

# **On The Dissemination Of Novel Chemistry And The Process Of Optimising Compounds In Drug Discovery Projects**

**Name: Stephanie Kay Ashenden Bsc (Hons) Msc (Res)**

**College: Darwin College**

**Submission Date: September 2018**

**University: University Of Cambridge**

**Funding: BBSRC And AstraZeneca**

**This Dissertation Is Submitted For The Degree Of Doctor Of Philosophy**

## **Preface**

This dissertation is the result of my own work and includes nothing, which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

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

It does not exceed the prescribed word limit for the relevant Degree Committee. For more information on the word limits for the respective Degree Committee.

## **Acknowledgements**

I would like to thank my academic supervisor Dr Andreas Bender as well as my current industrial supervisor Dr Ola Engkvist and previous industrial supervisor Dr Thierry Kogej, for all their time, effort and support. Furthermore, I would like to thank the members of the Quantitative Biology team at AstraZeneca including; Dr Claus Bendtsen, Dr Ian Barrett, Dr Mike Firth, Dr Stanley Lazic, Dr Maria Luisa Guerriero, Dr Piero Ricchiuto, Dr Aurelie Bornot, Dr Yin Hai Wang and Dr Jonathan Cairns. Additionally, I would like to thank Dr Garry Pairaudeau, Dr Peter Várkonyi and Clare Gregson.

I would like to thank the members of the Bender Group, especially, Kathryn Giblin, Ben Alexander-Dann, Erin Oerton, Fatima Baldo, Dr Christoph Schaufner, Dr Georgios Drakakis, Azedine Zoufir, Dr Avid Afzal. Furthermore, I would like to thank the help and support I have received from the Department of Chemistry, particularly that of Dr Rachel MacDonald, Dr Nick Bampos and Mrs Susan Begg as well as my college Darwin College.

Finally, I would like to dedicate this work to my parents, Paul and Tania Ashenden, as well as Rory Parsons for their continued support, patience and care. Thank you for all you have done and taught me.

## List of Written Publications

Ashenden, S.; Kogej, T.; Engkvist, O.; Bender, A. Innovation in Small-Molecule-Druggable Chemical Space: Where are the Initial Modulators of New Targets Published?, *JCIM*, 2017, 10.1021/acs.jcim.7b00295

Mason, D.; Stott, I.; Ashenden, S.; Weinstein, Z et al. Prediction of Antibiotic Interactions Using Descriptors Derived from Molecular Structure. *J. Med. Chem.*, 2017, 60 (9), pp 3902–3912

## Summary

Optimising the drug discovery process remains one of the largest challenges in medicine. Learning from previous compound-target associations as well as the process of optimising compounds will allow for a more targeted and knowledge-based approach. The aim of the first research chapter of this thesis is to understand where novel chemistry is first published. It is well established that the number of publications of novel small molecule modulators, and their associated targets, has increased over the years. This work focuses on publishing trends over the years with a focus on the comparison between patents and scientific literature, which is accessible via the ChEMBL and GOSTAR databases. More precisely, the patents and scientific literature associated with bioactive molecules and their target annotations have been compared to identify where novelty (in the meaning of the first modulator of a protein target) originated. Comparing the published date of the first small molecule modulator published in literature and patents for a target (with the modulators having either identical or different structures) shows that modulators are usually published in both scientific literature and in patents (45%), or in scientific literature alone (51%), but rarely in patents only. When looking at the time when first modulators are published in both sources, 65% of the time they are disseminated in literature first. Finally, when analysing just the novel small molecule modulators, regardless of the protein targets they have been published with, those structures representing novel chemistry tend to be published in patents first (61% of the time). It is concluded that novel chemistry, when associated with a target, is primarily published in the literature, therefore, when exploring known chemistry for a specific known target, this should be identified from the literature.

Following this, it is important to understand how chemists optimise compounds, and we use matched molecular pair analysis (MMPs) to this end, which allows us to compare the properties of two compounds that differ by only one chemical transformation and are important for the compound to be success as a drug. In this part of the thesis, we statistically analyse the most frequently observed MMPs within drug discovery projects by using the compound registration dates to determine the order in which compounds were made within projects and aggregate the findings over all internal projects in AstraZeneca. For those MMPs that are commonly observed in projects, we compare this frequency to the frequency of reverse change in structure, to determine if there are preferences in the chemical changes made in projects over time. Furthermore, we analyse the neighbouring environments for the position where the molecule has changed. 957 unique MMPs were found to occur at least 100 times across projects, comprising 81 unique molecular fragments as starting points and 197 unique molecular fragments as end points of MMPs. The most frequently occurring MMPs as well as

the most frequently occurring atomic environments differ between aliphatic and aromatic systems. Overall, this study provides a data-driven method to analyse the order in which molecular fragments are incorporated into molecules in drug discovery projects. This knowledge can be used to help guide decisions in future compound design.

Finally, relating these MMP findings to the measured assay results allows an overview to be made about the how the compounds themselves evolve throughout the project. MMPs are used when designing of new compounds to exploit existing knowledge of the effect of a molecular transformation on compound properties (such as binding, solubility, logD etc) and apply this to new compounds with the expectation of seeing the same outcome. The effect on physicochemical properties as measured in assays, from transformations on specific atomic environments since the year 2000, have been analysed *via* a time course analysis. This allows us to observe the effect of the transformations over time. In total 453 unique transformations were analysed. It highlights that even when just comparing between aromatic and aliphatic systems on a higher level, changes can be observed and shows that when designing a compound, consideration of the atomic environment is essential. These results can be used to identify the structural change that would improve a compound profile going through the design process; saving time, resources and money. Additionally, specific examples have been extracted for discussion. Notably, those examples that are considered extreme outliers, which generally refer to transformations involving a very large property change of the compound ( $\pm 4$  standard deviations). These extreme outliers highlight the need to always consider outliers in the analysis as they may be of importance but retaining them within a study may obscure additional results. Therefore, it is suggested to acknowledge these outliers, but not include them in the main study. Furthermore, case studies are given that show unexpected changes in property values when the logD increases such as solubility also increasing and is shown to be the result of surrounding chemistry of the atomic environment.

## Contents

|   |    |
|---|----|
| Preface .....   | 2  |
| Acknowledgements .....  | 3  |
| List of Written Publications .....  | 4  |
| Summary .....   | 5  |
| Contents .....  | 7  |
| 1 Introduction.....   | 10 |
| 1.1 Drug Discovery Process.....   | 10 |
| 1.1.1 High-Throughput Screening (HTS) .....   | 10 |
| 1.1.2 Lead Optimisation .....   | 11 |
| 1.1.3 Assessing biological activity and ADMET properties .....  | 12 |
| 1.1.3.1 Biological Activity (On and Off-Target) .....   | 13 |
| 1.1.3.2 Absorption.....   | 15 |
| 1.1.3.3 Distribution.....   | 16 |
| 1.1.3.4 Metabolism .....  | 16 |
| 1.1.3.5 Excretion.....  | 17 |
| 1.1.3.6 Toxicity.....   | 17 |
| 1.1.4 Assay Test Results .....  | 18 |
| 1.2 Compound Sources and Comparison .....   | 20 |
| 1.3 Dissemination of Chemistry .....  | 21 |
| 1.4 Matched Molecular Pairs.....  | 22 |
| 1.5 Maximum Common Substructure.....  | 25 |
| 2 Innovation in Small-Molecule-Druggable Chemical Space: Where are the Initial Modulators of New Targets Published?.....  | 28 |
| 2.1 Introduction .....  | 28 |
| 2.2 Materials and Methods.....  | 29 |
| 2.2.1 Extraction and organisation of the GOSTAR dataset .....   | 29 |
| 2.2.2 Extraction and organisation of the ChEMBL 21 dataset .....  | 30 |
| 2.2.3 Visualisation of data.....  | 32 |
| 2.2.4 Comparison of patent vs. public datasets and the distribution of the years difference and number of targets for each year's difference .....  | 32 |
| 2.2.5 Comparison of patent vs. public datasets and the distribution of the years difference and number of molecular frameworks, topological frameworks and compounds for each year's difference ..... | 33 |
| 2.2.6 Comparison of patent vs. public datasets and for each target class and each year bin when and where was an annotation published first.....  | 33 |
| 2.2.7 Comparison of patent vs. public datasets for observing trends in annotations that were only published in patents or only published in literature .....  | 34 |
| 2.2.8 Statistical Validation.....   | 34 |

|         |   |    |
|---------|---|----|
| 2.2.9   | Compound Novelty .....  | 34 |
| 2.2.10  | Compound Similarity.....  | 34 |
| 2.3     | Results and Discussion.....   | 35 |
| 2.3.1   | Number of Unique Compound-Target annotation analysis.....   | 35 |
| 2.3.2   | Time based analysis of the source for new compound-target annotations.....  | 38 |
| 2.3.3   | Analysis of the number of target annotations that were only published in either a patent or in scientific literature via time course analysis .....                         | 45 |
| 2.3.4   | Case Studies.....   | 46 |
| 2.3.5   | Analysis of where the novel bioactive structures (compounds, molecular and topological frameworks) were first published .....   | 47 |
| 2.3.6   | Analysis of the number of structures that were <i>only</i> published in either a patent or in scientific literature as a function of time .....                             | 50 |
| 2.4     | Chapter overview .....  | 55 |
| 3       | Analysing the Matched Molecular Pair Transformations in Drug Discovery Projects as a Function of Time and Molecular Environment: – Frequency of Molecular Transformations . | 57 |
| 3.1     | Introduction .....  | 57 |
| 3.2     | Materials and Methods.....  | 60 |
| 3.2.1   | Compilation of the dataset .....  | 60 |
| 3.2.2   | Determination of the atomic environment of the compound .....   | 60 |
| 3.2.3   | Determining the matched molecular pairs .....   | 62 |
| 3.2.4   | Parameters used to process the data .....   | 62 |
| 3.2.5   | Calculation of time difference and ratio between MMP transformations and their opposite   | 63 |
| 3.3     | Results and Discussion.....   | 63 |
| 3.3.1   | Analysis of the most frequently observed molecular fragments found in transformations as a function of to time .....  | 63 |
| 3.3.1.1 | Aromatic Systems.....   | 65 |
| 3.3.1.2 | Aliphatic Systems .....   | 68 |
| 3.3.2   | Analysis of the most frequently observed MMP transformations and their inverse transformations.....   | 69 |
| 3.3.2.1 | Aromatic Systems.....   | 72 |
| 3.3.2.2 | Aliphatic Systems .....   | 75 |
| 3.3.3   | Analysis of the most frequently observed environments and the MMP transformations performed on them .....   | 78 |
| 3.3.3.1 | Aromatic Systems.....   | 80 |
| 3.3.3.2 | Aliphatic Systems .....   | 81 |
| 3.4     | Chapter overview .....  | 82 |
| 4       | Analysing the Matched Molecular Pair Transformations in Drug Discovery Projects as a Function of Time and Molecular Environment: – Effects on Compound Properties.....      | 83 |
| 4.1     | Introduction .....  | 83 |
| 4.2     | Materials and Methods.....  | 85 |



|         |  |     |
|---------|--|-----|
| 4.2.1   | Data compilation .....   | 85  |
| 4.2.2   | Assay properties analysed.....   | 85  |
| 4.2.3   | Determining a significant increase, decrease or a minimal change in the log property values.....   | 86  |
| 4.2.4   | Outliers.....  | 87  |
| 4.3     | Results and Discussion.....  | 87  |
| 4.3.1   | Analysis of the physicochemical properties and assay result change over the course of a project.....   | 87  |
| 4.3.2   | Analysis of the most frequently observed MMP transformations and their effect on compound properties.....  | 93  |
| 4.3.2.1 | Aromatic Systems.....  | 94  |
| 4.3.2.2 | Aliphatic Systems .....  | 96  |
| 4.3.3   | Analysis of both the proportion of significant property changes as well as the quantitative amount of change, as a function of performing transformations on different atomic environments ..... | 98  |
| 4.3.3.1 | Aromatic Systems.....  | 104 |
| 4.3.3.2 | Aliphatic Systems .....  | 109 |
| 4.3.4   | Analysis of extreme outliers ( $\pm 4$ Standard Deviations).....   | 112 |
| 4.3.4.1 | Aromatic Systems.....  | 112 |
| 4.3.4.2 | Aliphatic Systems .....  | 115 |
| 4.4     | Case Studies of Transformations Performed on Atomic Environments of Which Affected Property Changes Unexpectedly when Increasing the LogD .....  | 120 |
| 4.5     | Chapter overview .....   | 124 |
|         | Conclusion.....  | 126 |
|         | References .....   | 128 |

# 1 Introduction

## 1.1 Drug Discovery Process

Drug discovery is a long, expensive and complex process<sup>1</sup>, only a small proportion of molecules that are identified as a candidate drug are approved as new drugs each year<sup>2</sup>. It was found that of the 98 companies analysed, that had only launched one drug within the decade, \$350 million was the median cost of approved drugs developed by a company<sup>1</sup>. Whereas, when looking at companies that approve between eight and 13 drugs over 10 years, the costs per drug were calculated as much as \$5.5billion<sup>1</sup>.

A general overview of a typical drug discovery process (Figure 1) is split up into several different stages<sup>3</sup>.



Figure 1: A general overview of a drug discovery process from target identification to the clinical testing

While there is no one definite way to arrive at a novel drug, starting with natural products provides one convenient source of novel drug leads<sup>4</sup> or *via* phenotypic screening methods<sup>5</sup>. Today, drug discovery is being led by techniques such as high-throughput screening and empirical screening which involves screening libraries containing chemicals against targets in a physical way. However virtual screening, which screens libraries computationally for compound chemicals that target known structures and having them tested experimentally, has become a leading method to predict new structures<sup>6</sup>. Experimental testing confirms that interactions between the known target and the desired compound is therefore optimised in order to maintain or improve favourable properties<sup>3</sup> including biological activity, whilst reducing or eliminating negative properties (such as toxicity).

### 1.1.1 High-Throughput Screening (HTS)

A hit compound is a compound that has been shown to have activity against a particular target<sup>7</sup>. A frequently used approach for the identification of a hit is *via* means of high-throughput screening (HTS). HTS refers to the process of screening and assaying compounds

against targets on a large scale<sup>8</sup>. Following the development of an assay, high throughput screening is one of the most commonly applicable methods that allow for the identification of a lead compound.

HTS utilises robotics and automatized technologies that allow for rapid tests, such as pharmacological tests, to be conducted. The fact that large numbers of compounds can be screened in small assays against biological targets at the same time has made HTS a powerful tool in the combat of discovering new medicines. Before the birth of this technique, the approach was done manually and only allowed for between 20-50 compounds to be analysed each week<sup>9</sup>. However, with new and improved techniques for identifying potential targets began emerging, it became clear that this methodology could not be sustainable and more efficient technologies and methodologies that were cost effective would need to be introduced.

The future of HTS was discussed recently by Mayr and Fuerst<sup>10</sup>. This paper notes that over time, particularly in the last 20 years, HTS has adapted to the needs and requirements of lead discovery such as improved quality whereas, previously the focus had been on quantity by implementing miniaturisation techniques. Yet, in recent years there has been some disagreement between achieving 'quantity' and consideration of the relevance of the data. Mayr *et al.*, argues that with the implementation of plates with larger numbers of wells being used, such as 384-well plates to conduct the assays, focus will move away from miniaturisation and towards increasing the relevance of each hit-finding strategy<sup>10</sup>. An essential ingredient to the successful improvements for this technique will be the curation of adequate chemical libraries that contain good diversity and drug-like properties<sup>10,11</sup>. Therefore, knowing where to find relevant compounds and where they are published to incorporate into screening libraries is highly important to improve HTS.

### 1.1.2 Lead Optimisation

Lead optimisation refers to the process of designing and improving a pre-identified lead compound, and involves manipulation of multiple parameters of the compound<sup>12</sup>, relying on chemical modifications to the compound. The purpose is to improve the compound properties to the best they can be in terms of the biological activity with respect to on and off target; absorption, distribution, metabolism, excretion and toxicity (ADMET) profiles of a compound are the focus of the lead optimisation efforts. Using drug metabolism and pharmacokinetics (DMPK) parameters in lead optimisation (both *in vitro* and *in vivo*) is a focus of research organisations to aid in producing compounds that fall with an acceptable range in terms of these properties<sup>13</sup>.

Due to the magnitude of different effects a compound can have in the body, there are several criteria that are considered when producing new medicines. Several factors are of focus in the discovery of new drugs including absorption (ability of the compound to be taken up into the blood stream), distribution (how the compound is moved around the body to its desired site), metabolism (how well the compounds are broken down once in the body), excretion (how the compounds and any metabolites are removed from the body) and toxicity (any negative effect the compound will have on the body) (ADMET). However, these properties are not only considered during the lead optimisation phase and are monitored throughout earlier drug development stages<sup>14</sup>. These properties are explored in the following sub sections and are also summarised in Table 1.

These properties are assessed through various assays discussed below as well as in section 1.1.4, which allow for changes in the property values to be recorded as the compound is modified.

### 1.1.3 Assessing biological activity and ADMET properties

Hits that are intended to become leads are assessed for the chemical, synthetic and functional behaviours by using structure-activity-relationships (SAR) as well as their physicochemical and potential toxicology profiles by analysing the compounds absorption, distribution, metabolism, excretion and toxicity (ADMET) properties. The biological activity of the lead compound can be considered one of the most important properties and during the lead optimisation process is carefully considered<sup>15</sup>. These profiles should be determined as early as possible to prevent failure later on which amounts to increased costs<sup>16</sup>. Experimentally, these properties are tested and measured in various assays as seen below (Table 1).

Table 1: Summary of important assays used to test suitability of compounds

| <b>Test</b>          | <b>Applications</b>   | <b>Limitations</b>  | <b>Advantages</b>   |
|----------------------|---|---|---|
| Caco-2 <sup>17</sup> | Assess the permeability and absorption and model the intestinal barrier | Does not tolerate organic solvents<br>Sensitive to excipients<br>Sensitive to different environments and cultures<br>Takes 21 days for the cells to differentiate | Reduces need for animal studies<br>Aid in understanding of transport mechanism and drug pathway |

|                                     |   |   |  |
|-------------------------------------|---|---|--|
| LogD <sup>18</sup>                  | Measure of lipophilicity of the compound specifically, the dissociation of weak acids and bases | Mathematical and experimental derived values cannot be easily compared  | Takes into consideration the pH dependence of the molecule in an aqueous solution  |
| Solubility <sup>19</sup>            | Assess the saturation concentration of a solute in a solvent                                    | Depends on the presence of other species in the solvent   | Can be characterised into levels of solubility, regardless of solvent used.  |
| Microsomal Metabolism <sup>20</sup> | Used to investigate compound metabolism and clearance   | Microsomes do not exist in healthy, human cells. Expression and activity of CYP enzymes is variable depending on factors such as genetics and environmental aspects.  | Can observe CYP (cytochrome P) enzymes.  |
| Hepatocyte Metabolism <sup>21</sup> | Used to investigate compound metabolism in hepatocytes  | Limited by donor availability<br>Limitations in adult differentiated hepatocytes proliferate in culture   | The liver is the primary place that drug metabolism takes place  |
| hERG IC <sub>50</sub> <sup>22</sup> | Used to assess cardiac toxicity by inhibition of hERG.  | The binding assay does not have agonistic or antagonistic effects information<br>The binding assay cannot identify when a compound only binds to one state or other sites of the channel<br>Manual methods, are technically difficult | The binding assay is low cost and can be used in high-throughput<br>Can use automated or manual methods depending on individual requirements/ aims |

### 1.1.3.1 Biological Activity (On and Off-Target)

The biological activity of a compound can be described as the ability to cause an effect in the biological process<sup>23</sup>.

Activity is quantified by a dose-response relationship and often the experiment will be repeated at varying doses and then assessed for what the drug actually does against what the drug is and how much is present<sup>23</sup>. The focus of the observation can lead to differences in how the activity is measured and therefore, different equations and different definitions exist for understanding biological activity.

$$A = cf$$

Equation 1: Definition of a biological activity of an entity proposed by Jackson et al<sup>23</sup>.

Jackson *et al.*<sup>23</sup> proposed a relationship between the thermodynamic activities of a solute with its concentration (Equation 1) *via* an activity coefficient. In this equation, A is the activity, c is the concentration of substance and f is the inherent activity, therefore both biological and chemical considerations are maintained.

Activity concentration is measured by several values including, EC<sub>50</sub>, IC<sub>50</sub>, K<sub>i</sub> and K<sub>d</sub>. The K<sub>d</sub> is the dissociation constant<sup>24</sup> that measures how much a large object dissociates reversibly into smaller constituents (Equation 2). Where [A<sub>x</sub>B<sub>y</sub>] is the complex concentration and [A] and [B] represent the subunit concentrations.

$$K_d = \frac{[A]^x[B]^y}{[A_xB_y]}$$

Equation 2: The dissociation constant

The value of K<sub>i</sub> represents the concentration required to produce half the maximum inhibition<sup>25</sup> (Equation 3), where [S] is the substrate concentration and K<sub>m</sub> is the substrate concentration (without an inhibitor) at the half-maximal velocity of the reaction:

$$K_i = \frac{IC_{50}}{1 + \frac{[S]}{K_m}}$$

Equation 3: The inhibition constant

While IC<sub>50</sub> is the concentration required to cause an inhibitory effect by 50%, and can also be computed using the above equation, once rearranged. Alternatively, it can be calculated *via* linear regression techniques such as *via* Equation 4.

$$IC_{50} = \frac{(0.5 - b)}{a}$$

Equation 4: The half maximal inhibitory concentration equation

EC<sub>50</sub> corresponds to the half maximal effective concentration, and can be described by many different equations, such as in Equation 5, where Y is the observed value, Z is the lowest observed value, A is the highest observed value as well as the Hill coefficient (largest absolute value of the slope of the curve, (as shown as Equation 6)). The Hills equation<sup>26-28</sup> describes

an important relationship in biochemistry/pharmacology since ligand binding is often enhanced in the presence of other ligands that are bound on the same target and so the equation allows us to understand the fraction of the target that is saturated by the ligand as a function of the ligand concentration. In the Hills equation,  $\theta$  is the fraction of the target concentration that is bound to the ligand.  $[L]$  represents the unbound ligand concentration. The  $K_d$  and  $K_A$  values refer to the dissociation constant and the ligand concentration that occupies half the binding sites, respectively. Finally,  $n$  represents the Hills coefficient which describes ligand binding in terms of cooperativity where a positive cooperative binding score refers to the affinity of ligands increasing given another bound ligand.

$$Y = (Z) + \frac{(A) - (Z)}{1 + \left(\frac{x}{EC_{50}}\right)^{-(Hill\ coefficient)}}$$

Equation 5: Equation to derive  $EC_{50}$

$$\theta = \frac{[L]^n}{K_d + [L]^n} = \frac{[L]^n}{(K_A)^n + [L]^n} = \frac{1}{\left(\frac{K_A}{[L]}\right)^n + 1}$$

Equation 6: Different ways to describe the Hills equation.

The off-target biological activity of a compound is the activity that occurs that is not related to the intended biological target. Any off-target interactions, can be high risk in terms of negative side-effects. Alternatively though, off-target interactions can be beneficial, especially, as a therapeutic agent for a different condition<sup>29</sup>. Off-target toxicity is discussed in more detail further on. Another key advantage to such polypharmacological methods is that in instances such as cancer where drug resistance is frequently observed, drugs targeting multiple-targets which are part of a greater biological process are less likely to develop such resistance<sup>29-31</sup>.

