer available to permit such a comparison to be made, it has not been possible to perform the confirmatory experiment.

In general, while not all of the generated compounds may be good drug candidates, (due to high molecular weight or poor solubility), they contain structural motifs that imply that they would be good starting points for drug optimisation and testing, if they were synthesised and tested for real. The fact that such a high proportion of the generated molecules in each class contain these scaffolds is due to the starting materials used as inputs. As the starting materials in these cases are examples of inhibitors in their own right that contain these scaffold, it is logical that the results with the highest predicted activities will retain these features, and therefore be structurally very similar. The SciFinder searches showed that very few of the generated compounds had been reported in literature, with no therapeutic information or activity data available. However, the fact that these molecules have not been reported suggests that the analogues produced are more interesting, with the RSV method capable of producing diverse analogues. In addition, the most drug-like compounds from each class tend to be derived from the lightest starting materials. Given that a key criterion of drug-likeness is a low molecular weight (below 500 g mol^{-1}), and that the general trend in the application of RSVs to molecules is to increase the molecular weight, it stands to reason that starting from a lighter molecule is more likely to result in molecules remaining below the threshold.

8.2.1 Assessment of synthetic accessibility

To determine the likelihood that the new products are synthetically feasible, they were processed using the retrosynthetic accessibility (RSynth) tool in the MOE suite (Chemical Computing Group Inc., 2015). This gives an estimate of the ability to synthesise a given molecule, based on the fraction of heavy atoms within it that can be resolved back to simple starting materials via retrosynthesis. The feasibility score is the fraction of these atoms that can be resolved, so a value of '1' represents something readily synthesised (where all non-hydrogen atoms could be traced back to readily

available start points). Conversely, a value of '0' represents something considered synthetically impossible, as this would suggest no non-hydrogen atoms could be traced back. The number of molecules generated for each inhibitor class with a score of 0.9 or above is shown in Table 8.3.

Inhibitor class	Number of molecules with RSynth score ≥0.9	Percentage of generated molecules with RSynth score ≥0.9	Number of molecules with RSynth score ≥ 0.9 and predicted pIC_{50} ≥ threshold
Ace	5688	32.0%	1927
Bzr	7244	52.1%	3736
Cox2	4588	53.4%	754
Dhfr	6818	55.5%	2902
Gpb	1425	46.4%	31
Therm	5487	39.7%	943

Table 8.3: Summary of RSynth scores for the generated inhibitors. The Gpb set contains no generated molecules with pIC_{50} scores above 4.5, so 4.0 was used as the threshold score instead.

While for some of the inhibitor classes the majority of the generated molecules produced have synthetic accessibility values below the threshold, the proportion in each inhibitor class that are readily accessible is more than sufficient for study in most cases. Taking just the Ace inhibitors as an example, a PCA plot was produced from the generated molecules identified as having RSynth scores above 0.9, a Tanimoto coefficient greater than 0.8 relative to at least one of the molecules in the original inhibitor class, and a predicted pIC_{50} of 7.0 or above, as shown in Figure 8.7. The descriptors used were identical to those used in Section 7.3, and summarised in Appendix B.

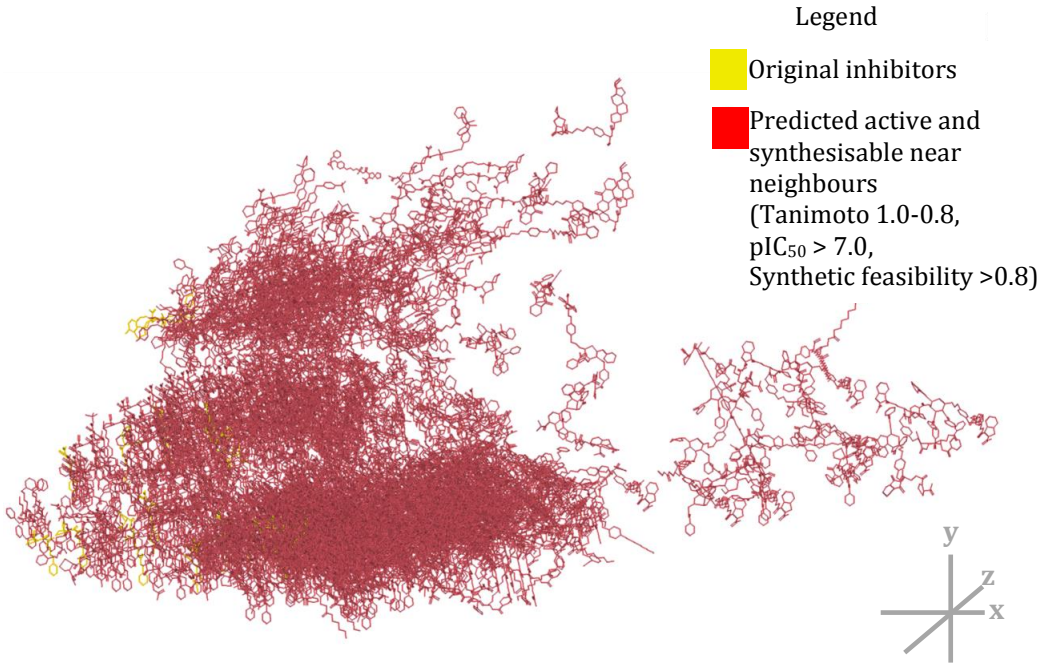


Figure 8.7: PCA plot of products generated by the Ace inhibitor class. (Sutherland et al., 2004, Wallace, 2015)

It appears that the generated molecules in this case expand, rather than fill in gaps in the PCA property space around the existing inhibitors. Two different distinct regions of near neighbours can be observed, the denser left region showing those molecules similar in property space to the originals, and the right region indicating molecules that are more distant in property space. A close examination of the denser region shows it largely consists of molecules with identical scaffolds to the known inhibitors but with different substituents. It should also be noted that none of these molecules have higher activity values than the known starting materials. While the activity range for the generated molecules is comparable to the original inhibitors, the novel reactions in the RSV knowledge base result in interesting expansions of the core structure. Some examples of these are shown in Figure 8.8, with the original molecule scaffold highlighted. As one of the aims of this project is to provide new, useful molecules that are synthetically accessible, these results suggest the tool is of worth for molecular suggestion. The less dense section of the PCA plot represents more diverse compounds and should not be discounted, since these are likely to be widely different to existing research efforts, and inspire new synthetic routes.

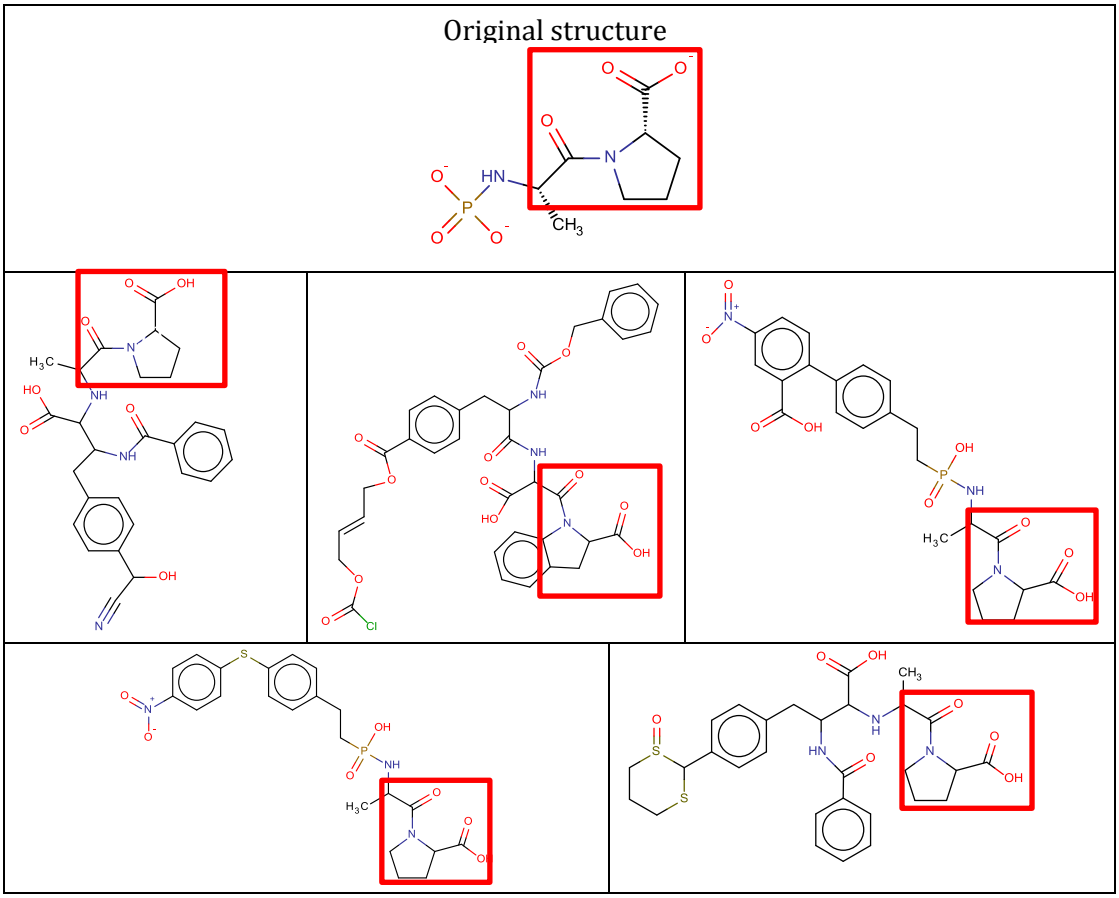


Figure 8.8: Examples of expansions to Ace inhibitor scaffolds.(Wallace, 2015)

As an additional measure of synthetic accessibility, examples of the *de novo* products for each inhibitor class covering the full range of calculated RSynth scores were presented to two medicinal chemists for analysis. The aim was to determine if the RSynth scores are an accurate representation of the ease of synthesis, as well as highlighting any problematic molecules that are generated by the tool. It was determined that using an RSynth threshold value of 0.6 is sufficient to limit the results to those that can be readily synthesised, with some molecules with RSynth scores as low as 0.2 also being considered feasible. This apparent disparity between the automated analysis and manual inspection can be explained via reference to the RSynth algorithm. Given that algorithm relies on comparison between the fragments produced from non-hydrogen atoms and the stored list of starting materials, should any of the fragmentation points be part of charged groups or complex ring structures, it is unlikely that a match will be found. These will be recorded as a retrosynthesis failure, leading to an RSynth score penalty. To avoid these false negatives compromising the quality of the results, a threshold of 0.8 may be more appropriate for assessing new compounds.

The expert analysis of the results also identified a few examples within the data set of structural classes that would be impossible to make in real world conditions, due to clashes in geometry or limitations of the RSV application method. Some of these structures are illustrated in Figure 8.9. In both of these cases, the RSynth retrosynthetic accessibility score is recorded at 0.5.

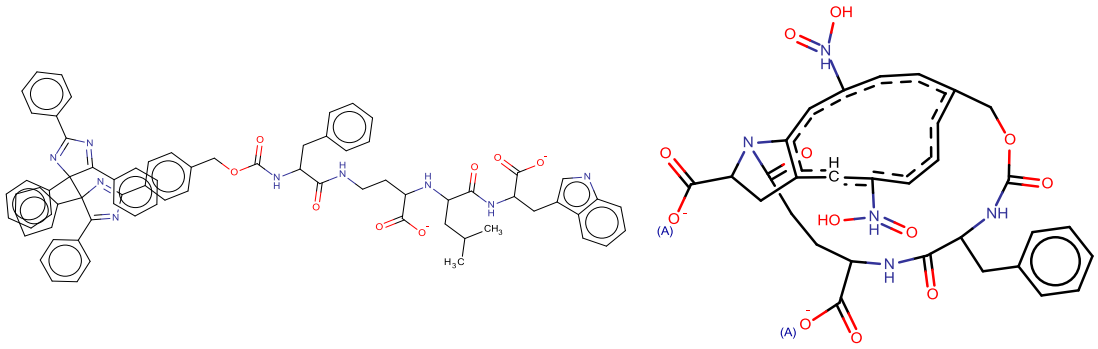


Figure 8.9: Examples of molecules produced that are considered impossible to synthesise. (Wallace, 2015)

These molecules highlight a number of limitations with the reaction vector format as currently implemented. One of the key problems is that, while the vector carries information regarding the nature of the bonds, including membership of ring systems, the size of the rings to which the atoms are members is not recorded in the atom pair lists. This can lead to transformations recorded for one ring size being applied to structures featuring a different ring size. This would in reality lead to structural incompatibility, due to the differences in electron density, but this is ignored by the structure generation tool. An example reaction for which this occurs, along with an example where the transformation is applied to a different sized ring is shown in Figure 8.10. The top example is a reaction in the JMCRoughley dataset from which a reaction vector is derived. The reaction vector encodes the addition of 1-methyl imidazole to an imidazole ring with one of the aromatic nitrogen atoms in the 5-membered ring identified as the atom which is acted upon. As the reaction vector does not encode ring size it can also be applied to pyrimidine as shown in the bottom of the figure, even though this is a quite different reaction to that from which the reaction vector was derived.

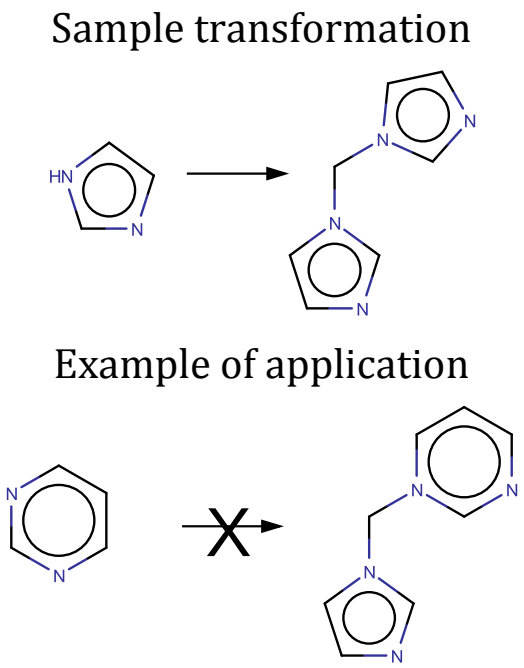


Figure 8.10: An example of a transformation that can be incorrectly applied in the structure generator. (Wallace, 2015)

An additional issue reported during the analysis of the products relates to the encoding of the environment of a reaction. In RVs and RSVs, the encoding is limited to bonds that are adjacent to those that are added or removed in the reaction. There is currently no support in the vector format for specifying the effects of electron withdrawing or donating groups which are more than one bond away from the reaction centre, or those that remain unchanged over the course of the reaction groups, thus preventing direct implementation of mesomeric effects or electron induction. The ARChem synthesis prediction tool (Johnson et al., 2008) attempts to resolve these problems by incorporating these additional environmental factors within its transformation rules database. However, these factors make the rules considerably more complex to generate than the RVs and RSVs, requiring considerable manual input.

8.3 Comparison of RSVs and RVs for *de novo* design

The previous RV method was designed to perform multi-objective *de novo* design (Gillet et al., 2014). This led to a number of issues, such as increased execution time and complexity relative to single reaction analyses, as well as the fact that the intermediate steps in such optimisation pathways may not score well in property evaluation methods. As a result, selection of the best pathways may be difficult to achieve, and

potentially useful results may be discarded. As an RSV represents a pathway in its entirety, this problem can be effectively worked around utilising the new structure generation method. To compare the two approaches directly, a series of experiments was carried out using the inhibitor data sets from the Sutherland molecule collection, as described in Section 8.1.

8.3.1 Single starting materials

8.3.1.1 Angiotensin converting enzyme (Ace) inhibitors

A simple multi-objective model was created using KNIME, consisting of three separate scores each of which is calculated on a continuous scale. As the SVM model used for the previous work was only available when working at the Lilly site, an alternative method was required to perform these experiments. The approach taken was to build a Bayesian model, using the existing Ace inhibitor set. Each molecule in the set was assigned as active or inactive based on its activity value, using the pIC_{50} value of 7.0 as a threshold; molecules at or above this threshold were classed as active and those below classed as inactive.

Each molecule was then represented using structural fingerprints generated by the RDKit software (Landrum), with the model generated using the Bayesian implementation included within KNIME. In this method, each substructure fragment is assigned a weight based on the frequency with which it occurs in an active molecule relative to its frequency in inactives. For this implementation the fragment weighting method 'R2' was used with the individual weights for a fragment calculated by:

$$P_{final}(A|j) = A_j + 1/\left(T_j \frac{N_A}{N_T} + 1\right)$$

Equation 8.1: Calculation of fragment weight for the Bayesian activity model.

where $T_j$represents the total number of compounds in the training set containing the fragment j, $A_j$represents the number of these that are active, N_A represents the total number of actives and N_T represents the total number of compounds in the training set. When the model is applied to a test molecule (in this case generated by the *de novo* design tool) the sum of the log values of the weights for the fragments in the molecule gives a numerical value (referred to here as a Bayesian score) that represents the likelihood of the molecule being active, with higher values indicating a greater likelihood (Hert et al., 2006).

In addition to scoring using the Bayesian model, generated molecules were also scored on molecular weight and logP value. In the latter two cases, scores were assigned based on whether the molecule fits into the range of properties associated with drug-like molecules (Lipinski et al., 1997), namely molecular weight between 0.0 and 500 g mol^{-1}, and predicted logP value between 0.0 and 5.0, as calculated using the RDKit library. For these properties, the value is converted into a numerical score: if the properties are within the range, the score is set to the maximum to indicate that the criterion has been satisfied. However, if the property is outside of the given range, the score is reduced *pro rata* based on the difference between the property value and the range boundaries. This is scaled such that the larger the difference between the property and the range boundary, the lower the score. The reduction is performed in such a manner as to treat being under or over the range equally.

In order to compare the RV and RSV approaches directly, attempts at a full enumeration of solution space were carried out using each method, with the inhibitor from the Ace set with the lowest molecular weight used as the starting material (as shown in Figure 8.11). This has a pIC_{50} score of 2.96, suggesting that it is not particularly active itself, as it lacks the proline residue common to many inhibitors. The Bayesian model gives this molecule a Bayesian score of 0.92, which can be used to measure the relative performance of the RV and RSV approaches in terms of their ability to optimise the molecule. For the RSV case, the structure generation experiment was rerun in the same manner as in Section 8.2, using the JMCRoughley RSV database, representing sequences from one to eleven steps in length. The activity and molecular property values were used to give Pareto rankings to the entire population.

In the RV case, a simple iterative loop was set up using the RVs from the JMCRoughley database. The unique products generated from the first experiment were used as the starting materials for the next iteration and so on using the same RVs, with eleven iterations attempted in order to match the range of sequence lengths in the RSV experiment.

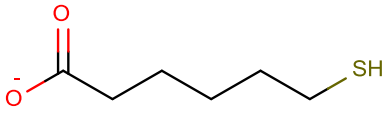


Figure 8.11: Starting material for the RV comparison experiment. (Sutherland et al., 2004)

However, the RV experiment had to be abandoned after three iterations, as the full enumeration of the products became too large to process. For the first generation, 890 products were generated, which led to 357,972 for the second generation, and 816,925 for the third generation. As the number of molecules increases, the execution time for each iteration increases dramatically, with the second iteration taking over 24 hours to complete on the i7 Linux workstation (running Red Hat 6.4, with eight cores running at 3.40 GHz and 16 Gb of available memory) running KNIME, and the third iteration taking 36 hours. From this, it is clear that a complete enumeration of the solution space for an inhibitor set using RVs would be impossible.

In order to produce a meaningful set of results, a revised RV experiment used an optimisation and sampling method in order to limit the number of products generated. This is based on a multi-objective optimisation approach, as previously discussed in Section 3.2.5.3. An initial population of results is generated from the given starting material as before, with each result molecule scored, with the best results used as starting materials to produce a further generation. In this case, each iteration utilises a sampling method based on tournament selection, rather than enumerating all of the possibilities. In the tournament selection process, a number of molecules from the starting population are selected at random from the pool, according to the specified tournament size. From this subset, the highest scoring molecule according to the fitness criteria (a particular property for example, or the Pareto ranking) is selected to use as the input molecule for RV application. A randomly selected RV is then chosen from those applicable to this molecule and used to generate a new product. This process of tournament selection and RV application was repeated until a new population of molecules was produced that is equal to the required population size (in this case, as close to 200 molecules as is possible). A KNIME workflow for this process is shown in Figure 8.12. Careful selection of the size of the tournaments is required to ensure good result quality. If the tournament size is too large, it is likely that the selection will become elitist, as the best scoring molecules are more likely to be selected multiple times. Similarly, a sampling frame that is too small will not adequately explore the solution space, leading to wide variations in the data between runs and a lack of reproducibility.

In this instance, the first step for the tournament-based optimisation process consisted of a full enumeration of all products using the same single starting material as input and all applicable RVs from the 26,235 reactions (including the reagent data) in the

JMCRoughley set. This resulted in 890 molecules, as before. These molecules were then scored, firstly only using the Bayesian scores, with the 200 best molecules selected as starting materials for the tournament selection processes. A series of experiments was carried out to determine the optimal tournament size, with this varied from five molecules up to 175 molecules with three iterations carried out. Three complete runs were carried out for each tournament size and results are reported for the run which produced the molecule with highest Bayesian score. Between individual runs, the highest Bayesian scores were largely consistent, with relatively low degrees of deviation. The data for all of the repeated runs is shown in Appendix C.1.1, Table C-1.

As the RV approach in this case is based on a single objective optimisation rather than Pareto ranking, it is possible to make comparisons between the RV and RSV approaches by sorting the generated molecules on activity. The results are summarised in Table 8.4. For each experiment, the number of unique molecules produced is listed, as well as the ranges of the molecular properties observed. For the RSV case, the results are presented based on the lengths of sequences used, between one and eleven steps. For example, RSV 1-3 represents the results from RSVs representing sequences between one and three steps in length, and is therefore equivalent to the number of iterations used in the tournament selection method.

While none of the sampling methods (either in the RV or RSV cases) match the full enumeration of the RV database in terms of the highest Bayesian score, the tournament selection method is an effective approach to find high Bayesian scoring molecules with a short execution time (taking an average of 15 minutes to complete each run, on the i7 workstation). Both of the methods using RVs (full enumeration and tournament sampling) perform better at finding high Bayesian scoring molecules. However, a significant proportion of these results have higher molecular weight and logP values than would be considered drug-like, leaving these unsuitable for therapeutic use.

Sampling experiment	Number of products	Highest Bayesian score	Bayesian score range	MW range / g mol^{-1}	LogP range
Full RV enumeration (3 generations)	816,925	23.95	-1.78 to 23.95	242 to 2,026	-1.81 to 24.19
Tournament size 5	200	10.87	-8.32 to 10.87	236 to 1,166	0.78 to 14.58
Tournament size 75	200	10.51	-9.63 to 10.51	254 to 1,118	2.17 to 14.36
Tournament size 150	200	8.43	-8.71 to 8.43	260 to 810	1.63 to 9.69
Tournament size 10	200	7.71	-8.05 to 7.71	239 to 1,062	0.98 to 14.51
Tournament size 20	200	7.66	-8.29 to 7.66	240 to 958	0.67 to 12.18
Tournament size 30	200	6.79	-7.92 to 6.79	240 to 1,233	1.50 to 18.46
Tournament size 100	200	6.64	-8.71 to 6.64	260 to 1,117	1.63 to 14.36
Tournament size 15	200	6.51	-8.50 to 6.51	268 to 1,289	0.81 to 15.90
Tournament size 50	200	6.15	-9.14 to 6.15	284 to 1,327	0.67 to 16.17
RSV 1-2	178	5.88	-13.42 to 5.88	172 to 1,235	-0.66 to 15.48
RSV 1-3	239	5.88	-13.42 to 5.88	120 to 1,235	-0.66 to 15.48
RSV 1-4	243	5.88	-13.42 to 5.88	120 to 1,235	-0.66 to 15.48
RSV 1-5	244	5.88	-13.42 to 5.88	120 to 1,235	-0.66 to 15.48
RSV 1-6	246	5.88	-13.42 to 5.88	120 to 1,235	-0.66 to 15.48

Table 8.4: Summary of Bayesian scores for RV enumeration, tournament selection and RSV approach. The results from RSV 1-6 to RSV 1-11 are identical, and so have been omitted.

It should be noted that, for each experiment there is a large improvement over the starting material in terms of Bayesian score, as this had a value of 0.92. However, in the sampling experiments there are a number of results with much lower Bayesian scores than the starting material. This is particularly the case with the RSV experiments, where there are significant deviations from the molecule scaffold. The fact that so few compounds are produced via the RSV enumeration as opposed to the RV approach shows the limitations of the solution space available, when transformations are restricted to sequences. However, the experiments still produce good results in terms of Bayesian scores relative to the starting material.

In order to determine how many drug-like molecules are found by each method, the results were filtered by removing any molecule with a molecular weight over 500g mol^{-1} or a logP value greater than 5. This alters the relative ranking of experiments slightly, as illustrated in Table 8.5, but the general trends in Bayesian

scores are still observed in that the Bayesian scores are higher for the full RV enumeration, with tournament selection outperforming the RSV method.

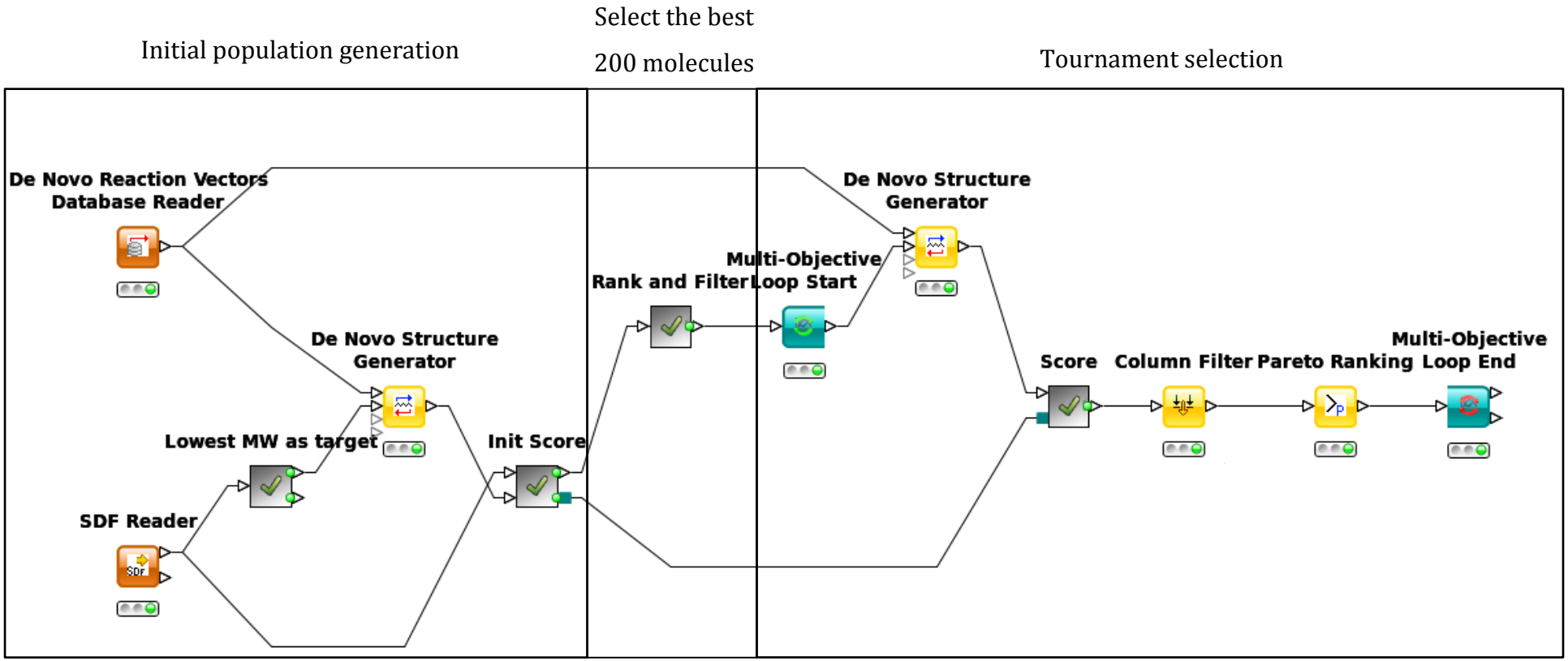


Figure 8.12: KNIME workflow for the tournament selection process.