The ADMET properties are important in the lead optimisation process as they dictate how well a compound will succeed within the body. Lead optimisation also addresses changes that will positively improve these properties (ADMET), as well as activity.

### 1.1.3.2 Absorption

Absorption, in terms of pharmacology, refers to the ability of a compound to move from the target site and into the bloodstream<sup>32</sup>. For a drug to be absorbed it must be in solution, meaning that in solid forms, such as tablets (oral drugs), the tablet must be able to dissolve under the relevant conditions<sup>33</sup>. Generally, oral drugs need to be able to cross cell membranes

and can do so in a variety of ways, such as passive diffusion, facilitated passive diffusion, active transport and pinocytosis<sup>33</sup>.

The key physicochemical properties for absorption are both solubility and permeability<sup>34</sup>. These properties are affected by several physical properties, including the molecular size and lipophilicity, and therefore can be modelled to estimate oral absorption. Such models have been published in attempts to increase the absorption including in 1997 the well-cited Lipinski's Rule of Five<sup>35</sup>.

During the lead optimisation process, 'rules' provide rough guidelines to ensure that a particular compound falls within a range that is likely to have favourable ADME properties. These properties are assessed through various assays. Generally, chemists aim to reduce clearance by human microsomal metabolism, human hepatocyte metabolism and rat hepatocyte metabolism. It is also favoured to reduce hERG inhibition (less risk of cardiotoxicity) and human Caco-2 efflux ratio so that the drug stays in the cell long enough to have an effect. Whereas, chemists aim to increase the aqueous solubility and Caco-2 intrinsic permeability as they aid in the drug's ability to get to its desired location in the body. LogD tends to have a neutral preference of property value change, as it is a physical property that correlates with many other endpoints<sup>36</sup>.

#### 1.1.3.3 Distribution

Distribution is formally concerned with the movements of a drug between the blood and tissue<sup>37</sup>, although how the drug is distributed within the tissue has also been assessed<sup>32</sup>.

In the drug discovery landscape, plasma protein binding is assessed experimentally as it may give insight into the drugs behaviour particularly, its distribution<sup>38</sup>. In addition, to predict how much a drug has been distributed throughout the body, the administered dose is divided by the concentration of plasma<sup>39</sup> (Equation 7) .

$$\text{Volume of Distribution} = \frac{\text{Administered Dose (mg)}}{\text{Concentration of Plasma } \left(\frac{\text{mg}}{\text{L}}\right)}$$

Equation 7: Equation to assess the extent of drug distribution throughout the body

#### 1.1.3.4 Metabolism

Metabolism refers to the process of the metabolic breakdown of a drug within the body and is a two-phase process, with phase zero and phase three, of which are frequently cited in the



literature, referring to entry into the cell and export. In the first phase, the drugs may be oxidised, reduced, or hydrolysed for the purpose of introducing a reactive group, while in the second phase reactions may include sulfation, acetylation or methylation to conjugate with polar moieties<sup>40</sup> as these are easier to excrete.

The family of cytochrome P450 enzymes are important in the role of drug metabolism, for which the expression of these enzymes is influenced by a variety of factors including genetic, polymorphisms, hormones and the general metrics of the individual (such as age, sex etc.)<sup>41</sup>. Differences in drug response amongst patients is also an important consideration<sup>40</sup>.

#### 1.1.3.5 Excretion

There can be confusion amongst the terms drug excretion, drug elimination and drug clearance, with many individuals citing drug elimination and excretion as the same thing<sup>42</sup>. However, they can refer to separate processes in different publications. Such as drug excretion<sup>43</sup> is the loss of the drug from the body, drug elimination<sup>44</sup>, describes the removal of the drug from the body or is described as the loss of the drug from the site of measurement within the body. Drug clearance<sup>44</sup> refers to drug elimination without the identification of the mechanism of the process.

The kidneys, as well as the liver, are the major organs, although not restricted to, in the body that deal with drug elimination. When excreted from the kidneys, drugs are excreted by glomerular filtration and by active tubular secretion. The more bound the drugs are to plasma proteins, the less likely they are to be filtered by glomerular filtration, whereas, this factor is independent for tubular secretion<sup>45</sup>.

#### 1.1.3.6 Toxicity

The toxicity of around one-third of drug candidates is estimated to be the reason for their attrition and contributes heavily to the high costs of drug discovery<sup>46</sup>. The toxicity of a drug can be caused by several reasons including chemical, on and off target toxicity. On target toxicity is caused by adverse effects at the desired target, whereas, off target toxicity refers to those caused by other targets that are modulated by the drug, often undesired<sup>47</sup>.

It is extremely important to ensure that compounds target only the proteins they need to and not also other proteins that may cause adverse side effects. For this reason, when searching for new drug targets, in diseases caused by a pathogen such as a parasite or a bacterium,

targets that are found in the microorganism but not the human host are desirable. This means the drug can only interact with the pathogen and not the human host, thereby increasing selectivity (and decreasing toxicity) in this way in many cases.

#### 1.1.4 Assay Test Results

A variety of different assays are performed by chemists to assess how a compound will fair within the body, particularly in areas such as permeability and solubility. Generally, *in vitro* assays on their own do not represent a whole human system; therefore, the predictions of adverse effects must be carefully considered.

To assess permeability and absorption, chemists analyse Caco-2 cell monolayer in and out of the cell, allowing them to understand and model the intestinal barrier (most notable the human small intestinal mucosa). Caco-2 cells originate from human colorectal adenocarcinoma and growth in monolayer epithelial cells.

Chris Lipinski discussed<sup>48</sup> screening doses on Caco-2 permeability assays and explained that if the dose is too low, then you can overestimate the importance of efflux transporters but if the drug concentration is too high, the transporters in which the compound translocate via can become saturated. Due to the heterogeneous nature of the cells, a variety of transports (absorptive and efflux) may be present and results will vary from laboratory and laboratory. Caco-2 cells develop into small intestinal cells, however, cell differentiation takes at least 21 days<sup>49</sup>. They are sensitive to excipients, which means that the original potency and functionality may not be well represented. Additionally, they do not tolerate organic solvents and so when studying in this situation, caco-2 cells are not ideal<sup>50</sup>.

LogD (distribution coefficient) is the log of the partition of a compound between the lipid and aqueous phases and represents a compound's lipophilic nature and is used to predict the *in-vivo* permeability<sup>51</sup> (Equation 8).

$$\log D_{\frac{Oct}{Wat}} = \log \left( \frac{[Solute]_{octanol}^{ionised} + [Solute]_{octanol}^{un-ionised}}{[Solute]_{water}^{ionised} + [Solute]_{water}^{un-ionised}} \right)$$

Equation 8: Equation to calculate logD

Frequently, logP is used in place of logD; however, logP only considers neutral compounds rather than ionizable compounds. LogD accounts for a molecules pH dependence in aqueous solution<sup>51</sup>. This is an important factor as the human body does not maintain a single pH,

instead it changes throughout the body as well as whether the individual has fasted or been fed<sup>52</sup>. Comparisons of mathematical and experimental calculations are not easily made as logP varies in range between the two methods<sup>53</sup>.

Solubility is another important property that is measured. It is the property of a solute to dissolve in a solvent. It has been suggested that over 40% of new chemical entities are insoluble<sup>19</sup>. Solubility is measured as the saturation concentration so when adding more solute, the concentration in the solution does not increase<sup>54</sup>. There are several techniques that formulation scientists can use to improve the solubility of the compounds. These methods can include changing the pH and using buffers or even the use of novel excipients<sup>19</sup>. Solubility can be influenced by the species in the solvent, however, the USP (United States Pharmacopeia)<sup>55</sup> and BP (British Pharmacopoeia)<sup>56</sup> have classified solubility, regardless of the solvent used. It is based on part of solvent that is required per part of solute. The lower the part of solvent required per part solute, the more soluble it is considered.

Following this, microsomal metabolism is analysed in both human and animals. After death, human livers are extracted as soon as possible<sup>57</sup>, whereas for animals, they are bred for preclinical studies. Microsomes are not present in living, healthy cells, but are reformed from parts of the endoplasmic reticulum during laboratory procedures to break up cells<sup>58</sup>. Microsomes are used, to assess the metabolism of compounds and they can express key enzymes in the drug metabolizing process, most notably cytochrome P enzymes. It is estimated that 60% of marketed drugs are substrates for CYP enzymes<sup>20</sup>. However, it is important to note that the expression and activity of these enzymes varies greatly from individual to individual as they are affected by a wide range of genetic and environmental factors<sup>41</sup>.

As well as microsomal metabolism, hepatocyte metabolism is also assessed in vitro. Hepatocytes, make up approximately 80% of the livers mass<sup>59</sup>. However, this analysis is limited by hepatocyte donors as well as the ability of adult differentiated hepatocytes to proliferate in culture<sup>60</sup>.

Finally, functional scientists analyse hERG (human Ether-à-go-go-Related Gene), which contributes to the hearts electrical activity and is used in assays to assess cardiac toxicity. Due to this contribution, it is a common reason that drugs fail preclinical testing when the compounds interact with this target, as interruptions to the hearts electrical activity can be fatal. Therefore, the risk is weighed up against the disease the compound is being used to treat. For example, a compound, that is being used to treat a non-life-threatening illness, that interacts with hERG that could potentially induce torsade de pointes arrhythmia. Cisapride is

a drug that was removed from the market for this reason and it was shown that it only induced torsade de pointes arrhythmia in approximately 1 out of 120,000 patients<sup>61</sup>. The hERG binding assay, is not a functional test and therefore does not highlight the agonistic or antagonistic effects of the compound and it does not identify which state the compound binds to on the channel, nor which sites it binds to<sup>22</sup>. Therefore, it requires compounds to be followed up with a functional assay test such as by use of a patch-clamp assay. This measures the interactions between a compound and hERG and can be automated or manually performed. Manual methods, are low throughput and technically difficult, but good for determining the IC<sub>50</sub> of compounds. Automated methods, can be performed at high throughput and is cheaper, but offers less flexibility<sup>22</sup>.

## 1.2 Compound Sources and Comparison

Generally, there are different types of compound sources available, such as publicly available datasets such as, ChEMBL 21<sup>62-64</sup>, datasets that require paid access to such as GOSTAR of which also includes data from patents<sup>65,66</sup> and internally available data that is only accessible to those within a particular company such as AstraZeneca.

ChEMBL<sup>62-64</sup> a database that is public and contains a wealth of information for, but not limited to, bioactive and drug-like compounds. This information includes binding, functional and ADMET data and has been manually extracted from literature regularly as well as additional sources. As described by the release notes, ChEMBL 21 (ChEMBL version 21) has been used in the first chapter of this thesis and was prepared on the 1<sup>st</sup> of February 2016 containing:

1,929,473 compound records

1,592,191 compounds

13, 968,617 activities

1,212,831 assays

11,019 targets

62,502 documents

This data is accumulated from 23 different bioassay sources (with much of data coming from scientific literature) and 7 compound-only data sources. Gaulton and colleagues note that the journals selected have been done so because they aim to capture a large amount of high-

quality data. However, there are still gaps and incomplete labelling that continues to be addressed. Furthermore the dataset covers a diverse set of reporting such as targets and bioactivities<sup>62</sup>.

GOSTAR (GVK Bio)<sup>65,66</sup> has been manually curated from scientific literature and patents, chemical data including a compound and its associated target with its reported activity. They have extracted data from approximately 2.2 million patents and 336,426 journals. The aim is to be able to capture as much of the biological and chemical space as possible whilst providing data such as SAR, ADMET, preclinical, clinical and structural<sup>65,66</sup>.

AstraZeneca<sup>67</sup> is a large pharmaceutical company of which employees over 59,000 individuals. It was formed in 1999 when Astra AB and the Zeneca Group plc. merged together. Since the merger, AstraZeneca has been involved in several acquisitions including Cambridge Antibody Technology, MedImmune, Spirogen and Definiens. AstraZeneca, throughout its history, has focused on several disease areas including, anaesthetics, cardiovascular, diabetes, gastrointestinal, infectious disease, neuroscience, oncology and respiratory and inflammatory disease.

### 1.3 Dissemination of Chemistry

Industry has pushed to create compound collections through the methods such as purchase of combinatorial synthesis<sup>68</sup>. Publicly, available data can also be incorporated into a screening library, to aid in an increase in screening library diversity. This data can be found at both scientific literature and patents.

A patent is something that protects novel inventions meaning to use such an invention requires payment to the inventor. Scientific literature is where novel findings are published to aid the wider scientific community. Occasionally, findings will be published in patents exclusively (from private companies); however, publishing in scientific journals usually increases the exposure of the data that might lead to collaboration and further funding opportunities, and it represents additional value both for researchers in companies, as well as being crucial in academia and research institutes. What is communicated depends on where the information is being published; for example, a patent will not necessarily have all the biological activity information such as the activity type but a journal publication may not depict the molecular structures<sup>69</sup>. For instance, it has been shown that patents actually contain more chemical information than publication, and it has even been suggested that they may contain the information up to decades before they appear in literature<sup>70</sup>. Thus, during a drug discovery

program, accessing all the published scientific knowledge around a biological target available through both scientific literature and patents seems crucial.

Time is a tremendously important parameter in pharmaceutical development and numerous studies have been made to measure the time needed for drug discovery and development. Among those, the difference between the launch of a drug and publication dates (the date the drug was published in either a patent or in scientific literature) for oral drugs has been investigated.

The decision *via* which route to publish a protein modulator is dependent on several factors. These can include the need to protect the intellectual property of the compound structure (as in the case of patents), or to spread novel findings that can be used by the scientific community (as in the case of scientific publications). Moreover, without contradicting the observation made in reference<sup>70</sup>, a protein modulator could be first found in scientific literature rather than in patent since the first published bioactive compound to a given target in either a patent or scientific literature may differ. However, it is conceivable that a novel structure has been first published in literature and then later patented as part of a formulation (a mixture such as an active compound and other ingredients found in a tablet) rather than the compound on its own. Additionally, it is worth mentioning that due to formulation patents, a compound can appear in multiple patents<sup>71</sup>. Furthermore, a compound can have already been disclosed in a previous patent if the use is different and is not mentioned in the old patent.

The dissemination of chemistry is an important aspect for scientists to aid in finding relevant information and chemistry, to support their work in the lead optimisation process. Particularly, as the chemistry found in patents is likely to be the type of chemistry identified at the end of the lead optimisation process (as worth the investment to protect the intellectual property).

Improving compounds in the lead optimisation process, is a key focus of this thesis, most notable through the means of matched molecular pairs.

#### 1.4 Matched Molecular Pairs

Matched molecular pairs (MMPs) can be described as two compounds that are identical with exception of a molecular fragment that differs in the same position<sup>72</sup>. An example of a MMP is shown in Figure 2, showing two compounds that are identical with exception of a chlorine whereas atom (on the left) being replaced with a fluorine (on the right). The term was first used in the book *Cheminformatics in Drug Discovery* written by Kenny and Sadowski<sup>72</sup>. Matched

molecular pairs are useful for observing step by step how the changes in the compounds chemistry affect properties.

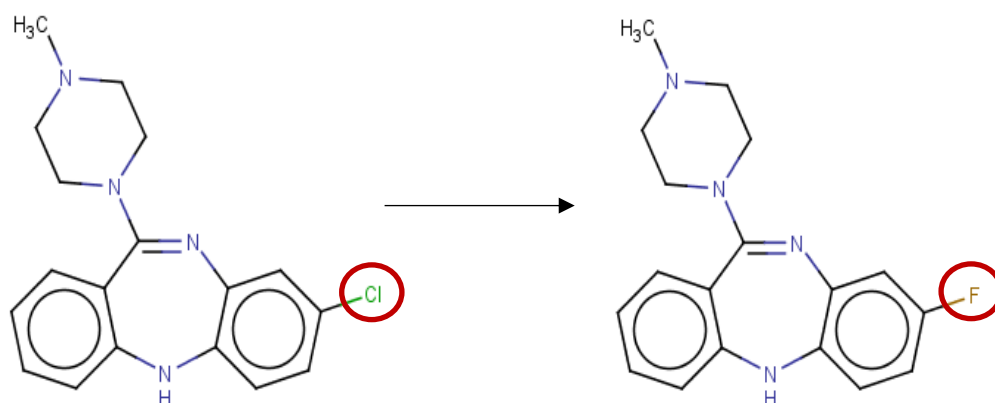


Figure 2: Visual example of a matched molecular pair, the molecular fragment transformation is from a chlorine to a fluorine

MMP algorithms can be split into two categories, namely supervised and non-supervised methods. The difference between the two categories is that in supervised methods the transformations are predefined (the transformations that make the MMP) whereas in unsupervised methods, an algorithm is used to identify all the potential pairs<sup>73</sup>.

Unsupervised methods frequently use maximum common substructure algorithms, whereas, supervised methods such as fragment-based methods rely on known transformations. Therefore, for all algorithms that are available, the advantages and dis-advantages can be described by these two methods. For example, un-supervised methods can identify new MMPs whereas, supervised methods have precise control of what the MMP is.

Table 2: Key MMP algorithms available

| Algorithm                   | Based upon  | Method type    | Advantages   | Disadvantages  |
|-----------------------------|---|----------------|--|--|
| Huassain-Rea <sup>74</sup>  | Hussain-Rea Fragmentation   | Non-supervised | Computationally efficient  | Does not allow for considerations of the environment |
| Gleeson et al <sup>75</sup> | Requires a partial definition of the transformation to identify MMPs and then | Supervised     | Can identify MMPs where the substructure that is specified is a core or a terminal group | Computationally inefficient                          |

|                            |  |                |  |  |
|----------------------------|--|----------------|--|--|
|                            | performs multiple substructure searches  |                |  |  |
| Hajduk et al <sup>76</sup> | Pairwise comparison of compound using findsubs routine (Daylight) – uses specified transformations | Supervised     | Uses specified substructure for a more targeted identification process | Limited to terminal or side group changes only                               |
| ThricePairs <sup>77</sup>  | Specified transformations and SMARTS   | Supervised     | Yields good Tanimoto scores of which suggests chemical diversity       | Yields a low number of transformations                                       |
| WizePairs <sup>78</sup>    | Maximum common substructure and SMIRKS   | Non-supervised | Can capture the local single site environment                          | Can be applied to larger datasets however, the authors example is very small |

The algorithms involved have all been summarised and discussed<sup>73</sup> very recently and shows that a range of methods are available, although, some are proprietary. Table 2 highlights the key MMP algorithms that have been developed.

The Hussain-Rea<sup>74</sup> methodology takes all molecules that are inputted and enumerates all acyclic single cuts and then indexes each fragment into the start and ending fragments. This therefore allows identification of the transformation. However, this algorithm also allows for double cuts to be made and the methodology follows the same as single cuts. The Hussain-Rea fragmentation method forms the basis of several applications of MMPs by such as that of Matsy<sup>79</sup>, a knowledge-based methodology to predict R groups that are likely to improve biological activity.

The Gleeson et al<sup>75</sup> methodology on the other hand requires the specification of a substructure (x) and then the method performs multiple substructure searches. To be more precise, it starts with the identification of all compounds that contain substructure x and then to remove x from one of the identified compounds (Y). Therefore, Y with x removed is itself used as a substructure query to identify new compounds Z. Those compounds (Z) that are identified are a MMP with Y. Therefore, the starting molecular fragment is x and the ending molecular fragment (y) is the result of Z-(Y-x).

An example of unsupervised methods include the WizePair algorithm<sup>78</sup>. This method is based on the maximum common substructure approach to identify the potential MMPs of which are



then verified to ensure that the MMPs are located at a single site and encoded in SMIRKS reaction notation, and is able to capture the local single site environment<sup>78</sup>. The method can be applied to large datasets. The authors use this method on some set 11-histone deacetylase inhibitors where the system could be used to apply medicinal chemistry knowledge from one project to the next. In addition the method allows for common bioisosteres identification<sup>78</sup>. An example of a supervised algorithm is that of ThricePairs which has defined transformations and SMARTS<sup>77</sup>. The ThricePairs method is an in house proprietary software and due to the defined transforms, yields a low number of matched pairs, but those that did had a desirable mean Tanimoto score, suggesting chemical diversity. The authors use their method on a large dataset to assay in vitro human liver microsomal turnover assay results<sup>77</sup>.

MMPA can be further extended to a match molecular series where instead of just comparing two compounds you extend this to a series of compounds<sup>79</sup>. Again, each of these compounds in the series is identical with exception of a single molecular fragment in the same location. This type of analysis allows for medicinal chemists to analyse trends in activity over a project<sup>79</sup>. The limitations of many methods is their time-consuming nature relating to computational efficiency and is often a problem with calculating MMPs<sup>74</sup>. Whilst MMPs have great use in understanding activity cliffs<sup>80</sup> and differences between compounds in terms of similarity, the methodology is not without its limitations<sup>73</sup>. Furthermore, it is advised to include contextual information when using MMPs as it has been shown that in cases of predicting hERG inhibition, solubility and lipophilicity, the prediction ability is enhanced<sup>81</sup>. The algorithms that derive the MMPs can themselves be a source of limitations. Those using a set of predefined molecular transformations of which are used as a starting point are limited by the fact that transformations that differ from those in the starting source will not be identified<sup>81</sup>. This limitation can be avoided by using Most Common Molecular fragment algorithms however, despite being shown they can work well on large datasets<sup>82</sup>, they can be computationally exhaustive<sup>81</sup>.

## 1.5 Maximum Common Substructure

The Maximum Common Substructure (MCS) is used to detect the largest identical substructures between two compounds and thanks to its unsupervised approach can lead to the identification of novel matched pairs<sup>83</sup>.

A 2D structure is represented as atoms (vertices) and bonds (the edges) whereas a 3D structure again represents the atoms as vertices, but the edges as represent the geometric

distance between the vertices. It can be shown that 2D graphs, due to their sparseness, have approximately equal number of edges and vertices whereas for 3D chemical graphs, there is an edge between each pair of vertices and so the number of vertices is approximately the square of the number of edges (Equation 9 and Equation 10) <sup>84</sup>.

$$|E(G) \approx O|V(G)|$$

Equation 9: Equation showing the number of edges in proportion to the number of vertices in 2D representations

$$|E(G) \approx O|V(G)|^2$$

Equation 10: Equation showing the number of edges in proportion to the number of vertices in 3D representations

MCS can donate two types of graph subtypes, the maximum common induced subgraph (MCIS) and the maximum common edge subgraph (MCES). These represent a graph with the largest number of vertices (with edges in between) and edges, respectively, common to the two graphs being compared. The MCS can also be split into whether it is connected (there is only one subgraph – each pair of vertices forms the endpoints of a path) or disconnected (multiple subgraphs) and therefore, MCSs can lack uniqueness in terms of a number of substructures can be determined from two compounds<sup>84</sup>. In short connected MCSs are composed of only one fragment whereas disconnected MCSs can have many fragments.

The advantage is that the MCS able to screen all compounds in a collection against each other allowing for the potential identification of novel transformations not previously known. Furthermore, there is that there is a clear define between the two compounds in that only the smallest change will be observed as the difference i.e. as a transformation X to Y. Essentially the backbone is left of the compounds that are identical and a single molecular fragment that differs between the two compounds at the same position. It is likely that this backbone is chemically important, particularly in terms of compound activity.

Improving how MMPs are computed have been discussed where the study highlights that maximum common substructure (compares two molecules and identifies the largest possible substructure that is identical between the two) and the fragment and index (cleaves the acyclic single bonds and compares all possible fragments between the two molecules) methods are the most prevalent methods at finding rules. It is suggested that combining the two methods increases the effectiveness of finding rules<sup>85</sup>.

It has been shown that using a MCS based similarity measure is more effective than atom-pair based methods when it comes to searching chemical databases and compliments the atom-pair based methods well. The authors suggest using a hybrid of the two methods for prediction models, namely, bioactivity prediction models<sup>83</sup>.

## 2 Innovation in Small-Molecule-Druggable Chemical Space: Where are the Initial Modulators of New Targets Published?

### 2.1 Introduction

An increasing number of novel druggable targets have been identified over the years as well as a plethora of compounds being identified and published. Analysing this data in a time course manner can allow researchers to understand preferred modes of publishing modulators of protein targets, as well as to identify trends over time. This study aims to achieve this goal by examining compounds and their associated targets over time in the two main avenues of dissemination, namely patents and peer-reviewed scientific literature.

The main objective of this study is to try to understand where pharmaceutical innovations in the form of new modulators of protein targets reported as a function of time. For achieving this, we investigated whether the first bioactive compound (a compound that has been shown to have activity on a target) for a novel target tends to be primarily published, either in patents or in scientific literature. In the remaining thesis, we refer to a protein modulator (a compound and the target it has been associated with by a measured activity), as a compound that has a bioactivity ( $IC_{50}$ ,  $EC_{50}$ ,  $K_i$  and  $K_d$ ) ( $\leq 1\mu M$ ) on a particular biological target (ENTREZ\_GENE IDs) and a bioactive compound ( $\leq 1\mu M$ ) as a compound that has activity on a target (identified as ENTREZ\_GENE ID).

In one study, the authors noted that the earliest publication date for oral drugs usually corresponds to a patent<sup>86,87</sup>. Nevertheless, the analysed dataset size was small (592 drugs), mainly because it was restricted to launched drugs for which all necessary information could be identified. Additionally, a previous study analysed a small number of protein modulators and considered the delay of the publication of these annotations in scientific literature, after having been published in a patent. In this study the authors found that on average there is a four year delay between publishing a patent to scientific literature for compound-target interactions which also highlighted the need for scientists to be able to search patents reliably<sup>88</sup>.

Thus, not restricted to approved drugs, our work will cover a much larger number of protein modulators than previous work, namely *all* first modulators of protein targets, independently of whether this resulted in an approved drug later or not.

The sources for scientific literature and patents used in this work are on the one hand ChEMBL a large open access bioactivity database, in which those the protein modulators that were

published in scientific literature where studied<sup>63,64</sup> and on the other hand GOSTAR which is a family of commercial databases manually curated from publicly available scientific literature as well as from patents<sup>66</sup>.

In the first section of this work a comparison of the analysed scientific literature and patents datasets is presented. Following this, we analyse from which publication sources (patent or scientific literature) novel protein modulators could have been found over time. In addition, we will also investigate whether the result has been affected by the 18-month delay in patent between filing and publishing. Going into more detail, we next analysed from which publication sources novel protein modulators have been identified, depending on target class and year bin. Overall, we were hence aiming to highlight and understand where novel chemistry is first published allowing for a knowledge-based approach to identify information as and when required.

## 2.2 Materials and Methods

### 2.2.1 Extraction and organisation of the GOSTAR dataset

Data was extracted from GOSTAR (GVK Bio)<sup>66</sup> using SQL via SQL Developer (Version 4.0.1.14)<sup>89</sup>. The dataset was curated (see below) with the use of KNIME (Version 2.11.2)<sup>90</sup>. The SMILES were standardised using an in-house program<sup>91</sup>. A year bin was assigned to each published year where the data was analysed every year from 1990 to 2014 with all data originating from before 1990 being assigned as historic. The target class was added to the dataset based on the EGID (target class annotations to EGIDs had been previously assigned in house with exception of epigenetic target classes and the full list is published in the supplementary of the corresponding manuscript<sup>92</sup>). One Uniprot ID can have multiple EGIDs due to having different family members as an example, however, one EGID was assigned to one uniprot in this analysis – duplicate uniprot IDs were randomly removed. The epigenetic target class were also added to the dataset, matching the EGIDs for the labelled epigenetic protein families<sup>93</sup> to the EGIDs in the file after duplicates were removed. The “other” target class comprises all targets that did not fall within the other target class labels or the ENTREZ-GENE ID had not been assigned to a UNIPROT name. Kinases were separated out from enzymes due to their high therapeutic interest for analysing their trends. Therefore, kinases and enzymes are treated as two separate target classes. Only human targets were retained, and the earliest instance of the compound-target being recorded was retained.

Rows where the Micro\_Molar\_Prefix (of which determines whether the recorded activity value was greater than, less than, equal to etc) was set to equals were retained to maximise the accuracy of the activity value. Compounds with an activity of  $\leq 1\mu\text{M}$  were retained (any activities  $<0$  were removed) and any data that was from a source of "Other" was removed due to low numbers. Furthermore, activity types that were reported as  $K_i$ ,  $IC_{50}$ ,  $EC_{50}$  and  $K_d$  were retained.

The GOSTAR dataset used in this study has been then split into two parts depending on the source of the data, namely GOSTAR Patent and GOSTAR Journal. Finally, our dataset has been further subdivided using the protein target classes, eight of which will be distinguished and investigated here, namely enzymes, epigenetic targets, G protein-coupled receptors (GPCRs), kinases, ion channels, nuclear hormone receptors (NHRs), transporters and "other" targets. It is important to note that the sources we used are not exhaustive, and hence the analysis presented is meant to show trends and preferences in publishing bioactivity information, as opposed to representing in every case numerically comprehensive results.

The Molecular Weight was calculated using RDKit<sup>94</sup>. A small number of compounds were removed due to failure to calculate their molecular weight. Compounds with a molecular weight larger than 900Da were removed. This left the dataset with a total of 221,429 and 338,093 unique protein modulators for GOSTAR Journal and GOSTAR Patent respectively. Duplicate were removed during the pre-processing of the data whilst retaining the first instance of a protein modulator only.

## 2.2.2 Extraction and organisation of the ChEMBL 21 dataset

The ChEMBL 21 <sup>63,64</sup> file was extracted using Toad for MySQL 5.0.0345<sup>95</sup>. A total of 3,504,431 rows were collected containing the following fields, Accession, ID (compound), Canonical SMILES, Activity (standardised values), Activity units, Activity Relations, Year, Activity type as well as all reference columns (where it was published, reference, volume number, issue number and title). The standard value is not null, and the polymer flag is = 0. Additionally, the assay type needed to be 'B' or 'F' and the assay confidence score had to be  $\geq 8$ . The confidence score of  $\geq 8$  includes homologous single protein target assigned, and this matched protein target level that had been extracted from GOSTAR. As only human targets were used, this will have had little effect on the results and was selected to capture a complete picture of what is being published and where. We also believe we have struck a good balance by including bioactivity data, when there was no species information given, since in most cases the protein target studied has been the human orthologue. Additionally, we have included data where the protein is human, but the organism is non-human. However, we did not want to go

below a confidence score of 8 to minimise the chance of inaccuracies. When extracting directly from ChEMBL, there is a difference of 326 accessions between extracting a confidence score of 8 and 9 or 9 only. These 326 accessions are a confidence score and make up ~12% of the total accessions extracted from ChEMBL. A column called REFERENCE was added to the dataset showing where the target annotation was published. Any missing values were removed. The ChEMBL 21 SMILES were standardised using an in-house method at AstraZeneca<sup>91</sup> and the two files were joined together, on the compound ID, using the Joiner node in KNIME (Version 2.11.2)<sup>90</sup>. The SMILES standardised using the in-house program were used in the study for consistency with the other datasets. Data published after 2014 in ChEMBL 21 was removed from the analysis to ensure the year had been adequately captured and updated in all data sources. To label the target classes, firstly a file containing the accession numbers was uploaded to <http://www.UNIPROT.org/uploadlists/>. From the drop-down list Uniprot KB AC/D to EGID was selected and just the EGID was taken after using it to select the UniprotKB column, the following information was extracted: Entry, Your List, EntryName, Rev/UnRev, Organism ID. Only human data was used. The file was sorted by reviewed/ unreviewed. Duplicates were removed from the reviewed based on what was first integrated into UniprotKB/ Swiss Prot. Where the date was the same, the one with the highest number of publications (including additional computationally mapped reference) was retained. Duplicates were also removed from unreviewed EGIDs so each EGID was only represented once. The gene annotations were joined together with the EGID to give the target class. Target class had been previously annotated to EGIDs and included all of the classes included with exception of epigenetic which was compiled separately<sup>93</sup> of which was added to the file after duplicates had been removed. As with the GOSTAR data, kinases were separated out from enzymes. Duplicate protein modulators were removed using a shell script. Before being written out to a csv file the file was split into two GOSTAR Patent and GOSTAR Journal and duplicate protein modulators were removed from each dataset and concatenated back together before being read back into KNIME<sup>90</sup>. Duplicate results were removed during the pre-processing of the data whilst retaining the first instance of a protein modulator. Only rows where the units were nM were kept. Additionally, activity values of  $\leq 1\mu\text{M}$  (any row with an activity reported as  $\leq 0$  was removed), instances where the activity relation was “=” to the value and only those values that were reported as  $K_i$ ,  $K_d$ ,  $IC_{50}$  and  $EC_{50}$  were retained. The molecule weight (using parallel chunk nodes to calculate molecular weights in parallel) was calculated, calculated very small amount of compounds were removed due to failure to calculate the molecular weight, using RDKit nodes<sup>94</sup> and those compounds with a molecular

weight of  $\leq 900$ Da were retained. This was read out to a csv file. In total 276,650 rows were used for analysis.

### 2.2.3 Visualisation of data

The output file was read into TIBCO SPOTFIRE (Version 6.5.2.26)<sup>96</sup> where all visualisations were produced.

### 2.2.4 Comparison of patent vs. public datasets and the distribution of the years difference and number of targets for each year's difference

A comparison of the first compound that is active on a target that was published in a patent compared to the first compound that is active on the same target that was published in a patent was performed to see where the first compounds for a target are published first. KNIME (Version 2.1.1.2)<sup>90</sup> was used to manipulate the data as follows. First the data was split into GOSTAR Journal, GOSTAR Patent and ChEMBL21. For each dataset, the data was sorted by EGID and then the year. This allowed the retrieval of the first compound to be published for each target. Duplicate EGIDs were then removed leaving the first instance. GOSTAR Journal data was merged with GOSTAR Patent and ChEMBL 21 data was merged with GOSTAR Patent to allow for the comparison. The journal published year was taken away from the patent year to get the years difference and any duplicate EGIDs were removed. A year difference of 0 indicates that the annotations were published in the same year. To show the effect of the 18-month difference between filing date and published date of a patent, the year's difference had one and a half years subtracted from it to demonstrate the effect on when and where a target annotation was published. This was repeated for four activity bins ( $\leq 10\mu\text{M}$ ,  $\leq 1\mu\text{M}$ ,  $\leq 0.1\mu\text{M}$  and all available activities) and three datasets (enzyme, kinase and GPCR). It was also used to test how the filtering had affected the result by using two different filtering methods and can be used to understand how errors may affect the results. The first had no filtering for either patent data or public data; the second had filtering applied to it only for the public data. The reason for testing this was that a target may have been published in 2003 but because of the activity filtering it was not recorded as being first published until 2009 as an example. Therefore, it was important to explore the effect that such filtering has on the results observed.

This testing (to see the effect that the filtering has on the results) was completed as follows: was removed (including the prefix of the result having to be equal to the result etc.) from both datasets (public (from literature) and patent data). As the standard relations for activities from ChEMBL that were extracted were ('~', '=', '<', '<=', '<<', '<<<' or '<') this did not include



standard relations that were '>', '>>' or '>>>' and therefore the ChEMBL file was extracted again but this time allowed any type of standard relation to capture more data. Additionally, any type of activity type was extracted ( $K_i$ ,  $K_d$  etc.).

#### 2.2.5 Comparison of patent vs. public datasets and the distribution of the years difference and number of molecular frameworks, topological frameworks and compounds for each year's difference

When applied to molecular frameworks, topological frameworks and compounds, the set up was performed slightly differently. As with all previous analysis annotations with an activity of  $\leq 1\mu\text{M}$  was used. The data was split into GOSTAR and ChEMBL and the SMILES were cast as SMILES to be used in the RDKit node Find Murcko<sup>94,97</sup>. A Murcko Scaffold (molecular framework) removes side chain atoms and retaining the central ring structure with some exceptions (non-ring systems that are required to connect two ring systems together as well as the first atom that has been branched off from the straight chain via a double bond), the topological framework is the generic structure of the framework<sup>94,97</sup>. Molecular or Topological Frameworks that do not have a ring structure are written as NA and treated as one. GOSTAR data was then split into patent and literature. Duplicate structures were filtered using the Filter Duplicates node from MOE<sup>98</sup> KNIME nodes, by comparing Standard InChI keys<sup>99</sup>. The public datasets were concatenated together to enable the public data to be read out separately for future analysis (as was the patent data). The datasets were joined together on their molecular frameworks / compounds / topological frameworks to identifying overlapping molecular frameworks and then these were concatenated together. Additional duplicates were filtered out the years difference was calculated (patent – public published year), and molecular frameworks that could not be calculated (those without a ring structure) were removed.

#### 2.2.6 Comparison of patent vs. public datasets and for each target class and each year bin when and where was an annotation published first

To observe when an annotation had been published regardless of whether it was in a patent first or a publication first the data used the output from the previous analysis (determining whether their distribution of years difference and number of targets for each year's difference) for  $\leq 1\mu\text{M}$ . This was split into patents first or journals and year bins were used (<1990, 1990-1994, 1995-1999, 2000-2004, 2005-2009, 2010-2014 and 2015). 2015 was excluded from visualisations due to the minimal amount of data captured in this year bin.

### 2.2.7 Comparison of patent vs. public datasets for observing trends in annotations that were only published in patents or only published in literature

To see trends in the publication of annotations in either patents or literature, but not both, first the literature data was concatenated together and duplicate EGIDs or structures were removed. The first instance of each EGID annotation or structure was retained for each dataset (as performed in the previous analysis to get the year's difference and the number of targets for each year's difference). EGIDs and structures that were unique to each dataset were retrieved and retained and everything was concatenated back together and these were read out to a csv file which was loaded into Spotfire<sup>96</sup>.

### 2.2.8 Statistical Validation

Prop-tests<sup>100</sup> and pairwise-prop-tests<sup>101</sup> with p-value adjustment method of Bonferroni were performed in RStudio – Version 0.98.1103 to confirm significance of findings. The alternative hypothesis used was two sided<sup>102</sup>.

### 2.2.9 Compound Novelty

The molecular fragments from the originally extracted compounds from ChEMBL and GOSTAR (prior to any processing) were compared to those that had been published in patents first or patents for understanding the novelty of the chemistry by determining whether the molecular framework had originally occurred before those observed in the study. The structures from the originally extracted compounds were not standardised however the standard InChI keys<sup>99</sup> were used to make the comparisons calculated in rdkit<sup>94</sup>.

### 2.2.10 Compound Similarity

The similarity of compounds was assessed by measuring the most similar compound, in terms of its structure, to each compound in the set analysed. The similarity was represented by a Tanimoto score and was compared by Morgan fingerprints in RDKit with a radius of 2 and a bit vector of 2048. When analysing the similarity between the first compounds to be published in literature first against a target against the first compound to be published associated with the same target but in a patent, a distance matrix was calculated first, after which the pairs of compounds that had the same target were extracted from the matrix to give the Tanimoto distance of each compound (published in either source) that were associated with the same target. Finally, in order to get the Tanimoto similarity 1- Tanimoto distance was calculated.

## 2.3 Results and Discussion

### 2.3.1 Number of Unique Compound-Target annotation analysis

We first analysed the trends the publications of compound-target annotations over time and by data source. The number of unique compound-target annotations, which is defined as the number of unique targets having at least one reported bioactive compound, have steadily, grown over the years supporting previous work<sup>103</sup> for all sources studied here (Figure 3(A)). These increases may be the result of a series of factors such as the progress of screening automation (e.g. HTS) that allowed for a greater number of compound-target annotations to be discovered, the increased investments in drug discovery in academia, as well as the generally increasing number of scientific publications<sup>103</sup> and patents<sup>104,105</sup>. On the other hand, there appears to be a difference between the number of annotations abstracted in GOSTAR journals, over the last few years, and the ones reported in ChEMBL by about 10,000 instances. The gap between the cumulative sum of unique compound-target annotations in ChEMBL and GOSTAR patent widens from 1993, with ChEMBL containing more unique compound-target annotations than GOSTAR Patent from that date. However, this difference is significantly reduced by 2014 showing that, as time passes, more compound-target annotations are being published in patents (as abstracted in the respective databases).

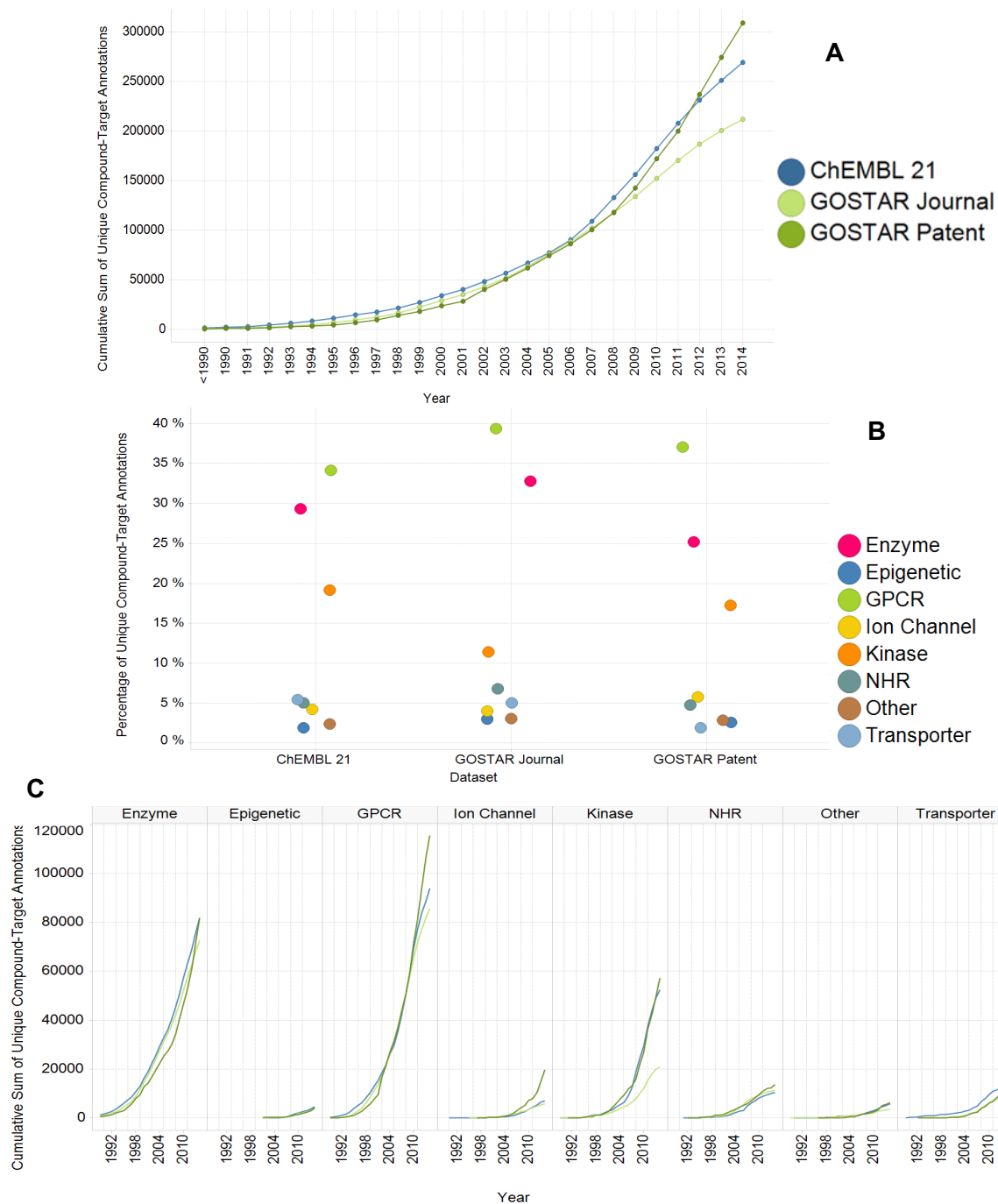


Figure 3: Protein modulators from different data sources as a function of time. Unique protein modulators have steadily increased over the years for all datasets with the target class preference in each dataset varying. Additionally, the strongest increase in unique protein modulators over the years has occurred for enzymes, GPCR and kinases. The numbers of unique protein modulators published are presented over the years (A) and for each target class (normalised to 100% for each dataset across target class) (B), the points have been jittered for easier viewing. (C) Shows for each year the cumulative sum of unique-compound target annotations published for each target class. Protein modulators presented have an activity of  $\leq 1\mu\text{M}$ .

In Figure 3(B), the percentage of bioactive compounds for each target class with respect to the dataset is displayed. Although the trends in slopes for ChEMBL and GOSTAR Journal are similar, there are several differences which are likely due to how the databases curate their

journals and which journals are covered. For example, according to the ChEMBL FAQ the literature coverage in ChEMBL focuses on approximately 47 journal papers, however this changes as new versions are released<sup>106</sup>. It can be observed that the GOSTAR Journal dataset has a higher percentage of compounds being associated with epigenetic targets (2% of the dataset) compared to the other datasets as well as the highest number of compounds associated with enzymes (comprising 34% of the dataset) and NHRs (5%). The percentage of compounds associated with epigenetic targets is low compared to the other target classes in all three datasets, which likely reflects the novelty of the class in terms of therapeutic interest. A significant difference between the percentage of bioactive compounds associated with enzymes in GOSTAR Journal and GOSTAR Patent (approximately 12%) can be observed, while this difference is small between GOSTAR Journal and ChEMBL (approximately 5%). Overall, this suggests that compounds associated with an enzyme target seem to have preferably been reported in scientific journal rather than in patents.

ChEMBL and GOSTAR Patent have similar and high percentages of modulators for kinases (18-19%). This is probably related to both higher target promiscuity but also the high therapeutic relevance of this target class. The reason why there are fewer kinase associations in GOSTAR Journal than ChEMBL is likely due to differences in the curation of information such as which journals are abstracted. Compounds annotated as being bioactive against ion channels are more represented in GOSTAR Patent (6% of the dataset) compared to 3% in both literature sources.

In the next phase of our study we analysed the cumulative sum of unique compound-target annotation binned per year for each target class is shown in Figure 3(C). It is observed that the increase in unique compound-target annotations for a given target class in patents follow the trend observed in scientific journals in preceding years. This supports the understanding that academic labs primarily investigate the biology on a target and any disease implications (basic research). Once this groundwork has been done, either industry becomes interested (which leads to patents) or academia needs to do more groundwork (such as identifying modulators of the target) before industry becomes interested, which leads to publications in journals first. Despite slight differences, generally the curve trends appear similar. The similarity in curves of the number of unique compound-target annotations between patents and scientific literature is likely due to an increase in published data and the curves represent the cumulative increase. A striking example of such similarity between patent and scientific literature cumulative curves can be found in the case of ion channel ligands, where the first annotations were captured in ChEMBL in 1990 and in GOSTAR Journal in 1993 while it was not until 1997, that an annotation appeared in GOSTAR Patent. The number of epigenetic compound-target annotations, increases at a slow rate throughout history. The number of

bioactive compounds targeting this protein class is likely to increase further as pharmaceutical companies and academia work together to understand underlying biology better, with the aim to generate novel therapies<sup>107</sup>. In 2013, the question whether GPCR targets are still a source of new targets has been raised<sup>108</sup>. It seems from the data analysed here that GPCRs appear to still be of significant interest (Figure 3(C)) both in patents and scientific literature. The authors from reference<sup>108</sup> found that marketed drugs often target bio aminergic receptors which accounts for only ~ 10% of targets in the GPCR family. Therefore, the reason for the on-going interest in GPCRs may be due to the diverse nature of GPCRs<sup>108</sup> and possibly further exploration into the five main human GPCR families<sup>109</sup>. Interestingly, many compounds associated with GPCR targets (in general) were published in patents which may reflect that the related screening collection were more diverse than estimated in the article.