Sampling experiment	Number of products	Highest Bayesian score	Bayesian score range	MW range / g mol^{-1}	LogP range
Full RV enumeration (3 generations)	1,901	17.63	2.95 to 17.63	242 to 499	-0.21 to 5.00
Tournament size 5	52	6.03	-7.93 to 6.03	273 to 498	0.96 to 4.99
Tournament size 75	33	5.72	-9.23 to 5.72	254 to 490	2.17 to 4.87
Tournament size 150	22	5.72	-8.71 to 5.72	260 to 474	0.94 to 4.85
Tournament size 100	90	5.39	-8.71 to 5.39	260 to 489	1.63 to 4.72
Tournament size 50	50	5.45	-5.88 to 5.45	284 to 497	0.67 to 4.96
Tournament size 10	60	5.16	-6.70 to 5.16	280 to 499	1.59 to 4.99
Tournament size 15	66	4.86	-8.50 to 4.86	268 to 497	0.81 to 4.98
Tournament size 20	69	4.41	-8.29 to 4.41	240 to 496	0.67 to 4.97
Tournament size 30	69	4.41	-7.93 to 4.41	240 to 491	1.51 to 4.94
Tournament size 175	97	3.75	-8.71 to 3.75	259 to 469	0.84 to 4.59
RSV 1-2	131	2.07	-13.42 to 2.07	172 to 495	-0.66 to 4.96
RSV 1-3	189	2.07	-13.42 to 2.07	120 to 495	-0.66 to 4.96
RSV 1-4	194	2.07	-13.42 to 2.07	120 to 495	-0.66 to 4.96
RSV 1-5	195	2.07	-13.42 to 2.07	120 to 495	-0.66 to 4.96
RSV 1-6	200	2.07	-13.42 to 2.07	120 to 495	-0.66 to 4.96

Table 8.5: Summary of Bayesian scores for RV enumeration, tournament selection and RSV approach, filtered for drug-likeness. The results from RSV 1-6 to RSV 1-11 are identical, so only RSV 1-6 is shown.

Comparing these values with Table 8.4 shows that the vast majority of molecules produced in the full enumeration are non-drug-like, indicating that the vast proportion of the time spent on exploring the solution space leads to the generation of unusable products. Drug-like compounds with Bayesian scores for each method (RV enumeration, tournament selection and RSV methods) are shown in Tables 8.6, 8.7 and 8.8 respectively.

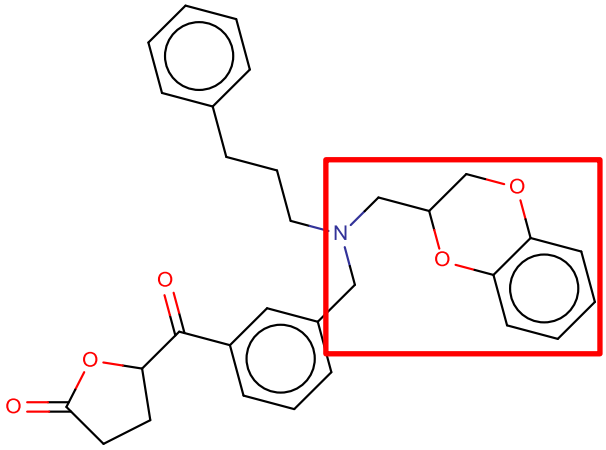
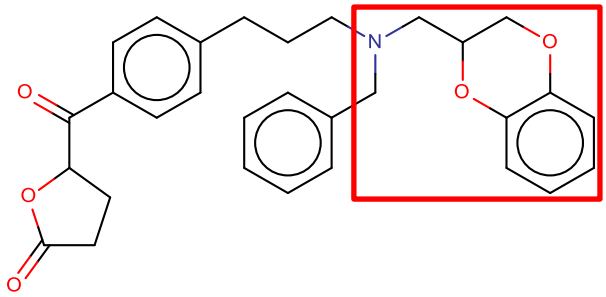
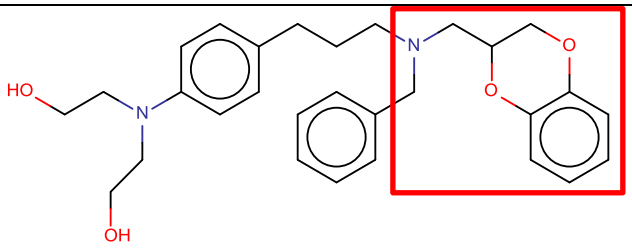
RV enumeration			
Structure (common inhibitor scaffold highlighted)	LogP	Molecular weight / g mol^{-1}	Bayesian score
	4.85	485	17.63
	4.85	485	16.60
	3.75	476	15.38

Table 8.6: Drug-like compounds with highest Bayesian scores from the RV enumeration method for the initial Ace experiment (sorted by predicted activity). (Wallace, 2015)

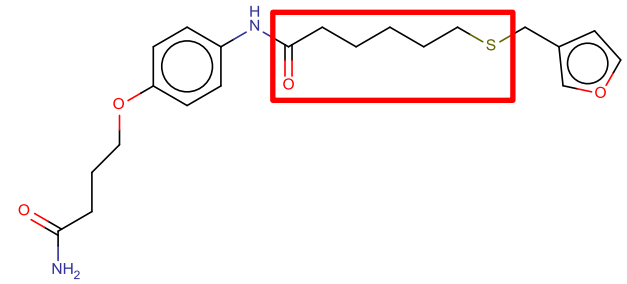
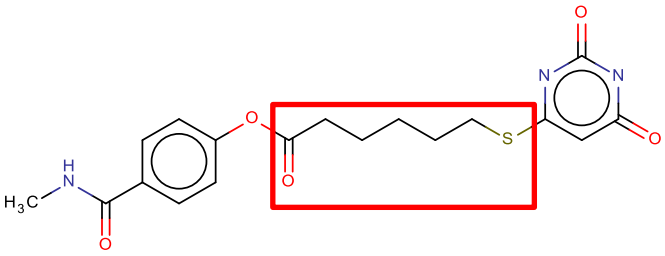
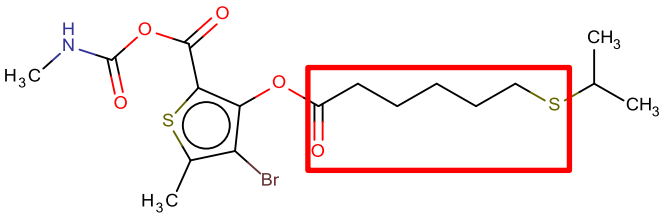
Tournament selection			
Structure (common inhibitor scaffold highlighted)	LogP	Molecular weight / g mol^{-1}	Bayesian score
	4.36	404	5.58
	1.25	389	4.26
	4.92	465	3.84

Table 8.7: Drug-like compounds with highest Bayesian score produced from tournament selection for the initial Ace experiment (sorted by activity). (Wallace, 2015)

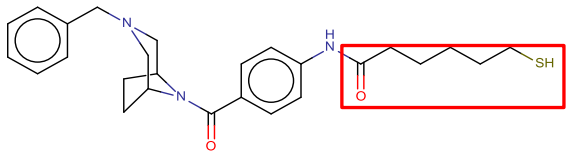
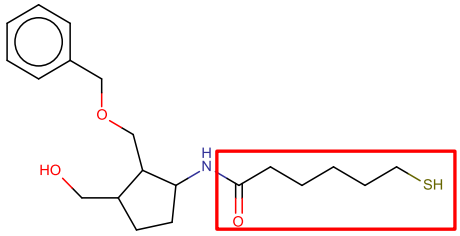
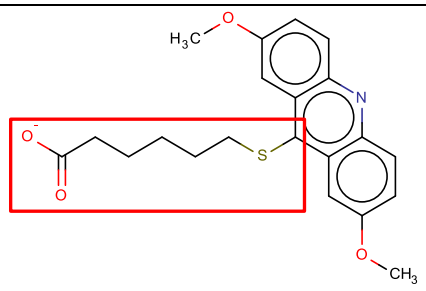
RSV enumeration			
Structure (common inhibitor scaffold highlighted)	LogP	Molecular weight / g mol^{-1}	Bayesian score
	4.60	451	2.07
	3.20	365	1.82
	3.81	384	0.71

Table 8.8: Drug-like compounds with highest Bayesian score produced from RSV enumeration for the initial Ace experiment (sorted by activity). (Wallace, 2015)

After the initial study, scoring based on the molecular weight and logP values for the molecules were added, turning the method into a multi-objective optimisation. As discussed previously, the scores for these are based on how well the molecules fit into the range of drug-likeness parameters, with the three scores used to build a Pareto rank used to evaluate the results. In addition, Pareto ranking is used when choosing between candidates in the tournament selection method. In theory, by adding these parameters to the scoring, the results of the sampling experiments will favour the formation of compounds that are more drug-like, improving result quality. The complete results for this experiment are presented in Appendix C.1.2, Table C-2, and summarised in Table 8.9. As the scoring of the results of the RV enumeration occurs independently of the sampling, the overall results are unaffected, and so are not duplicated. However, it should be noted that the tournament selection results are different to the previous experiment, due to the difference in scoring function and the random nature of the sampling process. The best drug-like molecules (according to the Pareto ranking) for the RV enumeration and tournament selection methods are

illustrated in Table 8.10 and 8.11. Again, the best results for the RSV enumeration are identical to those for the activity only case shown in Table 8.8 and so they are not duplicated here. As only drug-like results are shown, sorting by Pareto rank in this instance is effectively the same as sorting by Bayesian score, as the other two parameters are guaranteed to be in range.

As expected, the use of the Lipinski parameters for scoring has resulted in a larger proportion of the results being drug-like, in comparison to scoring solely by Bayesian score. However, the molecules with the highest Bayesian scores for the tournament selection method are identical to those shown in Table 8.7, with the total number of drug-like compounds remaining relatively low. Looking at the Pareto rankings, 64 molecules from the full RV enumeration are part of the non-dominated front, compared with 5 in the RSV case. When filtered for drug-likeness each of the molecules with the highest Bayesian score was assigned to a unique Pareto front. It should be noted that only results for tournament sizes between five and fifteen are shown here, as larger tournament sizes did not produce any noticeable increases in drug-like molecules with high Bayesian score. In addition, while a tournament size of five produces the molecule with the highest Bayesian score, a tournament size of fifteen actually produces more molecules designated as drug-like.

Sampling experiment	Number of products	Highest Bayesian score	Bayesian score range	MW range / g mol^{-1}	LogP range
Full RV enumeration (3 generations)	816,925	23.95	-1.78 to 23.95	242 to 2,026	-1.81 to 24.19
Tournament size 5	200	10.35	-6.54 to 10.35	204 to 1,211	1.15 to 16.74
Tournament size 5 (drug-like compounds)	49	7.59	-1.69 to 7.59	281 to 492	1.28 to 4.92
Tournament size 10	200	7.41	-10.33 to 7.41	240 to 1,019	1.55 to 12.50
Tournament size 10 (drug-like compounds)	57	5.86	-7.12 to 5.86	307 to 499	0.67 to 4.82
Tournament size 15	200	7.36	-3.46 to 7.36	217 to 1,390	-0.03 to 12.38
Tournament size 15 (drug-like compounds)	65	4.82	-3.46 to 4.82	217 to 499	0.23 to 4.97

Table 8.9: Summary of the best performing runs of the structure generation using the lightest Ace inhibitor as the starting material, using Pareto ranking.

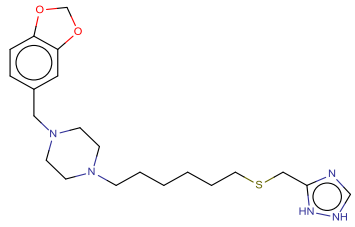
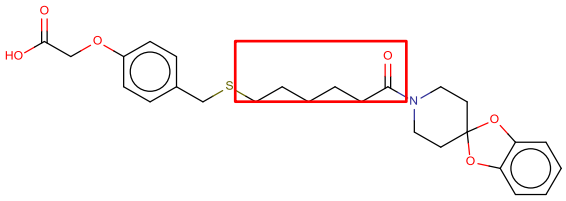
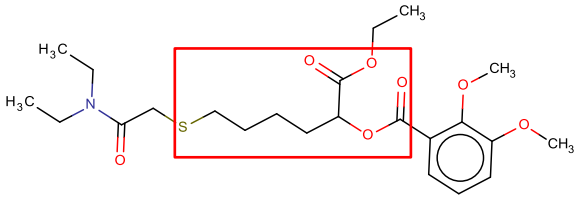
RV enumeration				
Structure (common inhibitor scaffold highlighted)	LogP	Molecular weight / g mol^{-1}	Bayesian score	Pareto Ranking
	2.93	416	4.91	1
	4.73	485	4.68	2
	3.56	469	4.26	3

Table 8.10: Drug-like compounds with highest Bayesian score from the RV enumeration method for the revised Ace experiment (sorted by Pareto ranking). (Wallace, 2015)

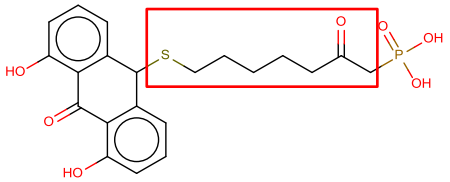
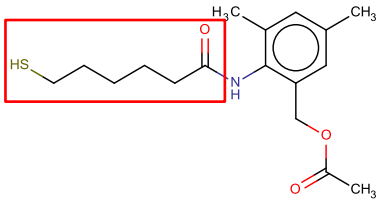
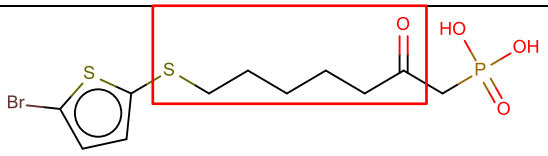
Tournament selection				
Structure (common inhibitor scaffold highlighted)	LogP	Molecular weight / g mol^{-1}	Bayesian score	Pareto Ranking
	3.77	450	7.59	1
	3.80	323	4.35	2
	3.91	386	4.24	3

Table 8.11: Drug-like compounds with highest Bayesian score from the tournament selection method for the revised Ace experiment (sorted by Pareto ranking). (Wallace, 2015)

Searching for the highest Bayesian scoring drug-like compounds in each case in SciFinder did not return any data on the specific molecules, but many of the results contain scaffolds common to Ace inhibitors, as seen with the initial starting material and discussed in Section 8.1. As in the previous study, these results have structures similar to amino acids such as L-Leucine, which are known to be effective Angiotensin I inhibitors, although they do not contain any direct amino acid functionality, such as proline scaffolds.

As the results from each tournament size represent the best performing run out of three repetitions, an understanding of the degree of reproducibility can be determined through statistical comparisons of these repetitions. The results for this analysis are shown in Table 8.12.

Tournament size	Mean (Highest Bayesian score)	Standard Deviation (Highest Bayesian score)	Variance (Highest Bayesian score)
5	8.81	1.41	1.98
10	6.84	0.85	0.72
15	6.48	1.01	1.02

Table 8.12: Summary of statistical analysis of the tournament selection method for the revised Ace experiment.

Considering the mean values for the Bayesian score alone, it can be seen that the same relative relationship between tournament size and Bayesian score is present as with the consideration of the highest Bayesian score alone, with the smaller sizes producing a higher overall value. As with the highest Bayesian score data, there is little to distinguish the results for the tournament sizes of ten and fifteen in terms of the mean Bayesian score. However, looking at the variance and standard deviation (indicating the degree of spread around the mean value), it appears that the tournament size of ten shows lower variance in the data over the runs, with the tournament size of fifteen showing a smaller spread than the tournament size of five. While larger tournament sizes are more likely to select the same high performing molecules repeatedly, it is unlikely that this elitism leads to the smaller spread of results in this case. It is more likely that the mean Bayesian scores for the tournament size of five is due to high performing outliers, skewing the distribution towards higher scores. In the other two

cases, these outliers do not exist, and the two activity distributions are closer in nature. To see if these trends hold for other examples, further studies would be needed.

Alternative starting materials

In order to determine if optimal tournament sizes identified above are suitable for use in other examples, the structure generation experiment was repeated with a second molecule from the Ace set as starting material. The second lowest molecular weight compound (illustrated in Figure 8.13) was chosen which has a pIC_{50} value of 5.62, making it more active than the previous starting material, but still below the activity threshold itself. The Bayesian score in this case is 2.00, which is higher than the previous starting material. The same Pareto ranking scoring system was used to evaluate the results, which are summarised in Table 8.13, with the molecules with highest Bayesian score illustrated in Table 8.14, Table 8.15 and Table 8.16.

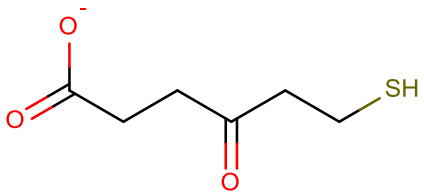


Figure 8.13: Second lightest molecule in the Ace inhibitor set. (Sutherland et al., 2004, Wallace, 2015)

In these results, there are fewer products generated than in the previous experiment. While the difference between the two starting materials is very small (an additional carbonyl group), it significantly reduces the number of unique compounds that can be created by disrupting the straight carbon chain, leading to a smaller set of unique results. (As reported previously (Table 6.8) a long alkyl chain such as that in the first Ace inhibitor can result in a large number of products.) For this particular starting material, the larger tournament size produces the most active molecule, in contrast to the first Ace inhibitor studied. As before, each experiment produces results with Bayesian scores higher than the starting material, but with some molecules in each collection representing a significant backwards step. It should be noted that the RSV sampling experiments do not produce any drug-like compounds with comparable Bayesian scores to the starting materials, and the full RV enumeration gives the best results.

Sampling experiment	Number of products	Highest Bayesian score	Bayesian score range	MW range / g mol^{-1}	LogP range
Full RV enumeration (3 generations)	361,081	14.84	-5.84 to 14.84	119 to 1,585	-3.51 to 19.36
Full RV enumeration (drug-like compounds)	190,356	10.14	-5.29 to 10.14	119 to 499	-2.51 to 5.00
Tournament size 15	200	9.44	-1.39 to 9.44	268 to 1,125	1.39 to 15.00
Tournament size 15 (drug-like compounds)	66	6.55	-0.79 to 6.55	293 to 495	0.19 to 4.85
Tournament size 10	200	8.56	-0.96 to 8.56	292 to 1,276	0.68 to 15.02
Tournament size 10 (drug-like compounds)	52	7.92	-0.72 to 7.92	314 to 490	0.06 to 4.87
Tournament size 5	200	8.03	-10.08 to 8.03	308 to 1,123	-0.95 to 11.93
Tournament size 5 (drug-like compounds)	63	4.88	-7.62 to 4.88	308 to 499	0.12 to 4.91
RSV 1-2	165	4.13	-15.53 to 4.13	175 to 1,248	-1.48 to 14.66
RSV 1-2 (drug-like compounds)	127	-0.04	-15.53 to -0.04	186 to 478	0.09 to 4.88
RSV 1-3	225	4.13	-15.53 to 4.13	134 to 1,248	-1.48 to 14.66
RSV 1-3 (drug-like compounds)	180	-0.04	-15.53 to -0.04	134 to 483	0.09 to 4.88
RSV 1-4	229	4.13	-15.53 to 4.13	134 to 1,248	-1.48 to 14.66
RSV 1-4 (drug-like compounds)	184	-0.04	-15.53 to -0.04	134 to 483	0.09 to 4.88
RSV 1-5 & RSV 1-6	230	4.13	-15.53 to 4.13	134 to 1,248	-1.48 to 14.66
RSV 1-5 & RSV 1-6 (drug-like compounds)	185	-0.04	-15.53 to -0.04	134 to 483	0.09 to 4.88
RSV 1-7	232	4.13	-15.53 to 4.13	134 to 1,248	-1.48 to 14.66
RSV 1-7 (drug-like compounds)	185	-0.04	-15.53 to -0.04	134 to 483	0.09 to 4.88

Table 8.13: Summary of the best performing runs of the structure generation for the second lightest Ace inhibitor, using Pareto ranking. The results from RSV 1-7 to RSV 1-11 are identical, duplicates are omitted.

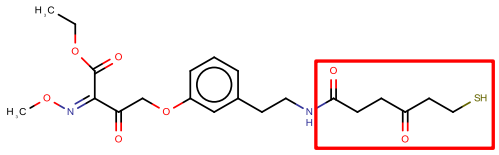
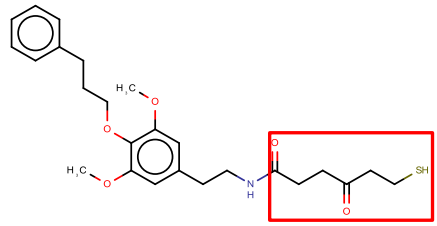
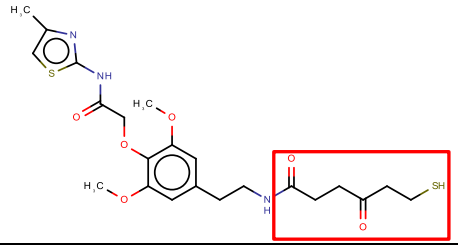
RV enumeration				
Structure (common inhibitor scaffold highlighted)	LogP	Molecular weight / g mol^{-1}	Bayesian score	Pareto Ranking
	2.35	453	10.14	1
	4.86	460	9.66	2
	3.64	496	9.50	3

Table 8.14: Drug-like compounds with highest Bayesian score from the RV enumeration method from the second lightest Ace inhibitor (sorted by Pareto ranking). (Wallace, 2015)

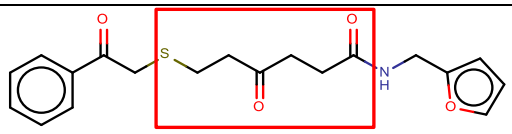
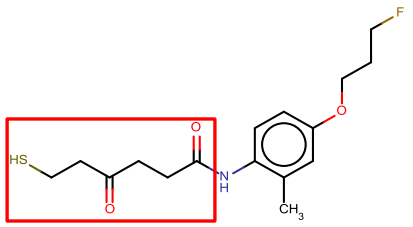
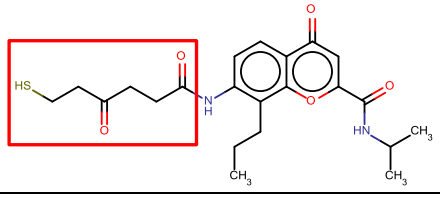
Tournament selection				
Structure (common inhibitor scaffold highlighted)	LogP	Molecular weight / g mol^{-1}	Bayesian score	Pareto Ranking
	3.25	359	6.11	1
	3.34	327	5.74	2
	3.49	432	5.29	2

Table 8.15: Drug-like compounds with highest Bayesian score from the tournament selection method from the second lightest Ace inhibitor (sorted by Pareto ranking). (Wallace, 2015)

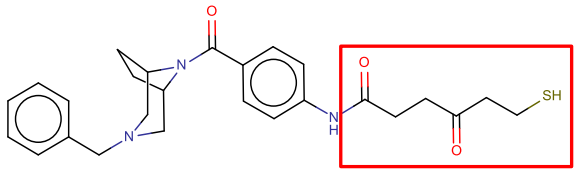
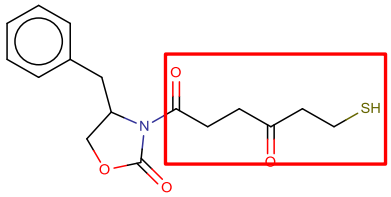
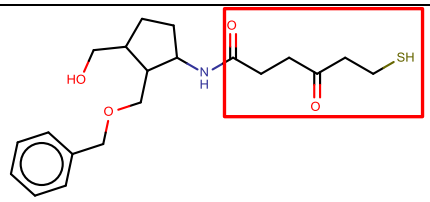
RSV enumeration				
Structure (common inhibitor scaffold highlighted)	LogP	Molecular weight / g mol^{-1}	Bayesian score	Pareto Ranking
	3.78	465	-0.04	1
	2.25	321	-0.16	2
	2.38	379	-0.29	3

Table 8.16: Drug-like compounds with highest Bayesian score from the RSV enumeration method from the second lightest Ace inhibitor (sorted by Pareto ranking). (Wallace, 2015)

Both the RV and RSV methods produce the same category of amino acid derivatives known to be Ace inhibitors. A search in SciFinder for the molecules with highest Bayesian scores does not reveal any specific literature activity data for any of these molecules. In the RV enumeration, a second compound class is seen among the results, with considerably higher Bayesian scores. These are molecules containing methanamine (methylammonium) groups, which are known to be strong angiotensin II receptor antagonists when combined with the typical Ace inhibitor scaffolds already discussed (Bessa Belmunt, 2008).

Looking at the Pareto rankings for the full results of each experiment, 18 molecules from the RV enumeration are not dominated, as opposed to three in the RSV case. In general, the additional molecules seen in the RV enumeration on this front are those with high molecular weights and logP values well outside of the desired range. In these situations, the very high Bayesian activity scores are balanced by the penalties due to the non-drug-like structural parameters.

A tournament size of fifteen appears to give the best results in terms of quality and highest Bayesian scores, which is very different to the previous example, where smaller sizes were better. As before, a statistical analysis of the individual runs shows that the larger tournament sizes show a lower degree of variance between runs, but with a higher peak and mean Bayesian score. These results are summarised in Table 8.17. Given the relative speed at which the sampling experiments can be conducted, it would be possible to run a number of experiments with these different parameters and choose the best results as appropriate, although larger sizes appear to be superior. In this case, the best for both starting materials tournament sizes of ten or fifteen would be appropriate, as they have the best compromise between predicted activity and result reproducibility.

Tournament size	Mean (Highest Bayesian score)	Standard Deviation (Highest Bayesian score)	Variance (Highest Bayesian score)
5	6.12	1.66	2.77
10	8.26	0.32	0.10
15	8.51	0.82	0.67

Table 8.17: Summary of statistical analysis of the tournament selection method for the revised Ace experiment (second starting material).