### 2.3.2 Time based analysis of the source for new compound-target annotations

To identify where the *first* bioactive compounds have been published for a *novel* target we next analysed the difference in the publication years between patents and scientific literature. Figure 6 shows the number of targets that have a published bioactive compound associated with it in both a patent and scientific literature with respect to the time difference between the literature and patents publication dates. Note that this analysis does not pay attention to the compound structure, it only considers the fact that a *modulator of a protein target* has been published in a given location at a given time point. It can be noticed that novel compound-target information is more often published in literature prior to it being published in a patent (547 out of 848 cases, 65%). Patents have an 18-month delay in being published, which can be considered as a significant difference compared to the scientific literature corresponding process of submission-publication. To try to mitigate this in our analysis, Figure 2 also depicts an adjusted curve based on an 18-months period.

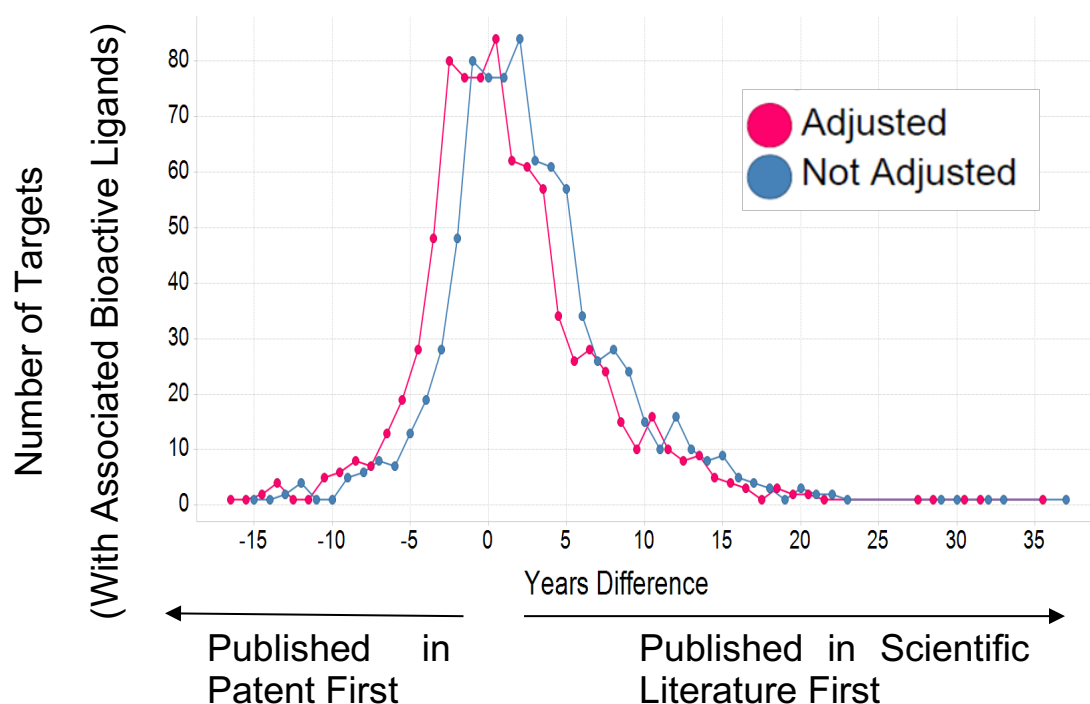


Figure 4: Number of targets (with an associated bioactive ligand) for which the first ligand has been published in a journal or a patent, respectively. Note that compound structures of both instances do *not* need to match in this analysis. The figure shows the difference between the raw dates ('Not Adjusted'), as well as an adjusted value ('Adjusted'), which considers the ca. 18-month time gap between the filling of a patent and its publication. Positive numbers indicate publication first in a journal, negative numbers publication first in a patent. Protein modulators are more frequently published in journals prior to being published in patents regardless of whether taking into account the 18 month gap between the filling of a patent and publishing the patent.

Comparing the distribution to that of the 18-month delay distribution, an increase from 26% of target annotations being first published in patents to 45% is observed, hence resulting in a nearly equal number of first ligands of targets reported *via* either dissemination route. This result is independent of the activity cut-off, with Figure 5 showing the results for various activity cut-offs, hence not supporting the hypothesis that patents more frequently contain more active ligands which are more likely to show activity in an *in vivo* setting. This is confirmed by applying the prop test<sup>100</sup>, which is used to determine that the proportions of protein modulators that are published in patents first, for each activity cut off, are significantly different or not. In this case there is no statistical significant difference between the proportions as the p-value is > 0.05. Therefore, there is a preference to publish in scientific literature prior to publishing in patents. In this figure, the earliest year that a compound was published in a patent was 1980, whereas it was 1960 for journals, which explains the difference in tail length between what was published in a patent first and what was published in literature first.

The data published in literature, can be used in the design and synthesis cycle. This has been shown to be the case in the molecular design cycle where compounds are brought together

from various sources, such as high-throughput screening, fragment screening or from those published to be synthesised and tested to gain knowledge<sup>110</sup>.

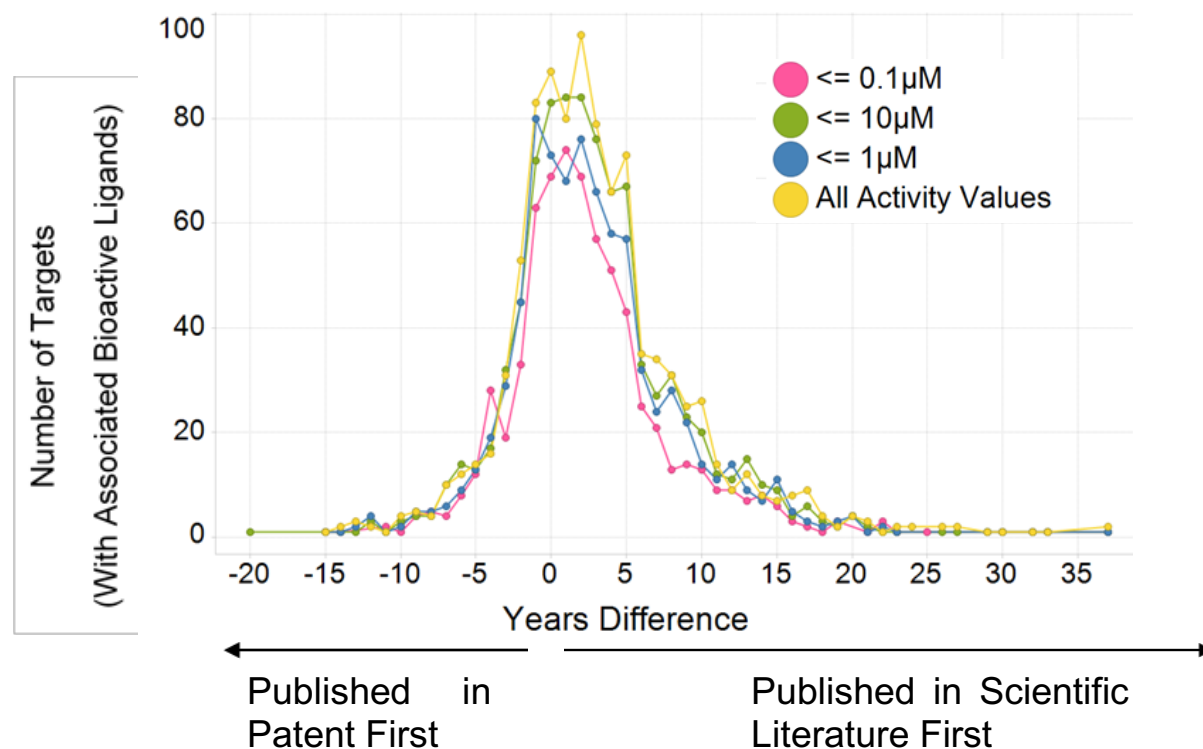


Figure 5: Number of novel targets with an associated ligand for each time difference between publication of a bioactive compound in a journal and a patent, respectively. (Positive numbers indicate publication first occurred in a journal, negative numbers publication first in a patent.) The first bioactive compound for a novel target is most frequently published in journals prior to being published in patents. The plot represents four different activity cut-offs that were explored. All activity values,  $\leq 0.1\mu\text{M}$ ,  $\leq 1\mu\text{M}$  and  $\leq 10\mu\text{M}$ .



We then analysed the effect of removing all filtering (Figure 6). It can be seen that the tail is not significantly cut off on the patent side. This suggests that when a compound that is published with a particular target in a patent first, the same target (although likely associated with a different compound) will be published in scientific literature faster than the reverse (compound associated with a particular target is published in literature first will appear in a patent later, but not as quickly as it does the other way around).

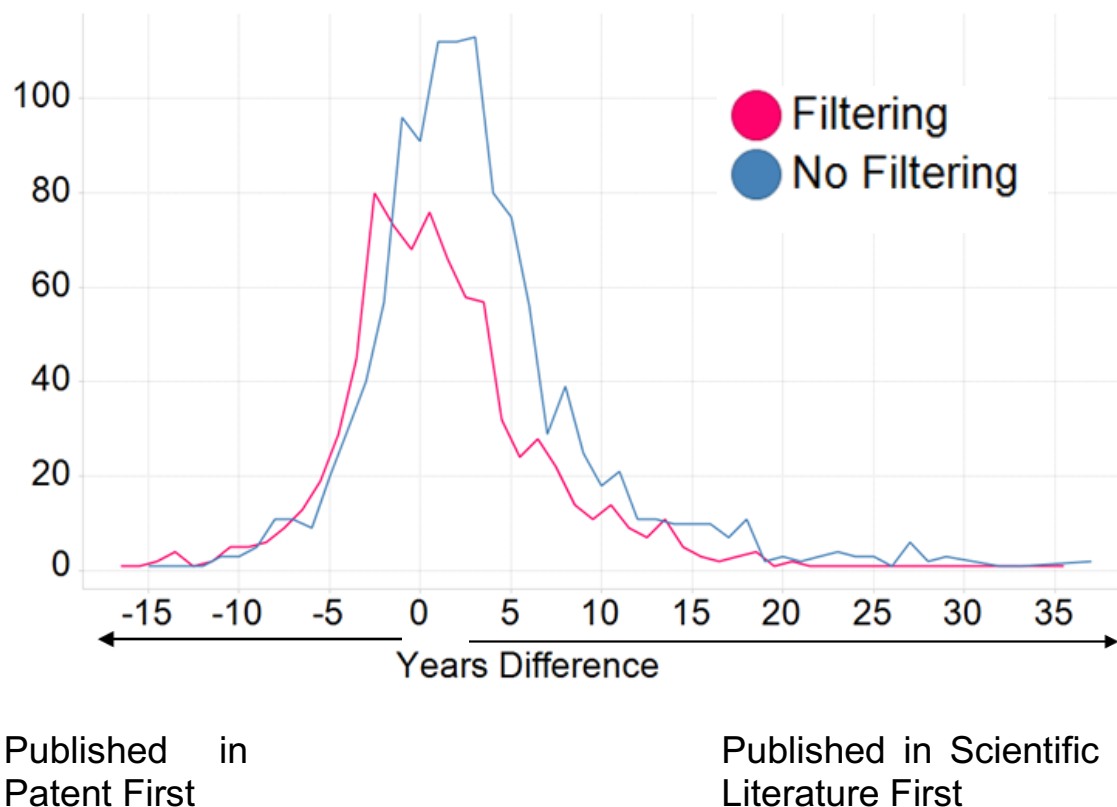


Figure 6: Reducing the strictness of the filtering (removing activity constraints such as allowing all activity types (not just  $K_i$ ,  $K_d$  etc.) and all activity relations (not just values reported equal to an amount)) either on both patent and public data or only on patent data does not affect the trends observed: The number of targets with an associated ligand for each years difference for annotations with no filtering and filtering only on the patent dataset. The years difference is defined by the patent year – the journal year and therefore a positive years difference number denotes that the annotation was published in a journal prior to being published in a patent.

However, not all targets have ligands for them published in both literature and patents, as shown in Figure 7. Here, we explore the proportion of ligands for targets that were published either in both patents and literature, literature only or in a patent only. Thus targets that have been pursued to find patentable bioactive chemical matter are in almost all cases also of scientific interest to publish (45% and 51% of compound-target annotations were published in either patents and scientific literature or literature only, respectively), but there exists many targets where there is scientific interest (as evidenced by scientific publications), but where

there hasn't so far been any interest to identify novel drug candidates (as evidenced by a lack of a patent for a ligand for that protein target). On the other hand, cases where there are ligands patented, but where no ligands have been reported in literature yet, is rather small (only 4%).

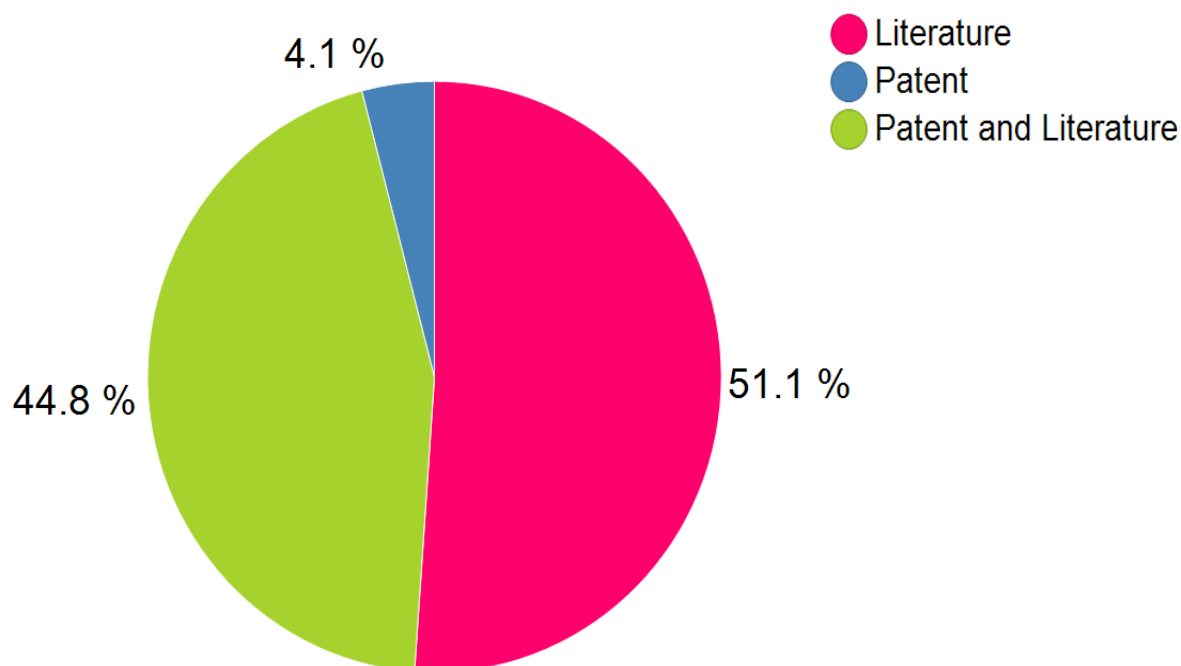


Figure 7: The number targets with associated bioactive compounds that are published in either literature only, patent only or in both patent and literature. Targets are mostly published either in literature only or in both patent and literature; targets with patented ligands, but no ligands reported in literature, are on the other hand rather rare.

The impact of the target class on the route of dissemination was investigated, the results of which are shown in Figure 8. For compounds that are active on enzymes, kinases or GPCRs, the three most frequently published target classes; it can be observed that 25% (49 out of 198) of GPCR targets and their first associated ligand, were published in patents before journal publications. This is approximately 6% more target annotations and 9% less target annotations with their first associated ligand, than the result shows for enzymes (19% (52 out of 277)) and kinases 34% (60 out of 176), respectively. This suggests that the target class impacts when and where the target annotation is published, and is confirmed statistically, where a prop test leads to a p-value of 0.001133 is derived. This is likely the result of changing interests in different therapeutic trends<sup>111</sup>.

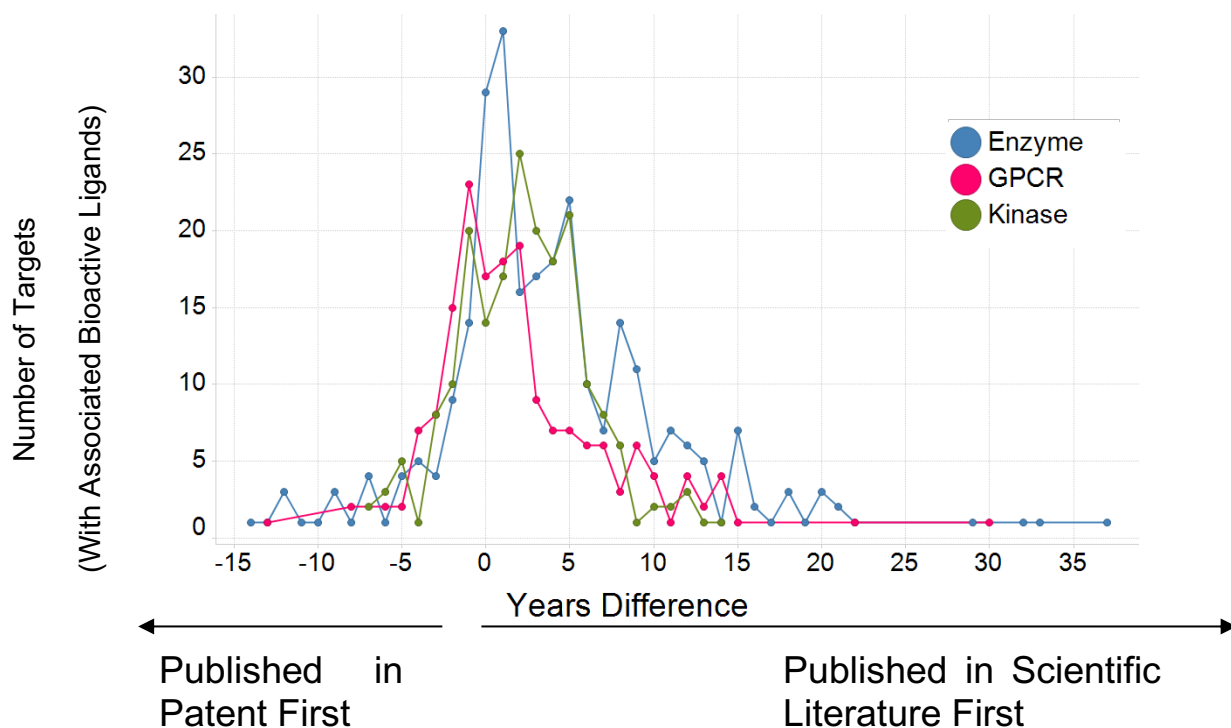


Figure 8: Number of targets (with associated bioactive ligands) for each time difference between publication of an active compound in a journal and a patent, respectively, for Enzymes, Kinases and GPCRs. (Positive numbers indicate publication first in a journal, negative numbers publication first in a patent). It can be seen that when and where a target annotation is published depends to a certain extent on its target class.

Figure 9 captures whether the publication year, in addition to the target class, affects when and where a target annotation is published. The result shows that this is indeed the case, with 39% of kinase target annotations being published in patents first between 2000 and 2004, whereas the percentage drops to 14% between 2005 and 2009 (a p-value of 0.0006569 is observed). The therapeutic relevance, interest and focus of a target or target class at that point in time hence contributes significantly in terms of where information is disseminated.



Figure 9: Number of novel targets (that are associated with a bioactive compound) published in journals (pink) and patents (blue) or both at the same time (green), as a function of target class and time. These pie charts show that the number of first ligands (at an activity of  $\leq 1\mu\text{M}$ ) kinases are increasing over time but decreasing for GPCRs. Individual pie charts are sized based on the absolute number of targets they represent.

In both Figure 8 and Figure 9 it is shown that the target class affects when and where the first active compounds are published for a novel target, for the more exploited target classes. Annotations are usually published in literature prior to patents with exception of those compounds associated with the GPCR (and for some years for NHRs) target class. The GPCR target class has an increasing percentage of target annotations published in patents prior to being published in literature throughout history. No historic compound-target annotations (those published before 1990), for GPCRs, were observed as being published in patents before being published in scientific literature. The number of compound-target annotations that were published in patents prior to being published in scientific literature increased to 14%, 33%, 54%, 61% and 63% in the years 1990-1994, 1995-1999, 2000-2004, 2005-2009 and 2010-2014 respectively. A detailed analysis of GPCR drug targets that have been published prior in patents versus publications in the timespan 1995-2005 revealed several targets related to inflammation like for instance CCR1, CCR2 and CCR3 as well as metabolic diseases like for instance NPY2, MCHR1 and FFAR1. While several small molecules for these targets has reached the clinic, no drug has so far reached the market. The NHR target class, and the

compounds that are associated with them, also shows that in each year bin a large portion of annotations are published in patents first with the highest percentage being 100% in 2010-2014 (where only one novel target was published) and 67% in 1995-1999. However, there are fewer novel targets published in each year bin than compared to the GPCR target class possibly due to the target class size. It has been shown by examining Google trends, that disease trends are observable and can be used to predict potential upcoming disease instances<sup>111,112</sup>.

It is also possible to see how the number of unique targets (published with an associated ligand) has increased for target classes such as kinases (increasing from 7 in 1990-1994 to 91 in 2005-2009) but decreased for others in the same time span such as GPCRs (36 in 1990-1994 and 18 in 2005-2009, with an increase to 52 in 1995-1999) (Figure 5). We can also see the steady increase and any potential plateaus of novel targets (and the first ligand associated with it) for the other target classes. As an example, ion channels saw changes in the year bins, 1990-1994, 1995-1999, 2000-2004, 2005-2009 and 2010-2014, of the number of novel targets and the first ligand associated with it total 1, 14, 15, 15 and 3 respectively. There are years where no novel targets (with an associated ligand) were observed, for example, NHR between 2005 and 2009 and transporters 2010-2014. Finally, trends in target class interest over time. Compounds that are associated with enzymes or GPCRs have increased in interest over time followed by kinases and then followed by those associated with the epigenetic target classes.

### 2.3.3 Analysis of the number of target annotations that were only published in either a patent or in scientific literature via time course analysis

We next investigated the number of target annotations that were *only* published in one of the sources (patent or literature) by observing the number of targets with an associated bioactive compound over time by target class. The result is shown in Figure 10 where it can be observed that a total of 967 target annotations are published in literature only and a total of 77 target annotations are published in patents only.