It should be noted that for each starting material, the tournament selection using RVs provide results superior to those determined by RSV enumeration in terms of the Bayesian scores. The molecules with the best Pareto rank produced via the RSV method in these cases have small, negative Bayesian scores, suggesting that they are highly unlikely to be active. While compounds with positive Bayesian scores exist in the data set, these are associated with non-drug-like properties. On the other hand, the RV-based methods produce drug-like molecules that have high Bayesian scores, making RV enumeration or sampling better approaches. Given that the main difference between the RV and the RSV method is the restriction of the latter to previously established sequences, this implies that the limited diversity in the collection of transformations in the RSV database leads to the lack of high Bayesian scoring compounds. In addition, although sequences from two to eleven steps in length were used, the molecules with highest Bayesian scores come from the sequences between two and five steps in length. It is therefore unlikely that long reaction sequences will be of benefit when generating drug-like molecules from these simple start points.

8.3.1.2 Benzodiazepine receptors (Bzr) inhibitors

The performance of the RSVs was then compared with RVs for the benzodiazepine inhibitors using an appropriate Bayesian model for likelihood of activity. The starting molecule was the lowest molecular weight molecule in the set. Summary results are presented below with additional results in Appendix C.2, Table C-3.

The Bzr inhibitor set (Figure 8.14, pIC_{50} value of 6.46) shows the same trends as the Ace inhibitors, but with lower Bayesian scores overall. The starting material in this case is below the suggested threshold for activity for the class (pIC_{50} = 7.52, Bayesian score = -4.82), and consists of a simple benzodiazepine scaffold. Summary results are given in Table 8.18, with the molecules with the highest Bayesian scores shown in Table 8.19, Table 8.20 and Table 8.21. These scores indicate that the majority of produced molecules are not likely to be active, much like the starting material itself, with the tournament selection approach providing the only molecule with a positive Bayesian score outside of the full RV enumeration. However, each experiment does give results with significant improvements over the starting material in terms of predicted activity. It should also be noted that the results of the RV enumeration bear little resemblance to the benzodiazepine starting material, suggesting that those features associated with higher scores in the model may not be the key features of the benzodiazepine scaffold.

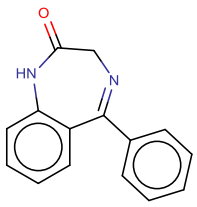


Figure 8.14: Lowest molecular weight molecule in the Bzr inhibitor set. (Sutherland et al., 2004)

In this case, the training set contains molecules that exhibit two different mechanisms for inhibition – the more selective inhibitors and the traditional benzodiazepines. The more selective compounds have higher Bayesian scores than the original benzodiazepines, and represent a considerable proportion of the active class in the training set. This leads to higher weights being associated with the fragments from these molecules (including functionalised piperidines and pyridines) in the Bayesian model, with fragments derived from the original benzodiazepines scoring considerably lower. As a consequence, generated molecules that resemble the benzodiazepines have relatively low Bayesian scores, despite the fact that benzodiazepines have reported

activity in this class. As with the Ace inhibitor data set, the smallest tournament size gives results with the highest Bayesian score, but a tournament size of ten gives the smallest variance.

Sampling experiment	Number of products	Highest Bayesian score	Bayesian score range	MW range / g mol^{-1}	LogP range
Full RV enumeration (3 generations)	2,026,928	1.67	-13.80 to 1.67	251 to 1,900	-2.23 to 21.76
Full RV enumeration (drug-like compounds)	245,035	1.31	-12.24 to 1.31	251 to 499	-1.40 to 5.00
Tournament size 5	176	0.30	-8.94 to 0.30	322 to 1,027	0.15 to 11.77
Tournament size 5 (drug-like compounds)	44	-1.98	-8.10 to -1.98	322 to 499	0.15 to 4.87
Tournament size 10	178	-1.13	-9.76 to -1.13	348 to 1,173	-0.49 to 14.47
Tournament size 10 (drug-like compounds)	27	-2.47	-8.05 to -2.47	348 to 498	0.49 to 4.94
Tournament size 15	181	-1.31	-9.63 to -1.31	322 to 1,180	0.39 to 13.58
Tournament size 15 (drug-like compounds)	38	-1.20	-7.31 to -1.20	322 to 496	0.39 to 4.77
RSV 1-2	3,123	-0.88	-8.33 to -0.88	249 to 1,332	1.09 to 13.85
RSV 1-2 (drug-like compounds)	373	-1.96	-6.88 to -1.96	249 to 499	1.09 to 4.99
RSV 1-3	3,189	-0.03	-8.33 to -0.03	249 to 1,332	1.09 to 13.85
RSV 1-3 (drug-like compounds)	417	-0.03	-7.03 to -0.03	249 to 499	1.09 to 4.99
RSV 1-4	3,210	-0.03	-8.33 to -0.03	249 to 1,332	1.09 to 13.85
RSV 1-4 (drug-like compounds)	427	-0.03	-7.03 to -0.03	249 to 499	1.09 to 4.99
RSV 1-5	3,280	-0.03	-8.33 to -0.03	249 to 1,332	1.09 to 13.85
RSV 1-5 (drug-like compounds)	461	-0.03	-7.03 to -0.03	249 to 499	1.09 to 4.99

Table 8.18: Summary of the best performing runs of the structure generation for the lowest molecular weight Bzr inhibitor, using Pareto ranking. The results from RSV 1-5 to RSV 1-11 are identical, and so only RSV 1-5 is shown.

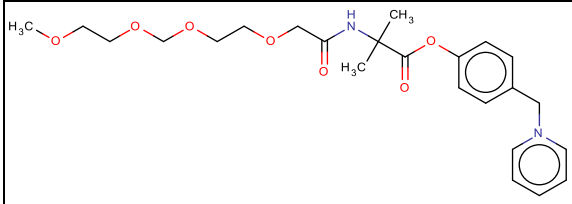
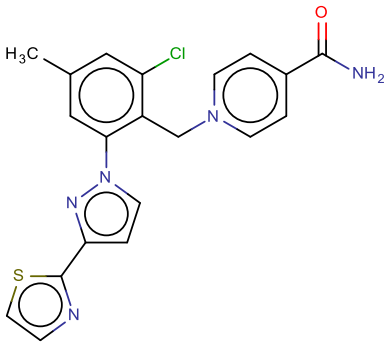
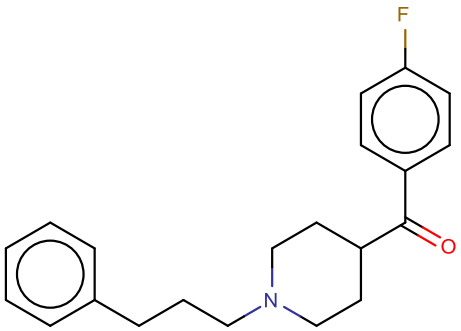
RV enumeration				
Structure	LogP	Molecular weight / g mol^{-1}	Bayesian score	Pareto Ranking
	2.27	161	1.31	1
	4.19	410	1.07	2
	4.35	325	0.72	3

Table 8.19: Drug-like compounds with the highest Bayesian score from the RV enumeration method for the Bzr experiment (sorted by Pareto ranking). (Wallace, 2015)

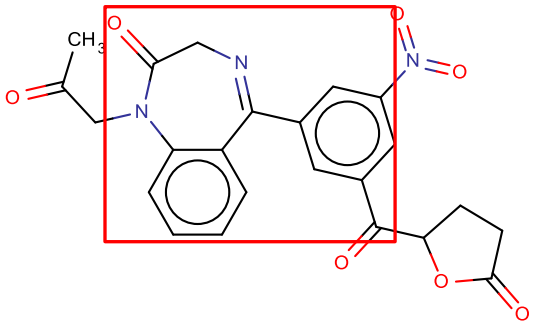
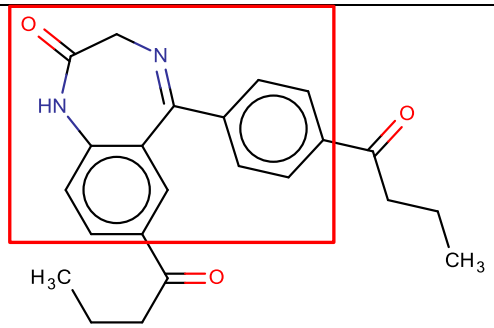
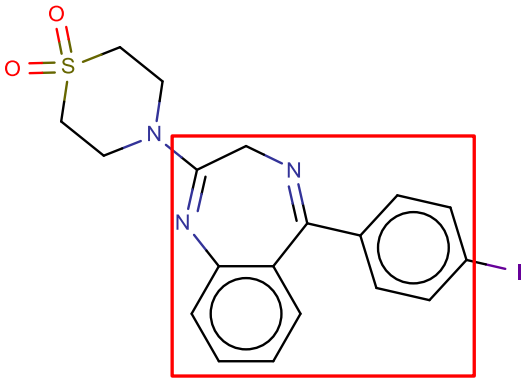
Tournament selection				
Structure (common inhibitor scaffold highlighted)	LogP	Molecular weight / g mol^{-1}	Bayesian score	Pareto Ranking
	2.25	449	-1.98	1
	4.44	376	-2.62	2
	2.90	479	-2.69	3

Table 8.20: Drug-like compounds with the highest Bayesian score from the tournament selection method for the Bzr experiment (sorted by Pareto ranking). (Wallace, 2015)

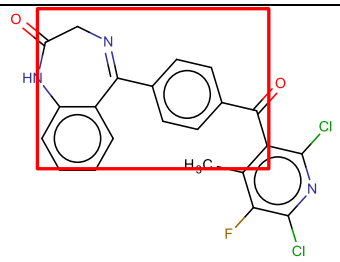
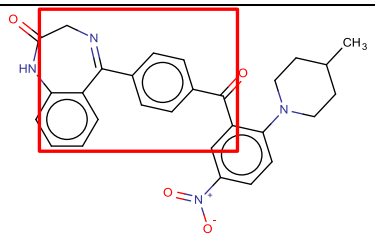
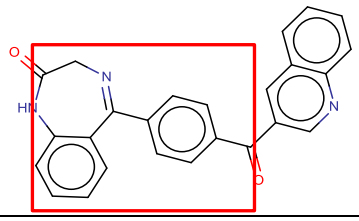
RSV enumeration				
Structure (common inhibitor scaffold highlighted)	LogP	Molecular weight / g mol^{-1}	Bayesian score	Pareto Ranking
	4.86	441	-1.73	1
	4.85	482	-1.96	2
	4.25	391	-2.00	3

Table 8.21: Drug-like compounds with the highest Bayesian score from the RSV enumeration method for the Bzr experiment (sorted by Pareto ranking). (Wallace, 2015)

Looking at the Pareto rankings for the generated molecules, there are six molecules in the full RV enumeration that are part of the non-dominated front, compared with two in the RSV case. Searching the SciFinder database for the 20 compounds identified as having the highest Bayesian scores in each case did not return specific entries for any of them. The RSV and tournament selection methods generate products that are closer to benzodiazepines in structure, while the full RV enumeration also includes functionalised 4-pyridines and 4-piperidines with scaffolds very similar to known drugs. These latter compounds have been reported as more selective benzodiazepine receptor analogues, used for the treatment of migraines (Burgey et al., 2006).

Overall, the RSV approach and tournament selection produce collections of drug-like molecules that are predicted to be highly unlikely to be active, albeit with scores above that of the starting material. This appears to be due to a limitation of the Bayesian model when dealing with multiple chemical classes. In these cases, as weights are

assigned to the structural fragments from the whole training set, confidence values are generally reduced as molecules that rely on one class will not have the features associated with the other. This affects the tournament selection method more than the RSV approach, as the lower scores compromise the selection of candidates for optimisation. However, the limit to using existing sequences in the RSV method reduces the amount of solution space that can be accessed; hence the lack of active results in this case.

8.3.1.3 Cyclooxygenase-2 (Cox-2) inhibitors

The Cox-2 inhibitor used as the starting material in this experiment is shown in Figure 8.15, with the results presented in Appendix C.3, Table C-4 and summarised in Table 8.22. This starting material has a pIC_{50} score of 7.22, above the activity threshold for the Bayesian model, which is set at 6.0. However, the Bayesian score for the starting material is -1.77, as there are few active examples in the data set, leading to lower confidence values for the model. This means that, in this case, when comparing the results of the sampling experiments, results with low negative scores may still be active. The molecules with the highest Bayesian score are shown in Table 8.23, Table 8.24 and Table 8.25. Unlike the other examples, while a tournament size of 10 gives the result with the highest Bayesian score, there is a significant variance in the data set. Considering mean peak Bayesian scores over the relevant runs, a tournament size of five gives more consistent results (a smaller variance), but these have lower scores than the products of the RV or RSV enumerations. For the tournament size of fifteen, all of the results have lower Bayesian scores than the starting material, indicating that in this case the optimisation has not been successful.

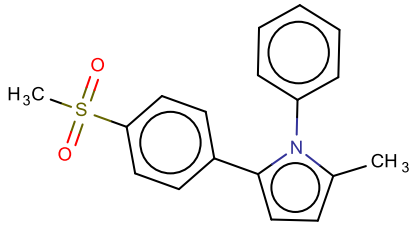


Figure 8.15: Lowest molecular weight molecule in the Cox-2 inhibitor set. (Sutherland et al., 2004)

Sampling experiment	Number of products	Highest Bayesian score	Bayesian score range	MW range / g mol^{-1}	LogP range
RV (full enumeration, 3 generations)	276,820	1.61	-12.78 to 1.61	317 to 1,959	-0.41 to 23.38
RV (drug-like compounds only)	2,925	0.39	-8.20 to 0.39	317 to 499	1.36 to 5.00
Tournament size 10	125	-1.17	-11.43 to -1.17	403 to 1,107	3.31 to 13.40
Tournament size 10 (drug-like compounds only)	5	-1.17	-4.07 to -1.17	403 to 445	3.32 to 4.64
Tournament size 5	121	-1.27	-10.03 to -1.27	447 to 1,058	2.86 to 11.73
Tournament size 5 (drug-like compounds only)	7	-1.27	-6.20 to -1.27	447 to 484	3.84 to 4.61
Tournament size 15	122	-2.47	-11.37 to -2.47	430 to 1,173	2.42 to 12.24
Tournament size 15 (drug-like compounds only)	5	-2.87	-7.03 to -2.87	430 to 494	4.64 to 4.87
RSV 1-2	3,101	0.29	-13.19 to 0.29	326 to 1,407	2.47 to 15.23
RSV 1-2 (drug-like compounds only)	44	0.01	-5.19 to 0.01	326 to 499	2.47 to 4.98
RSV 1-3	3,163	0.29	-13.19 to 0.29	326 to 1,407	2.47 to 15.23
RSV 1-3 (drug-like compounds only)	52	0.01	-5.19 to 0.01	326 to 499	2.47 to 4.98
RSV 1-4	3,184	0.29	-13.19 to 0.29	326 to 1,407	2.47 to 15.23
RSV 1-4 (drug-like compounds only)	59	0.01	-5.19 to 0.01	326 to 499	2.47 to 4.98
RSV 1-5	3,254	0.29	-13.19 to 0.29	326 to 1,407	2.47 to 15.23
RSV 1-5 (drug-like compounds only)	62	0.01	-5.19 to 0.01	326 to 499	2.47 to 4.98

Table 8.22: Summary of the best performing runs of the structure generation for the lowest molecular weight Cox-2 inhibitor, using Pareto ranking. The results from RSV 1-5 to RSV 1-11 are identical, and so only RSV 1-5 is shown.

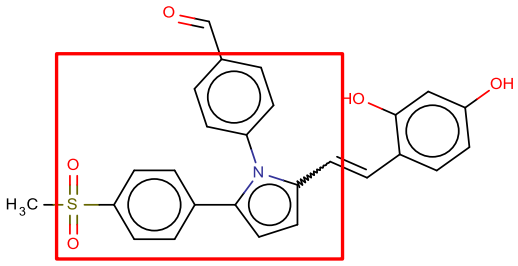
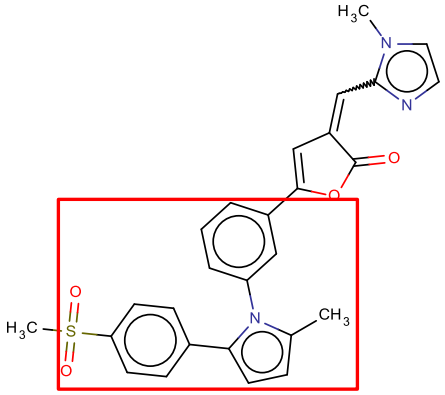
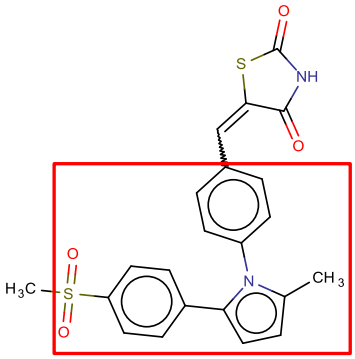
RV enumeration				
Structure (common inhibitor scaffold highlighted)	LogP	Molecular weight / g mol^{-1}	Bayesian score	Pareto Ranking
	4.94	459	0.39	1
	4.57	485	0.38	2
	4.18	438	0.21	3

Table 8.23: Drug-like compounds with the highest Bayesian score from the RV enumeration method for the Cox-2 experiment (sorted by Pareto ranking). (Wallace, 2015)

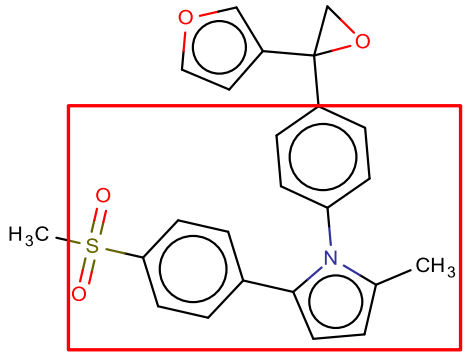
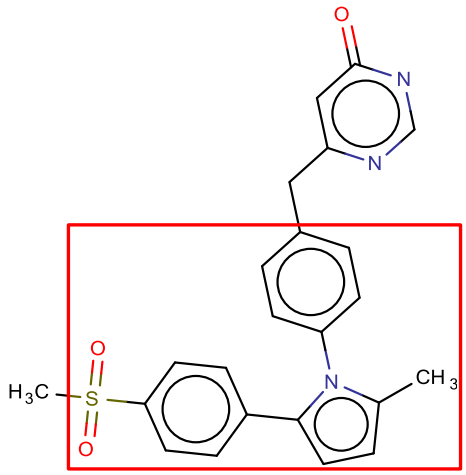
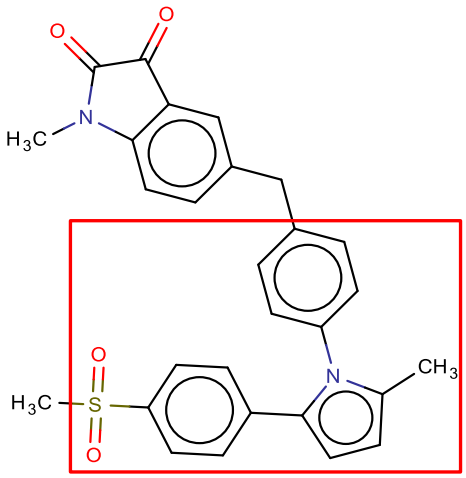
Tournament selection				
Structure (common inhibitor scaffold highlighted)	LogP	Molecular weight / g mol^{-1}	Bayesian score	Pareto Ranking
	4.72	419	-1.10	1
	3.32	418	-1.17	2
	4.61	484	-1.35	3

Table 8.24: Drug-like compounds with the highest Bayesian score from the tournament selection method for the Cox-2 experiment (sorted by Pareto ranking). (Wallace, 2015)

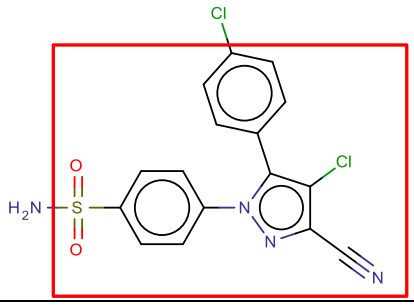
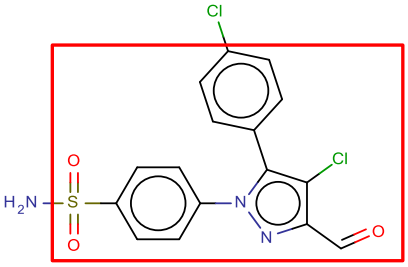
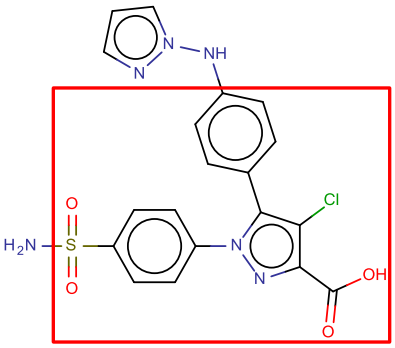
RSV enumeration				
Structure (common inhibitor scaffold highlighted)	LogP	Molecular weight / g mol^{-1}	Bayesian score	Pareto Ranking
	3. 37	391	0.07	1
	3.31	394	0.06	2
	2.61	458	0.05	3

Table 8.25: Drug-like compounds with the highest Bayesian score from the RSV enumeration method for the Cox-2 experiment (sorted by Pareto ranking). (Wallace, 2015)

The non-dominated Pareto fronts for the RV and RSV enumeration experiments are very similar, containing 12 and 15 molecules respectively. However, the results from the RSV approach are more drug-like, with smaller molecular weight and logP values than the equivalents in the RV case. Unusually, the tournament selection approach is outperformed by the RSV enumeration in terms of Bayesian scores in this case, with few drug-like compounds with positive Bayesian score generated overall. This is partly due to the model. As the starting material used is already highly active, but perceived by the model as having a low likelihood of activity, transformations performed on it to produce the initial population of results can distort the scaffold and lead to lower Bayesian scores. This leads to similar problems with scoring as seen with the Bzr case,

affecting the RV tournament selection method. It should be noted that the most active molecules in each group all possess the functionalised pyrrole scaffold associated with the literature examples of Cox-2 inhibitors. As with the previous classes, none of the highest scoring structures for each experiment are identified specifically as known inhibitors, with SciFinder containing no specific records for any of these results.

8.3.1.4 General trends in RV and RSV usage

Similar experiments to the ones summarised above were carried out for the Dihydrofolate reductase, Gpb and Thermolysin inhibitor classes in the Sutherland set. These generally followed the same trends as the others regarding the relative performance of the RV and RSV methods, with the RV enumeration and tournament selection producing molecules with higher scores than the RSV enumeration methods. The data for these runs are included in Appendix C, Sections C.4, C.5 and C.6. In all cases, molecules were generated that had higher Bayesian scores than the relevant starting material. For the RSVs, in the vast majority of cases, despite having sequences between two and eleven steps in length, the best results come from two or three step sequences. Given that many of the starting materials are active inhibitors in their own right, very few transformation steps are required to make changes in the predicted likelihood of activity through simple scaffold modification. In addition, as further reactions are performed on the molecule, any potential increases in Bayesian score are negated by increases in molecular weight and scaffold distortion, making the results less drug-like and potentially affecting the overall score. Although the shorter sequences were most effective in these cases, this trend may not be true if starting from typical reagents such as those in Section 6.3.3, where longer sequences may be of benefit due to the small sizes of the reagents.

Of particular interest is that fact that, in these examples, the RV approaches outperform the RSVs in terms of the likelihood of activity, in all but one case. This exposes a limitation of the RSV process, namely the restriction on chemical diversity that can be achieved. As all RSVs are generated from connected paths in the reaction network, the potential combinations of transformations that can be used is restricted to those that are known to directly follow on from one another, or operate on the same molecule. While RSVs can be applied iteratively, this has a negative effect on the accessibility of the method, as the combination of multiple unconnected reaction pathways makes planning the resulting synthesis more difficult. On the other hand, for the RV enumeration, each step operates independently of the previous steps, with the only

limitation being the applicability checks for individual vectors. While the restriction of diversity is a limitation of the RSV approach, it is also an integral part of its design. By limiting the combinations of reactions to those that have a real world synthetic precedent, this increases the likelihood of a molecule generated from RSV application being synthetically accessible. When combining random reactions, such as in the RV method, no consideration is made as to whether a previous reaction has any electronic or resonance effect on the structure that could prevent the current reaction from working correctly. In these circumstances, the resulting molecule may appear to be of interest in the *in silico* result list, however the resulting real world synthesis may be challenging due to structural incompatibility. For this reason, the RSV approach remains a worthwhile companion to the standard RV approaches. It should also be noted that the results overall for the RSV cases tend to plateau after considering sequences of approximately six steps in length with no further molecules being generated at longer sequences. As there are fewer sequences present at these lengths, the likelihood of there being transformations applicable to the starting material is greatly reduced, meaning that no new results will be produced.

When using Pareto ranking to analyse the results of the full enumerations of the RV and RSV solution spaces for each inhibitor, it can be seen that there are relatively few molecules at the first Pareto front, representing the non-dominated solutions. This is surprising, but given the sensitivity of the Pareto ranking method to changes in logP and molecular weight, and the relatively wide range of molecular properties observed, the Pareto fronts are relatively small. It should also be noted that for these data sets there does not appear to be a direct connection between molecular weight and activity. The results from the RV enumeration do not necessarily represent drug-like compounds, due to the structures not conforming to the Lipinski rules, but the RSV examples tend to be closer to the accepted values. Performing the same analysis after filtering out the non-drug-like molecules shows very different results, with only one molecule being present in the lead front in each class. In this case, the ranking is effectively based on the Bayesian scores alone, as the other properties will all be within the ranges defined in the Lipinski rules.

Considering the RV tournament selection method, it appears that different tournament sizes give the best results for different inhibitor classes. Statistical analysis of the reported results of each individual run (as shown in Appendix C.7) indicates that there is a fairly wide deviation between the different tournament sizes and the Bayesian

scores. In general, it appears that larger tournament sizes will lead to results with higher scores, but the correlation is not strong. However, tournament sizes that give higher Bayesian scores are associated with wider variance in the data, and thus show less consistency in the data. The logical explanation for this is that some of these higher scoring results are detached from the bulk of the distribution, resulting in a significant skew, as seen by the variance. It is therefore necessary to find the best compromise finding the best results, while controlling result consistency reproducibility. Since the larger tournament sizes tend to have smaller variance figures, and a tournament size of fifteen consistently gives results with reasonably high likelihood of activity, it is recommended that this size should be used.

8.3.2 Multiple starting materials

To determine the effect of larger initial populations on the tournament sampling method, a second series of experiments was performed using multiple starting materials for each data set. For each set of inhibitors, ten starting materials were selected at random, representing the full range of activity and molecular properties. These starting materials were used to perform tournament selection over three iterations, using a tournament size of fifteen, as well as a full RSV enumeration. It should be noted that a full RV enumeration for these classes was impossible due to the very large numbers of molecules produced in each iteration. Consequently, the RV tournament selection method was compared directly with the full RSV enumeration for sequences up to three steps in length, as these two methods represent an equivalent number of reactions.