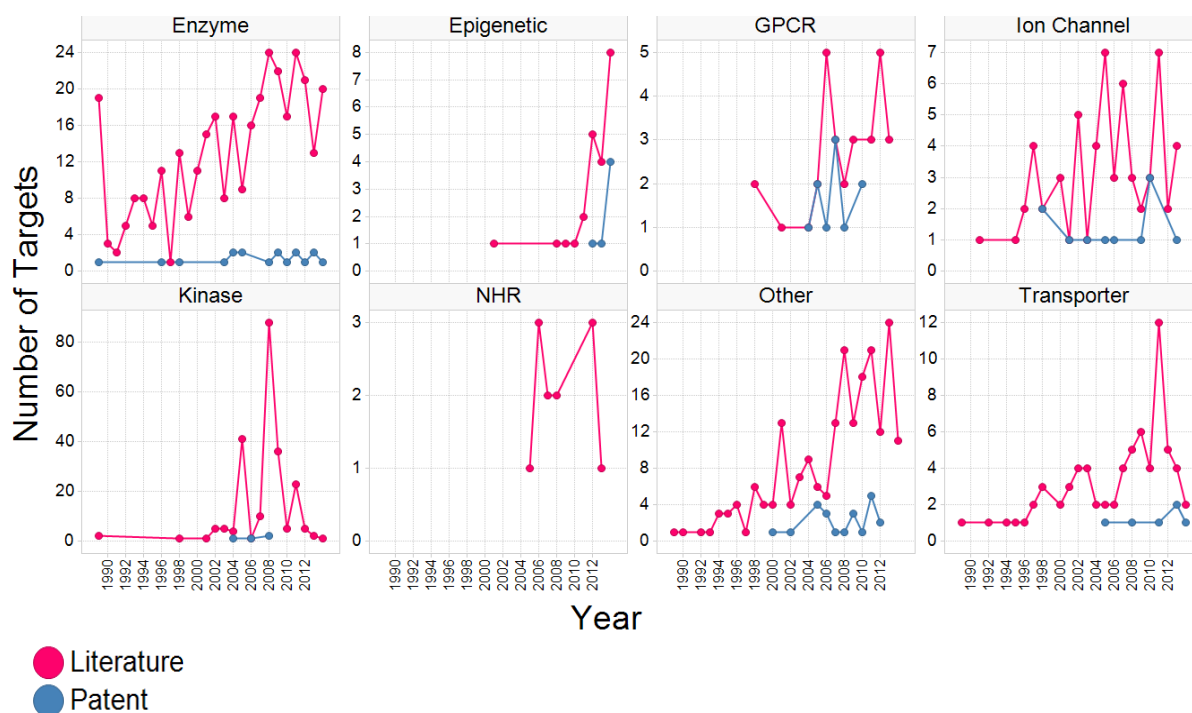


Figure 10: Number of novel targets (with an associated bioactive ligand) published in either only literature or patents, as a function of target class and time. The number of kinase targets in 2008 for in public data was substantially higher than those in patents due to the publishing of the first large scale selectivity panel data. As the years proceeds, for all target classes (patent or public) the number of targets increases or remains steady. The number of targets in each year bin for each target class that were published only in literature or only in patents. The results are for target annotations  $\leq 1\mu\text{M}$ .

The first ligands for enzymes see an increase in targets over the years that were published in literature only (although this does fluctuate between the years), with 3 targets that are associated with their first ligand in 1990 and 21 in 2012 (Figure 10). Another large increase in the number of target annotations can be seen in 2008 for kinases (and their associated ligands) from scientific publications This correlates with the publication of the first large scale kinase selectivity panel comprising an interaction map for 317 kinases with 38 kinase inhibitors<sup>113</sup>. In addition, the overall sales of kinase inhibitors in 2008 were nearly at \$14 billion and increased in subsequent years which emphasises the importance of this target class<sup>114</sup>.

### 2.3.4 Case Studies

We will now give examples of the first ligands from different target classes forming part of this analysis, namely BACE1 (published in literature first), GSK3b (published in literature first) and LRRK2 (published in patent first). BACE1 is an enzyme, first reported in 2000 in the Journal of the American Chemical Society in a study on the design of inhibitors for this target<sup>115</sup>. It was then later published as part of a patent detailing a method of screening for inhibitors for this

gene<sup>116</sup>. The article that GSK3b was published in was focused on identifying a novel compound class that were inhibitors of GSK3b via scaffold hopping<sup>117</sup>. A year later, the target appeared in a patent disclosing pyrrole-2, 5-dione derivatives and their uses as GSK3b inhibitors<sup>118</sup>. Finally, LRRK2, a kinase, was published in literature as part of a kinase inhibitor selectivity analysis<sup>113</sup>. However, this was after it had been published in a patent (of which was looking at compositions and methods for treating Parkinson's disease<sup>119</sup>). There is a wide variety of why compounds are published and patented and often do not result in approved drugs. To our knowledge these genes do not have an approved marketed drug. The number of targets for approved therapeutic drugs is debated but recently the number 324 is used, across target classes<sup>120</sup>.

### 2.3.5 Analysis of where the novel bioactive structures (compounds, molecular and topological frameworks) were first published

We next investigated when and where novel bioactive structures were first published, now also explicitly taking the structure of the compounds into account. This was performed on three different levels of structural diversity, namely compound structure, molecular framework, and topological framework<sup>97</sup>. The compound structure is the most specific and topological framework is the most generic descriptor. An example is shown in Figure 11.

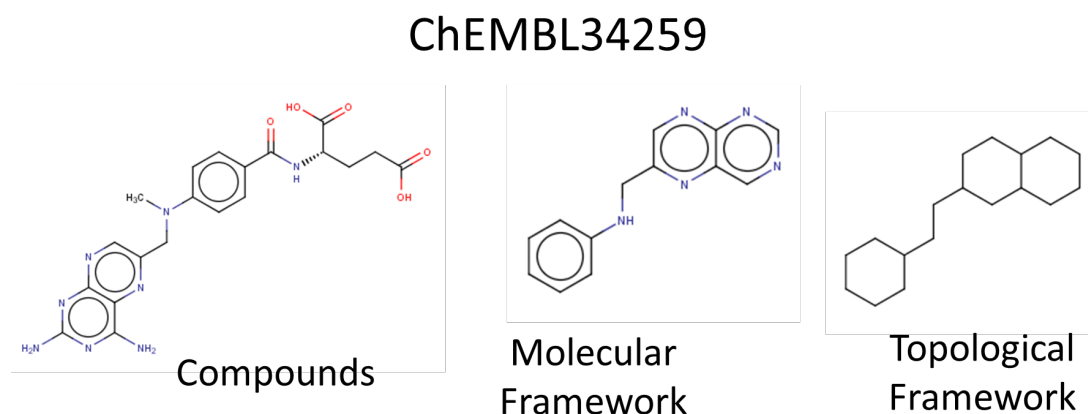


Figure 11: An example of a compound's computed molecular and topological framework. The methodology used to calculate the frameworks is the Bemis-Murcko Scaffold.

In total there are 18,751 compounds published in *both* patents and in the scientific literature, which formed the basis of all subsequent analyses. The distributions seen in Figure 12 are reminiscent of the analysis of the first active compound published for a given target (Figure 4). However, the distribution is shifted to the left, with 61% of compounds having been published in patents first (11,464 out of 18,751 compounds). The percentage of molecular frameworks published in patents prior to being published in literature is 61% (6,670 out of 10,982 molecular frameworks) while the percentage of novel topological frameworks published in patents first drops to 54% (5,065 out of 9,356 topological frameworks). Novel compounds as well as more abstract molecular representations, like molecular and topological frameworks, are published first in patents, which is likely related to the large compound collection comprising novel chemical matter available to the pharmaceutical industry, which frequently result in publication *via* patents. It also demonstrates that protecting novel structures and chemistry is important. The trend is further emphasised when considering the 18 months publication delay for patents. Taking this into account, even 79% of novel structures (14,787 out of 18,751, Figure 12). When considering the adjusted value for molecular frameworks, this represents 74% of the data points (8,176 out of 10,982) as well as 65% (6,122 out of 9,356).

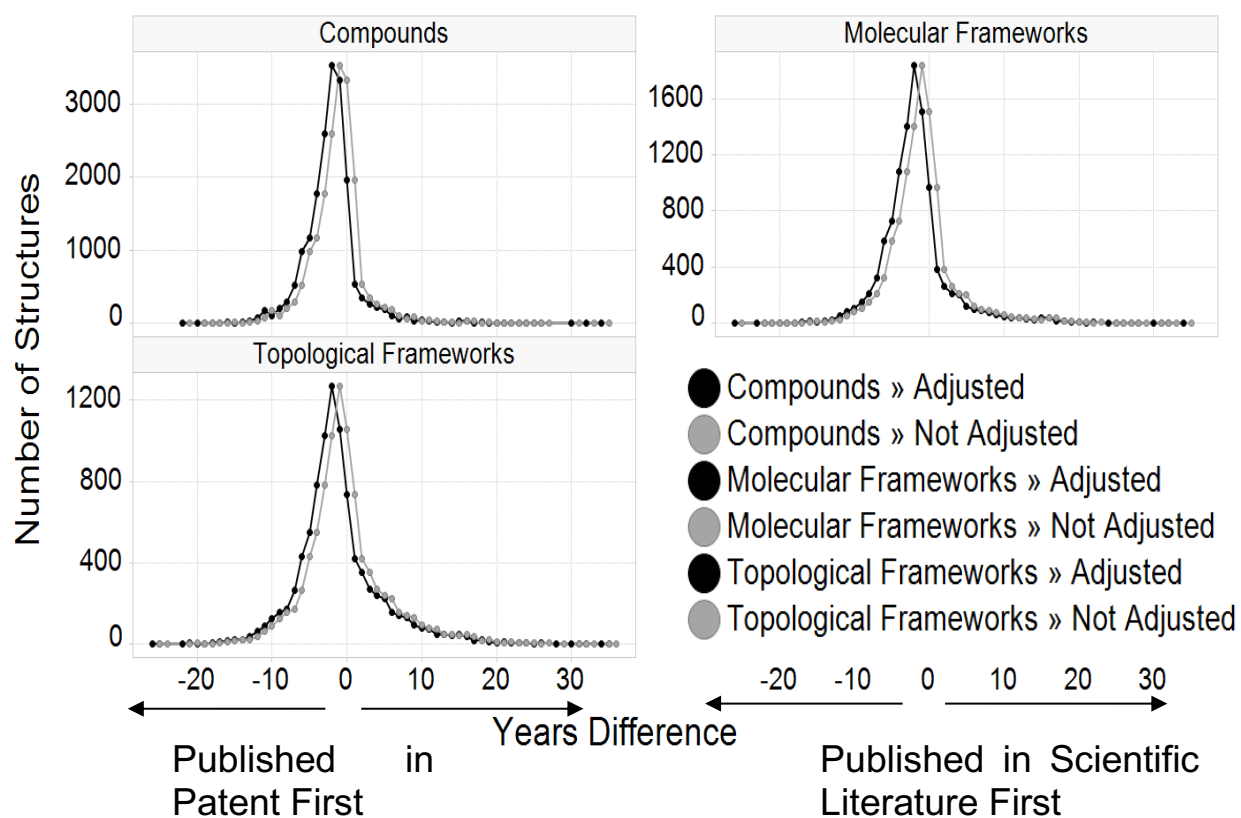




Figure 12: Number of novel bioactive compounds, molecular framework and topological frameworks published in both literature and patents with adjusted and not adjusted values. Adjusted values consider the 18-month time gap between filling of a patent and its publication (Positive numbers indicate publication first in a journal, negative numbers publication first in a patent). The compounds, molecular frameworks and topological frameworks for each year difference showing that most of novel bioactive structures are first published in patents (61%).

We performed a pair wise prop test<sup>101</sup> and adjusted the p-values using the Bonferroni correction method. The three tests (compounds and molecular frameworks, compounds and topological frameworks and molecular and topological frameworks) all gave highly significant p-values of  $<2e-16$ , with exception of compounds against molecular frameworks, which gave a p-value of 1. Therefore, one can conclude that even though the percentage of structures published in patents first varies only slightly (61%, 61% and 54% for compounds, molecular frameworks and topological frameworks respectively), the trend is consistent and significant: the novelty of a structural diversity of a structure, in terms of a full compound or its molecular/topological frameworks (as defined by Bemis and Murcko<sup>97</sup>) influences where they are first published (with exception of compounds and molecular frameworks). The more generic the structure is, the more likely (in relative terms) it is to be published first in scientific literature compared to a patent.

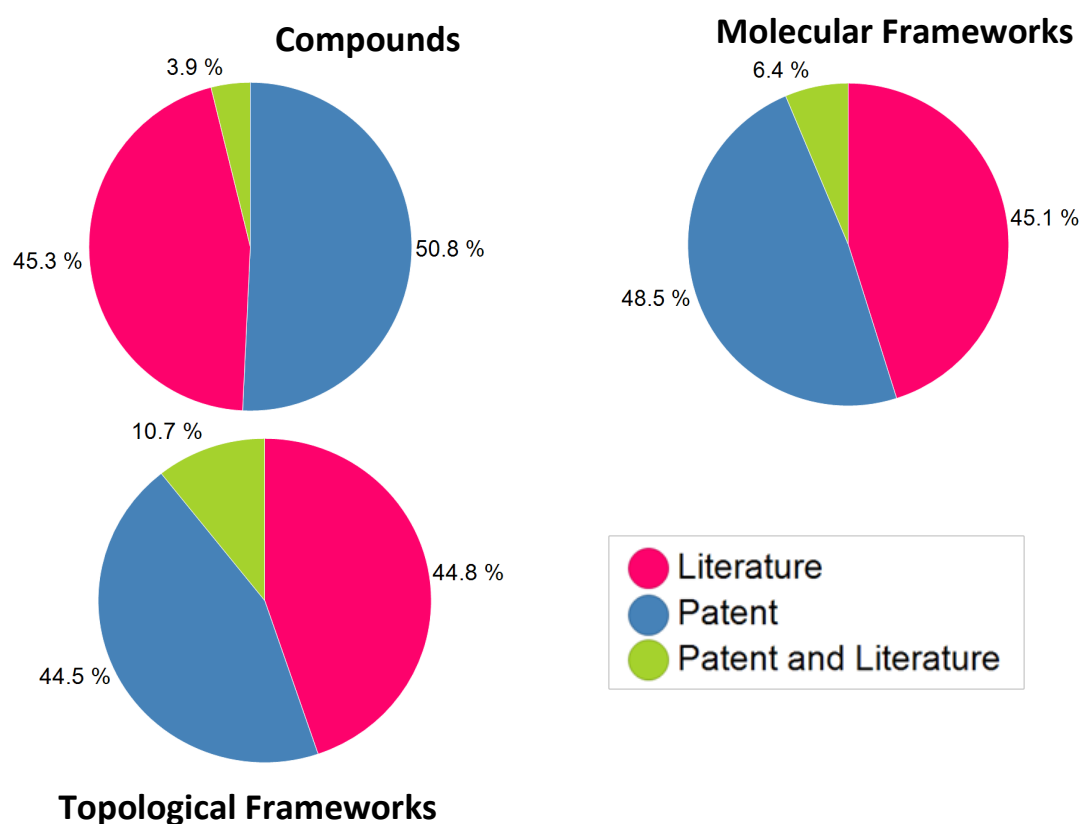


Figure 13: The percentage of compounds, molecular frameworks and topological frameworks that are published in either literature only, patent only or in both patent and literature.

The number of structures (compounds, molecular and topological frameworks) that were published either both in patents and literature, literature only or a patent only is shown in Figure 13. It shows that for all structures there are a slightly higher proportion that are published in literature alone, however, the percentage difference is very small between patent alone and scientific literature alone. Surprisingly the overlap i.e. the number of structures and frameworks published in both patents and literature is rather low. This analysis illustrates that chemical space published in literature and patents is highly complementary, and hence both information sources need to be taken into account when judging the novelty of a given structure as has been previously been discussed<sup>121</sup>.

### 2.3.6 Analysis of the number of structures that were *only* published in either a patent or in scientific literature as a function of time

We also investigated the number of compounds, molecular frameworks and topological frameworks that were *only* published in literature and those that were *only* published in patents (Figure 14 to Figure 16). In all three-structure types, the general trend observed is that the number of compounds published for each target class has been increasing over the years with a slight decrease most recently. In total there are 216,493 compounds published only in literature and 242,586 that are only published in patents (Figure 14) where as there are 18,751 compounds published in both patents and scientific literature. There are 77,603 molecular frameworks that are only published in literature compared to 83,397 that are only published in patents (Figure 15) and 10,982 that were published in both sources. Finally, for topological frameworks there were 39,304 published only in literature compared to 39,060 that were published only in patents (Figure 16) and 9,356 that were published in both data sources. These numbers have been spread across all analysed target classes. Increases in the number of structures for each target class over time. This shows that for all structures being published in only literature or only in a patent, there are published roughly in equal amounts for either source based on the data sources we considered in the analysis performed here.

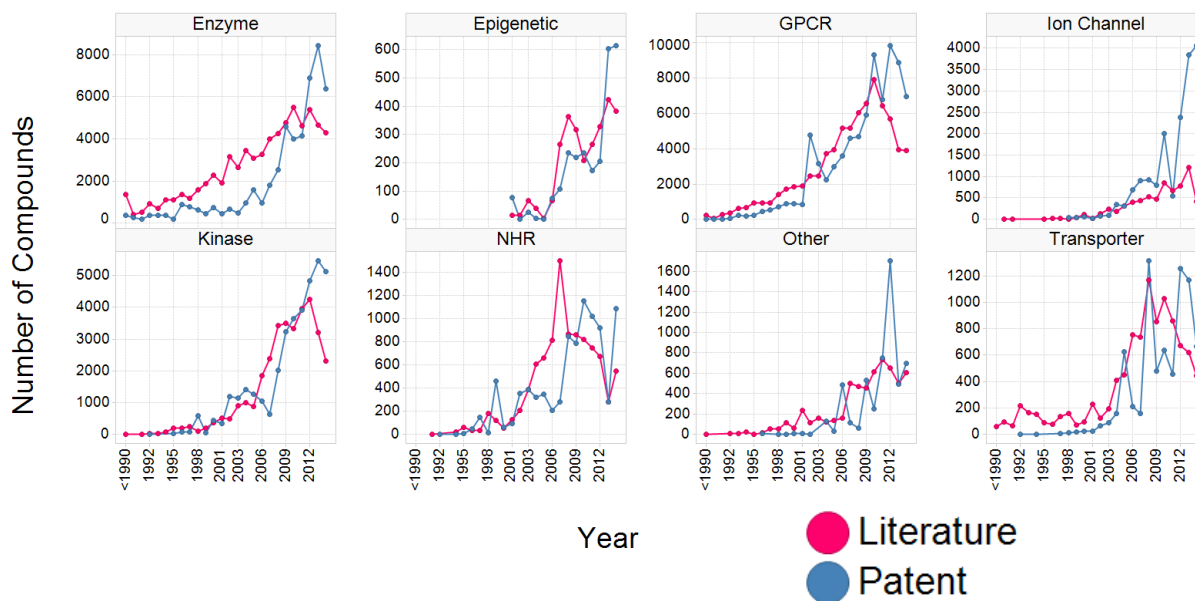


Figure 14: Number of novel compounds published in either only literature or patents, as a function of target class and time. The number of compounds increases throughout the years especially for GPCRs both in literature of patents, despite the decrease in recent years. The year bins included are <1990, 1990-1994, 1995-1999, 2000-2004, 2005-2009 and 2010-2014.

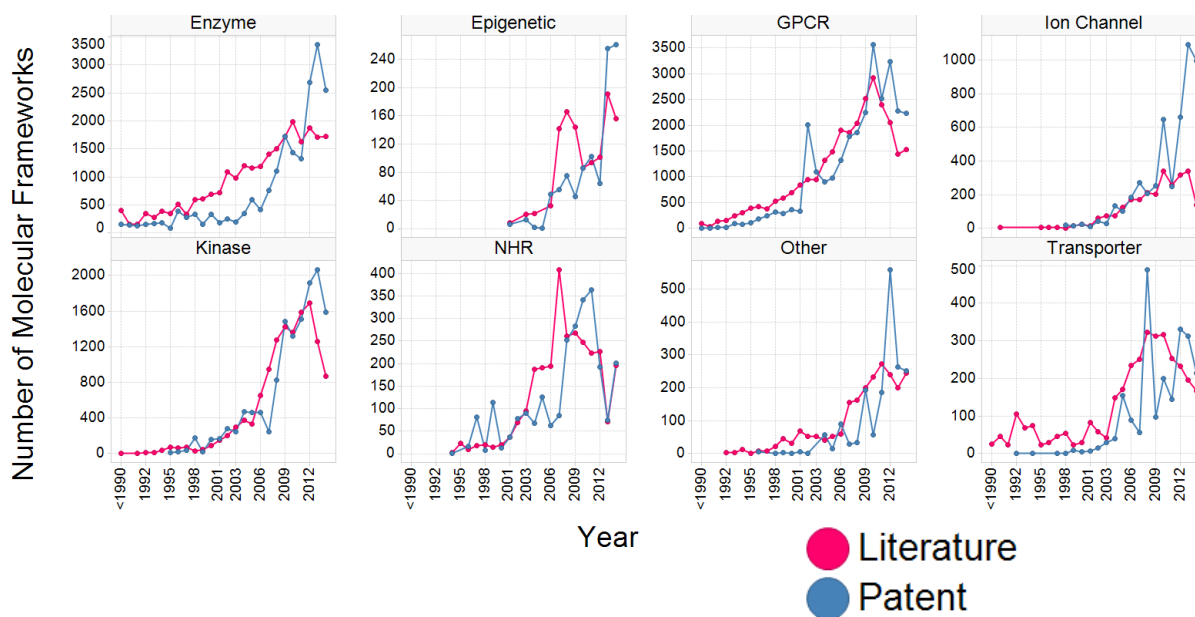


Figure 15: Number of novel molecular frameworks published in either only literature or patents, as a function of target class and time. The number of molecular frameworks increases throughout the years especially for GPCRs both in literature of patents. The year bins included are <1990, 1990-1994, 1995-1999, 2000-2004, 2005-2009 and 2010-2014.

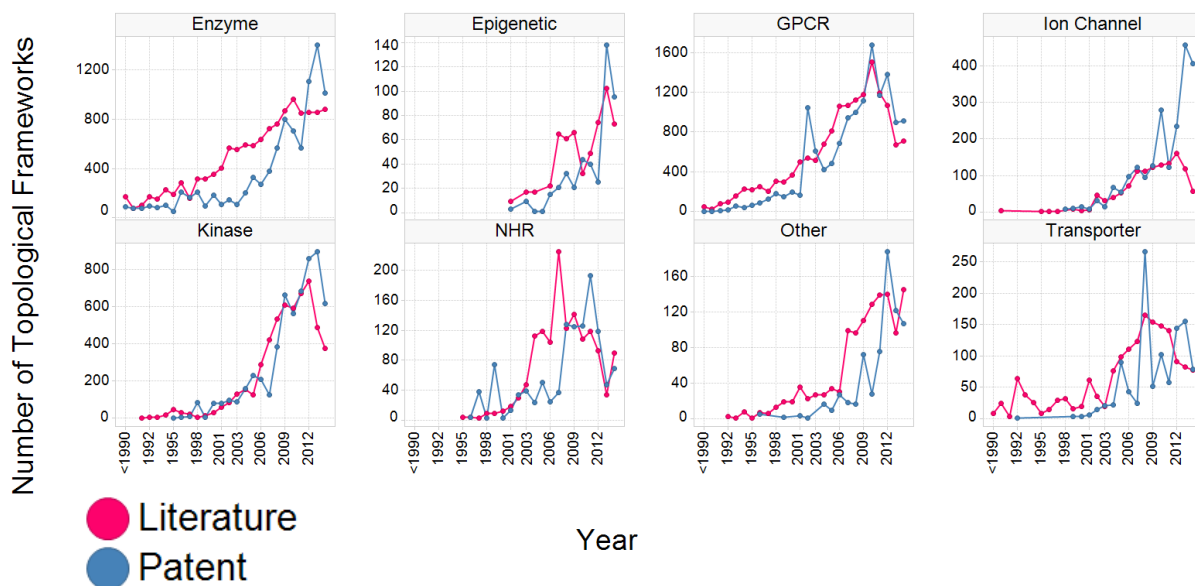


Figure 16: Number of Topological Frameworks published in either only literature or patents, as a function of target class and time. The number of topological frameworks increases throughout the years especially for GPCRs both in literature or patents. Additionally, it can be seen a large increase in recent years for Enzymes in patents. The year bins included are <1990, 1990-1994, 1995-1999, 2000-2004, 2005-2009 and 2010-2014. The activity cut-off used is  $\leq 1\mu\text{M}$ .