For the tournament selection process, each starting material was considered separately as an independent experiment, with three runs of the tournament method used for each. As before, the run containing the result with the highest Bayesian score was selected for each experiment, with the best results pooled to give an overall result for the data set (giving a maximum of 2000 result molecules for each collection). Each of these runs took approximately 15 minutes to complete on the i7 workstation, making this process particularly time consuming (approximately 7 hours 30 minutes per compound class). Considering that the full RSV enumeration for the data set only takes 20 minutes per class in total under the same conditions, it is clear that, in evaluating more potential combinations of reactions and products the RV approach is significantly slower. The results of the tournament selection approach and the RSV enumeration for the Ace inhibitor sets are summarised in Table 8.26 below, with the drug-like

molecules with the highest Bayesian scores shown in Table 8.27 and Table 8.28. It should be noted that, for the RV tournament selection case, fewer than 2,000 molecules are reported, as some of these were duplicates that were filtered out. As expected, the use of additional starting materials results in the Bayesian scores increasing relative to the single material case, with a considerably larger set of results for the RSV enumeration. In this case, the starting material in the data set with the greatest Bayesian score had a value of 8.32, showing that both approaches offered apparent improvements over the original collection.

Sampling experiment	Number of products	Highest Bayesian score	Bayesian score range	MW range / g mol^{-1}	LogP range
RV (tournament selection, 3 generations)	1,954	13.48	-4.16 to 13.48	217 to 1,735	-4.48 to 15.00
RV (drug-like)	174	10.97	-6.98 to 10.97	252 to 499	0.09 to 4.94
RSV 1-2	502	11.22	-4.08 to 11.22	271 to 968	-3.89 to 6.96
RSV 1-2 (drug-like)	73	8.01	-2.51 to 8.01	271 to 499	0.16 to 3.09
RSV 1-3	16,840	15.55	-4.08 to 15.55	271 to 1,627	-3.89 to 12.51
RSV 1-3 (drug-like)	366	8.57	-2.85 to 8.57	271 to 499	0.02 to 4.44

Table 8.26: Summary of the best performing runs of the structure generation for the Ace inhibitor set, using Pareto ranking.

There is a considerable increase in the number of molecules produced via RSV enumeration in this case, compared to the when individual starting materials were considered. Of particular interest is the much greater number of products derived from three-step sequences, an increase of over 16,000. This is due to some of the starting materials with higher molecular weights being more suitable to the application of RSVs than the smaller materials previously considered. As the number of RSVs increases, yet more products can be generated, hence the dramatic increase in population size. However, because these molecules are derived from high molecular weight starting materials, the likelihood of them being drug-like is very low, and as such the filtered results do not show a significant increase.

Searching for the 20 highest scoring compounds from each experiment on SciFinder indicated that the results are close in nature to derivatives of L-Proline and other amino

acids as seen with the previous experiments with this data set (Section 8.2). However, the structures are significantly different to those produced from the lighter starting materials used in Section 8.3.1.1, in that the mercaptohexanoic acid scaffolds are not present. This would suggest that using a single starting material for this type of study may be unreliable, especially if the starting material does not contain functionality that is key for activity, such as the proline residue that most Ace inhibitors possess. As before, no specific entries in SciFinder were found for any of these result molecules. To confirm the general trends, further experiments were carried out using the other inhibitor classes, as summarised in Table 8.29 and Table 8.30. These show the results from the RV tournament selection mode for each inhibitor class, compared with the results of the RSV enumeration in each case.

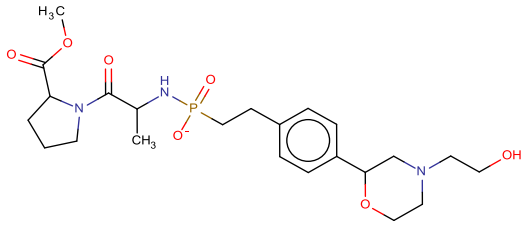
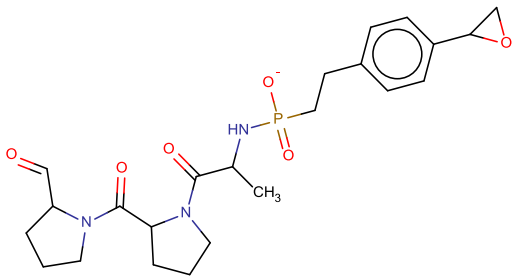
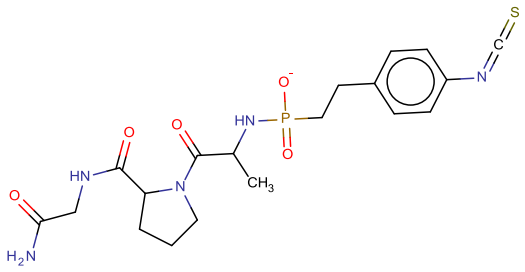
Tournament selection				
Structure	LogP	Molecular weight / g mol^{-1}	Bayesian score	Relative Pareto Ranking
	0.29	496	10.97	1
	1.01	476	8.98	2
	0.09	466	8.66	3

Table 8.27: Drug-like compounds with the highest Bayesian scores from the tournament selection method for the Ace inhibitor set (sorted by Pareto ranking). (Wallace, 2015)

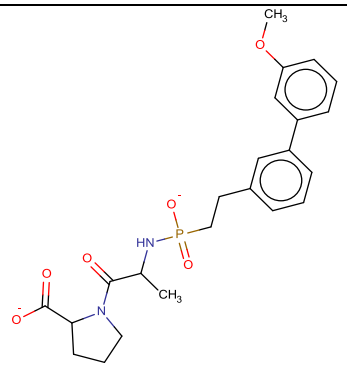
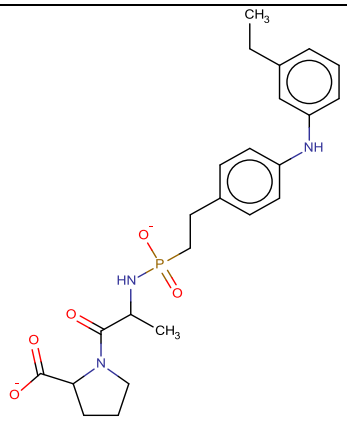
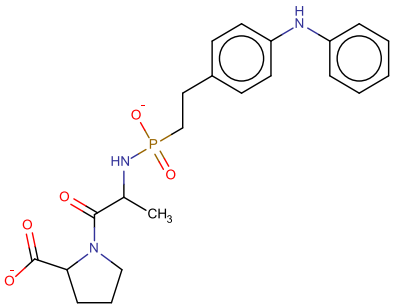
RSV enumeration				
Structure	LogP	Molecular weight / g mol^{-1}	Bayesian score	Pareto Ranking
	1.18	458	8.57	1
	1.81	471	8.01	2
	1.25	443	7.70	3

Table 8.28: Drug-like compounds with the highest Bayesian scores from the RSV enumeration for the Ace inhibitor set, using multiple starting materials (sorted by Pareto ranking). (Wallace, 2015)

Inhibitor class	Number of products	Highest Bayesian score	Bayesian score range	MW range / g mol^{-1}	LogP range
Ace	1,954	13.48	-4.16 to 13.48	217 to 1,735	-4.48 to 15.00
Ace (drug-like)	174	10.97	-6.98 to 10.97	252 to 499	0.09 to 4.94
Bzr	1,777	1.24	-10.35 to 1.24	369 to 1,706	-0.05 to 16.87
Bzr (drug-like)	43	-0.30	-5.35 to -0.30	376 to 498	2.20 to 4.49
Cox-2	1,919	0.09	-12.26 to 0.09	421 to 1,747	1.51 to 24.70
Cox-2 (drug-like)	21	0.09	-7.61 to 0.09	421 to 499	2.25 to 4.80
Dhfr	1,976	1.67	-23.18 to 1.67	305 to 1,671	-1.00 to 18.89
Dhfr (drug-like)	84	0.28	-20.59 to 0.28	305 to 499	0.01 to 4.89
Gpb	1,983	3.09	-1.38 to 3.09	180 to 1,806	-3.91 to 14.75
Gpb (drug-like)	77	1.40	-1.21 to 1.40	276 to 478	0.08 to 5.00
Therm	2,000	7.03	-4.99 to 7.03	389 to 1,823	-3.91 to 18.54
Therm (drug-like)	146	3.45	-3.27 to 3.45	292 to 499	0.22 to 4.77

Table 8.29: Summary of the best performing runs for all activity classes from the RV based tournament selection.

Inhibitor class	Number of products	Highest Bayesian score	Bayesian score range	MW range / g mol^{-1}	LogP range
Ace	16,840	15.55	-4.08 to 15.55	271 to 1,627	-3.89 to 12.51
Ace (drug-like)	366	8.57	-2.85 to 8.57	271 to 499	0.02 to 4.44
Bzr	10,774	2.61	-8.87 to 2.61	316 to 1,448	1.86 to 15.83
Bzr (drug-like)	122	1.57	-6.28 to 1.57	316 to 499	1.86 to 5.00
Cox-2	64	0.33	-9.89 to 0.33	352 to 1,112	2.89 to 11.61
Cox-2 (drug-like)	9	-0.04	-4.70 to -0.04	352 to 478	3.16 to 4.45
Dhfr	6,114	0.01	-27.42 to 0.01	213 to 1,444	-0.87 to 15.72
Dhfr (drug-like)	394	0.01	-17.23 to 0.01	213 to 499	0.32 to 5.00
Gpb	27	1.96	-1.10 to 1.96	265 to 632	-4.11 to 6.64
Gpb (drug-like)	2	-0.45	-0.45 to -0.45	318 to 325	0.30 to 0.55
Therm	26,104	8.25	-3.85 to 8.25	309 to 1,792	-2.64 to 13.10
Therm (drug-like)	355	4.33	-1.95 to 4.33	309 to 499	0.05 to 4.80

Table 8.30: Summary of the best performing runs for all activity classes from the RSV enumeration (sequences from 1 to 3 steps in length).

It should be noted that the Bzr, Cox-2 and Dhfr results generate particularly low numbers of molecules and these had low Bayesian scores overall for both RVs and RSVs. In the case of Cox-2 these scores are lower than the scores assigned to the starting molecules (1.01), implying the optimisation gives worse results than the original material. As both the RSV and RV approaches are affected in the same manner, this problem is not due to any issues with the population sampling required for tournament selection. The Bayesian scores were also very low in the previous single starting material experiments. This seems to be due to a lack of active examples in the respective data sets, adversely affecting the scoring and analysis.

Comparing the two approaches shows that the RSV enumeration outperforms the RV tournament selection in terms of the highest Bayesian score for each inhibitor class in all but two cases. However, if only considering drug-like molecules this is reversed, with the tournament selection being superior in four of the six classes studied. Given that the RSVs are derived from sequences that can be up to length 11 and the approach is based on full enumeration, there is a tendency for this approach to generate molecules with large molecular weights where these subsequences are part of a long process. These larger molecules are more likely to contain high scoring fragments, hence the higher Bayesian scores. When these are removed, the highest Bayesian score drops below that which is achieved using the RVs. The greater diversity of compounds that is accessible using RVs leads to results with better Bayesian scores within the drug-like region, even using the tournament sampling. Theoretically a full enumeration using RVs may lead to further increases in the number of high scoring molecules, however, this is computationally prohibitive especially with multiple starting materials. Instead, using the RSV approach as part of a pilot study can determine what is feasible for a given set of starting materials. The best start points can then be used with a full RV enumeration or tournament selection to provide an optimal set of results, without considering starting materials that are inappropriate. For example, in the Ace case discussed above, a number of the starting materials have high molecular weights and complex structures, making it unlikely that these would generate active, drug-like results on further reactions. While identifying molecules like this in a real world evaluation would be straightforward, other, more subtle factors that affect the applicability of starting material may be present, leading to wasted time if these molecules are used to generate large quantities of unsuitable results.

An additional point to consider is that both experiments only consider sequences up to three steps in length. For the RSV enumeration, longer sequences lead to larger molecules with a higher Bayesian score in most cases. As adding further iterations to the RV method (full enumeration or tournament selection) would cause more issues with execution time and system memory due to the increased number of structures to consider, a speculative evaluation with the RSV method may be preferable to determine the optimal sequence length prior to a full study. Longer sequences may also lead to a greater proportion of results that are not drug-like, limiting any potential gain in result quality.

8.3.3 Structure generation from simple starting materials

In order to study the relative capabilities of the RV and RSV methods to generate active structures from typical reagent molecules, experiments were carried out to produce analogues of thrombin inhibitors, starting from a pool of starting materials as shown in Figure 8.19. These materials are not inhibitors themselves, unlike the examples in the Sutherland collection, which all had some level of inhibition activity. However, they do all contain the benzamidine group required for thrombin inhibition.

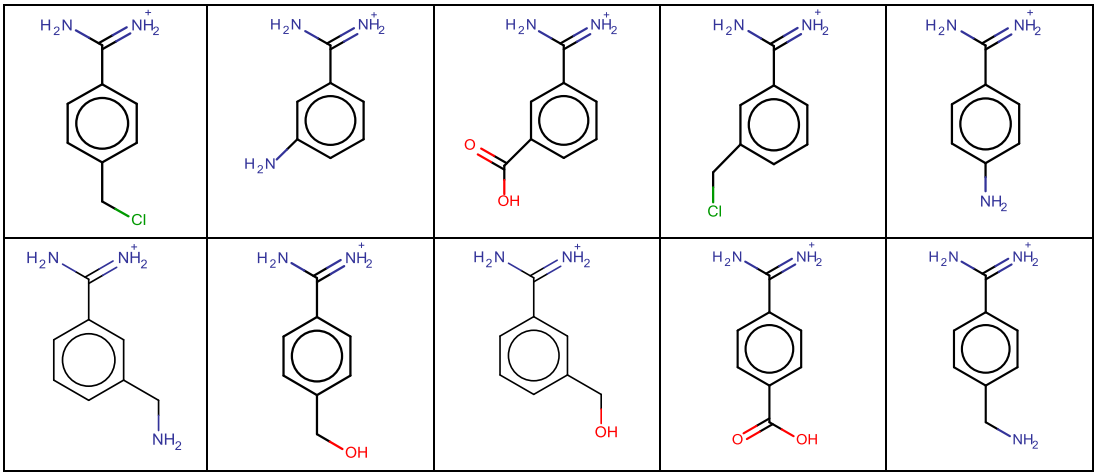


Figure 8.16: Starting materials used for the thrombin structure generation experiment. (Wallace, 2015)

In this instance, the result molecules were Pareto ranked by logP value, molecular weight and similarity value. LogP and molecular weight were scored in accordance with the Lipinski rules as before, and similarity was based on fingerprint similarity to four known thrombin inhibitors, as illustrated in Figure 8.20.

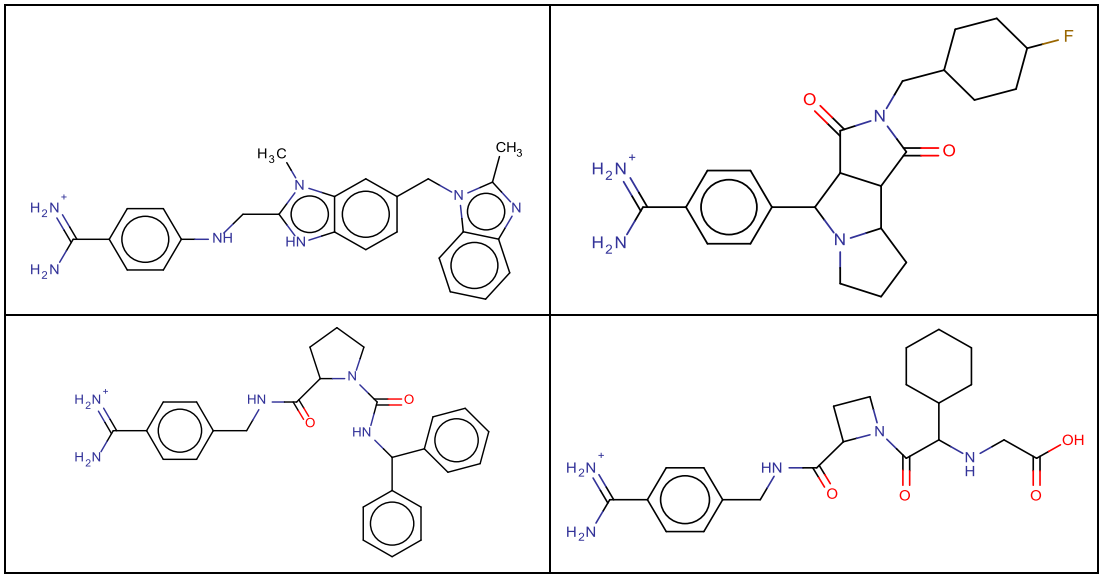


Figure 8.17: Known thrombin inhibitors used to generate fingerprints. (Wallace, 2015)

A composite fingerprint was achieved by generating RDKit structural fingerprints for each inhibitor. A new fingerprint was then produced, with each bit set if the corresponding bit is set for any of the known inhibitors. Structures were then evaluated based on similarity to this fingerprint, with structural similarity implying activity in accordance with the similar property principle.

In this experiment, the starting materials were used to generate structures, using the RSVs from the JMCRoughley database that represent sequences between one and four steps in length. Due to the same issues with execution time and system memory experienced in previous experiments with multiple starting materials, full enumerations of the RV database were not possible. Instead, the RSV results were compared to the results of a tournament selection sampling method, carried out using the same method as in Section 8.3.2. This involves carrying out three separate runs for each starting material with a tournament size of fifteen, and a population size of 500, with results reported for the run which gave the largest number of drug-like results and the highest similarity scores. Each run was carried out over four iterations of the RV process to be comparable with the sequence lengths from the RSV enumeration. As a result, each iteration took approximately 20 minutes to complete, with the full experiment taking ten hours on the i7 workstation. The results at each iteration of the tournament selection were considered separately, permitting comparisons for sequences from two to four steps in length.

The results of the RSV and RV approaches are summarised in Tables 8.31 and 8.32. In total, 2,542 unique molecules were produced from the application of RSVs to the

original starting materials, with 1,585 of these being drug-like. Of the remainder, 914 had molecular weights above 500g mol^{-1}; 337 molecules had logP values outside of the required range; and 294 molecules had properties outside of both thresholds. The RV tournament selection, on the other hand, produced 4,565 unique molecules after combining the results for all starting materials, with 970 of these being drug-like. The other molecules consisted of 3,191 that had molecular weights above 500g mol^{-1}; 2,800 molecules had logP values outside of the range; and 2,396 molecules were outside of both ranges.

RSV Enumeration								
Sequence length (number of reactions)	Number of unique products	Similarity range	LogP range	MW range/ g mol^{-1}	Number of drug-like molecules	Lipinski rule violations		
						Molecular weight >500g mol^{-1}	LogP out of range	Both
2	2156	0.06 to 0.29	-4.28 to 9.48	136 to 1,291	1,427	766	183	220
3	205	0.10 to 0.24	-2.46 to 7.87	179 to 889	94	64	93	46
4	181	0.10 to 0.30	-2.08 to 7.62	178 to 911	64	84	61	28

Table 8.31: Summaries of the results of the RSV enumeration approach for suggesting Thrombin analogues.

RV Tournament Selection								
Number of reaction iterations	Number of unique products	Similarity range	LogP range	MW range/ g mol^{-1}	Number of drug-like molecules	Lipinski rule violations		
						Molecular weight >500g mol^{-1}	LogP out of range	Both
2	1591	0.09 to 0.26	-2.99 to 11.90	165 to 1465	650	733	461	253
3	1598	0.05 to 0.22	-3.03 to 14.44	192 to 1363	265	1186	1074	927
4	1376	0.14 to 0.18	-3.17 to 17.50	176 to 1398	55	1272	1265	1216

Table 8.32: Summaries of the results of the RV tournament selection approach for suggesting Thrombin analogues.

Using the similarity to the known inhibitors as the main measure of performance, it can be seen that the RSV enumeration approach produces molecules that have higher similarity values, with a greater proportion of the results for each sequence length

being drug-like than the equivalent results for the tournament selection. Studying the best performing results from the two methods shows that they both contain results with similar structural features, with the RSV method having simpler scaffolds. Examples of these are shown in Tables 8.33 and 8.34.

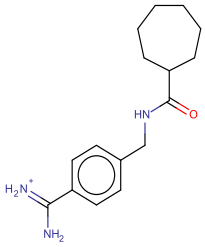
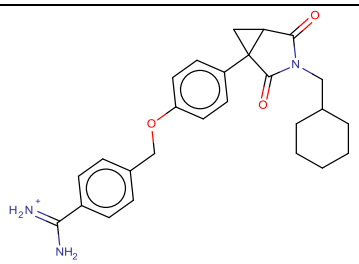
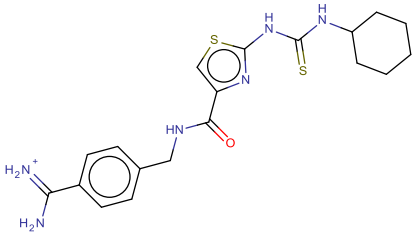
Structure	Similarity	LogP	Molecular weight / g mol^{-1}	Pareto Ranking
	0.30	0.74	274	1
	0.27	1.94	432	2
	0.26	1.16	417	3

Table 8.33: Examples of the best scoring, drug-like results for the RSV experiment. (Wallace, 2015)

The similarity scores for these results are relatively low, compared to the composite fingerprint. Looking at the literature examples in isolation, it can be seen that a wide range of functional groups are represented on the inhibitor scaffold, leading to a composite fingerprint with a high number of bits set. With such a diverse collection of functionalities, it is unlikely that a single molecule will contain the majority of these, and as such, the similarity coefficient will be low. It is important to note, however, that the main inhibitor functionality is present in the vast majority of the result molecules, in the form of the benzenecarboximidamide group.

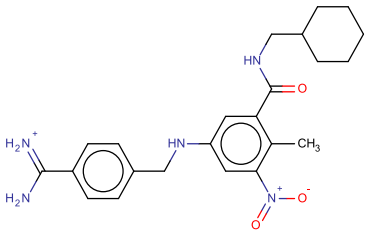
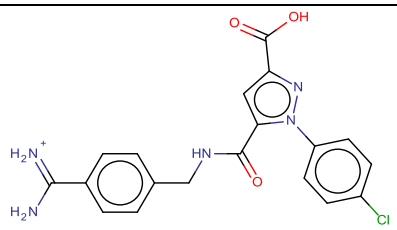
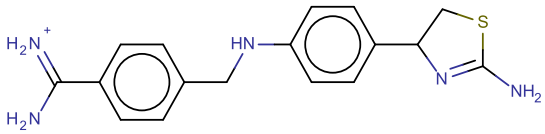
Structure	Similarity	LogP	Molecular weight / g mol^{-1}	Pareto Ranking
	0.26	2.29	424	1
	0.23	0.62	398	2
	0.22	0.87	326	3

Table 8.34: Examples of the best scoring, drug-like results for the RV Tournament selection experiment. (Wallace, 2015)

Benzenecarboximidamide has been reported as the key component of factor Xa inhibitors used for treatment of thrombin related disorders (Dorsch et al., 1999), and as such its presence in the generated structures indicates a high likelihood of activity. The remainder of the structure serves to optimise the structure in terms of bioavailability and binding affinity, hence the wider variety of functionality.

Comparing the best scoring drug-like molecules at each sequence length (Table 8.35) shows that the two different approaches generate different, but largely comparable molecules. As with the general summary, the RSV approach leads to results that are closer in similarity to the existing inhibitors with the similarity score dropping with each iteration of the RV approach. Once four reaction iterations have been completed, the results for the RV approach no longer contain the benzenecarboximidamide group, significantly reducing the activity. In this case, as the similarity scores are relatively low, so that selecting the best candidates for optimisation at each step is difficult. When poor candidates have been selected, further iterations do not lead to score improvements, resulting in greater deviation in the results, and ultimately poorer solutions. In the RSV case, since there is no scoring of the results until the enumeration is completed, these limitations have no effect, and as such the similarity scores remain largely consistent as the sequence length increases.

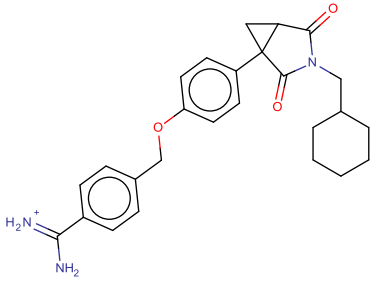
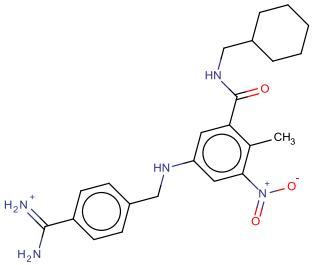
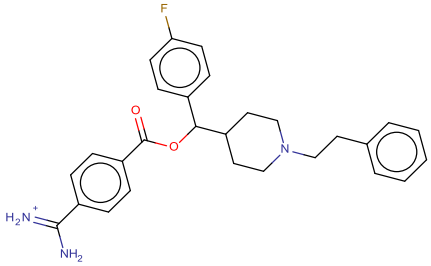
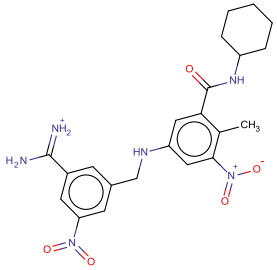
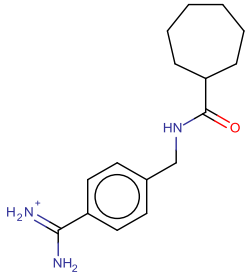
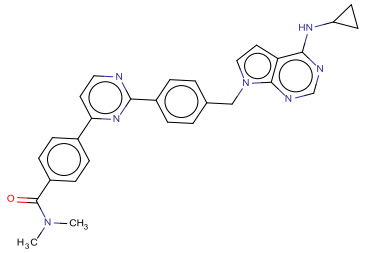
	RSV Enumeration				RV Tournament Selection			
Sequence length	Molecule	Tanimoto similarity to known actives	LogP	Molecular weight / g mol^{-1}	Molecule	Tanimoto similarity to known actives	LogP	Molecular weight / g mol^{-1}
2		0.27	1.94	432		0.26	2.29	424
3		0.24	3.14	460		0.21	1.95	455
4		0.30	0.74	274		0.16	4.88	489

Table 8.35: Best performing drug-like result molecules for the RV and RSV approaches, for Thrombin analogues. (Wallace, 2015)

Overall, the greatest proportion of drug-like molecules with reasonable similarity to the existing inhibitors is found with the RSV enumeration, with the tournament selection tending towards compounds with greater deviation from the original inhibitor structures. In situations where the scoring of compounds is complex, or there are other issues with the optimisation of solutions, full enumeration approaches such as the RSV method will ultimately be of benefit. A full RV enumeration would most likely provide the best solutions due to the wider range of chemistry available; however, as noted previously, full enumerations using multiple starting materials are computationally infeasible.

8.4 Conclusions

In this chapter, a study of the effect of using a database of RSV material with starting materials unconnected to the original reactions demonstrated that meaningful results could be obtained despite the increase in specificity in the RSVs. A significant proportion of the generated molecules have been assessed as synthetically accessible, through both retrosynthetic analysis and review by experts. However, this review also highlighted some limitations of the vectors when used for structure generation such as the treatment of rings in certain circumstances. Overcoming these limitations would require a reworking of the vector format itself and in some cases it would be relatively easy to simply filter out any unintended molecules.