Some increases of structures are noted such as the number of compounds associated with NHRs that are only published in patents. In early 2014 a review showed that ROR (Retinoic acid receptor-related Orphan Receptors) and REV-ERB (Nuclear Receptor subfamily 1, group D, member1) were suitable drug targets<sup>122</sup> suggesting that efforts were being made into exploring this target class for novel druggable targets. Additionally, a gradual increase in the number of unique protein modulators is observed over the years for enzymes, regardless of structure type, that were only published in literature where the target is not published in a patent. On the other hand, a greater increase is observed for those only published in patents in recent years, suggesting that more enzyme targets (with bioactive structures) are published more frequently in patents only than literature only (in recent years). A sharp rise in structures active on kinases that are only published in patents is also observed for all three types of structural descriptions but less so in literature only, suggesting that the target class has remained of therapeutic interest and therefore structures associated with kinases, are frequently being patented to address this medically relevant area.

The similarity between compounds published in each source (patent only, literature only and both patent and literature) was analysed (Figure 17(A)) as well as the similarity between the first compound to be published in a journal in association with a particular target and the first

compound to be published in a patent in association with the same target ((Figure 17 (B)). The aim is to show how similar the compounds are across different sources. It can be observed that generally, the compounds in each source have a low similarity to those in another source (Figure 8) We have previously shown that the majority of compounds are either published in either scientific literature or a patent rather than both sources (Figure 13). The results in (Figure 17 (A)) are asymmetrical because for each compound in each source, it is compared to the all the compounds in another source and for each compound for the maximum similarity is reported. For example, in the analysis of Patents Only and Literature Only, each compound in the Patents Only source has the maximum similarity reported out of all the compounds in Literature Only. This explains the difference in curves where the sources that are compared to those compounds that are published in both scientific literature and patents as there are fewer compounds in the source. However, there are some compounds that have a very high similarity as well as some compounds being identified with a Tanimoto score of 1 of which normally suggests that the two compounds are identical, however, they can also differ by their stereochemistry as in this case (Figure 8(A)) where the compounds have been observed as being published in either literature only, patents only or both sources. When comparing the Tanimoto similarity between the first compound to be published in literature for a given target against the first compound to be published against the same target Figure 17(B), but in a patent, shows that the two compounds often differ significantly in terms of their structure. Despite this, there are still 28 targets where their associated compounds (first published in literature and first published in patent to that compound) that have a similarity of 1. This suggests that for these 28 targets, the first compounds to be published in either source for that target were very similar (may differ in their stereochemistry) or the same compound.

Figure 17 (A) shows differences in the curves depending on the source of the compounds and what source the compounds are being compared to. The curves where compounds have been compared from patent only or literature only to those compounds published in both sources show a peak at a Tanimoto score around 0.80 suggesting that there are compounds published in either patents or literature only are like those published in both sources. This may be due to the compounds that are published in both sources have been disseminated further and therefore their chemical space is more readily available. For example, compounds only published in patents may not have their chemical space yet published, therefore you would expect to not find many similar compounds associated with them. However, the same is not found for compounds that were compared from both sources to only one source. The reason for this is likely due to the high proportion of compounds being published in only one source compared to both sources. This will also explain why the curve is shifted to the right for comparisons between compounds published in only one source.

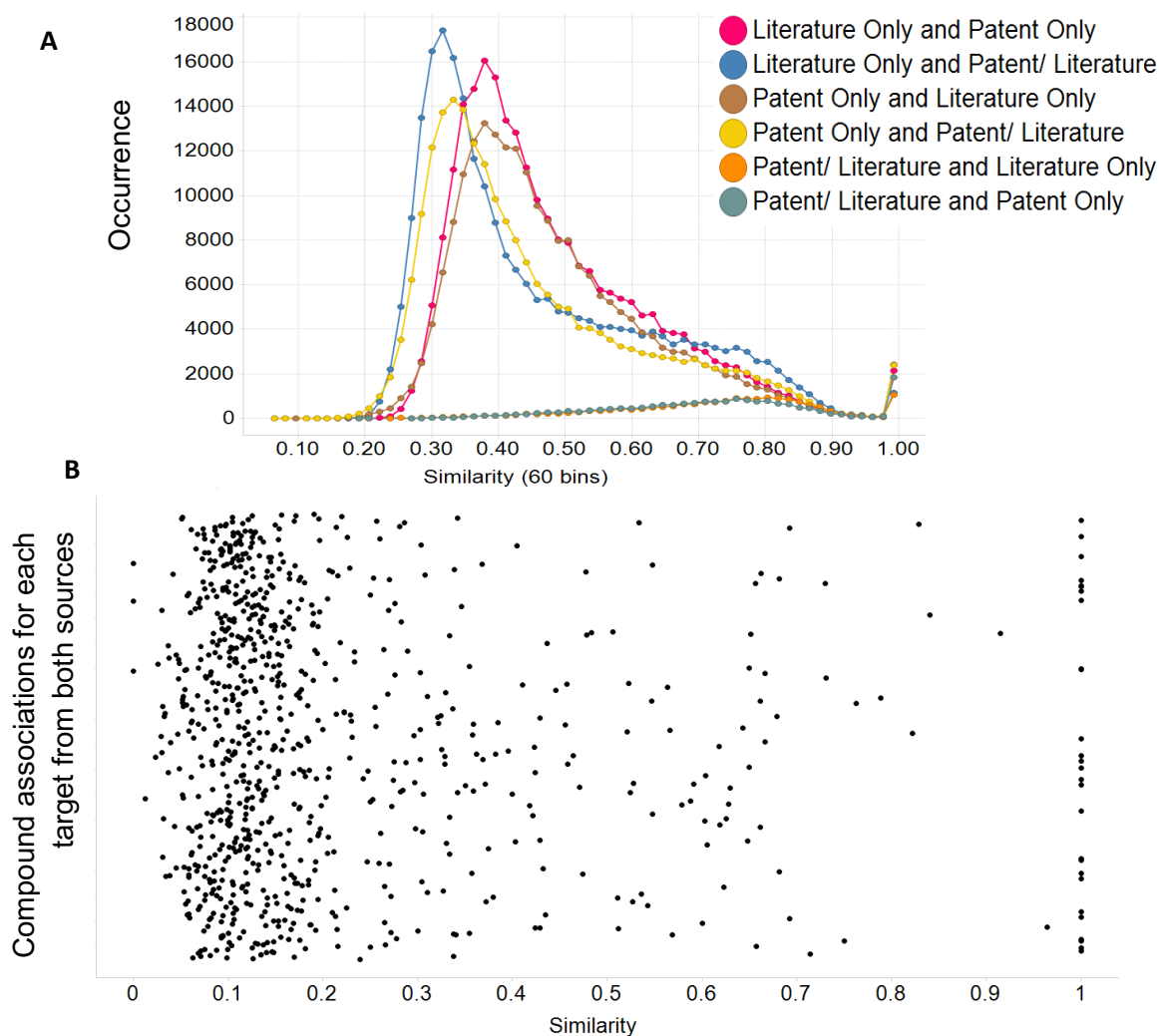


Figure 17: Tanimoto similarity between compounds published in each source (A) and Tanimoto similarity between the primary compound published in literature for a given target and the primary compound published in a patent for the same given target (B). With regards to part (A), for each compound in each source, the maximum similarity was calculated (for example for the label Patent Only and Literature Only, each compound published in the Patents Only source reports the maximum Tanimoto score of the most similar compound to it from the Literature Only source). Most compounds have a low similarity (less than a Tanimoto score of 0.45) to any other compound in the sources. The Tanimoto score has been binned into 60 portions. With regards to (B) The first compound to be associated with a target and the first compound to be associated with the same target but published in a patent shows that the two compounds tend to differ structurally.

To further the analysis, this study considered how many of the compounds published in patents first or in patents only, regardless of the target, had molecular frameworks that had already been previously published. This would suggest that the structure was completely novel and had not originated from a previous compound. To this end, the molecular frameworks of the 227,957 compounds that were published in patents first or a patent only were extracted, which were found to comprise 86,577 unique molecular frameworks (Those without ring systems were excluded). These were joined with matching InChI Keys from the first occurrence of the molecular fragment from the originally extracted data from GOSTAR and

ChEMBL and resulted in 224,931 unique compounds and 85,450 unique molecular frameworks to compare. This considered all compounds that were originally extracted from ChEMBL and GOSTAR and their first published year. It was found that only 1% (4,233 out of the 224,931 unique compounds) published in patents first or patents only had molecular frameworks that had been identified previously. However, the remainder were published in the same year. This highlights that novel chemistry (the molecular framework) is an important factor in patenting compounds.

## 2.4 Chapter overview

As described in this chapter, the number of published novel protein modulators has grown cumulatively over the years. There has been a steeper increase for the number of compounds active on kinases over the years showing that kinases have continued to be a prioritised target class whereas for patents, the number of compounds active on GPCRs has decreased over the years. The number of unique compound-target annotations appears to tail off in recent years, but the same trend is not observed for patents. The size of the target class may also be an important factor to consider as more targets suggests more opportunity for starting new drug discovery projects and therefore more bioactive compounds being produced for these target classes.

We have analysed bioactivity data from patents and scientific literature and found that there is a preference of first bioactive compounds for a novel target to get published in scientific literature earlier than in patents, but structures tend to get published in patents prior to being published in scientific literature. This study takes the first bioactive compound for a novel target published in either scientific literature or patents and therefore the two compounds are likely to be different. This explains why they can be published in literature prior to being published in patents. Target class and publication year have an influence on where target annotations are published. Additionally, when analysing different publication sources (patents only, literature only or both sources) for compounds (and their associated targets), it has been shown that bioactive compounds for a novel target tends to be published in literature only or in both patents and literature but not in patent only. Whereas, structures are likely to be published in either only a patent or only in literature rather than in both sources. These results reflect the fact that patenting is crucial for protecting the intellectual property of the finding, but publishing allows for the discovery to be available to other scientists in the field. This might reflect that for many targets the first molecules discovered are used to study the biology of a target not necessarily for pursuing a drug discovery project.

A caveat with the type of analysis presented here is that there is no guarantee that all active compounds in the scientific literature and patents are covered in the used databases. The addition of other datasets may yield different results. An example of an additional data source that could be used is SureChEMBL<sup>123</sup> of which is a text-mined patent database. This analysis focused on manually curated sources of which is why SureChEMBL was not included, however the incorporation of SureChEMBL would be interesting to look at in the future. This was observed using GOSTAR and ChEMBL where a large amount of data is captured and represented from many patents and scientific literature. The inclusion of more data from these sources (as shown when analysing the effects of the filtering applied to the analysis) demonstrated the effect on the result was small.



### **3 Analysing the Matched Molecular Pair Transformations in Drug Discovery Projects as a Function of Time and Molecular Environment: – Frequency of Molecular Transformations**

#### 3.1 Introduction

ADMET (absorption, distribution, metabolism, excretion and toxicity) properties as well as binding and potency at the biological target are the key focus during lead optimisation<sup>3</sup>. In general, properties and potency will increase<sup>124</sup> from a hit molecule to a candidate molecule. The most important property to modulate during optimisation is lipophilicity as this has an effect on almost all of the other measured properties. The value of modulating lipophilicity is emphasised by the discontinuation of the development of lipophilic compounds<sup>125</sup>. Generally, compounds will need to become more ligand efficient, more permeable, more soluble, less cleared and more bioavailable<sup>126</sup> and those compounds synthesised nearer the end of drug discovery projects are more likely to reaching these criteria. This type of analysis could provide useful insight for future decision-making, and in this study, we explore this using Matched Molecular Pairs (MMPs) and compound registration date.

Previous studies have used matched molecular pairs (MMPs) to understand ADMET rules and aid in compound optimisation. In a recent article, MMPs were combined with machine learning techniques to allow for development of novel compounds<sup>127</sup>. To identify MMPs previous reports have fragmented the compounds based on retrosynthetic rules. The fragments were defined as the smallest possible component, which cannot be fragmented any further. Overall, it was reported that combining MMP analysis with machine learning techniques, in particular deep neural networks, was effective at automating SAR decomposition and prediction.<sup>127</sup> Therefore, the results indicate that using MMP in collaboration with machine learning techniques, can support the compound optimisation process. The validation sets considered two concepts, new fragments and new static core which implicitly considering binding sites. In another study, the authors found that the most frequently used automated matched pair identification methods, were synergistic with each other<sup>85</sup>. These methods were the maximum common substructure and the fragment and index methods and when combined, they were the most effective at transformation identification.

A third study compared the MMPs between different companies to learn ADMET rules has been previously performed<sup>128</sup>. In this study, matched molecular pairs and atomic environments were compared from multiple pharmaceutical companies. They show that the three companies share a total of 58,000 rules, yet each company has a greater number of rules that are only

used by themselves (139k, 84k and 70k rules used only by company A, B and C, respectively). The study highlighted that companies sharing of MMP rules leads to a larger rule set to implement in compound design. Furthermore, the study compared the properties amongst the MMPs<sup>128</sup> and showed that by combining rules across companies there is still good correlation between properties such as log D, solubility, in-vitro clearance and plasma protein binding. Examples identified that the MMP effect on properties were unexpected, such as the case in which a tertiary dimethyl alcohol was replaced with a primary alcohol leads to a reduction in log D without having a significant effect on solubility.

Another study<sup>129</sup> analysed the relationship between experimental uncertainty and MMPA and also showed most common MMP transformations that have been found to have a significant effect on hERG activity from within the public dataset ChEMBL21 were in agreement within in-house Novartis data when the results are statistically significant<sup>129</sup>. An additional paper that is related to this work<sup>130</sup> analyses the types of synthesis used in three different pharmaceutical companies. Here they defined the reaction and the involved molecular fragments however they do not compare atomic environments that each reaction is performed on. All those analyses can be used to determine how to optimize a compound, i.e. to decide which transformation to perform on a molecule to achieve the desired effect.

More recently it has been reported that due to the large amount of data that is available machine learning and matrix analysis can be highly important as a tool to aid medicinal chemists. As chemical intuition plays such an important part in any development, there is great value in ensuring that biases and downfalls of the methods need to be accurately communicated<sup>131</sup>. In addressing the influence of medicinal chemist's intuition, one report discussed how chemists simplify problems, the amount of agreement between chemists on the criteria used, and the accuracy of reporting the relevant criteria<sup>132</sup>. To do this, chemists were asked to select chemical fragments from a set of approximately 4,000. The findings showed that chemists greatly simplify the problem by using few criteria and generally, although there is agreement on what parameters should be used, there was no strong agreement between chemists on how the parameter preferences were determined and thus what constituted undesirable parameters. Overall, the study highlighted that here is a low consensus between chemists<sup>132</sup>. In another study<sup>133</sup>, the authors assessed how consistent the medicinal chemists' opinion is and investigated this along with a compound acquisition program that was conducted as Pharmacia. This particular report also showed a lack of consistency between chemists and highlighted the danger of declaring a compound as undesirable as it is then excluded from further assessment (as well as structurally similar compounds). What we identify is a conflict between chemists' designation of the undesirability of a compound, which inevitably can influence the related computational models.

Following on from these previous studies, in this work also take the time into account, and assess the effect of frequently occurred transformations on frequently occurring environments. To optimise a compound, structural modifications are made in different positions on the molecules to reveal structural-activity relationships (SAR) as well as physicochemical properties. Several MMP transformations can be made between two matched pairs and therefore they can vary in size depending on where the structural change is considered to be occurring<sup>134</sup>. Previous reviews of MMPA in drug discovery have shown that MMPA is able to aid in multiple parameter optimisation by cutting down the number of required design cycles<sup>135</sup>. They highlighted that due to the flexibility of the technique; it can lead to chemists being able to make informed decisions about compounds without having to make them. An automated closed-loop optimization platform has been shown to be successful in the identification of inhibitors for hepsin. This process combines artificial intelligence and automated synthesis as well as biological assays for hepsin<sup>136</sup>.

As multiple parameters need to be optimised simultaneously, there is a need to improve automated systems. There have generally been a number of improvements in experimental procedures, particularly in the automatised area, such as improved technology and robotics, parallelization and miniaturization methodologies, as well as artificial intelligence that are able to help improve a design hypothesis<sup>110</sup>. These improvements, including the potential to reduce the risk of false positives as well as high speed and reduced costs.

A general statistical analysis of the matched molecular pairs used in a pharmaceutical company including information about registration dates and the atomic environment has not, to our knowledge, been published. Our study derived all the matched pairs since 2000 for 453 internal AstraZeneca projects to analyse the most common MMPs in the forward direction, where the second compound was registered after the first, whilst also considering the atomic environment (aliphatic or aromatic) attached to the functional group. In our work, we have also applied the maximum common substructure algorithm to ensure that the smallest and most likely transformation that the chemists intended to perform has been performed and that the backbone of the matched molecular pair is the largest possible similar structure between the two compounds.

## 3.2 Materials and Methods

### 3.2.1 Compilation of the dataset

The data analysed in this study was extracted from internal AstraZeneca sources. For each project the starting compound was registered before the ending compound. The project data, including the compounds, registration dates and project titles from internal AstraZeneca databases were extracted with Perl and the matched molecular pairs were calculated using Perl<sup>137</sup>, and the output files were read into KNIME<sup>90,138</sup>. Standardised SMILES format was used to represent the compounds, and were standardised using an in-house method<sup>91</sup>. The attachment point is represented by a \* when representing the transformations and environments as smiles/ fingerprints.

### 3.2.2 Determination of the atomic environment of the compound

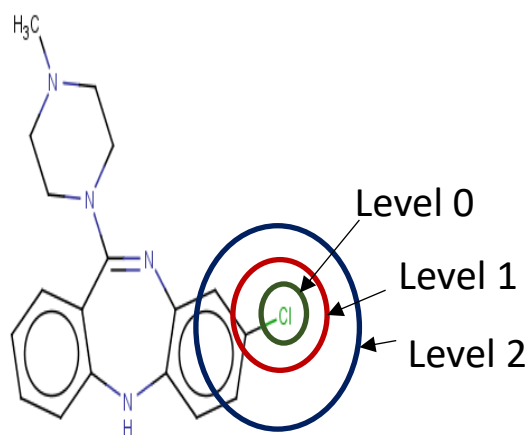
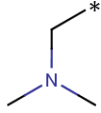

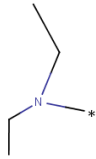
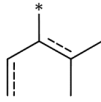
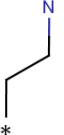
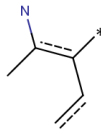
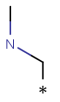

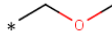




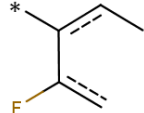
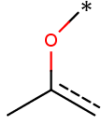

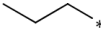

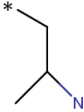



Figure 18: Example of atomic environment levels as defined by signature fingerprints

The atomic environments levels 1-3 are presented as signature fingerprints (Table 3) (note that the molecular fragments represent atomic environment level 0 (Figure 18) as this is the part that changes). The system type that the MMP transformations are performed on is represented as either aliphatic or aromatic. In total 311,782 rows of data were analysed.

Table 3: Table showing the top 10 atomic environments for aromatic and aliphatic systems. The \* represents the attachment point. The signature fingerprints are also shown.

|      | Aliphatic Systems   |                            | Aromatic Systems   |                                |
|------|---|----------------------------|--|--------------------------------|
| Rank | Atomic Environment  | Signature Fingerprint      | Atomic Environment   | Signature Fingerprint          |
| 1    |    | [*]([C]([N]([C][C])))      |    | [*]([c]([c]([c])[c]([c])))     |
| 2    |    | [*]([N]([C]([C])[C]([C]))) |    | [*]([c]([c]([c])[c]([c][C])))  |
| 3    |    | [*]([C]([N]([C])))         |    | [*]([c]([c]([c])[c]([c][N])))  |
| 4    |   | [*]([C]([C]([N])))         |    | [*]([c]([c]([c])[n]([c])))     |
| 5    |  | [*]([C]([O]([C])))         |   | [*]([c]([c]([c])[c]([c][O])))  |
| 6    |  | [*]([N]([C]([C]=[O])))     |  | [*]([c]([c]([c])[c]([n])))     |
| 7    |  | [*]([C]([N]([C]=[O])))     |  | [*]([c]([c]([c])[c]([c][F])))  |
| 8    |  | [*]([O]([C]([c][c])))      |  | [*]([c]([c]([c])[c]([c][Cl]))) |
| 9    |  | [*]([C]([C]([C])))         |   | [*]([c]([c]([c][C])[n]([c])))  |

|    |   |                       |  |                            |
|----|---|-----------------------|--|----------------------------|
| 10 |  | [*]([C]([C]([C][N]))) |  | [*]([c]([n]([c])[n]([c]))) |
|----|---|-----------------------|--|----------------------------|

### 3.2.3 Determining the matched molecular pairs

The matched molecular pairs were identified using an internal code that determined the matched molecular pairs using a maximum common substructure method<sup>83</sup>. The method involved cutting every bond that is not within a ring and recording all pairs of fragments. Implicit hydrogens are recorded and therefore considered. For each fragment that has been generated, a unique identifier was assigned of which is used as a normalisation step as it is possible to write the same fragment in different ways (when written as a structure format such as signature fingerprints). Searching for all compounds, which have at least one fragment of a size of at least 9 heavy atoms in common, creates the table of matched pairs. The two compounds should differ at only one position, and then all possible matched pairs are compared for any pair of compounds and only the one corresponding to the biggest fragment in common. The attachment point is then detected in both molecules at the point where the two compounds are differing. This ensures that the smallest possible MMP transformations (and most likely, in terms of what the chemists intended) are analysed, as this is likely to reflect the chemists design ideas. One compound example needed manual alteration due to not being able to be captured in the program.

The 20 most frequently occurring MMP transformations that are performed on aliphatic and aromatic systems well as the top 10 atomic environments for each system were visualised in ChemAxon Marvin View KNIME nodes<sup>139</sup>.

### 3.2.4 Parameters used to process the data

Where information, such as the registration date, on the compound pairs was missing, that compound pairs information was removed from further processing. The minimum time from the first registered compound and the last compound in a project needed to be at least 90 days and projects that had fewer than 100 unique compounds were removed. The registration date between compound 1 (transformed from) to compound 2 (transformed to) was analysed to see the days taken between the two compounds to register them. These dates were binned.

Only compounds registered from 2000 onwards were retained in the dataset, which represents the year of the formation of AstraZeneca. The MMP transformations ( $X \gg Y$ ) as well as the opposite MMP transformations ( $Y \gg X$ ) and the starting molecular fragment as well

as the end molecular fragment, were ranked in terms of their frequency of occurrence. Additionally, the percentage of each MMP transformations for the total MMP transformations in total was calculated. The percentage difference and ratio between the MMP transformations and its opposite MMP transformations is calculated. All these calculations were done for different systems (aliphatic or aromatic systems). Only MMP transformations that occur at least 100 times, in total, are included in the analysis.

### 3.2.5 Calculation of time difference and ratio between MMP transformations and their opposite

The time difference between the start of a project and the starting compound of each transformation was calculated and the median time was represented. Furthermore, the time difference between compound 1 and compound 2 was calculated by taking the difference of registration dates between subsequent compounds registered for a project.

We next calculated the inverse transformation as the count occurrences of the MMP transformations divided by the count occurrences of its opposite MMP transformations ( $((\text{Count}(X \gg Y) / (\text{Count}(Y \gg X)))$ ), and the difference in the percentage differences.

the resulting figures from this analysis were produced using Spotfire software<sup>96</sup>. When exploring the correlation between MMP transformations and atomic environment was produced in Python in Anaconda Navigator<sup>140</sup>.

## 3.3 Results and Discussion

### 3.3.1 Analysis of the most frequently observed molecular fragments found in transformations as a function of to time











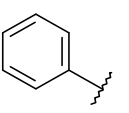





We first analysed the frequency of the starting point and end of the MMPs occurring chronologically. We separate these findings based on molecular environment (aromatic and aliphatic systems). The results of this analysis are displayed in Figure 4.

There are 957 unique transformations in total analysed in this study. The unique transformations are made up of 81 unique start points and 197 unique end points overall. Splitting this into the different systems, aromatic and aliphatic systems we find that, for aromatic systems, 935 out of the total of 957 unique transformations are performed on aromatic systems. These unique transformations are made up of 76 unique start points and 186 end points. For aliphatic systems, of the total 957 unique transformations, 944




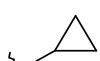





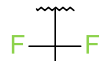

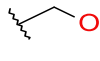




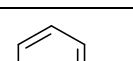


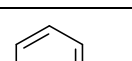
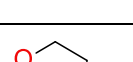
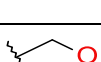
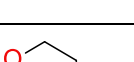
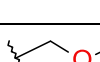
transformations are performed on aliphatic systems. 80 of the unique start points are observed on aliphatic transformation all 197 unique end points.

Methyl groups, the second most frequently occurring starting point for both aromatic and aliphatic systems, respectively, (and second and first most frequently occurring end point, for aromatic and aliphatic systems, respectively), are able to modulate the properties of a compound, in both the biological and physical sense<sup>73</sup>. Despite the large number of MMPs that result in a methyl group being replaced by another group (37,765 instances), there are a larger number (37,862 instances) of MMPs, which involve changing a functional group to a methyl group. A methyl group can be used to introduce a twist in the compound to aid in better binding<sup>141,142</sup> due allowing the biological conformation to be adopted in solution<sup>143</sup>. This effect can also influence solubility; moving ring systems out of the plane can interfere with stacking interactions (planarity and symmetry) and increase the solubility<sup>144</sup>. Another reason for adding a methyl group is to remove a donor from a molecule<sup>145</sup> in order to reduce permeability. Examples, of this occurring are replacing oxygen with a methyl or a nitrogen with a methyl group. Therefore, the high frequency of occurrences of the addition of a methyl group observed may be an effort to improve the binding of the compound to the target or optimise properties such as solubility and permeability. Additionally, a recent study highlighted the importance of methyl groups, and summarised 22 beneficial cases that have been observed on different important areas of the optimisation process, notable, potency, selectivity, solubility, metabolism as well as their PK/PD properties<sup>146</sup>. The authors also expect an increase in the number of methyl containing therapeutics.

Table 4: The occurrences of the top 10 most frequently occurring molecular fragment starting and end points performed on aromatic and aliphatic systems. The A marks the attachment point of the molecular fragment.

| Aromatic systems  |        |   |        | Aliphatic systems   |        |   |        |
|---|--------|---|--------|---|--------|---|--------|
| Change From   | Count  | Change To   | Count  | Change From   | Count  | Change To   | Count  |
|  | 39,620 |  | 16,292 |  | 43,768 |  | 22,747 |
|  | 12,312 |  | 13,301 |  | 21,512 |  | 15,179 |
|  | 11,783 |  | 10,642 |  | 7,290  |  | 4,766  |
|  | 10,914 |  | 10,642 |  | 5,246  |  | 4,436  |



|   |       |   |        |   |       |   |       |
|---|-------|---|--------|---|-------|---|-------|
|  | 7,558 |  | 10,227 |  | 4,450 |  | 4,094 |
|  | 4,063 |  | 6,459  |  | 4,144 |  | 3,948 |
|  | 3,585 |  | 4,585  |  | 3,968 |  | 3,885 |
|  | 1,946 |  | 2,447  |  | 3,564 |  | 3,536 |
|  | 1,423 |  | 1,737  |  | 2,769 |  | 2,934 |
|  | 1,216 |  | 1,484  |  | 2,337 |  | 2,886 |



















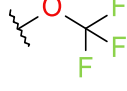

### 3.3.1.1 Aromatic Systems



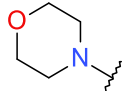
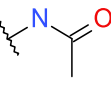
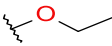



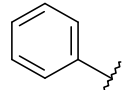
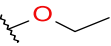
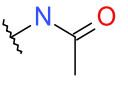
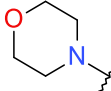



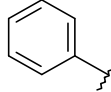


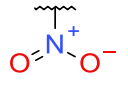


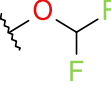

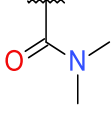



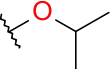
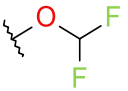

Halogen atoms (specifically, fluorine, bromine and chlorine) are prevalent amongst the most frequently observed starting and end points in the analysed MMPs (55,435 instances in total), when performed on aromatic systems (Table 4). The introduction of halogen atoms within drug design is often used to increase potency and or selectivity due to their electronic and steric effects by modulating the ligand-protein interactions<sup>147</sup>. A consequence of this is that the DMPK properties can be changed by the halogens, increase in lipophilicity<sup>148</sup>. Halogenated FDA approved drugs from between 1988 and 2006 show that the fluorine is the most commonly incorporated atom, followed by chlorine, bromine and then iodine<sup>147</sup>. The addition of fluorine in drug candidates has been extensively discussed in scientific literature, in comparison to the other halogen atoms, due to its atomic properties, such as its high electronegativity and its small size<sup>149–151</sup>. As fluorine is the most electronegative element, its introduction into a molecule can alter electron distribution, which can impact on the pKa, dipole moment and even the chemical reactivity and stability of neighbouring functional groups<sup>152</sup>. Additionally, fluorine can be used to block sites of oxidative metabolism by cytochrome P450 monooxygenases, when substituted place of hydrogen on an aromatic ring, whilst leaving potency unaltered due to being a comparable size to hydrogen<sup>152</sup>. This would explain the high frequency of a hydrogen atom being transformed to a fluorine on aromatic systems. Fluorination also affects the acidity/basicity of neighbouring groups in the molecule, for example making carboxylic acids more acidic and lowering the basicity of amines<sup>152</sup>. The many uses of fluorine are highlighted in several review articles<sup>151,153,154</sup>. As with fluorine,

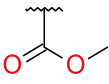


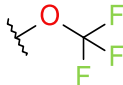

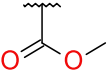
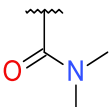
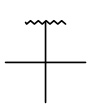
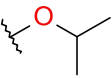


chlorine also is frequently observed in MMP transformations and molecular fragments<sup>147</sup>. Chlorine is larger in size than fluorine and is a moderate hydrogen bond acceptor. This is useful because potency increases due to picking up additional interactions with the protein.

Of the most frequently occurring MMPs found for aromatic systems, both fluorine and chlorine were observed as the starting point of the MMPs as well as ending point of the MMPs. There are 34 unique end points that halogenated atoms are transformed to (Table 5). This represents ~17% of the total number of transformed to molecular fragments showing that replacing a halogenated atom is a common practise in the design process. Halogenated atoms do not enter the 20 most frequently occurring MMP transformations for those performed on aliphatic systems, which may be because halogenated atoms (with exception of fluorine) are leaving groups and are therefore reactive<sup>155</sup>.

Table 5: List of the molecular fragments that the halogenated atoms are transformed to.

| <i>Aromatic</i> |   | <i>Aliphatic</i> |   |
|-----------------|---|------------------|---|
| <i>Count</i>    | <i>Fragment</i>   | <i>Count</i>     | <i>Fragment</i>   |
| 5,395           |  | 540              |  |
| 3,216           |  | 190              |  |
| 2,871           |  | 175              |  |
| 2,192           |  | 99               |  |
| 2,179           |  | 60               |  |
| 2,115           |  | 50               |  |
| 1,729           |  | 47               |  |
| 1,199           |  | 46               |  |
| 587             |  | 44               |  |
| 527             |  | 36               |  |

|     |   |    |   |
|-----|---|----|---|
| 473 |    | 36 |    |
| 363 |    | 31 |    |
| 357 |    | 30 |    |
| 344 |    | 26 |    |
| 336 |    | 22 |    |
| 331 |    | 20 |    |
| 329 |    | 17 |    |
| 316 |  | 17 |   |
| 285 |  | 12 |  |
| 285 |  | 11 |  |
| 282 |  | 10 |  |
| 275 |  | 9  |  |
| 251 |  | 9  |  |
| 236 |  | 9  |  |
| 236 |  | 8  |  |

|     |   |   |   |
|-----|---|---|---|
| 224 |  | 7 |  |
| 120 |  | 4 |  |
| 110 |  | 1 |  |
| 92  |  | 1 |  |
| 91  |  | 1 |  |
| 77  |  |   |   |

Other groups commonly introduced into aromatic systems include acids and amides, which can make specific interactions with amino acids in the protein target as well as modulate the pKa of the compound<sup>156</sup>, as well as trifluoromethyl groups (12,781 instances on aromatic systems) which have the advantage of being similar to isopropyl or ethyl groups in terms of size. Trifluoromethyl groups do however tend to be smaller than isopropyl groups and are considered more similar to ethyl groups<sup>150,157,158</sup>.

### 3.3.1.2 Aliphatic Systems

Cyclopropyl groups appear in the 20 most frequently occurring MMP transformations groups for aliphatic systems (16<sup>th</sup> most frequently observed transformation, with respect to time). Cyclopropyl groups are increasingly being incorporated into compounds as a replacement for methyl groups because of the reported increase in metabolic stability<sup>159</sup>. It has been shown that the addition of a cyclopropyl ring directly influences physicochemical properties, target specificity and potency in a favourable way as well as pharmacokinetics<sup>159</sup>. It is however important to note that how the properties are affected will be dependent on the surrounding chemistry of the compound. MMP transformation has been shown to increase several factors including potency and stability<sup>135</sup>. This MMP transformation was the 66<sup>th</sup> most frequently observed MMP transformations on aromatic systems but ranked 16<sup>th</sup> on aliphatic systems. This section shows that the most frequently observed molecular fragments have been extensively studied and reviewed in the chemical literature for their impact on drug properties

and that chemists use this knowledge in design during compound optimisation. Here to the effects of these molecular fragments on compound properties will be dependent on the surrounding compound chemistry. While many transformations make sense according to literature, it is still overall difficult to make sense of them all, because they depend on context, and therefore; it is difficult to fully rationalize transformations.

### 3.3.2 Analysis of the most frequently observed MMP transformations and their inverse transformations

We next analysed the most frequent MMP transformations on different systems. This highlights the most frequently occurring MMP transformations that are performed on the different atomic environments.

Removing molecular fragments to leave a hydrogen atom or replacing a hydrogen atom is the most frequent MMP transformations made in the 20 most frequently. The frequency of adding or removing a hydrogen atom, likely shows the effects of expanding and growing or conversely truncating a molecule in the drug design process to optimise interactions<sup>160</sup>. Alternatively, to develop SAR of which moieties in a molecule are important for potency and properties.

The 20 most frequently occurring MMP transformations for either system all involve small (few number of heavy atoms involved) molecular fragments as both starting and endpoints; for example, replacing a hydrogen atom with fluorine. This may suggest that such small changes are often made to fine-tune properties by the medicinal chemist in lead optimization. Figure 19 also shows that the more frequently a MMP transformation occurs, the more projects it is involved with, which shows that the MMP transformations have global applicability across project contexts, for different target classes. It has been shown that there are overall relatively few reactions used in chemistry, which may be the result of commercial availability and selectivity<sup>161</sup>.

The fact that there are more instances of hydrogens being replaced by heavy atoms rather than the other way around (within the 20 most frequently occurring MMP transformations) relates to the fact that molecular weight (compound size) tends to increase throughout the compound optimisation process<sup>125,160</sup>.

There are only four MMP transformations that appear in both 20 most frequently occurring MMP transformation lists. These are hydrogen to methyl, methyl to hydrogen, hydrogen to ether, and methyl to ether. However, they do vary in terms of their rank in frequency. Hydrogen to methyl and methyl to hydrogen are the most and second most frequently occurring MMP transformations in the 20 most frequently occurring MMP transformations that are performed

on aliphatic systems. They appear as the 2<sup>nd</sup> and 6<sup>th</sup> most frequently occurring MMP transformations for those performed on aromatic systems, respectively. Hydrogen to ether ranks position 4<sup>th</sup> for those performed on aromatic systems and 7<sup>th</sup> when performed on aliphatic systems. Whereas, methyl to ether occurs at the 17<sup>th</sup> and 18<sup>th</sup> positions for MMP transformations performed on aromatic and aliphatic systems. What this shows it that these transformations are key attempts made regardless of the system, likely as an attempt to grow the compound.

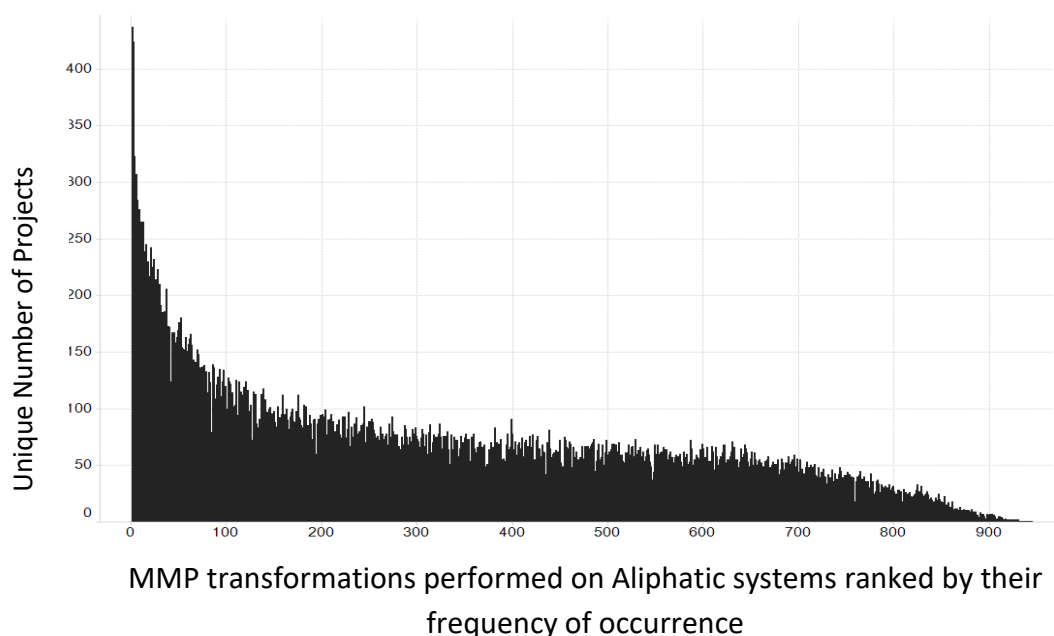
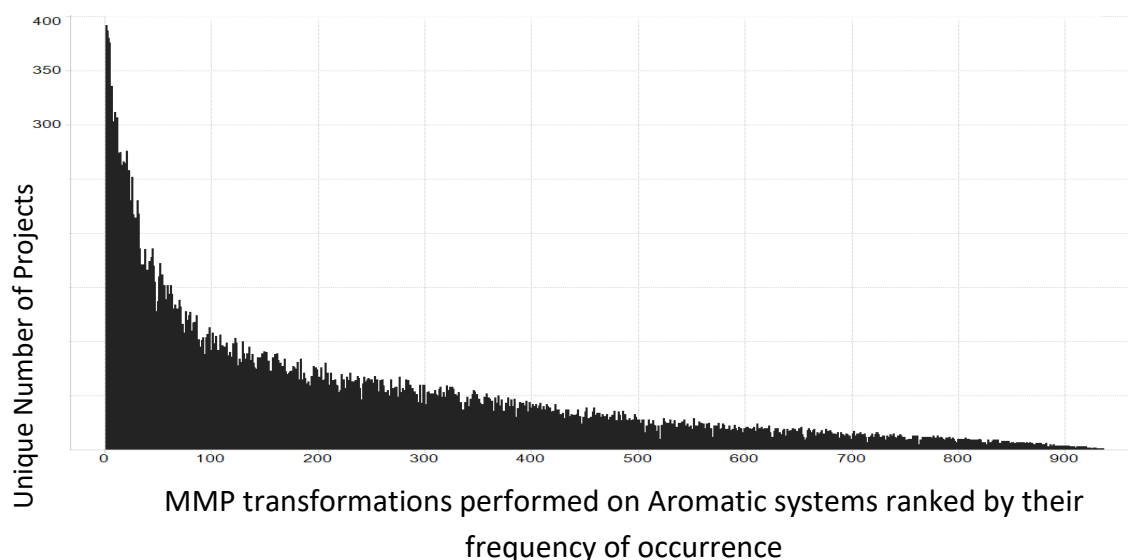


Figure 19: For each system (A) aromatic and (B) aliphatic a bar chart showing the portion of unique different projects of each MMP transformations is observed in. The higher the rank of the MMP transformations, i.e. the more occurrences of MMP transformations, the more projects that MMP transformations appears in. In this presentation, the X axis represents the rank of the frequency of the transformations performed on each environment. MMP transformations can be very commonly occurring but only appear in a couple of projects.



We next considered the MMP transformations in relation to their reverse MMP transformations for both aromatic (Table 6) and aliphatic systems (Table 7). The frequency of the reverse MMP transformations are calculated and compared against the forward MMP transformation occurrence frequency to suggest why particular MMP transformations may be performed, but the opposite might be less frequent. Some of these MMP transformations and reverse MMP transformations occur very close in rank in terms of their frequency of occurrence for both aromatic and aliphatic systems. For example, chlorine to a methyl group MMP transformations and a methyl to a chlorine group MMP transformations are the two transformations that are performed in similar proportions to each other. Being around the same size substituents, these are commonly switched for SAR reasons<sup>162</sup>.

### 3.3.2.1 Aromatic Systems

50% of the MMP transformations that occur on aromatic systems (listed in Table 6: The top 20 MMP transformations performed on aromatic systems and their statistics. There are two instances in the top 10 where a MMP transformations reverse MMP transformations do not occur in the 20 mos) involve a hydrogen atom of which suggests that the most common comparison made is to an un-substituted framework from different substituents. In fact, the 20 most frequently occurring MMP transformations that occur on aromatic systems, involve only 7 different unique molecular fragments.

























































The fourth most frequently occurring transformation, hydrogen to methoxy is an interesting one in that when bonded to a benzene ring is considered an electron-donating group. Another key example is the introduction of pyridine of which is likely to be a MMP transformations for growing the molecule and pyridines are often found in approved drugs<sup>164</sup>. Pyridines provide an increase in lipophilicity, an opportunity for interactions with amino acid residues through  $\pi$ -stacking interactions<sup>165</sup>, and have the advantage that they are less lipophilic than benzene and display differing electronics and reactivity which makes them amenable to further synthetic modification<sup>166</sup>.

Only 4 of the inverse MMP transformations performed on aromatic systems, do not occur in the 20 most frequently occurring MMP transformations along with their forward transformation. These are nitrile group to a hydrogen atom (position 41), trifluoromethyl to a hydrogen (position 46, several FDA approved drugs contain a trifluoromethyl group including Prozac<sup>167</sup>, ether to a methyl group (position 22), an ether to a chlorine (position 25), and the MMP transformations



positions were 9, 14, 17 and 20 respectively. Hydrogen atom as a starting point occurs 6 times and as an ending point occurs 4 times. Only two MMP transformations that involve a hydrogen atom as a starting point do not have their reverse MMP transformations in the 20 most frequently occurring (nitrile group to a hydrogen (position 41) and trifluoromethyl to hydrogen (position 46)).

Table 6: The top 20 MMP transformations performed on aromatic systems and their statistics. There are two instances in the top 10 where a MMP transformations reverse MMP transformations do not occur in the 20 most frequently occurring.

| Rank (A) | Transformation  | Count (B) | Rank Inverse (C) | Reverse Transformation   | Count Inverse (D) | C-A | Ratio ((B)/(D)) |
|----------|---|-----------|------------------|--|-------------------|-----|-----------------|
| 1        |  →      | 8,951     | 5                |  →      | 2,839             | 4   | 3               |
| 2        |  →      | 7,400     | 6                |  →      | 2,620             | 4   | 3               |
| 3        |  →      | 5,895     | 7                |  →      | 2,556             | 4   | 2               |
| 4        |  →      | 5,608     | 11               |  →      | 1,940             | 7   | 3               |
| 5        |  →  | 2,839     | 1                |  →  | 8,951             | -4  | 0               |
| 6        |  →  | 2,620     | 2                |  →  | 7,400             | -4  | 0               |
| 7        |  →  | 2,556     | 3                |  →  | 5,895             | -4  | 0               |
| 8        |  →  | 2,192     | 10               |  →  | 2,179             | 2   | 1               |
| 9        |  →  | 2,192     | 41               |  →  | 549               | 32  | 4               |
| 10       |  →  | 2,179     | 8                |  →  | 2,192             | -2  | 1               |
| 11       |  →  | 1,940     | 4                |  →  | 5,608             | -7  | 0               |
| 12       |  →  | 1,800     | 13               |  →  | 1,674             | 1   | 1               |
| 13       |  →  | 1,674     | 12               |  →  | 1,800             | -1  | 1               |
| 14       |  →  | 1,599     | 46               |  →  | 449               | 32  | 4               |

|    |  |       |    |  |       |    |   |
|----|--|-------|----|--|-------|----|---|
| 15 |  | 1,526 | 19 |  | 1,391 | 4  | 1 |
| 16 |  | 1,467 | 18 |  | 1,416 | 2  | 1 |
| 17 |  | 1,417 | 22 |  | 1,173 | 5  | 1 |
| 18 |  | 1,416 | 16 |  | 1,467 | -2 | 1 |
| 19 |  | 1,391 | 15 |  | 1,526 | -4 | 1 |
| 20 |  | 1,345 | 25 |  | 1,055 | 5  | 1 |

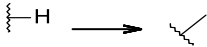
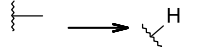
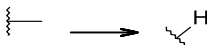
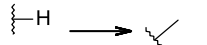
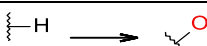
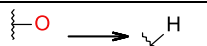
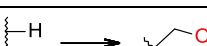
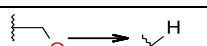
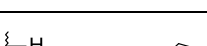
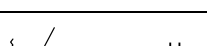
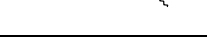

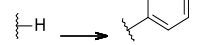
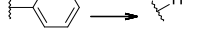
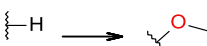

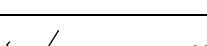
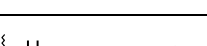
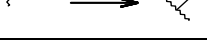
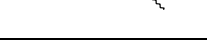

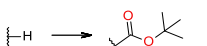
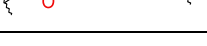
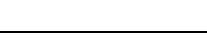

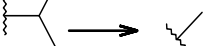
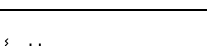
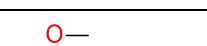

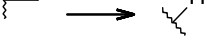
### 3.3.2.2 Aliphatic Systems

The top 20 transformations that are performed on aliphatic systems are shown in Table 7 where we find that hydrogen is transformed to an alcohol group very frequently, occurring as the fourth most frequently observed transformation. This is because the addition of an alcohol functional group onto a molecule introduces the electron rich oxygen (due to its electronegativity) and therefore the covalent bonds are polarized leading to improved solubility.