Comparisons between the RSV method and multi-objective optimisation utilising the RV approach indicate that the latter method is considerably more effective at producing molecules predicted to be active for simple multi-objective studies from single starting materials. This is to be expected, as the collected RVs represent a far greater diversity than the equivalent RSV collection, and so better results are more likely to be generated. In theory, this implies that the best results would be achieved via a full enumeration of the sample space via RVs, evaluating the population after generation. However, when multiple starting materials and/or multiple iterations are required as part of the enumeration, the many different combinations of possible reactions and starting materials results in a combinatorial explosion. As a result, sampling methods are still required to keep the results manageable.

Sampling approaches for RVs such as tournament selection appear to be effective where the scoring of results is well defined i.e. where there is a smooth progression between starting material and product. However, where the scoring of intermediate

results is problematic, the tournament approach suffers from the same problems as other similar approaches in terms of identification of the ideal candidates to select for optimisation. In these cases, the best course of action is to evaluate the compounds *post hoc*, requiring a full enumeration of the sample space. The RSV method is a good compromise in this situation, as the restrictions in the diversity of the data enables rapid sampling of the covered solution space quickly. While these results are not as complete as with RV enumeration, they are ideal for pilot studies, selecting areas for further analysis via the more complete RV data, enabling a more focussed study.

In this chapter, two different models were used to predict the activity of the result molecules; an in-house SVM model, and a simpler Bayesian predictor. The SVM approach offers more precise prediction of pIC_{50} values in addition to classification as active and inactive, with the Bayesian approach only offering confidence values for the class assignment. This may be sufficient for the majority of cases, but where there are multiple mechanisms of action within the same inhibitor set (such as with the benzodiazepine case), the Bayesian model is less effective. In these situations, the activity predictions are based on the structural features of active examples taken from all of the mechanisms used, which can be mutually exclusive. As the predicted activity score is based on the proportion of these features that are present in the molecule being evaluated, this can result in lower predicted activity scores, and possibly incorrect categorisation of the molecule in question. In these circumstances, a more robust model is required that can distinguish between potential mechanisms, in order to more accurately classify the molecule.

Chapter 9:
Conclusions and Future Work

9.1 Conclusions

This thesis details the creation and evaluation of a new method for encoding and applying reaction sequence information for the purposes of *de novo* design. The ability to encode and represent entire sequences in this form is of particular importance in drug design applications to avoid many of the problems associated with multi-objective optimisation of result molecules. In particular, the new approach removes the pitfalls associated with sequences that proceed through intermediates that cannot be accurately scored and evaluated relative to the desired end product.

Chapter 4 describes the Reaction Vector format as originally used for individual reactions, and the methods by which novel structures can be generated through application of reaction vectors. Two different methods were discussed, the original approach designed by Patel (Patel et al., 2009) is explained in Section 4.3.1, and a more efficient revised approach by Hristozov (Hristozov et al., 2011) is covered in Section 4.3.2.

Chapter 5 describes the first new experiments carried out for this study, with the creation of a test database of reaction sequences outlined in Sections 5.2 and 5.3. These were collated into a network form that connects all the known molecules in the database according to the reactions that transform them, as shown in Section 5.4. After establishing that this approach could be used to obtain sequence information (via the use of network path finding algorithms), the same method was used to provide sequence information for a larger set of reactions previously collated from research papers published in J. Med. Chem., as discussed in Section 5.5. A reaction network was created for these reactions, and a collection of sequences generated and profiled for use in *de novo* studies. The reaction sequence vectors (RSV) are described in Chapter 6 and are based on computing the difference between the start and end points of a reaction sequence. The Hristozov structure generation algorithm was extended to allow RSVs to

be applied to a starting material to generate the product of a reaction sequence and the algorithm was demonstrated to perform with between 80 and 99% success in reproducing the original sequence content. Section 6.3.3 contains a direct comparison between the RV and RSV approach in terms of the number of unique products generated. The RSV approach leads to considerably fewer results than the RV approach, due to the bypassing of the intermediate stages. In terms of the quality of results however, there is little difference between the two sets, with the RSV approach being far more efficient in execution. An analysis of the novelty of the generated results in Section 6.3.4 shows that, for all of the given sets of sequences tested, there is a distinct skew in terms of which starting materials produce the most results, as well as in terms of which reaction sequence vectors are used. The most popular starting materials are those with the most potential for functional group addition (such as those with basic ring structures). In Section 6.3.5 an analysis of the RSV usage was carried out which and showed that the requirement for unusual metal atoms or complex functional groups in the required starting material features reduces the likelihood of a sequence being applicable, while straightforward, those requiring aromatic functionality are more likely to be applied.

Chapter 7 illustrates the application of the RSV method and the *de novo* design tools for a variety of real world applications. Section 7.2 described the use of RSVs to identify where multiple routes exist between the same start and end points. Section 7.3, describes the use of RSVs to expand the potential products in literature-based SAR analyses based on cilomilast, hydroxamates, carboxamides and a substituted alkyne feedstock. In all of the examples, the original starting material leads to the production of a number of interesting analogues that are structurally similar to those already known, but have not been studied for activity. The developed tools enable analysis of these compounds in PCA plots, as well as the reporting of the reactions identified to synthesise them, in the form of interactive reaction networks.

Chapter 8 describes a more detailed analysis of the relative merits of the RV and RSV methods, taking a series of active drug compounds and using them as starting materials for structure generation processes. The results produced using reactions obtained from J. Med. Chem. complement the original results in terms of their predicted pIC_{50} values, demonstrating the ability of encoding transformations in a more generic form. Automatic and manual analysis of the synthetic accessibility of the compounds indicated that at least 34,250 of the 68,369 products of the RSVs were likely to be

synthesisable in real world conditions. Finally, a direct comparison is made between the RSV method and a multi-objective method based around the application of multiple RVs (Section 8.2). As the RSV method was designed specifically to address issues with this multi-objective approach, it was interesting to note that, for simple synthesis examples, the RV approach outperformed the RSV method in terms of result quality. However, the RV method is slower to run, and cannot be fully enumerated for large result sets. When considering more complex synthesis sequences, the RSV approach may be more effective at providing a summary of solution space, profiling all areas of solution space evenly.

9.2 Future work

There are a number of areas that are identified for potential improvements and additions to this work. The first issue concerns the reaction network form used to generate the reaction sequences. In the current approach, an algorithm is used to identify the key molecules of each reaction, which can sometimes lead to errors and misrepresentation. Some collections of reaction data, such as the NextMove collection of patent information (Lowe and Sayle, 2014) feature atom mapping information that can be used to identify the roles of the individual molecules. Adapting the network code to identify and use these roles would ensure the correct intentions of all reactions are selected in all cases, improving the quality of the data used for RSV generation. Adding further information to the reaction network is also potentially of benefit for making comparisons between potential routes of interest. Currently, information such as the molecule structure information and the original reaction reference is stored in the network, but potentially other properties associated with reactions could be added. Factors such as estimated cost of a particular process (financial or in terms of environmental impact), or any reported yields could be used to score the individual reactions and sequences, and made available to the user enabling an effective comparison to be made between routes.

Secondly, both the RV and RSV processes have issues with execution speed, in terms of the generation of the RV and RSV databases and the application of the vectors to produce new molecules. The RV approach in particular is affected by this, with full enumerations running too slowly to be performed for all but the simplest of examples. This is due to the approach used to retrieve vectors for application, which slows down when a large number of vectors are stored. When creating structures, every vector stored in the database has to be analysed to determine if it is applicable. For large

collections of reactions, this has a considerable impact on processing speed, particularly as this process is repeated for every iteration. An alternative approach to storing the vector content would be to use the negative AP2 content of each vector as an indexing key for the database, grouping the vectors on that basis. When searching for vectors to apply, AP2 content of the starting material would be compared to the various index values, with only the vectors in groups that match this initial step being analysed further. This approach could also be used to enhance the efficiency of the RSV process, enabling very large collections of sequences to be handled. This could be further enhanced via performing more thorough curation of the sequences to be stored, removing those that are overly complex, or irrelevant to the structure generation experiment in question. While this will not resolve the issues with processing large quantities of molecules due to the combinatorial explosion, it should be sufficient to improve the searching speed to a point that makes RV methods more appropriate.

Appendix A:

Frequency distribution analysis

A.1 Analysis of the first US patent set

This set was collated at random from a collection of US patent data representing chemical reactions used in an industrial context.

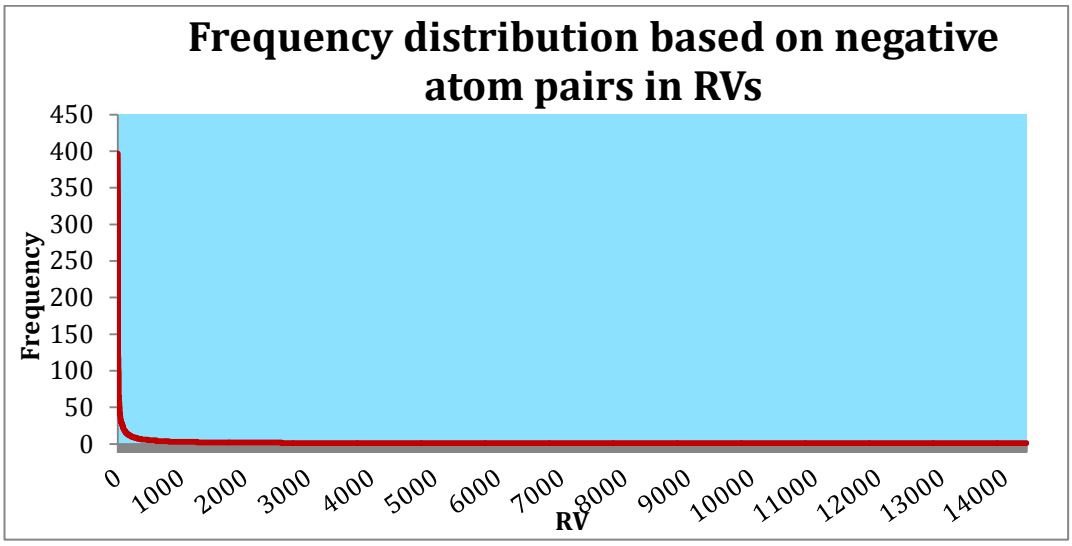


Figure A-1: Frequency distribution curve based on the negative AP2 content for the first random sample extracted from the US patent database.

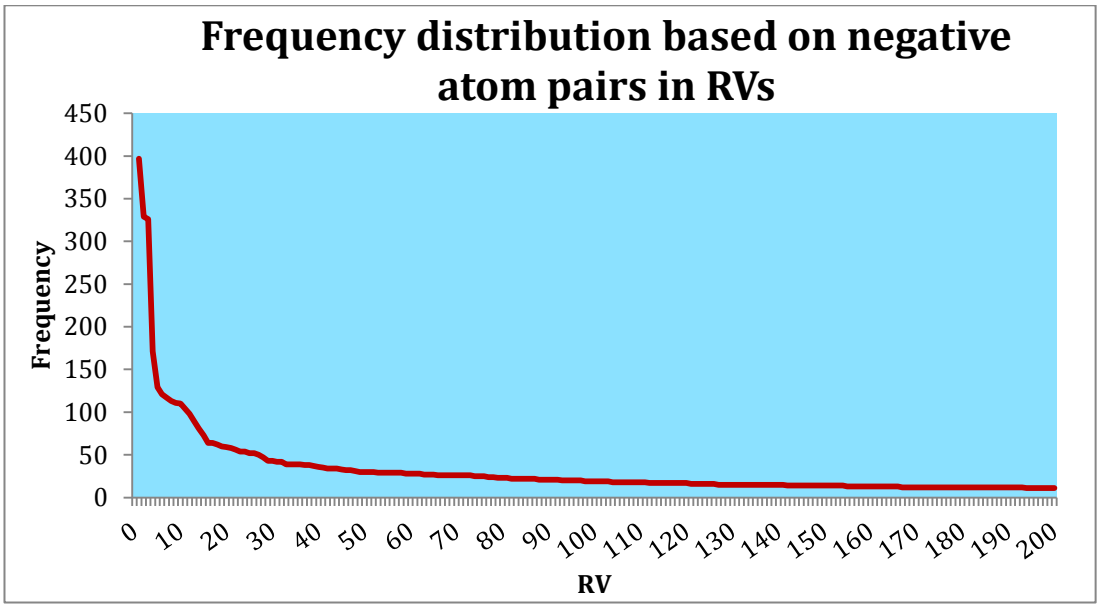


Figure A-2: Expansion of the first 200 entries in Figure A-1.

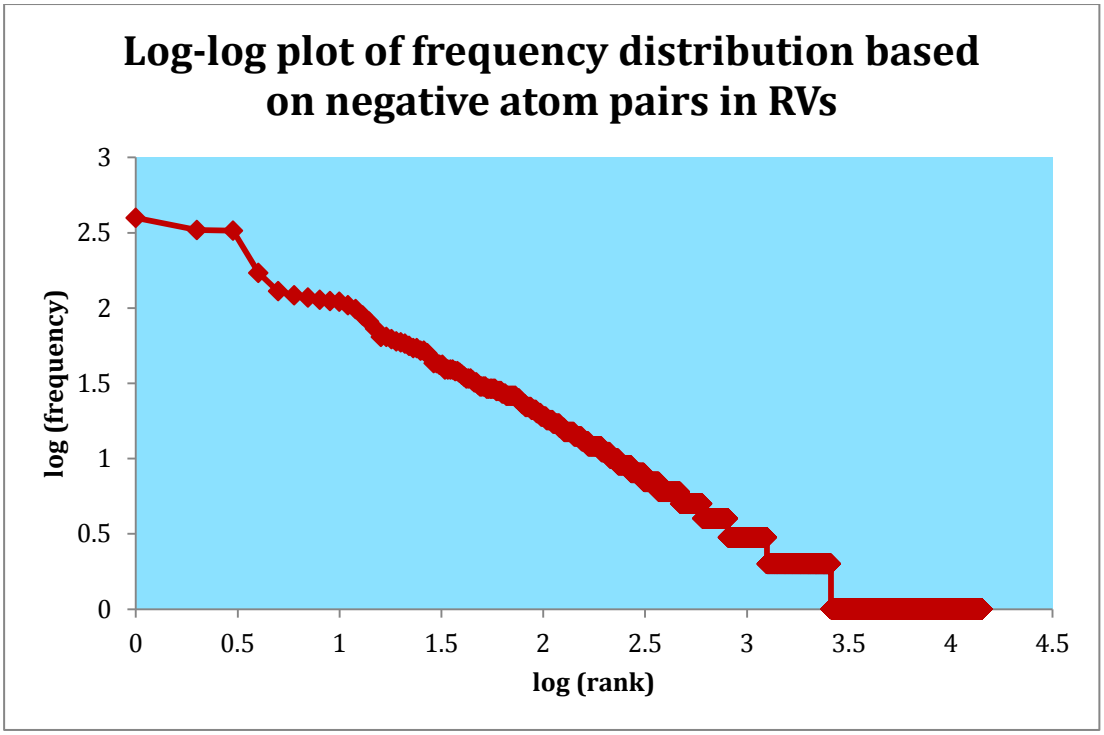


Figure A-3: Log-log plot of the frequency distribution of the negative AP2 content for the first random sample extracted from the US patent database.

A.2 Analysis of the second US patent set

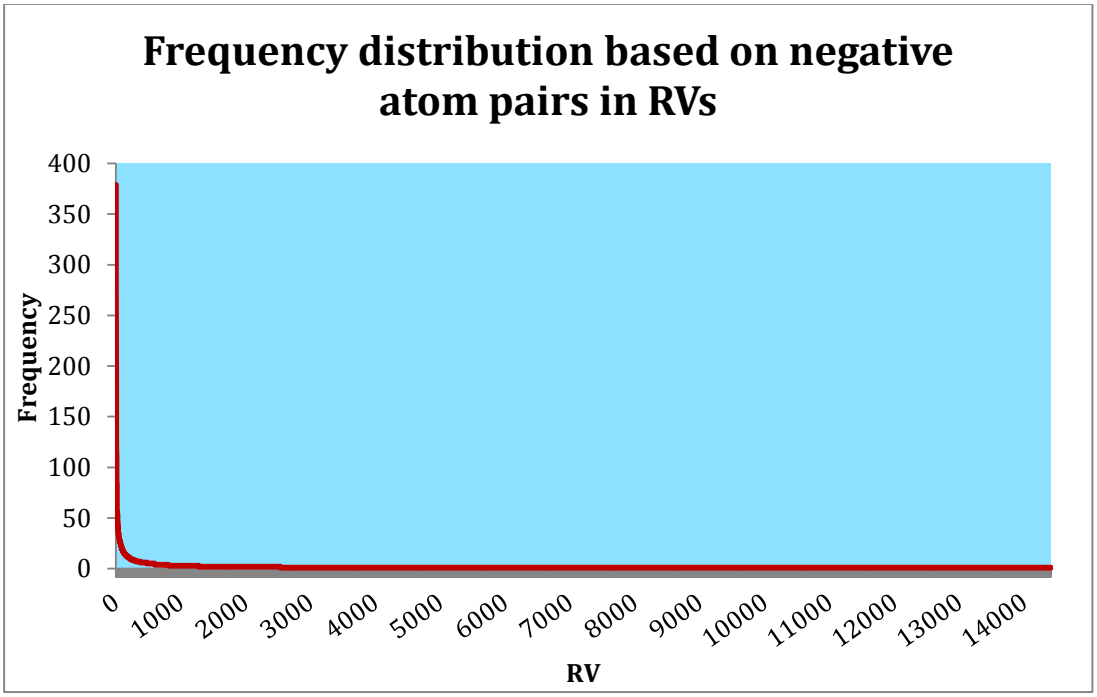


Figure A-4: Frequency distribution curve based on the negative AP2 content for the second random sample from the US patent database.

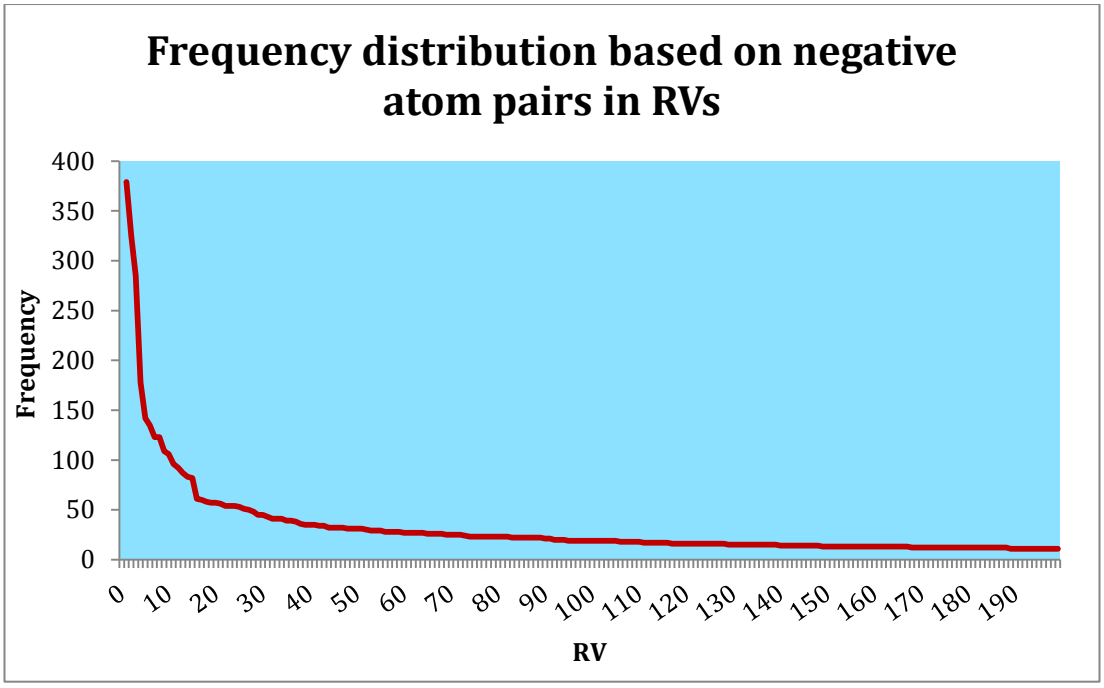


Figure A-5: Expansion of the first 200 entries in Figure A-4.

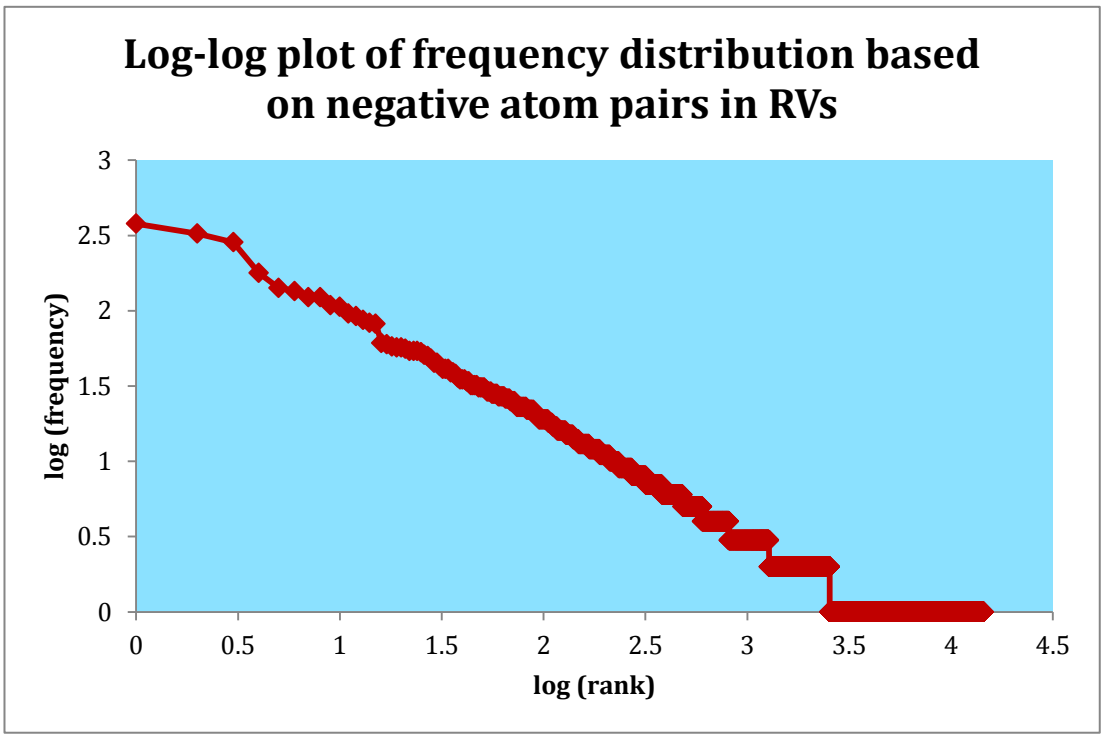


Figure A-6: Log-log plot of the frequency distribution of the negative AP2 content for the second random sample extracted from the US patent database.

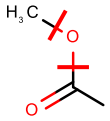
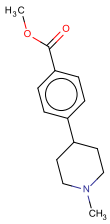
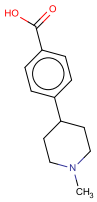
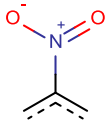
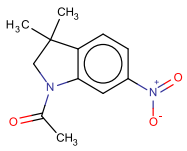
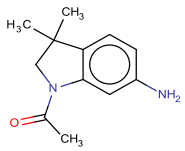
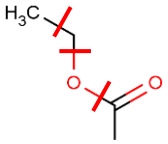
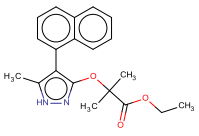
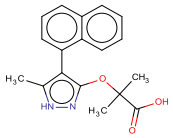
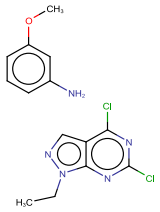
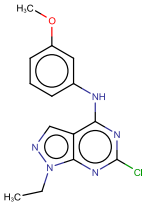
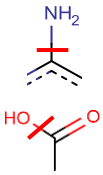
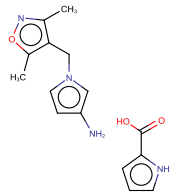
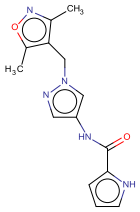
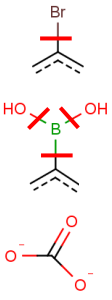
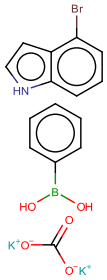
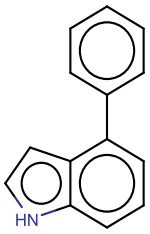
Negative pair grouping (duplicates indicate multiple entries)	Number of reactions represented	Reaction centre structure	Sample reactant(s)	Sample product
O(2,0,0)-2(1)-C(1,0,0) O(2,0,0)-2(1)-C(3,1,0)	379			
N(3,1,0)-2(1)-C(3,2,1) O(1,0,0)-2(1)-N(3,1,0) O(1,1,0)-2(2)-N(3,1,0)	325			
C(2,0,0)-2(1)-C(1,0,0) O(2,0,0)-2(1)-C(2,0,0) O(2,0,0)-2(1)-C(3,1,0)	285			
Cl(1,0,0)-2(1)-C(3,2,1) N(1,0,0)-2(1)-C(3,2,1)	178			
N(1,0,0)-2(1)-C(3,2,1) O(1,0,0)-2(1)-C(3,1,0)	142			
Br(1,0,0)-2(1)-C(3,2,1) C(3,2,1)-2(1)-B(3,0,0) O(1,0,0)-2(1)-B(3,0,0) O(1,0,0)-2(1)-B(3,0,0) O(1,0,0)-2(1)-C(3,1,0) O(1,0,0)-2(1)-C(3,1,0) O(1,1,0)-2(2)-C(3,1,0)	135			

Table A-1: Representation of the five largest groups of partial AP2 RVs in the second random sample from the US patent database. Where shown, the red lines indicate broken bonds in the reaction centre structure where ambiguity exists.(Wallace, 2015)

The last example in the above table has a more complex partial RV than the others, due to the presence of the carbonate group in the reaction centre in addition to the

standard Suzuki coupling reagents. In the patent database, the carbonate is specified as a reagent rather than a catalytic agent, and as such it is defined as being 'lost' over the course of the reaction. There are a number of examples of partial RVs for which only one reaction exists in the data set (seen in Figure A-7). It is at this point that the main differences between the data sets are noticed. There are considerably fewer examples of sulphate chemistry in this collection of singletons and over the set in general. Instead, the lowest represented reaction centres contain metal complexes that make the reaction centres unique. As a result, it is unlikely that these examples will be particularly relevant to *de novo* design, however.

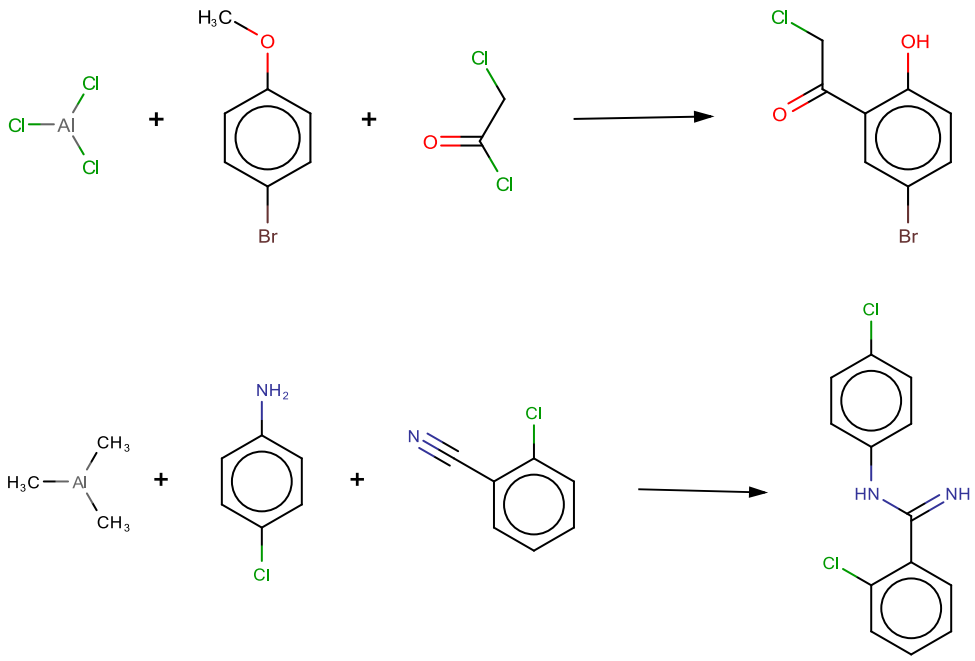


Figure A-7: Examples of reaction centres for which only single examples exist in the second US patent database. (Wallace, 2015)

A.3 Comparison of the US patent sets

To analyse the two data sets in more detail, it is possible to use the RV content to make direct comparisons between the atom pair groupings. By determining the overlap between the two data sets, it is possible to ascertain how similar they are overall. Depending on whether the comparison is made using the negative AP2 and AP3 groupings together, or using the negative AP2 grouping alone, differing degrees of overlap can be identified. For the AP2 and AP3 combined case, 3,759 of the 18,668 negative atom pair groups in the first set were present in the second set, seemingly suggesting an overlap of around 20% between the two collections. To confirm this is

not an artefact of the additional environment data provided via the AP3s, a comparison using AP2 content alone was performed. In this case, 3,686 out of 14,237 of these groups are present in both sets, suggesting 25.9% of the reaction centres are common between them. These groups coincide with the most popular groupings of negative atom pairs identified previously. Conversely, the areas of the sets with the lowest overlap are the groups that only contain one or two examples, representing more obscure chemistry, or unusual leaving groups such as metal atoms. Considering the nature of the patent collection, the total amount of data sampled in these collections is too small to be truly representative of the set as a whole. In addition, as the nature of a patented set of reactions is towards diversity due to the need to preserve exclusivity, the chance of finding sufficiently similar molecules to network together is smaller than for other databases of equivalent size.

Appendix B:

PCA descriptors

In the PCA plots used in this thesis, a number of topological and geometric descriptors were used to generate the components. These are listed below. Further information on these descriptors can be found in the QSAR.sf.net Descriptor Dictionary (Floris et al.).

B.4 Topological descriptors

Descriptor name	Description	Additional References
Moreau-Broto Autocorrelation descriptors	Measure based on charge, molecular weight, polarizability.	(Hollas, 2003)
Carbon types	Carbon hybridisation states	
Carbon Hybridisation Ratio	Ratio of the different carbon states	
Kier and Hall cluster, chain path and kappa molecular shape indices	Categorise different aspects of molecular shape	(Kier and Hall, 1986) (Hall and Kier, 2007)
Kier and Hall SMARTS Descriptors	Substructure counting	(Hall and Kier, 1995)
Eccentric Connectivity index	Measure of the separation between individual atoms, and those the furthest distance away	(Sharma et al., 1997)
Petitjean Number	Alternative measure of distance and eccentricity	(Petitjean, 1992)
Murcko framework	Count of ring systems (one or more rings sharing an edge in the molecular graph)	(Bemis and Murcko, 1996)
Fragment Complexity	Defined as C=|B^2-A^2+A|+(H/100) Where: C = complexity, A = number of non-hydrogen atoms, B = number of bonds, H = number of heteroatoms	(Nilakantan et al., 2006)
Molecular edge descriptors for Carbon, Nitrogen and Oxygen	Relationship between atomic distance and the edges of the adjacency of the graph.	(Liu et al., 1998)
Topological Polar surface area	The surface sum over all polar atoms in the molecule	(Prasanna and Doerksen, 2009)

Descriptor name	Description	Additional References
VABC Volume Descriptor	Van der Waals volume prediction	(Zhao et al., 2003)
Vertex adjacency information	Calculated as $1 + log_2\ b$ where b is the number of bonds between heavy atoms.	
Weighted path descriptors	Indicator of molecular branching.	(Randić and Basak, 1999)
Wiener path number	Equal to half of the sum of all bond distance matrix entries.	(Wiener, 1947)
Wiener polarity number	Computed in the same way as the path number, but only entries with a value of 3 are counted.	(Behmaram et al., 2012)
Zagreb Index	Sum of the square of the atom degrees for all heavy atoms.	(Gutman and Trinajstić, 1972)

Table B-1: Descriptions of the topological descriptors used to make PCA plots.

B.5 Geometric descriptors

Descriptor name	Description	Additional References
Charged partial surface areas	29 features based on molecule surface areas, obtaining partial charges using the Gasteiger-Marsilli algorithm.	(Ertl et al., 2000)
Gravitational Index	Molecular weight distribution of the molecule	(Katritzky et al., 1996)
Petitjean Shape indices	Alternative measure of distance and eccentricity	(Petitjean, 1992)

Table B-2: Descriptions of the geometric descriptors used to make PCA plots.

Appendix C:

Multi-objective drug design

Each multi-objective experiment was performed as three separate runs at each tournament size, with the run showing the highest activity values recorded. The complete results are listed for each compound class below, along with the starting material used.

C.1 Ace inhibitors

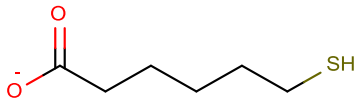


Figure C-1: Lowest molecular weight molecule in the Ace inhibitor set, used as starting material (pIC_{50} = 2.96, Bayesian score = 0.92). (Sutherland et al., 2004)

Sampling experiment	Number of products	Highest Bayesian score	Bayesian score range	MW range / g mol^{-1}	LogP range
RV (full enumeration)	816,925	23.95	-1.78 to 23.95	242 to 2026	-1.81 to 24.19
RV (druglike compounds)	1901	17.63	2.95 to 17.63	242 to 499	-0.21 to 5.00
Tournament size 5	200	8.33	-9.02 to 8.33	239 to 1284	0.36 to 14.70
	200	10.87	-8.32 to 10.87	236 to 1166	0.78 to 14.58
	200	8.58	-7.93 to 8.77	273 to 1207	0.96 to 16.25

Sampling experiment	Number of products	Highest Bayesian score	Bayesian score range	MW range / g mol^{-1}	LogP range
Tournament size 10	200	6.61	-8.50 to 6.61	281 to 1238	0.54 to 14.53
	200	7.30	-9.92 to 7.30	280 to 1158	1.45 to 16.28
	200	7.71	-8.05 to 7.71	239 to 1062	0.98 to 14.51
Tournament size 15	200	6.51	-8.50 to 6.51	268 to 1289	0.814 to 15.90
	200	5.78	-7.41 to 5.78	252 to 1162	0.39 to 13.93
	200	6.34	-8.21 to 6.34	290 to 1207	0.37 to 16.25
Tournament size 20	200	7.01	-8.83 to 7.01	286 to 1116	1.60 to 15.12
	200	5.15	-9.53 to 5.15	309 to 1265	1.24 to 15.90
	200	7.66	-8.29 to 7.66	240 to 958	0.67 to 12.18

Sampling experiment	Number of products	Highest Bayesian score	Bayesian score range	MW range / g mol^{-1}	LogP range
Tournament size 30	200	5.92	-10.04 to 5.92	239 to 1117	0.58 to 14.36
	200	4.66	-7.50 to 4.66	274 to 1118	1.24 to 14.44
	200	6.79	-7.92 to 6.79	240 to 1233	1.50 to 18.46
Tournament size 50	200	6.15	-9.14 to 6.15	284 to 1327	0.67 to 16.17
	200	5.34	0.01 to 5.34	339 to 1118	1.89 to 14.36
	200	5.25	-9.63 to 5.25	230 to 1118	0.67 to 14.36
Tournament size 75	200	5.89	-8.83 to 5.89	274 to 1059	1.28 to 13.73
	200	10.51	-9.63 to 10.51	254 to 1118	2.17 to 14.36
	200	6.15	-11.46 to 6.15	372 to 1247	1.39 to 14.72

Sampling experiment	Number of products	Highest Bayesian score	Bayesian score range	MW range / g mol^{-1}	LogP range
Tournament size 100	200	5.62	-8.71 to 5.62	232 to 1160	1.42 to 14.36
	200	5.82	-8.52 to 5.82	289 to 1117	1.45 to 14.36
	200	6.64	-8.71 to 6.64	260 to 1117	1.63 to 14.36
Tournament size 150	200	3.29	-9.63 to 3.29	232 to 858	1.72 to 9.03
	200	5.72	-9.63 to 5.72	260 to 858	0.94 to 9.03
	200	8.43	-8.71 to 8.43	260 to 810	1.63 to 9.69
Tournament size 175	200	5.62	-8.71 to 5.62	259 to 745	0.84 to 8.50
	200	5.92	-6.77 to 5.92	358 to 1117	2.57 to 14.36
	200	4.83	-8.71 to 4.83	260 to 840	0.75 to 8.50

Table C-1: Results of complete tournament selection runs for the first Ace inhibitor experiment.

Sampling experiment	Number of products	Highest Bayesian score	Bayesian score range	MW range / g mol^{-1}	LogP range
RV (full enumeration, 2 generations)	816,925	23.95	-1.78 to 23.95	242 to 2,026	-1.81 to 24.19
RV (drug-like compounds)	1901	17.63	2.95 to 17.63	242 to 499	-0.21 to 5.00
Tournament size 5	200	10.35	-6.54 to 10.35	204 to 1,211	1.15 to 16.74
	200	7.59	-1.68 to 7.59	281 to 1,011	0.61 to 14.23
	200	8.50	-7.37 to 8.50	239 to 1,197	1.30 to 19.33
Tournament size 10	200	7.41	-10.33 to 7.41	240 to 1,019	1.55 to 12.50
	200	7.25	-2.23 to 7.25	218 to 1,080	-1.00 to 12.33
	200	5.86	-7.12 to 5.86	307 to 499	0.67 to 4.82
Tournament size 15	200	6.71	-8.37 to 6.71	250 to 1,135	1.24 to 15.46
	200	7.36	-3.46 to 7.36	217 to 1,390	-0.03 to 12.38
	200	5.38	-10.06 to 5.38	255 to 1,174	0.89 to 14.88

Table C-2: Results of tournament selection runs for the Ace inhibitor experiment.

C.2 Bzr inhibitors

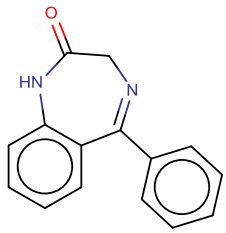


Figure C-2: Lowest molecular weight molecule in the Bzr inhibitor set, used as starting material (pIC_{50} = 6.46, Bayesian score = -4.82).(Sutherland et al., 2004)

Sampling experiment	Number of products	Highest Bayesian score	Bayesian score range	MW range / g mol^{-1}	LogP range
RV (full enumeration, 2 generations)	2,026,928	1.67	-13.80 to 1.67	251 to 1,900	-2.23 to 21.76
RV (drug-like compounds)	245,035	1.31	-12.24 to 1.31	251 to 499	-1.40 to 5.00
Tournament size 5	178	-1.43	-8.13 to -1.43	355 to 1,029	0.58 to 10.85
	170	-1.10	-8.98 to -1.10	362 to 953	1.16 to 12.19
	176	0.30	-8.94 to 0.30	322 to 1,027	0.15 to 11.77
Tournament size 10	178	-1.13	-9.76 to -1.13	348 to 1,173	-0.49 to 14.47
	192	-1.58	-8.63 to -1.58	318 to 1,249	0.91 to 12.31
	189	-1.95	-11.12 to -1.95	336 to 1,294	1.19 to 16.67
Tournament size 15	179	-2.29	-11.01 to -2.29	321 to 1,033	1.26 to 11.15
	181	-1.31	-9.63 to -1.31	322 to 1,180	0.39 to 13.58
	184	-0.40	-10.48 to -0.40	335 to 1,271	1.27 to 13.60

Table C-3: Results of tournament selection runs for the Bzr inhibitor experiment.

C.3 Cox-2 inhibitors

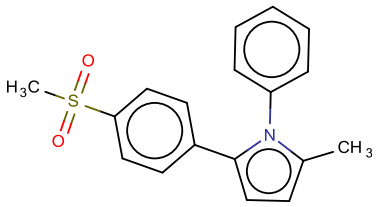


Figure C-3: Lowest molecular weight molecule in the Cox-2 inhibitor set, used as starting material (pIC_{50} = 7.22, Bayesian score = -1.77). (Sutherland et al., 2004)

Sampling experiment	Number of products	Highest Bayesian score	Bayesian score range	MW range / g mol^{-1}	LogP range
RV (full enumeration, 2 generations)	276,820	15.25	-4.17 to 15.25	317 to 1,959	-0.41 to 23.38
RV (drug-like compounds)	3,993	10.45	-1.79 to 10.45	317 to 499	1.36 to 5.00
Tournament size 5	126	-1.54	-9.22 to -1.54	430 to 1,142	3.14 to 14.22
	123	-1.94	-9.88 to -1.94	425 to 1,158	2.21 to 14.00
	121	-1.27	-10.03 to -1.27	447 to 1,058	2.86 to 11.73
Tournament size 10	125	-1.17	-11.43 to -1.17	403 to 1,107	3.31 to 13.40
	127	-3.27	-9.96 to -3.27	450 to 1,215	3.81 to 13.60
	127	-3.12	-10.27 to 3.12	431 to 1,227	0.68 to 13.61
Tournament size 15	122	-2.47	-11.37 to -2.47	430 to 1,173	2.42 to 12.24
	119	-2.58	-10.51 to -2.58	475 to 1,298	2.52 to 12.89
	125	-2.58	-11.57 to -2.58	418 to 1,127	2.68 to 12.32

Table C-4: Results of tournament selection runs for the Cox-2 inhibitor experiment.

C.4 Dhfr inhibitors

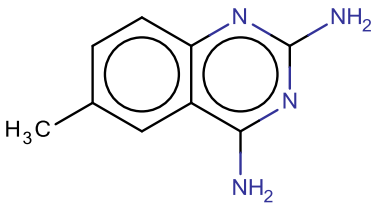


Figure C-4: Lowest molecular weight molecule in the Dhfr inhibitor set, used as starting material (pIC_{50} = 4.26, Bayesian score = -5.55). (Sutherland et al., 2004)

Sampling experiment	Number of products	Highest Bayesian score	Bayesian score range	MW range / g mol^{-1}	LogP range
RV (full enumeration, 2 generations)	234,283	3.83	-26.58 to 3.83	174 to 1,742	-2.32 to 17.23
RV (drug-like compounds)	60,590	3.47	-21.97 to 3.47	174 to 500	-2,32 to 5.00
Tournament size 5	171	-1.73	-22.45 to -1.73	188 to 1,721	0.61 to 13.06
	176	-1.64	-24.99 to -1.64	233 to 1,429	0.98 to 12.39
	181	-2.34	-22.61 to -2.34	299 to 1,010	1.72 to 12.00
Tournament size 10	174	-1.15	-23.57 to -1.15	309 to 897	1.56 to 11.69
	179	-0.81	-18.49 to -0.81	244 to 1,288	0.01 to 12.24
	181	-0.83	-23.01 to -0.83	325 to 1,054	1.72 to 10.99
Tournament size 15	176	-0.56	-19.32 to -0.56	305 to 1,143	1.77 to 12.18
	178	-0.51	-22.26 to -0.51	278 to 1,217	1.91 to 14.36
	192	-0.87	-21.11 to -0.87	294 to 1,013	1.03 to 10.74

Table C-5: Results of tournament selection runs for the Dhfr inhibitor experiment.

C.5 Gpb inhibitors

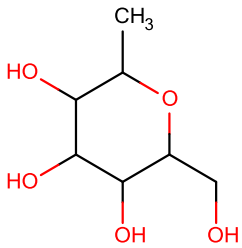


Figure C-5: Lowest molecular weight molecule in the Gpb inhibitor set, used as starting material (pIC_{50} = 1.3, Bayesian score = -0.55). (Sutherland et al., 2004)

Sampling experiment	Number of products	Highest Bayesian score	Bayesian score range	MW range / g mol^{-1}	LogP range
RV (full enumeration, 2 generations)	301,156	5.94	-2.02 to 5.94	164 to 2214	-8.56 to 19.51
RV (drug-like compounds)	43,846	4.86	-1.68 to 4.86	164 to 499	-5.44 to 5.00
Tournament size 5	181	1.32	-1.15 to 1.32	268 to 1,120	-0.69 to 8.38
	180	0.83	-1.33 to 0.84	276 to 1,088	-0.37 to 13.39
	190	1.85	-1.46 to 1.85	180 to 1,195	-1.17 to 11.11
Tournament size 10	182	1.59	-1.71 to 1.59	260 to 1,040	-0.11 to 9.51
	181	0.71	-1.26 to 0.71	178 to 820	-2.15 to 8.86
	177	1.37	-1.31 to 1.37	276 to 879	-1.11 to 9.07
Tournament size 15	185	2.18	-1.43 to 2.18	200 to 1,058	-2.04 to 10.30
	174	0.72	-1.40 to 0.72	180 to 1,006	-1.17 to 9.26
	183	0.85	-1.38 to 0.85	180 to 911	-1.17 to 9.60

Table C-6: Results of tournament selection runs for the Gpb inhibitor experiment.

C.6 Therm inhibitors

NOTE: Due to insufficient results being generated from the molecule with the lowest molecular weight, the molecule with the second lowest weight was used instead.

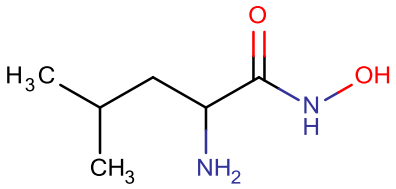


Figure C-6: Second lowest molecular weight molecule in the Thermolysin inhibitor set, used as starting material (pIC_{50} = 3.72, Bayesian score = 1.16). (Sutherland et al., 2004)

Sampling experiment	Number of products	Highest Bayesian score	Bayesian score range	MW range / g mol^{-1}	LogP range
RV (full enumeration, 2 generations)	154,414	7.62	-5.66 to 7.62	146 to 1,827	-3.48 to 19.37
RV (drug-like compounds)	45,237	5.72	-4.54 to 5.72	146 to 499	-3.04 to 5.00
Tournament size 5	194	3.86	-3.11 to 3.86	251 to 1,090	-0.18 to 9.90
	193	3.24	-3.23 to 3.24	279 to 1,138	-0.60 to 12.26
	197	3.21	-3.11 to 3.21	266 to 934	0.45 to 8.17
Tournament size 10	195	4.42	-5.65 to 4.42	296 to 1,310	-0.40 to 10.88
	195	3.63	-2.82 to 3.63	294 to 1,047	0.06 to 10.01
	194	3.34	-2.66 to 3.34	259 to 1,347	-0.87 to 12.24
Tournament size 15	194	3.62	-3.70 to 3.62	146 to 994	-1.19 to 9.21
	197	4.13	-5.27 to 4.13	311 to 1,005	-0.62 to 10.59
	192	4.74	-4.17 to 4.74	335 to 1,217	-0.31 to 9.74

Table C-7: Results of tournament selection runs for the Therm inhibitor experiment.

C.7 Statistical analysis of inhibitor runs

C.7.1 Ace inhibitors

Tournament size	Mean (Highest Bayesian score)	Standard Deviation (Highest Bayesian score)	Variance (Highest Bayesian score)
5	8.81	1.41	1.98
10	6.84	0.85	0.72
15	6.48	1.01	1.02

Table C-8: Summary of statistical analysis of the tournament selection method for the Ace inhibitor experiment.

Tournament size	Mean (Highest Bayesian score)	Standard Deviation (Highest Bayesian score)	Variance (Highest Bayesian score)
5	6.12	1.66	2.77
10	8.26	0.32	0.10
15	8.51	0.82	0.67

Table C-9: Summary of statistical analysis of the tournament selection method for the Ace inhibitor experiment (second starting material).

C.7.2 Bzr inhibitors

Tournament size	Mean (Highest Bayesian score)	Standard Deviation (Highest Bayesian score)	Variance (Highest Bayesian score)
5	-0.74	0.92	0.84
10	-1.55	0.41	0.17
15	-1.33	0.95	0.90

Table C-10: Summary of statistical analysis of the tournament selection method for the Bzr inhibitor experiment.

C.7.3 Cox inhibitors

Tournament size	Mean (Highest Bayesian score)	Standard Deviation (Highest Bayesian score)	Variance (Highest Bayesian score)
5	-1.58	0.34	0.11
10	-2.52	1.17	1.37
15	-2.54	0.06	0.00

Table C-11: Summary of statistical analysis of the tournament selection method for the Cox inhibitor experiment.

C.7.4 Dhfr inhibitors

Tournament size	Mean (Highest Bayesian score)	Standard Deviation (Highest Bayesian score)	Variance (Highest Bayesian score)
5	-1.90	0.38	0.15
10	-0.93	0.19	0.04
15	-0.65	0.20	0.04

Table C-12: Summary of statistical analysis of the tournament selection method for the Dhfr inhibitor experiment.

C.7.5 Gpb inhibitors

Tournament size	Mean (Highest Bayesian score)	Standard Deviation (Highest Bayesian score)	Variance (Highest Bayesian score)
5	-0.65	0.20	0.04
10	1.22	0.46	0.21
15	1.25	0.81	0.65

Table C-13: Summary of statistical analysis of the tournament selection method for the Gpb inhibitor experiment.

C.7.6 Therm inhibitors

Tournament size	Mean (Highest Bayesian score)	Standard Deviation (Highest Bayesian score)	Variance (Highest Bayesian score)
5	3.44	0.37	0.13
10	3.80	0.56	0.31
15	4.16	0.56	0.31

Table C-14: Summary of statistical analysis of the tournament selection method for the Therm inhibitor experiment.

Bibliography

ABDI, H. & WILLIAMS, L. J. 2010. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics,* 2**,** 433-459.

ADAMIC, L. 2011. Complex systems: Unzipping Zipf's law. *Nature,* 474**,** 164-165.

ALLU, T. K. & OPREA, T. I. 2005. Rapid Evaluation of Synthetic and Molecular Complexity for in Silico Chemistry. *Journal of Chemical Information and Modeling,* 45**,** 1237-1243.

APOSTOLAKIS, J., SACHER, O., KÖRNER, R. & GASTEIGER, J. 2008. Automatic Determination of Reaction Mappings and Reaction Center Information. 2. Validation on a Biochemical Reaction Database. *Journal of Chemical Information and Modeling,* 48**,** 1190-1198.

BACK, T., HAMMEL, U. & SCHWEFEL, H. P. 1997. Evolutionary Computation: Comments on the History and Current State. *IEEE Transactions on Evolutionary Computation,* 1**,** 3-17.

BAILEY, S., FISH, P. V., BILLOTTE, S., BORDNER, J., GREILING, D., JAMES, K., MCELROY, A., MILLS, J. E., REED, C. & WEBSTER, R. 2008. Succinyl Hydroxamates as Potent and Selective Non-peptidic Inhibitors of Procollagen C-proteinase: Design, Synthesis, and Evaluation as Topically Applied, Dermal Anti-scarring Agents. *Bioorganic & Medicinal Chemistry Letters,* 18**,** 6562-6567.

BAKER, D. J., TIMMONS, J. A. & GREENHAFF, P. L. 2005. Glycogen Phosphorylase Inhibition in Type 2 Diabetes Therapy: A Systematic Evaluation of Metabolic and Functional Effects in Rat Skeletal Muscle. *Diabetes,* 54**,** 2453-2459.

BARNARD, J. M. & DOWNS, G. M. 1992. Clustering of Chemical Structures on the Basis of Two-Dimensional Similarity Measures. *Journal of Chemical Information and Computer Sciences,* 32**,** 644-649.

BART, J. C. J. & GARAGNANI, E. 1977. Organic reaction schemes and general reaction-matrix types, II. Basic types of synthetic transformations. *Zeitschrift für Naturforschung B. A Journal of Chemical Sciences,* 32B**,** 465-468.

BASARAB, G. S., HILL, P. J., RASTAGAR, A. & WEBBORN, P. J. H. 2008. Design of Helicobacter Pylori Glutamate Racemase Inhibitors as Selective Antibacterial Agents: A Novel Pro-drug Approach to Increase Exposure. *Bioorganic & Medicinal Chemistry Letters,* 18**,** 4716-4722.

BEHMARAM, A., YOUSEFI-AZARI, H. & ASHRAFI, A. R. 2012. Wiener polarity index of fullerenes and hexagonal systems. *Applied Mathematics Letters,* 25**,** 1510-1513.

BEMIS, G. W. & MURCKO, M. A. 1996. The Properties of Known Drugs. 1. Molecular Frameworks. *Journal of Medicinal Chemistry,* 39**,** 2887-2893.

BERTHOLD, M. R., CEBRON, N., DILL, F., GABRIEL, T. R., KÖTTER, T., MEINL, T., OHL, P., SIEB, C., THIEL, K. & WISWEDEL, B. 2008. KNIME: The Konstanz Information Miner. In: *Data Analysis, Machine Learning and Applications, PREISACH, C., BURKHARDT, H., SCHMIDT-THIEME, L. & DECKER, R. (eds.)*, 319-326: Springer Berlin Heidelberg.

BESSA BELMUNT, J. 2008. *Process for Preparing an Angiotensin II Receptor Antagonist*. United States patent application 20080281097.

BIOVIA. *BIOVIA Available Chemicals Directory* [Online]. Available: http://accelrys.com/products/collaborative-science/databases/sourcing-databases/biovia-available-chemicals-directory.html [Accessed October 2015].

BIOVIA. *BIOVIA DiscoveryGate* [Online]. Available: http://accelrys.com/products/collaborative-science/databases/database-access/biovia-discoverygate.html [Accessed October 2015].

BIOVIA. *BIOVIA Draw*. BIOVIA. http://accelrys.com/products/collaborative-science/biovia-draw/ [Accessed October 2015].

BISHOP, K. J. M., KLAJN, R. & GRZYBOWSKI, B. A. 2006. The Core and Most Useful Molecules in Organic Chemistry. *Angewandte Chemie International Edition,* 45**,** 5348-5354.

BLAKE, J. E. & DANA, R. C. 1990. CASREACT: More Than a Million Reactions. *Journal of Chemical Information and Computer Sciences,* 30**,** 394-399.

BODA, K., SEIDEL, T. & GASTEIGER, J. 2007. Structure and reaction based evaluation of synthetic accessibility. *Journal of Computer-Aided Molecular Design,* 21**,** 311-325.

BOHACEK, R. S. & MCMARTIN, C. 1994. Multiple Highly Diverse Structures Complementary to Enzyme Binding Sites: Results of Extensive Application of a de Novo Design Method Incorporating Combinatorial Growth. *Journal of the American Chemical Society,* 116**,** 5560-5571.

BÖHM, H.-J. 1992. The Computer Program LUDI: A New Method for the de Novo Design of Enzyme Inhibitors. *Journal of Computer-Aided Molecular Design,* 6**,** 61-78.

BOITEN, J.-W., OTT, M. A. & NOORDIK, J. H. 1995. Automated Overlap Analysis of Reaction Databases. *Journal of Chemical Information and Computer Sciences,* 35**,** 115-120.

BOSTON UNIVERSITY. *CMLD - Synthesis Protocols* [Online]. Available: http://cmldprotocols.bu.edu/cmld/index.jsp [Accessed October 2015].

BROUGHTON, H. B., HUNT, P. A. & MACKEY, M. D. 2003. *Methods for Classifying and Searching Chemical Reactions*, US patent application US2003/0182094 A1.

BROWN, N., MCKAY, B., GILARDONI, F. & GASTEIGER, J. 2004. A Graph-Based Genetic Algorithm and Its Application to the Multiobjective Evolution of Median Molecules. *Journal of Chemical Information and Computer Sciences,* 44**,** 1079-1087.

BURGEY, C. S., STUMP, C. A., NGUYEN, D. N., DENG, J. Z., QUIGLEY, A. G., NORTON, B. R., BELL, I. M., MOSSER, S. D., SALVATORE, C. A., RUTLEDGE, R. Z., KANE, S. A., KOBLAN, K. S., VACCA, J. P., GRAHAM, S. L. & WILLIAMS, T. M. 2006. Benzodiazepine calcitonin gene-related peptide (CGRP) receptor antagonists: Optimization of the 4-substituted piperidine. *Bioorganic & Medicinal Chemistry Letters,* 16**,** 5052-5056.

CAREY, J. S., LAFFAN, D., THOMSON, C. & WILLIAMS, M. T. 2006. Analysis of the Reactions Used for the Preparation of Drug Candidate Molecules. *Organic & Biomolecular Chemistry,* 4**,** 2337-2347.

CARHART, R. E., SMITH, D. H. & VENKATARAGHAVAN, R. 1985. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences,* 25**,** 64-73.

CAVASOTTO, C. N. & PHATAK, S. S. 2011. Docking Methods for Structure-Based Library Design. In: *Chemical Library Design, ZHOU, J. Z. (ed.)*, 155-174: Humana Press.

CHEMAXON. *Marvin*. ChemAxon. http://www.chemaxon.com/products/marvin/ [Accessed October 2015].

CHEMICAL ABSTRACT SERVICES. *CAS Registry* [Online]. Available: http://www.cas.org/expertise/cascontent/registry/regsys.html [Accessed October 2015].

CHEMICAL ABSTRACT SERVICES. *CASREACT* [Online]. Available: http://www.cas.org/content/reactions [Accessed October 2015].

CHEMICAL ABSTRACT SERVICES. 2011. *CASREACT* [Online]. Chemical Abstract Services. Available: http://www.cas.org/expertise/cascontent/casreact.html [Accessed 5th December 2011].

CHEMICAL COMPUTING GROUP INC. 2015. *Molecular Operating Environment (MOE)*. Chemical Computing Group Inc., https://www.chemcomp.com/MOE-Molecular_Operating_Environment.htm [Accessed October 2015].

CHEN, L. 2008. Reaction Classification and Knowledge Acquisition. In: *Handbook of Chemoinformatics*, 348-390: Wiley-VCH Verlag GmbH.

CHEN, W. L., CHEN, D. Z. & TAYLOR, K. T. 2013. Automatic Reaction Mapping and Reaction Center Detection. *Wiley Interdisciplinary Reviews: Computational Molecular Science,* 3**,** 560-593.

CHERKASSKY, B., GOLDBERG, A. & RADZIK, T. 1996. Shortest paths algorithms: Theory and experimental evaluation. *Mathematical Programming,* 73**,** 129-174.

CHEUNG, M., KUNTZ, K. W., POBANZ, M., SALOVICH, J. M., WILSON, B. J., ANDREWS III, C. W., SHEWCHUK, L. M., EPPERLY, A. H., HASSLER, D. F., LEESNITZER, M. A., SMITH, J. L., SMITH, G. K., LANSING, T. J. & MOOK JR, R. A. 2008. Imidazo[5,1-f][1,2,4]triazin-2-amines as Novel Inhibitors of Polo-like Kinase 1. *Bioorganic & Medicinal Chemistry Letters,* 18**,** 6214-6217.

CLARK, D. E., FRENKEL, D., LEVY, S. A., LI, J., MURRAY, C. W., ROBSON, B., WASZKOWYCZ, B. & WESTHEAD, D. R. 1995. PRO_LIGAND: An Approach to de Novo Molecular Design. 1. Application to the Design of Organic Molecules. *Journal of Computer-Aided Molecular Design,* 9**,** 13-32.

CONGREVE, M., CARR, R., MURRAY, C. & JHOTI, H. 2003. A 'Rule of Three' for fragment-based lead discovery? *Drug Discovery Today,* 8**,** 876-877.

COOK, A., JOHNSON, A. P., LAW, J., MIRZAZADEH, M., RAVITZ, O. & SIMON, A. 2012. Computer-aided Synthesis Design: 40 Years on. *Wiley Interdisciplinary Reviews: Computational Molecular Science,* 2**,** 79-107.

COREY, E. J. & JORGENSEN, W. L. 1976. Computer-assisted Synthetic Analysis. Generation of Synthetic Sequences Involving Sequential Functional Group Interchanges. *Journal of the American Chemical Society,* 98**,** 203-209.

COREY, E. J., WIPKE, W. T., CRAMER, R. D. & HOWE, W. J. 1972. Computer-assisted Synthetic analysis. Facile Man-machine Communication of Chemical Structure by Interactive Computer Graphics. *Journal of the American Chemical Society,* 94**,** 421-430.

CORTES, C. & VAPNIK, V. 1995. Support-vector networks. *Machine Learning,* 20**,** 273-297.

CRABTREE, J. & MEHTA, D. 2009. Automated Reaction Mapping. *Journal of Experimental Algorithmics,* 13,1.15**,** 2-29.

CUSHMAN, D. W. & ONDETTI, M. A. 1991. History of the design of captopril and related inhibitors of angiotensin converting enzyme. *Hypertension,* 17**,** 589-592.

DANZIGER, D. J. & DEAN, P. M. 1989. Automated Site-Directed Drug Design: A General Algorithm for Knowledge Acquisition about Hydrogen-Bonding Regions at Protein Surfaces. *Proceedings of the Royal Society of London. B. Biological Sciences,* 236**,** 101-113.

DAYLIGHT CHEMICAL INFORMATION SYSTEMS, INC. *Daylight Theory Manual* [Online]. Available: http://www.daylight.com/dayhtml/doc/theory/index.html [Accessed October 2015].

DE LUCA, A., HORVATH, D., MARCOU, G., SOLOV'EV, V. & VARNEK, A. 2012. Mining Chemical Reactions Using Neighborhood Behavior and Condensed Graphs of Reactions Approaches. *Journal of Chemical Information and Modeling,* 52**,** 2325-2338.

DEGEN, J. & RAREY, M. 2006. FlexNovo: Structure-Based Searching in Large Fragment Spaces. *ChemMedChem,* 1**,** 854-868.

DESJARLAIS, R. L., SHERIDAN, R. P., SEIBEL, G. L., DIXON, J. S., KUNTZ, I. D. & VENKATARAGHAVAN, R. 1988. Using Shape Complementarity as an Initial Screen in Designing Ligands for a Receptor Binding Site of Known Three-Dimensional Structure. *Journal of Medicinal Chemistry,* 31**,** 722-729.

DEWITTE, R. S. & SHAKHNOVICH, E. I. 1996. SMoG: De Novo Design Method Based on Simple, Fast, and Accurate Free Energy Estimates. 1. Methodology and Supporting Evidence. *Journal of the American Chemical Society,* 118**,** 11733-11744.

DEY, F. & CAFLISCH, A. 2008. Fragment-Based de Novo Ligand Design by Multiobjective Evolutionary Optimization. *Journal of Chemical Information and Modeling,* 48**,** 679-690.

DIAS, R. & FILGUEIRA DE AZEVEDO JR., W. 2008. Molecular Docking Algorithms. *Current Drug Targets,* 9**,** 1040-1047.

DIJKSTRA, E. W. 1959. A note on two problems in connexion with graphs. *Numerische Mathematik,* 1**,** 269-271.

DISTRIBUTED CHEMICAL GRAPHICS, INC. *SYNLIB (Synthesis Library)*. Meadowbrook: Distributed Chemical Graphics, Inc.

DITTMAR, P. G., FARMER, N. A., FISANICK, W., HAINES, R. C. & MOCKUS, J. 1983. The CAS ONLINE Search System. 1. General System Design and Selection, Generation, and Use of Search Screens. *Journal of Chemical Information and Computer Sciences,* 23**,** 93-102.

DORSCH, D., JURASZYK, H., WURZIGER, H., BERNOTAT-DANIELOWSKI, S. & MELZER, G. 1999. *Benzamidine derivatives as factor Xa inhibitors*. World Intellectual Property Organization Patent WO/1999/016751.

DUGUNDJI, J. & UGI, I. 1973. An Algebraic Model of Constitutional Chemistry as a Basis for Chemical Computer Programs. In: *Computers in Chemistry*, 19-64: Springer Berlin / Heidelberg.

EISEN, M. B., WILEY, D. C., KARPLUS, M. & HUBBARD, R. E. 1994. HOOK: A Program for Finding Novel Molecular Architectures that Satisfy the Chemical and Steric Requirements of a Macromolecule Binding Site. *Proteins: Structure, Function, and Bioinformatics,* 19**,** 199-221.

EKINS, S., HONEYCUTT, J. D. & METZ, J. T. 2010. Evolving molecules using multi-objective optimization: applying to ADME/Tox. *Drug Discovery Today,* 15**,** 451-460.

ELSEVIER. *Reaxys* [Online]. Available: http://www.reaxys.com/info [Accessed October 2015].

EPAM LIFE SCIENCES. *Indigo*. http://lifescience.opensource.epam.com/indigo/ [Accessed October 2015].

ERTL, P., ROHDE, 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 Properties. *Journal of Medicinal Chemistry,* 43**,** 3714-3717.

FAULON, J.-L., COLLINS, M. J. & CARR, R. D. 2004. The Signature Molecular Descriptor. 4. Canonizing Molecules Using Extended Valence Sequences. *Journal of Chemical Information and Computer Sciences,* 44**,** 427-436.

FAYNE, D. 2013. Ligand-Based Molecular Design Using Pseudoreceptors. In: *De novo Molecular Design*, *SCHNEIDER, G., (ed)*, 227-244: Wiley-VCH Verlag GmbH & Co. KGaA.

FECHNER, U. & SCHNEIDER, G. 2005. Flux (1): A Virtual Synthesis Scheme for Fragment-Based de Novo Design. *Journal of Chemical Information and Modeling,* 46**,** 699-707.

FECHNER, U. & SCHNEIDER, G. 2007. Flux (2): Comparison of Molecular Mutation and Crossover Operators for Ligand-Based de Novo Design. *Journal of Chemical Information and Modeling,* 47**,** 656-667.

FEHER, M., GAO, Y., BABER, J. C., SHIRLEY, W. A. & SAUNDERS, J. 2008. The use of ligand-based de novo design for scaffold hopping and sidechain optimization: Two case studies. *Bioorganic & Medicinal Chemistry,* 16**,** 422-427.

FIALKOWSKI, M., BISHOP, K. J. M., CHUBUKOV, V. A., CAMPBELL, C. J. & GRZYBOWSKI, B. A. 2005. Architecture and Evolution of Organic Chemistry. *Angewandte Chemie International Edition,* 44**,** 7263-7269.

FLORIS, M., WILLIGHAGEN, E., GUHA, R., ROJAS, M. & HOPPE, C. *QSAR.sf.net Descriptor Dictionary* [Online]. Available: http://qsar.sourceforge.net/dicts/qsar-descriptors/index.xhtml [Accessed October 2015].

FOSCATO, M., OCCHIPINTI, G., VENKATRAMAN, V., ALSBERG, B. K. & JENSEN, V. R. 2014. Automated Design of Realistic Organometallic Molecules from Fragments. *Journal of Chemical Information and Modeling,* 54**,** 767-780.

FREELAND, R. G., FUNK, S. A., O'KORN, L. J. & WILSON, G. A. 1979. The Chemical Abstracts Service Chemical Registry System. II. Augmented Connectivity Molecular Formula. *Journal of Chemical Information and Computer Sciences,* 19**,** 94-98.

FULLER, P. E., GOTHARD, C. M., GOTHARD, N. A., WECKIEWICZ, A. & GRZYBOWSKI, B. A. 2012. Chemical Network Algorithms for the Risk Assessment and Management of Chemical Threats. *Angewandte Chemie International Edition,* 51**,** 7933-7937.

FUNATSU, K., ENDO, T., KOTERA, N. & SASAKI, S.-I. 1988. Automatic Recognition of Reaction Site in Organic Chemical Reactions. *Tetrahedron Computer Methodology,* 1**,** 53-69.

GANGJEE, A., YU, J., KISLIUK, R. L., HAILE, W. H., SOBRERO, G. & MCGUIRE, J. J. 2003. Design, Synthesis, and Biological Activities of Classical N-{4-[2-(2-Amino-4-ethylpyrrolo[2,3-d]pyrimidin-5-yl)ethyl]benzoyl}-l-glutamic Acid and Its 6-Methyl Derivative as Potential Dual Inhibitors of Thymidylate Synthase and Dihydrofolate Reductase and as Potential Antitumor Agents1. *Journal of Medicinal Chemistry,* 46**,** 591-600.

GARCÍA-DOMENECH, R., GÁLVEZ, J., DE JULIÁN-ORTIZ, J. V. & POGLIANI, L. 2008. Some New Trends in Chemical Graph Theory. *Chemical Reviews,* 108**,** 1127-1169.

GASTEIGER, J. & JOCHUM, C. 1978. EROS: A Computer Program for Generating Sequences of Reactions. In: *Organic Compounds*, 93-126: Springer Berlin / Heidelberg.

GASTEIGER, J., PFÖRTNER, M., SITZMANN, M., HÖLLERING, R., SACHER, O., KOSTKA, T. & KARG, N. 2000. Computer-assisted synthesis and reaction planning in combinatorial chemistry. *Perspectives in Drug Discovery and Design,* 20**,** 245-264.

GEHLHAAR, D. K., MOERDER, K. E., ZICHI, D., SHERMAN, C. J., OGDEN, R. C. & FREER, S. T. 1995. De Novo Design of Enzyme Inhibitors by Monte Carlo Ligand Generation. *Journal of Medicinal Chemistry,* 38**,** 466-472.

GHOSE, A. K., VISWANADHAN, V. N. & WENDOLOSKI, J. J. 1999. A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases. *Journal of Combinatorial Chemistry,* 1**,** 55-68.

GILLET, V., JOHNSON, A. P., MATA, P., SIKE, S. & WILLIAMS, P. 1993. SPROUT: A Program for Structure Generation. *Journal of Computer-Aided Molecular Design,* 7**,** 127-153.

GILLET, V., MYATT, G., ZSOLDOS, Z. & JOHNSON, A. 1995. SPROUT, HIPPO and CAESA: Tools for de Novo Structure Generation and Estimation of Synthetic Accessibility. *Perspectives in Drug Discovery and Design,* 3**,** 34-50.

GILLET, V. J., ALLEN, B., BODKIN, M., CHEN, B., COLE, J., HRISTOZOV, D., LIEBESCHUETZ, J. & PATEL, H. 2012. De novo design of synthetically accessible compounds: Application to fragment-based drug design. In: *243rd ACS National Meeting*, *March 25-29, 2012*. San Diego, California, USA. American Chemical Society, COMP-211.

GILLET, V. J., BODKIN, M. J. & HRISTOZOV, D. 2014. Multiobjective de novo design of synthetically accessible compounds. In: *De Novo Molecular Design*, 267-285: Wiley-VCH Verlag GmbH & Co. KGaA.

GILLET, V. J., PATEL, H., BODKIN, M. & CHEN, B. 2009. De novo design using reaction vectors: Application to library design. In: *237th ACS National Meeting*, *March 22-26, 2009.* Salt Lake City, Utah, USA. American Chemical Society, CINF-024.

GLEN, R. C. 2011. Connecting the Virtual World of Computers to the Real World of Medicinal Chemistry. *Future Medicinal Chemistry,* 3**,** 399-403.

GLEN, R. C. & PAYNE, A. W. R. 1995. A Genetic Algorithm for the Automated Generation of Molecules Within Constraints. *Journal of Computer-Aided Molecular Design,* 9**,** 181-202.

GLOBUS, A., LAWTON, J. & WIPKE, T. 1999. Automatic Molecular Design using Evolutionary Techniques. *Nanotechnology,* 10**,** 290-290.

GOODFORD, P. J. 1985. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *Journal of Medicinal Chemistry,* 28**,** 849-857.

GOTHARD, C. M., SOH, S., GOTHARD, N. A., KOWALCZYK, B., WEI, Y., BAYTEKIN, B. & GRZYBOWSKI, B. A. 2012. Rewiring Chemistry: Algorithmic Discovery and Experimental Validation of One-Pot Reactions in the Network of Organic Chemistry. *Angewandte Chemie International Edition,* 51**,** 7922-7927.

GREEN, G. A. 2001. Understanding NSAIDs: From aspirin to COX-2. *Clinical Cornerstone,* 3**,** 50-59.

GRETHE, G. & MOOCK, T. E. 1990. Similarity Searching in REACCS. A New Tool for the Synthetic Chemist. *Journal of Chemical Information and Computer Sciences,* 30**,** 511-520.

GRZYBOWSKI, B. A., BISHOP, K. J. M., KOWALCZYK, B. & WILMER, C. E. 2009. The 'wired' universe of organic chemistry. *Nature Chemistry,* 1**,** 31-36.

GUTLEIN, M., KARWATH, A. & KRAMER, S. 2012. CheS-Mapper - Chemical Space Mapping and Visualization in 3D. *Journal of Cheminformatics,* 4**,** 7.

GUTMAN, I. & TRINAJSTIĆ, N. 1972. Graph theory and molecular orbitals. Total ϕ-electron energy of alternant hydrocarbons. *Chemical Physics Letters,* 17**,** 535-538.

HALL, L. H. & KIER, L. B. 1995. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *Journal of Chemical Information and Computer Sciences,* 35**,** 1039-1045.

HALL, L. H. & KIER, L. B. 2007. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling. In: *Reviews in Computational Chemistry, Volume 2, LIPKOWITZ, K. B. & BOYD, D. B. (eds),* 367-422: John Wiley & Sons, Inc.

HAN, Y. & STEINBECK, C. 2004. Evolutionary-Algorithm-Based Strategy for Computer-Assisted Structure Elucidation. *Journal of Chemical Information and Computer Sciences,* 44**,** 489-498.

HANSCH, C. & FUJITA, T. 1964. p-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *Journal of the American Chemical Society,* 86**,** 1616-1626.

HART, P. E., NILSSON, N. J. & RAPHAEL, B. 1968. A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Transactions on Systems Science and Cybernetics,* 4**,** 100-107.

HARTENFELLER, M., PROSCHAK, E., SCHÜLLER, A. & SCHNEIDER, G. 2008. Concept of Combinatorial de Novo Design of Drug-like Molecules by Particle Swarm Optimization. *Chemical Biology & Drug Design,* 72**,** 16-26.

HARTENFELLER, M. & SCHNEIDER, G. 2011. Enabling Future Drug Discovery by de Novo Design. *Wiley Interdisciplinary Reviews: Computational Molecular Science,* 1**,** 742-759.

HARTENFELLER, M., ZETTL, H., WALTER, M., RUPP, M., REISEN, F., PROSCHAK, E., WEGGEN, S., STARK, H. & SCHNEIDER, G. 2012. DOGS: Reaction-Driven de Novo Design of Bioactive Compounds. *PLoS Computational Biology,* 8**,** e1002380.

HEER, J., CARD, S. K. & LANDAY, J. A. 2005. prefuse: a toolkit for interactive information visualization. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* Portland, Oregon, USA. ACM, 421-430.

HENDRICKSON, J. B. 1997a. The Application of Computers to Generate Organic Syntheses. *The Knowledge Engineering Review,* 12**,** 369-386.

HENDRICKSON, J. B. 1997b. Comprehensive System for Classification and Nomenclature of Organic Reactions. *Journal of Chemical Information and Computer Sciences,* 37**,** 852-860.

HENDRICKSON, J. B. & MILLER, T. M. 1990. Reaction Indexing for Reaction Databases. *Journal of Chemical Information and Computer Sciences,* 30**,** 403-408.

HERGES, R. & UGI, I. 1985. Synthesis of Seven-Membered Rings by [(σ2+π2)+π2] Cycloaddition to Homodienes. *Angewandte Chemie International Edition in English,* 24**,** 594-596.

HERT, J., WILLETT, P., WILTON, D. J., ACKLIN, P., AZZAOUI, K., JACOBY, E. & SCHUFFENHAUER, A. 2006. New Methods for Ligand-Based Virtual Screening: Use of Data Fusion and Machine Learning to Enhance the Effectiveness of Similarity Searching. *Journal of Chemical Information and Modeling,* 46**,** 462-470.

HICKS, M. G. 1990. Reactions in the Beilstein Information System: Nonaporic Organic Synthesis. *Journal of Chemical Information and Computer Sciences,* 30**,** 352-359.

HO, C. M. W. & MARSHALL, G. R. 1993a. FOUNDATION: A Program to Retrieve all Possible Structures Containing a User-defined Minimum Number of Matching Query Elements from Three-dimensional Databases. *Journal of Computer-Aided Molecular Design,* 7**,** 3-22.

HO, C. M. W. & MARSHALL, G. R. 1993b. SPLICE: A Program to Assemble Partial Query Solutions from Three-Dimensional Database Searches into Novel Ligands. *Journal of Computer-Aided Molecular Design,* 7**,** 623-647.

HOLLAS, B. 2003. An Analysis of the Autocorrelation Descriptor for Molecules. *Journal of Mathematical Chemistry,* 33**,** 91-101.

HÖLLERING, R., GASTEIGER, J., STEINHAUER, L., SCHULZ, K.-P. & HERWIG, A. 2000. Simulation of Organic Reactions: From the Degradation of Chemicals to Combinatorial Synthesis. *Journal of Chemical Information and Computer Sciences,* 40**,** 482-494.

HOLLIDAY, G. L., MURRAY-RUST, P. & RZEPA, H. S. 2005. Chemical Markup, XML, and the World Wide Web. 6. CMLReact, an XML Vocabulary for Chemical Reactions. *Journal of Chemical Information and Modeling,* 46**,** 145-157.

HONMA, T. 2003. Recent advances in de novo design strategy for practical lead identification. *Medicinal Research Reviews,* 23**,** 606-632.

HRISTOZOV, D., BODKIN, M., CHEN, B., PATEL, H. & GILLET, V. J. 2011. Validation of Reaction Vectors for *de Novo* Design. In: *Library Design, Search Methods, and Applications of Fragment-Based Drug Design*. 29-43: American Chemical Society.

HUANG, Q., LI, L.-L. & YANG, S.-Y. 2010. PhDD: A New Pharmacophore-based de Novo Design Method of Drug-like Molecules Combined with Assessment of Synthetic Accessibility. *Journal of Molecular Graphics and Modelling,* 28**,** 775-787.

INFOCHEM. *Classify* [Online]. Available: http://infochem.de/products/software/classify.shtml [Accessed October 2015].

INFOCHEM. *SPRESI* [Online]. Available: http://infochem.de/products/databases/spresi.shtml [Accessed October 2015].

JOHN WILEY & SONS. *e-EROS* [Online]. Available: http://onlinelibrary.wiley.com/book/10.1002/047084289X [Accessed October 2015].

JOHNSON, A. P., LAW, J., ZSOLDOS, Z., SIMON, A. & WILLIAMS, A. J. 2008. A new, automated retrosynthetic search engine: ARChem. In*: 236th ACS National Meeting*, *August 17-21*, *2008*, Philadelphia, Pennsylvania, USA. American Chemical Society, CINF-078.

JONES, C. D., ANDREWS, D. M., BARKER, A. J., BLADES, K., BYTH, K. F., FINLAY, M. R. V., GEH, C., GREEN, C. P., JOHANNSEN, M., WALKER, M. & WEIR, H. M. 2008. Imidazole Pyrimidine Amides as Potent, Orally Bioavailable Cyclin-dependent Kinase Inhibitors. *Bioorganic & Medicinal Chemistry Letters,* 18**,** 6486-6489.

JORGENSEN, W. L., RUIZ-CARO, J., TIRADO-RIVES, J., BASAVAPATHRUNI, A., ANDERSON, K. S. & HAMILTON, A. D. 2006. Computer-aided Design of Non-nucleoside Inhibitors of HIV-1 Reverse Transcriptase. *Bioorganic & Medicinal Chemistry Letters,* 16**,** 663-667.

KATRITZKY, A. R., MU, L., LOBANOV, V. S. & KARELSON, M. 1996. Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Inorganics. *Journal of Physical Chemistry,* 100**,** 10400-10407.

KENNEDY, J. & EBERHART, R. 1995. Particle Swarm Optimization. *IEEE International Conference On Neural Networks, 1995 Proceedings,* 4**,** 1942-1948.

KHAN, M. T. H., FUSKEVÅG, O.-M. & SYLTE, I. 2009. Discovery of Potent Thermolysin Inhibitors Using Structure Based Virtual Screening and Binding Assays. *Journal of Medicinal Chemistry,* 52**,** 48-61.

KHARKAR, P., DEODHAR, M. & KULKARNI, V. 2009. Design, synthesis, antifungal activity, and ADME prediction of functional analogues of terbinafine. *Medicinal Chemistry Research,* 18**,** 421-432.

KIER, L. B. & HALL, L. H. 1986. *Molecular connectivity in chemistry and drug research*: Academic Press.

KOWALIK, M., GOTHARD, C. M., DREWS, A. M., GOTHARD, N. A., WECKIEWICZ, A., FULLER, P. E., GRZYBOWSKI, B. A. & BISHOP, K. J. M. 2012. Parallel Optimization of Synthetic Pathways within the Network of Organic Chemistry. *Angewandte Chemie International Edition,* 51**,** 7928-7932.

KUHN, S., HELMUS, T., LANCASHIRE, R. J., MURRAY-RUST, P., RZEPA, H. S., STEINBECK, C. & WILLIGHAGEN, E. L. 2007. Chemical Markup, XML, and the World Wide Web. 7. CMLSpect, an XML Vocabulary for Spectral Data. *Journal of Chemical Information and Modeling,* 47**,** 2015-2034.

KUNTZ, I. D., BLANEY, J. M., OATLEY, S. J., LANGRIDGE, R. & FERRIN, T. E. 1982. A Geometric Approach to Macromolecule-Ligand Interactions. *Journal of Molecular Biology,* 161**,** 269-288.

KUTCHUKIAN, P. S., LOU, D. & SHAKHNOVICH, E. I. 2009. FOG: Fragment Optimized Growth Algorithm for the de Novo Generation of Molecules Occupying Druglike Chemical Space. *Journal of Chemical Information and Modeling,* 49**,** 1630-1642.

LANDRUM, G. *RDKit: Open-source cheminformatics*. http://www.rdkit.org [Accessed October 2015].

LAW, J., ZSOLDOS, Z., SIMON, A., REID, D., LIU, Y., KHEW, S. Y., JOHNSON, A. P., MAJOR, S., WADE, R. A. & ANDO, H. Y. 2009. Route Designer: A Retrosynthetic Analysis Tool

Utilizing Automated Retrosynthetic Rule Generation. *Journal of Chemical Information and Modeling,* 49**,** 593-602.

LEACH, A. R. & GILLET, V. J. 2003. *An Introduction to Chemoinformatics*: Springer.

LEACH, A. R. & KILVINGTON, S. R. 1994. Automated Molecular Design: A New Fragment-Joining Algorithm. *Journal of Computer-Aided Molecular Design,* 8**,** 283-298.

LEACH, A. R. & LEWIS, R. A. 1994. A Ring-Bracing Approach to Computer-Assisted Ligand Design. *Journal of Computational Chemistry,* 15**,** 233-240.

LEACH, M. *The Chemical Thesaurus* [Online]. Available: http://www.chemthes.com/ [Accessed October 2015].

LEDNICER, D. 2007. *The Organic Chemistry of Drug Synthesis*: John Wiley and Sons.

LEWELL, X. Q., JUDD, D. B., WATSON, S. P. & HANN, M. M. 1998. RECAP (Retrosynthetic Combinatorial Analysis Procedure):- A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *Journal of Chemical Information and Computer Sciences,* 38**,** 511-522.

LEWIS, R. A. & DEAN, P. M. 1989. Automated Site-Directed Drug Design: The Concept of Spacer Skeletons for Primary Structure Generation. *Proceedings of the Royal Society of London. B. Biological Sciences,* 236**,** 125-140.

LEWIS, R. A., ROE, D. C., HUANG, C., FERRIN, T. E., LANGRIDGE, R. & KUNTZ, I. D. 1992. Automated Site-Directed Drug Design using Molecular Lattices. *Journal of Molecular Graphics,* 10**,** 66-78.

LI, J. J., NAHRA, J., JOHNSON, A. R., BUNKER, A., O'BRIEN, P., YUE, W.-S., ORTWINE, D. F., MAN, C.-F., BARAGI, V., KILGORE, K., DYER, R. D. & HAN, H.-K. 2008. Quinazolinones and Pyrido[3,4-d]pyrimidin-4-ones as Orally Active and Specific Matrix Metalloproteinase-13 Inhibitors for the Treatment of Osteoarthritis. *Journal of Medicinal Chemistry,* 51**,** 835-841.

LIPINSKI, C. A. 2004. Lead- and Drug-like Compounds: The Rule-of-Five Revolution. *Drug Discovery Today: Technologies,* 1**,** 337-341.

LIPINSKI, C. A., LOMBARDO, F., DOMINY, B. W. & FEENEY, P. J. 1997. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Advanced Drug Delivery Reviews,* 23**,** 3-25.

LIU, H., DUAN, Z., LUO, Q. & SHI, Y. 1999. Structure-Based Ligand Design by Dynamically Assembling Molecular Building Blocks at Binding Site. *Proteins: Structure, Function, and Bioinformatics,* 36**,** 462-470.

LIU, S., CAO, C. & LI, Z. 1998. Approach to Estimation and Prediction for Normal Boiling Point (NBP) of Alkanes Based on a Novel Molecular Distance-Edge (MDE) Vector, λ. *Journal of Chemical Information and Computer Sciences,* 38**,** 387-394.

LLOYD, D. G., BUENEMANN, C. L., TODOROV, N. P., MANALLACK, D. T. & DEAN, P. M. 2003. Scaffold Hopping in De Novo Design. Ligand Generation in the Absence of Receptor Information. *Journal of Medicinal Chemistry,* 47**,** 493-496.

LOWE, D. M. & SAYLE, R. 2014. *Unleashing over a million reactions into the wild* [Online]. Available: http://nextmovesoftware.com/blog/2014/02/27/unleashing-over-a-million-reactions-into-the-wild/ [Accessed February 2015].

LUO, Z., WANG, R. & LAI, L. 1996. RASSE: A New Method for Structure-Based Drug Design. *Journal of Chemical Information and Computer Sciences,* 36**,** 1187-1194.

LYNCH, M. F. & WILLETT, P. 1978a. The Automatic Detection of Chemical Reaction Sites. *Journal of Chemical Information and Computer Sciences,* 18**,** 154-159.

LYNCH, M. F. & WILLETT, P. 1978b. The Production of Machine-Readable Descriptions of Chemical Reactions Using Wiswesser Line Notations. *Journal of Chemical Information and Computer Sciences,* 18**,** 149-154.

MATA, P., GILLET, V. J., JOHNSON, A. P., LAMPREIA, J., MYATT, G. J., SIKE, S. & STEBBINGS, A. L. 1995. SPROUT: 3D Structure Generation Using Templates. *Journal of Chemical Information and Computer Sciences,* 35**,** 479-493.

MCGREGOR, J. J. & WILLETT, P. 1981. Use of a Maximum Common Subgraph Algorithm in the Automatic Identification of Ostensible Bond Changes Occurring in Chemical Reactions. *Journal of Chemical Information and Computer Sciences,* 21**,** 137-140.

MCKAY, B. D. 1981. Practical Graph Isomorphism. *Congressus Numerantium,* 30**,** 45-87.

MCNAUGHT, A. 2006. The IUPAC International Chemical Identifier: InChl — A New Standard for Molecular Informatics. *Chemistry International,* 28**,** 12-14.

MEDINA-FRANCO, J. L. 2012. Interrogating Novel Areas of Chemical Space for Drug Discovery using Chemoinformatics. *Drug Development Research,* 73**,** 430-438.

MILLS, J. E., MARYANOFF, C. A., SORGI, K. L., SCOTT, L. & STANZIONE, R. 1988. REACCS in the Chemical Development Environment. 1. *Journal of Chemical Information and Computer Sciences,* 28**,** 153-155.

MOLECULAR NETWORKS GMBH. *SYLVIA*. http://www.molecular-networks.com/products/sylvia [Accessed November 2015].

MORGAN, H. L. 1965. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation,* 5**,** 107-113.

MURRAY-RUST, P. & RZEPA, H. S. 1999. Chemical Markup, XML, and the World Wide Web. 1. Basic Principles. *Journal of Chemical Information and Computer Sciences,* 39**,** 928-942.

MURRAY-RUST, P. & RZEPA, H. S. 2003. Chemical Markup, XML, and the World Wide Web. 4. CML Schema. *Journal of Chemical Information and Computer Sciences,* 43**,** 757-772.

MURRAY, C. W., CLARK, D. E., AUTON, T. R., FIRTH, M. A., LI, J., SYKES, R. A., WASZKOWYCZ, B., WESTHEAD, D. R. & YOUNG, S. C. 1997. PRO_SELECT: Combining Structure-based Drug Design and Combinatorial Chemistry for Rapid Lead Discovery. 1. Technology. *Journal of Computer-Aided Molecular Design,* 11**,** 193-207.

NACHBAR, R. B. 2000. Molecular Evolution: Automated Manipulation of Hierarchical Chemical Topology and Its Application to Average Molecular Structures. *Genetic Programming and Evolvable Machines,* 1**,** 57-94.

NICOLAOU, C. A., APOSTOLAKIS, J. & PATTICHIS, C. S. 2009. De Novo Drug Design Using Multiobjective Evolutionary Graphs. *Journal of Chemical Information and Modeling,* 49**,** 295-307.

NILAKANTAN, R., NUNN, D. S., GREENBLATT, L., WALKER, G., HARAKI, K. & MOBILIO, D. 2006. A Family of Ring System-Based Structural Fragments for Use in Structure−Activity Studies: Database Mining and Recursive Partitioning. *Journal of Chemical Information and Modeling,* 46**,** 1069-1077.

NISHIBATA, Y. & ITAI, A. 1991. Automatic Creation of Drug Candidate Structures Based on Receptor Structure. Starting Point for Artificial Lead Generation. *Tetrahedron,* 47**,** 8985-8990.

NIST. *NIST Chemistry Webbook* [Online]. National Institute of Standards and Technology,. Available: http://webbook.nist.gov/chemistry/ [Accessed November 2015].

OPENMOLECULES.ORG. *Webreactions* [Online]. Available: http://webreactions.net/index.html [Accessed October 2015].

ORGANIC SYNTHESES INC. *Organic Syntheses* [Online]. Available: http://www.orgsyn.org/ [Accessed October 2015].

OSTER, G. & PERELSON, A. 1974. Chemical reaction networks. *IEEE Transactions on Circuits and Systems,* 21**,** 709-721.

PARK, W. K. C., KENNEDY, R. M., LARSEN, S. D., MILLER, S., ROTH, B. D., SONG, Y., STEINBAUGH, B. A., SUN, K., TAIT, B. D., KOWALA, M. C., TRIVEDI, B. K., AUERBACH, B., ASKEW, V., DILLON, L., HANSELMAN, J. C., LIN, Z., LU, G. H., ROBERTSON, A. & SEKERKE, C. 2008. Hepatoselectivity of Statins: Design and Synthesis of 4-Sulfamoyl Pyrroles as HMG-CoA Reductase Inhibitors. *Bioorganic & Medicinal Chemistry Letters,* 18**,** 1151-1156.

PATEL, H. 2009. *Knowledge-Based De Novo Design using Reaction Vectors.* PhD, University of Sheffield.

PATEL, H., BODKIN, M. J., CHEN, B. & GILLET, V. J. 2009. Knowledge-Based Approach to de Novo Design Using Reaction Vectors. *Journal of Chemical Information and Modeling,* 49**,** 1163-1184.

PATEL, H., GILLET, V. J., CHEN, B. & BODKIN, M. 2008. Structure generation using reaction vectors. In: *235th ACS National Meeting*, *April 6-10, 2008,* New Orleans, Louisiana, USA. American Chemical Society, CINF-053.

PAUL, S. M., MYTELKA, D. S., DUNWIDDIE, C. T., PERSINGER, C. C., MUNOS, B. H., LINDBORG, S. R. & SCHACHT, A. L. 2010. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery,* 9**,** 203-214.

PEARLMAN, D. A. & MURCKO, M. A. 1993. CONCEPTS: New Dynamic Algorithm for de Novo Drug Suggestion. *Journal of Computational Chemistry,* 14**,** 1184-1193.

PEARLMAN, D. A. & MURCKO, M. A. 1996. CONCERTS: Dynamic Connection of Fragments as an Approach to de Novo Ligand Design. *Journal of Medicinal Chemistry,* 39**,** 1651-1663.

PEGG, S. C. H., HARESCO, J. J. & KUNTZ, I. D. 2001. A Genetic Algorithm for Structure-based de Novo Design. *Journal of Computer-Aided Molecular Design,* 15**,** 911-933.

PETITJEAN, M. 1992. Applications of the radius-diameter diagram to the classification of topological and geometrical shapes of chemical compounds. *Journal of Chemical Information and Computer Sciences,* 32**,** 331-337.

PIERCE, A. C., RAO, G. & BEMIS, G. W. 2004. BREED: Generating Novel Inhibitors through Hybridization of Known Ligands. Application to CDK2, P38, and HIV Protease. *Journal of Medicinal Chemistry,* 47**,** 2768-2775.

PRASANNA, S. & DOERKSEN, R. J. 2009. Topological polar surface area: a useful descriptor in 2D-QSAR. *Current Medicinal Chemistry,* 16**,** 21-41.

PROUDFOOT, J. R. 2013. Reaction Schemes Visualized in Network Form: The Syntheses of Strychnine as an Example. *Journal of Chemical Information and Modeling,* 53**,** 1035-1042.

RANDIĆ, M. & BASAK, S. C. 1999. Optimal Molecular Descriptors Based on Weighted Path Numbers. *Journal of Chemical Information and Computer Sciences,* 39**,** 261-266.

RAY, L. C. & KIRSCH, R. A. 1957. Finding Chemical Records by Digital Computers. *Science,* 126**,** 814-819.

REISEN, F. H., SCHNEIDER, G. & PROSCHAK, E. 2008. Reaction-MQL: Line Notation for Functional Transformation. *Journal of Chemical Information and Modeling,* 49**,** 6-12.

REUTERS, T. *Web of Science* [Online]. Available: http://wokinfo.com/ [Accessed October 2015].

REUTLINGER, M., RODRIGUES, T., SCHNEIDER, P. & SCHNEIDER, G. 2014. Multi-Objective Molecular De Novo Design by Adaptive Fragment Prioritization. *Angewandte Chemie International Edition,* 53**,** 4244-4248.

ROE, D. C. & KUNTZ, I. D. 1995. BUILDER V.2: Improving the Chemistry of a de Novo Design Strategy. *Journal of Computer-Aided Molecular Design,* 9**,** 269-282.

ROTSTEIN, S. H. & MURCKO, M. A. 1993a. GenStar: A Method for de Novo Drug Design. *Journal of Computer-Aided Molecular Design,* 7**,** 23-43.

ROTSTEIN, S. H. & MURCKO, M. A. 1993b. GroupBuild: a Fragment-based Method for de Novo Drug Design. *Journal of Medicinal Chemistry,* 36**,** 1700-1710.

ROUGHLEY, S. D. & JORDAN, A. M. 2011. The Medicinal Chemist's Toolbox: An Analysis of Reactions Used in the Pursuit of Drug Candidates. *Journal of Medicinal Chemistry,* 54**,** 3451-3479.

ROYAL SOCIETY OF CHEMISTRY. *Chemspider* [Online]. Royal Society of Chemistry. Available: http://www.chemspider.com/ [Accessed October 2015].

RUSU, T. & BULACOVSCHI, V. 2006. Multiobjective Tabu Search method used in chemistry. *International Journal of Quantum Chemistry,* 106**,** 1406-1412.

SALATIN, T. D. & JORGENSEN, W. L. 1980. Computer-assisted Mechanistic Evaluation of Organic Reactions. 1. Overview. *Journal of Organic Chemistry,* 45**,** 2043-2051.

SCHNEIDER, G. 2014. Future De Novo Drug Design. *Molecular Informatics,* 33**,** 397-402.

SCHNEIDER, G., LEE, M.-L., STAHL, M. & SCHNEIDER, P. 2000. De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. *Journal of Computer-Aided Molecular Design,* 14**,** 487-494.

SCHNEIDER, N., LOWE, D. M., SAYLE, R. A. & LANDRUM, G. A. 2014. Development of a Novel Fingerprint for Chemical Reactions and Its Application to Large-Scale Reaction Classification and Similarity. *Journal of Chemical Information and Modeling,* 55**,** 39-53.

SCHNEIDER, P. & SCHNEIDER, G. 2003. Collection of Bioactive Reference Compounds for Focused Library Design. *QSAR & Combinatorial Science,* 22**,** 713-718.

SCHNUTE, M. E., BRIDEAU, R. J., COLLIER, S. A., CUDAHY, M. M., HOPKINS, T. A., KNECHTEL, M. L., OIEN, N. L., SACKETT, R. S., SCOTT, A., STEPHAN, M. L., WATHEN, M. W. & WIEBER, J. L. 2008. Synthesis of 4-oxo-4,7-dihydrofuro[2,3-b]pyridine-5-carboxamides with broad-spectrum human herpesvirus polymerase inhibition. *Bioorganic & Medicinal Chemistry Letters,* 18**,** 3856-3859.

SCHÜLLER, A., SUHARTONO, M., FECHNER, U., TANRIKULU, Y., BREITUNG, S., SCHEFFER, U., GÖBEL, M. & SCHNEIDER, G. 2008. The Concept of Template-based de Novo Design From Drug-derived Molecular Fragments and its Application to TAR RNA. *Journal of Computer-Aided Molecular Design,* 22**,** 59-68.

SCHÜRER, S. C., TYAGI, P. & MUSKAL, S. M. 2005. Prospective Exploration of Synthetically Feasible, Medicinally Relevant Chemical Space. *Journal of Chemical Information and Modeling,* 45**,** 239-248.

SELLERS, P. 1967. Algebraic Complexes Which Characterize Chemical Networks. *SIAM Journal on Applied Mathematics,* 15**,** 13-68.

SHARMA, V., GOSWAMI, R. & MADAN, A. K. 1997. Eccentric Connectivity Index: A Novel Highly Discriminating Topological Descriptor for Structure−Property and Structure−Activity Studies. *Journal of Chemical Information and Computer Sciences,* 37**,** 273-282.

SIEGHART, W. 1994. Pharmacology of benzodiazepine receptors: an update. *Journal of Psychiatry and Neuroscience,* 19**,** 24-29.

SINANOGLU, O. 1975. Theory of chemical reaction networks. All possible mechanisms or synthetic pathways with given number of reaction steps or species. *Journal of the American Chemical Society,* 97**,** 2309-2320.

STEINBECK, C., HAN, Y., KUHN, S., HORLACHER, O., LUTTMANN, E. & WILLIGHAGEN, E. 2003. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *Journal of Chemical Information and Computer Sciences,* 43**,** 493-500.

STEVENS, K. L., RENO, M. J., ALBERTI, J. B., PRICE, D. J., KANE-CARSON, L. S., KNICK, V. B., SHEWCHUK, L. M., HASSELL, A. M., VEAL, J. M., DAVIS, S. T., GRIFFIN, R. J. & PEEL, M. R. 2008. Synthesis and Evaluation of Pyrazolo[1,5-b]pyridazines as Selective Cyclin Dependent Kinase Inhibitors. *Bioorganic & Medicinal Chemistry Letters,* 18**,** 5758-5762.

STEWART, K. D., SHIRODA, M. & JAMES, C. A. 2006. Drug Guru: A computer software program for drug design using medicinal chemistry rules. *Bioorganic and Medicinal Chemistry,* 14**,** 7011-7022.

SUTHERLAND, J. J., O'BRIEN, L. A. & WEAVER, D. F. 2004. A Comparison of Methods for Modeling Quantitative Structure−Activity Relationships. *Journal of Medicinal Chemistry,* 47**,** 5541-5554.

TALLEY, J. J., PENNING, T. D., COLLINS, P. W., ROGIER JR., D. J., MALECHA, J. W., MIYASHIRO, J. M., BERTENSHAW, S. R., KHANNA, I. K., GRANETO, M. J., ROGERS, R. S., CARTER, J. S., DOCTER, S. H. & YU, S. S. 2002. *Substituted pyrazolyl benzenesulfonamides for the treatment of inflammation*. United States patent application 6492411.

TEMKIN, O. N. & BONCHEV, D. G. 1992. Application of graph theory to chemical kinetics: Part 1. Kinetics of complex reactions. *Journal of Chemical Education,* 69**,** 544.

TEODORO, M. & MUEGGE, I. 2011. BIBuilder: Exhaustive Searching for De Novo Ligands. *Molecular Informatics,* 30**,** 63-75.

THIEME CHEMISTRY PUBLISHING. *Science of Synthesis* [Online]. Available: https://www.thieme.de/en/thieme-chemistry/science-of-synthesis-54780.htm [Accessed October 2015].

THOMPSON, D., ALDRIN DENNY, R., NILAKANTAN, R., HUMBLET, C., JOSEPH-MCCARTHY, D. & FEYFANT, E. 2008. CONFIRM: connecting fragments found in receptor molecules. *Journal of Computer-Aided Molecular Design,* 22**,** 761-772.

THOMSON REUTERS. *Current Chemical Reactions* [Online]. Available: http://thomsonreuters.com/en/products-services/scholarly-scientific-research/scholarly-search-and-discovery/current-chemical-reactions.html [Accessed October 2015].

TODA, K., GOTO, J. & HIRAYAMA, N. 2010. A novel target-based de novo ligand design by use of pseudomolecular probe. *MedChemComm,* 1**,** 349-354.

TODOROV, N. P. & DEAN, P. M. 1998. A Branch-and-Bound Method for Optimal Atom-type Assignment in de Novo Ligand Design. *Journal of Computer-Aided Molecular Design,* 12**,** 335-335.

TRIPOS. *EA-Inventor* [Online]. Available: http://tripos.com/data/SYBYL/EA_Inventor_072505.pdf [Accessed October 2015].

TRIPOS. *LeapFrog* [Online]. Available: http://tripos.com/data/SYBYL/LeapFrog_072505.pdf [Accessed October 2015].

TROTT, O. & OLSON, A. J. 2010. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry,* 31**,** 455-461.

TSCHINKE, V. & COHEN, N. C. 1993. The NEWLEAD Program: A New Method for the Design of Candidate Structures from Pharmacophoric Hypotheses. *Journal of Medicinal Chemistry,* 36**,** 3863-3870.

UCSF. *ChemViz* [Online]. Available: http://www.rbvi.ucsf.edu/cytoscape/chemViz [Accessed October 2015].

UGI, I., BAUER, J., BLEY, K., DENGLER, A., DIETZ, A., FONTAIN, E., GRUBER, B., HERGES, R., KNAUER, M., REITSAM, K. & STEIN, N. 1993. Computer-Assisted Solution of Chemical Problems—The Historical Development and the Present State of the Art of a New Discipline of Chemistry. *Angewandte Chemie International Edition in English,* 32**,** 201-227.

VINKERS, H. M., DE JONGE, M. R., DAEYAERT, F. F. D., HEERES, J., KOYMANS, L. M. H., VAN LENTHE, J. H., LEWI, P. J., TIMMERMAN, H., VAN AKEN, K. & JANSSEN, P. A. J. 2003. SYNOPSIS: SYNthesize and OPtimize System in Silico. *Journal of Medicinal Chemistry,* 46**,** 2765-2773.

VLÉDUTS, G. É. 1963. Concerning One System of Classification and Codification of Organic Reactions. *Information Storage and Retrieval,* 1**,** 117-146.

VRIELINK, A., OBEL-JORGENSEN, A. & CODDING, P. W. 1996. Hippuryl-l-histidyl-l-leucine, a Substrate for Angiotensin Converting Enzyme. *Acta Crystallographica Section C,* 52**,** 1300-1302.

WALLACE, J. E. A. 2015. All structures shown here are either trivial small proof of concept examples, taken directly from the literature as cited, or de novo generated by the procedures outlined from literature SAR sets without any input of Lilly intellectual property.

WALTERS, W. P., STAHL, M. T. & MURCKO, M. A. 1998. Virtual Screening—an Overview. *Drug Discovery Today,* 3**,** 160-178.

WANG, R., GAO, Y. & LAI, L. 2000. LigBuilder: A Multi-Purpose Program for Structure-Based Drug Design. *Journal of Molecular Modeling,* 6**,** 498-516.

WANG, T., LAMB, M. L., SCOTT, D. A., WANG, H., BLOCK, M. H., LYNE, P. D., LEE, J. W., DAVIES, A. M., ZHANG, H.-J., ZHU, Y., GU, F., HAN, Y., WANG, B., MOHR, P. J., KAUS, R. J., JOSEY, J. A., HOFFMANN, E., THRESS, K., MACINTYRE, T., WANG, H., OMER, C. A. & YU, D. 2008. Identification of 4-Aminopyrazolylpyrimidines as Potent Inhibitors of Trk Kinases. *Journal of Medicinal Chemistry,* 51**,** 4672-4684.

WANG, Z. 2010. Mignonac Reaction. In: *Comprehensive Organic Name Reactions and Reagents*: 436, 1945-1947, John Wiley & Sons, Inc.

WEININGER, D. 1988. SMILES. A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *Journal of Chemical Information and Computer Sciences,* 28**,** 31-36.

WEISGERBER, D. W. 1997. Chemical Abstracts Service Chemical Registry System: History, scope, and impacts. *Journal of the American Society for Information Science,* 48**,** 349-360.

WESTAWAY, S. M., BROWN, S. L., CONWAY, E., HEIGHTMAN, T. D., JOHNSON, C. N., LAPSLEY, K., MACDONALD, G. J., MACPHERSON, D. T., MITCHELL, D. J., MYATT, J. W., SEAL, J. T., STANWAY, S. J., STEMP, G., THOMPSON, M., CELESTINI, P., COLOMBO, A., CONSONNI, A., GAGLIARDI, S., RICCABONI, M., RONZONI, S., BRIGGS, M. A., MATTHEWS, K. L., STEVENS, A. J., BOLTON, V. J., BOYFIELD, I., JARVIE, E. M., STRATTON, S. C. & SANGER, G. J. 2008. The Discovery of Biaryl Carboxamides as

Novel Small Molecule Agonists of the Motilin Receptor. *Bioorganic & Medicinal Chemistry Letters,* 18**,** 6429-6436.

WHITE, D. & WILSON, R. C. 2010. Generative Models for Chemical Structures. *Journal of Chemical Information and Modeling,* 50**,** 1257-1274.

WIENER, H. 1947. Structural Determination of Paraffin Boiling Points. *Journal of the American Chemical Society,* 69**,** 17-20.

WILEY/FIZ CHEMIE BERLIN. *Cheminform* [Online]. Available: http://www.fiz-chemie.de/cheminform/ [Accessed October 2015].

WILLETT, P. 1980. The Evaluation of an Automatically Indexed, Machine-Readable Chemical Reactions File. *Journal of Chemical Information and Computer Sciences,* 20**,** 93-96.

WILLETT, P. 2011. Chemoinformatics: A History. *Wiley Interdisciplinary Reviews: Computational Molecular Science,* 1**,** 46-56.

WILLETT, P., BARNARD, J. M. & DOWNS, G. M. 1998. Chemical Similarity Searching. *Journal of Chemical Information and Computer Sciences,* 38**,** 983-996.

WILLETT, P. & WINTERMAN, V. 1986. A Comparison of Some Measures for the Determination of Inter-Molecular Structural Similarity Measures of Inter-Molecular Structural Similarity. *Quantitative Structure-Activity Relationships,* 5**,** 18-25.

XUE, C. X., ZHANG, R. S., LIU, H. X., LIU, M. C., HU, Z. D. & FAN, B. T. 2004. Support Vector Machines-Based Quantitative Structure–Property Relationship for the Prediction of Heat Capacity. *Journal of Chemical Information and Computer Sciences,* 44**,** 1267-1274.

YUAN, Y., PEI, J. & LAI, L. 2011. LigBuilder 2: A Practical de Novo Drug Design Approach. *Journal of Chemical Information and Modeling,* 51**,** 1083-1091.

ZBINDEN, P., DOBLER, M., FOLKERS, G. & VEDANI, A. 1998. PrGen: Pseudoreceptor Modeling Using Receptor-mediated Ligand Alignment and Pharmacophore Equilibration. *Quantitative Structure-Activity Relationships,* 17**,** 122-130.

ZHAO, Y. H., ABRAHAM, M. H. & ZISSIMOS, A. M. 2003. Fast Calculation of van der Waals Volume as a Sum of Atomic and Bond Contributions and Its Application to Drug Compounds. *Journal of Organic Chemistry,* 68**,** 7368-7373.