The 14<sup>th</sup> most frequently occurring transformation is that of a hydrogen being replaced by an acetyl group. This group can be added to increase the permeability and cross the blood brain barrier, and therefore acetylation would likely be incorporated in situations where the chemists are aiming to get the drug past the blood brain barrier. An example of such a drug is acetaminophen (paracetamol), or the acetylation of morphine to heroin (diacetylmorphine)<sup>168</sup>.

Deacetylation (the removal of an acetyl group) to a hydrogen group occurs considerably less frequently than the forward transformation. It is the 43<sup>rd</sup> most frequently occurring transformation creating a rank difference of 29 places. The forward transformation occurs over 3.5 times more frequently than deacetylation highlight the desire to get drugs past the blood brain barrier in many cases.

Table 7: The top 20 MMP transformations performed on aliphatic systems and their statistics. There are more inverse transformations that do not occur in the top 20 in comparison to those identified on aromatic systems.

| Rank (A) | Transformation  | Count (B) | Rank Inverse (C) | Reverse Transformation  | Count Inverse (D) | C-A | Ratio ((B)/(D)) |
|----------|---|-----------|------------------|---|-------------------|-----|-----------------|
| 1        |    | 16,840    | 2                |    | 9226              | 1   | 2               |
| 2        |    | 9,226     | 1                |    | 16840             | -1  | 1               |
| 3        |    | 2,676     | 13               |    | 1011              | 10  | 3               |
| 4        |    | 1,897     | 31               |    | 580               | 27  | 3               |
| 5        |    | 1,739     | 8                |    | 1418              | 3   | 1               |
| 6        |   | 1,536     | 20               |   | 853               | 14  | 2               |
| 7        |  | 1,532     | 36               |  | 536               | 29  | 3               |
| 8        |  | 1,418     | 5                |  | 1739              | -3  | 1               |
| 9        |  | 1,404     | 58               |  | 431               | 49  | 3               |
| 10       |  | 1,395     | 30               |  | 581               | 20  | 2               |
| 11       |  | 1,326     | 66               |  | 393               | 55  | 3               |
| 12       |  | 1,113     | 23               |  | 810               | 11  | 1               |
| 13       |  | 1,011     | 3                |  | 2676              | -10 | 0               |
| 14       |  | 1,011     | 43               |  | 489               | 29  | 2               |
| 15       |  | 992       | 57               |  | 433               | 42  | 2               |


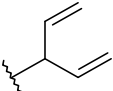
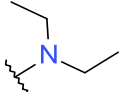
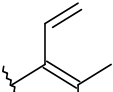
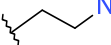
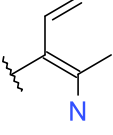
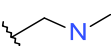
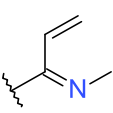
|    |  |     |    |  |      |     |   |
|----|--|-----|----|--|------|-----|---|
| 16 |  | 968 | 64 |  | 404  | 48  | 2 |
| 17 |  | 965 | 52 |  | 458  | 35  | 2 |
| 18 |  | 916 | 44 |  | 484  | 26  | 2 |
| 19 |  | 853 | 50 |  | 460  | 31  | 2 |
| 20 |  | 853 | 6  |  | 1536 | -14 | 1 |

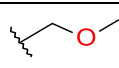
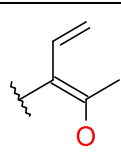
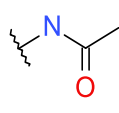
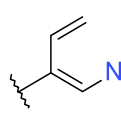
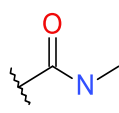
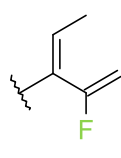
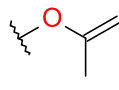
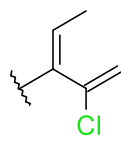
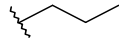
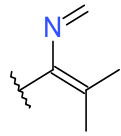
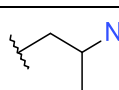
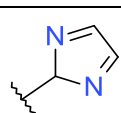
A large portion of the reverse MMP transformations performed on aliphatic systems do not occur in the 20 MMP transformations (60% (12 out of 20 reverse MMP transformations)). In comparison to the 20 most frequently occurring MMP transformations, 16 of the MMP transformations performed on aromatic systems appear in the 20 most frequently occurring MMP transformations overall. Only eight MMP transformations that occur on aliphatic systems appear in the 20 most frequently occurring MMP transformations overall hydrogen to a methyl, methyl to a hydrogen, hydrogen to an ether, hydrogen to an oxygen, hydrogen to an ether, methyl to an ether, hydrogen to a carbon-carbon bond and a hydrogen to a phenyl ring. This suggests that aliphatic systems are more sensitive to the types of transformations performed on them.

### 3.3.3 Analysis of the most frequently observed environments and the MMP transformations performed on them

In order to understand which environments are most frequently involved with transformations, we looked at the top 10 atomic environments in which the MMP transformations are applied to both aliphatic and aromatic systems (Table 8). The distribution of proportions, in terms of frequency of transformations being performed for each environment, is more evenly distributed in aliphatic systems than they are for aromatic systems. The proportion of the most frequently identified MMP transformations ( $X \gg Y$ ) that are performed on each of the top 10 local atomic environments for aromatic systems shows that approximately 55% of the transformations are performed on the most frequently occurring aromatic system, whereas, for aliphatic systems all of the top 10 most frequently occurring MMP transformations make up between 7% and 14%. This again supports the idea of aliphatic systems being more sensitive to the types of transformations that are performed on them (specific examples are presented in their relevant sections of this thesis).

Table 8: Table showing the top 10 atomic environments for aromatic and aliphatic systems. The \* represents the attachment point. The percent occurrence of the top 10 for each system is also shown.

| Rank | Aliphatic Systems   |                   | Aromatic Systems  |                   |
|------|---|-------------------|---|-------------------|
|      | Atomic Environment  | Percent of top 10 | Atomic Environment  | Percent of top 10 |
| 1    |  | 13%               |  | 55%               |
| 2    |  | 13%               |  | 16%               |
| 3    |  | 12%               |  | 7%                |
| 4    |  | 12%               |  | 6%                |

|    |   |     |   |    |
|----|---|-----|---|----|
| 5  |    | 11% |    | 4% |
| 6  |    | 9%  |    | 3% |
| 7  |    | 8%  |    | 3% |
| 8  |    | 7%  |    | 2% |
| 9  |    | 7%  |    | 2% |
| 10 |  | 7%  |  | 2% |

As with the MMP transformations the 20 most frequently occurring atomic environments and MMP transformations performed on them vary between aromatic and aliphatic systems as can be seen in Figure 20.

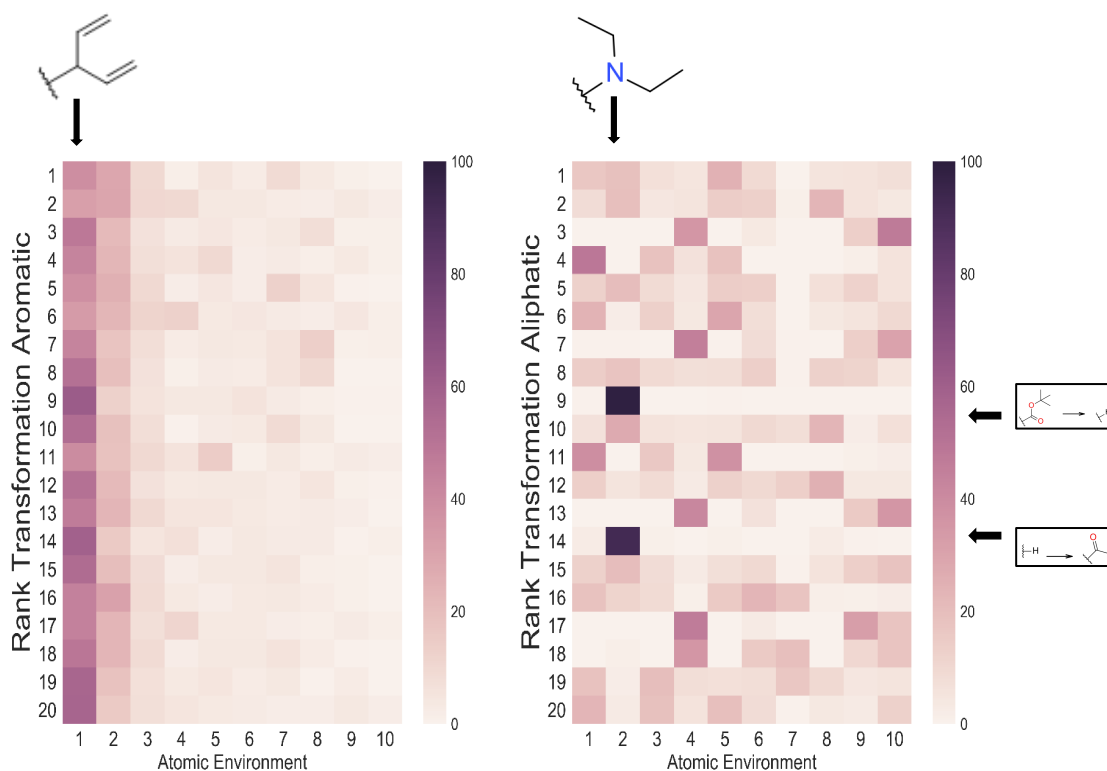


Figure 20: The percentage of each MMP transformations in the 20 most frequently occurring that are performed on (A) aromatic or (B) aliphatic systems that is performed in a given atomic environment (the top 10 most frequently occurring atomic environments overall). The most frequently analysed MMP transformations for aromatic systems tend to all be attached to one environment (where [\*] represents the attachment point). Whereas, the 20 most frequently occurring most frequently observed MMP transformations performed on aliphatic systems vary on the atomic environments. The shade of purple reflects the percentage of the top 10 environments a particular transformation is performed on. Therefore, each row equals 100%

### 3.3.3.1 Aromatic Systems

The most frequently occurring aromatic environments are shown in Table 14. The 4<sup>th</sup> most frequently occurring aromatic atomic environment is one that matches a pyrrole structure. Pyrrole is considered reactive (especially towards electrophilic groups), has a lone pair of electrons delocalized onto the aromatic ring, and its diverse activity features in a number of marketed drugs with a range of therapeutic properties<sup>169</sup> such as Lipitor which is used to lower LDL levels<sup>170</sup>. Alternatively, this environment could be a pyridine or a pyridazine, as is also the case with the 6<sup>th</sup> most frequently occurring atomic environment for aromatic systems. The 10<sup>th</sup> most frequently occurring aromatic atomic environment is an imidazole which is electron-rich and allows for derivatives to readily bind with a variety of targets such as enzymes<sup>171</sup>. Furthermore, imidazole is amphoteric and highly polar<sup>172</sup> and can readily accept or donate protons as well as participate in weak interactions.



In the next phase of our study we explored the most frequently occurring transformations that are performed on the most frequently occurring atomic environments as seen in Figure 20.

The highest observed instance of the 20 most frequently occurring MMP transformations which occurs on an aromatic system, is hydrogen to fluorine (22%), while the most frequently observed aromatic environment has the highest proportion of occurrences of transformations, which might not be unexpected as the most frequently occurring transformations are performed on the most frequently occurring atomic environment most often.

### 3.3.3.2 Aliphatic Systems

Following on with observations on aromatic systems, aliphatic systems are more sensitive to the transformations that are performed on them in comparison to those performed on aromatic systems, which is likely to be attributed to the difference in saturated and unsaturated structures. The MMP transformations that are performed on an aliphatic system are likely to provide different optimisation methods to those performed on aromatic systems.

The 9<sup>th</sup> and 14<sup>th</sup> most frequently occurring transformation are performed on the 2<sup>nd</sup> most frequently occurring aliphatic environments 100% of the time (out of the 10-top aliphatic atomic environments). The 9<sup>th</sup> most frequently occurring transformation is the removal of a boc-protecting group, protecting amines, allowing the nitrogen to be revealed. The boc-protecting group is often replaced in a molecule (the 9<sup>th</sup> most frequently occurring transformation involves a molecular fragment containing a boc-protecting group and is replaced by a hydrogen atom (Figure 20) and this occurs 1,404 times most likely as a result of the synthetic methods used to produce amine containing compounds<sup>173</sup>. The 14<sup>th</sup> most frequently occurring transformation depicts acetylation. Of the 10-top aliphatic atomic environments, acetylation occurs only on the 2<sup>nd</sup> most frequently occurring aliphatic atomic environment. It can potentially occur on other environments; however, this is not observed in this instance.

### 3.4 Chapter overview

In our study we have we have statistically analysed the MMP transformations applied during the drug optimisation process of various projects within AstraZeneca's collection. The decision of the design process is highly influenced by the atomic environment that the MMP transformations are performed on.

The most frequently occurring molecular fragments that act as starting and ending points in the transformations and the MMPs performed vary between aliphatic and aromatic systems. Furthermore, molecular fragments are favoured in the drug discovery process including halogenated atoms on aromatic atomic environments and nitrile groups in both aromatic and aliphatic environments. This is due to their known advantageous properties such as modulating potency and or selectivity for during and molecular recognition.

When analysing the frequency of the reverse transformation it is observed that aromatic reverse MMP transformations are more frequently occurring than those aliphatic systems, which is likely due to the frequency of use of aromatic rings in drug discovery and the understanding of synthetic chemistry around these scaffolds. Aliphatic systems are more sensitive to the types of MMP transformations that can be performed on them, with two MMP transformations out of the top 10 aliphatic atomic environments, being performed on a single environment, of the 10 most frequently occurring. These two examples are due to specific chemical restrains such as removing a boc-protecting group from the amine and the acetylation of the amine to an amide.

This analysis can help with compound optimisation by helping to understand which transformations are performed and how the compound transformation affects the assay results.

## 4 Analysing the Matched Molecular Pair Transformations in Drug Discovery Projects as a Function of Time and Molecular Environment: – Effects on Compound Properties

### 4.1 Introduction

In the previous chapter a statistical analysis was presented of the most frequently observed matched molecular pair transformations within internal AstraZeneca projects and the atomic environments they were performed. Extending on from that analysis, this chapter focuses on the observed effects of transformations on common properties of compounds identified for in lead optimisation and how these properties change over the course of a project; in particular hERG, Caco-2, logD, solubility, human hepatocyte metabolism, human and rat microsomal metabolism. Furthermore, the influence of the observed transformations on these properties need to be analysed in the context of the atomic environment, as the effects of the neighbouring atoms is crucial in determining the effect that the change will have on the properties due to factors including chemical reactivity, stability, electronics and sterics.

Previous examples of assessing the effect of the matched molecular pairs on compound properties have been discussed in the literature where it has been highlighted that comparing the structural changes and their influence on a particular property can lead to an understanding of the expected property change when applying these structural features to new molecules<sup>162</sup>. For example, matched molecular pair analysis has been performed on glycogen phosphorylase inhibitors to assess the effects on the properties (as well as activity) that occurs due to the change of one molecular fragment to another<sup>174</sup>.

In a previous study, the authors compared the change in property values for a handful of molecular fragments when adding them to aromatic rings and the methylation of heteroatoms. Specifically, they looked at methylating an amide, ROH group, ArOH group and an RR'NH group. In the case of methylating an amide the authors identified an average increase in the solubility, whereas the rest decreased on average. The properties tested were aqueous solubility, rat plasma protein binding, and oral exposure in an *in vivo* rat model. For solubility, the authors noted that when adding substituents on aromatic rings, the addition of heavier halogens correlates to how large their negative effect would be on the solubility. They also found that adding a bromine to an aromatic system decreased the solubility in 98% of the cases<sup>162</sup>. The study also explains that outliers can possibly indicate ways to avoid the general trend, which could be highly beneficial.

By analysing structural changes and the effect they have upon a property of interest it is possible to understand what is likely to happen when the transformation is performed on a new compound<sup>162</sup>. However, as well as considering the chemical change, it has been shown that the atomic environment of the chemical change must also be considered in order to optimise the predictive ability of MMPA analysis<sup>81</sup>. The variability of the physical effect as well as experimental error and experimental data should also be considered<sup>129</sup>.

To aid in finding the ideal medicinal chemistry of a compound, it has been suggested to use a 'nature' and 'nurture' process; nature to identify chemical starting points and nurture during the lead optimization<sup>175</sup>. In the 'nature' and 'nurture' study, they identify the issue of molecular obesity (compounds that are too large or too lipophilic (for absorption)) and that even though it is a well-known problem, compounds still tend to be lipophilic, the binding thermodynamics are related to this molecular obesity as potent compounds are often not aligned with optimal ADME profiles<sup>175</sup>. However, the term molecular obesity<sup>126</sup> shows that scientists have a tendency to increase the molecular weight, as well as the lipophilicity, as part of the desire to find compounds with desirable potency<sup>126</sup>. The increase in compound size throughout the course of a project is supported in the literature where it has been shown that throughout the course of a compound being built, from HTS collections to leads to patents, the median molecular weight of compounds increases as does the median cLogP from start to finish of optimisation pairs<sup>124</sup>. Our investigations will confirm the trends in properties that are observed across projects and highlight the matched molecular pairs which are commonly used to change these properties, as well as the local atomic environments the transformations are performed on.

In contrast to previous analyses, the assays analysed in our study is logD octanol at pH7.4, human microsomal metabolism, human hepatocyte metabolism (human hep), rat hepatocyte metabolism, solubility at pH7.4, hERG IC<sub>50</sub>, human Caco-2 efflux ratio, and Caco-2 intrinsic. Our work considers matched molecular pairs within a time-course analysis, allowing the observation of trends in the changes of properties over time in a project.

Overall, the matched molecular pairs have been analysed *via* a time course analysis allowing us to observe any trends in the changes of properties over time. The knowledge of functional group changes that influence compound properties, when performed on certain atomic environments, will allow chemists to make informed decisions when designing new compounds. This will benefit the design process by minimising the number of compounds that will require synthesis in a project, consequently, speeding up the DMTA cycle and reducing costs.

## 4.2 Materials and Methods

### 4.2.1 Data compilation

As outlined in the previous chapter, the matched molecular pairs were derived from internal AstraZeneca projects where the starting compound was the compound to be designed and registered within a project, and the ending compound is that which is potentially inspired by the starting compound. The starting and the ending compounds make up the matched molecular pair. Furthermore, on the forward transformation we need to have at least 100 instances of that transformation occurring in the dataset, and each project needed to have 100 unique compounds that could be included in this analysis.

The transformations in our study were ranked in terms of the frequency of both aromatic and aliphatic functionalities in addition to the atomic environments of the systems. This has allowed us to be able to observe the most frequently occurring transformations as a function of the most frequency occurring atomic environments. These atomic environments were defined by signature fingerprints where level zero is the molecular fragment that is involved in the transformation with subsequent levels representing the next set of attached bonds and atoms.

### 4.2.2 Assay properties analysed

Experimental assay data was extracted from the internal AstraZeneca data, and represent an accumulation of many tests and represent the mean value of repeated experiments. The tests analysed is logD in octanol at pH7.4 for human microsomal metabolism, human hepatocyte metabolism (human hep), rat hepatocyte metabolism, solubility at pH7.4, hERG Ic50 combined, human Caco-2 efflux ratio, and Caco-2 intrinsic. The data extracted varied for each row as not all data was available for every pair of compounds. Any empty entries for each property value were removed and each MMP transformation was analysed to assess whether each property value increased, decreased or for which there was no significant change. Results that was greater than or less than 4 standard deviations were removed. The resulting file which comprises of the project code, compound 1 and compound 2 their respective property change results, as well as standard deviation calculations, was merged with the previous file which contained the transformation information, ranks and atomic environment information as well as containing all the time data that had been constrained on the parameters mentioned previously. Not all compounds had assay information available and therefore the

total number of MMPs analysed that contained property values was 106,927 MMPs for aromatic systems and 133,736 MMPs for aliphatic systems.

Furthermore, the molecular weight, number of hydrogen bond donors and acceptors, number of rotatable bonds and the number of rings was calculated in KNIME using RDKit<sup>94</sup> for both class of compounds to allow observations to be made about the size of structural changes and when they are performed in a project.

The figures were generated in Spotfire<sup>96</sup> using the Hmisc<sup>176</sup> package and corrplot<sup>177</sup> package in R (version 3.4), and RStudio (Version 1.1)<sup>102,178</sup> was used to calculate the correlation coefficient and the significance value of trends. The median property analysed for each split (bin) of the project (up to a total of 10) was calculated and a spearman rank test applied.

#### 4.2.3 Determining a significant increase, decrease or a minimal change in the log property values

When logD increases, human hepatocyte metabolism, rat hepatocyte metabolism, human microsomal metabolism, and Caco-2 efflux ratio are expected to increase, whereas solubility, hERG IC50 and Caco-2 intrinsic decrease.

To determine if there is a significant change in properties for a MMP, the property change needed to increase or decrease by 0.3 log units is considered to be significant. Those values that do not change by 0.3 log value are considered a minimal change, while values within experimental would not be considered actionable within a project. To determine which values were significant, the effect noted (significantly increased, significant decreased or minimal change) for each transformation performed on an atomic environment needed to occur at least 5 times and fulfil the following formula (Equation 12) for each atomic environment group.

$$X_{Sig\_Inc(P)} > X_{Sig\_Dec(P)} + X_{Minimal\_Change(P)}$$

Equation 11: Relationship to assess whether a Property (P) significantly increases more times than when it is significant decreased or assessed as a minimum change for an atomic environment. X represents the number of occurrences. In addition, these X values would need to be more than 5 instances.

The cut-off is justified by comparison to a previous study<sup>81</sup> in which the authors analysed each set of property value data (hERG, Solubility and Lipophilicity) and the most frequent transformations. Following this, each of the most frequently occurring transformations are split into unfavourable, zero and favourable changes, and the authors compare how the

transformations performed on specific atomic environments affect the property change. However, only a few examples are used rather than all environments in the datasets for showing the predictive power of using MMP with contextual information. In our study all transformations against all environments are considered as well as quantifying, how much the transformations, on a given environment, change the property value.

#### 4.2.4 Outliers

In our analysis we consider the median difference for each property, for a given transformation, as a function of the atomic environment to gain an understanding of the property change without the outliers shifting the distribution drastically. As it is standard practise for the tests to be averaged for multiple repeats of the same assay, capturing outliers in each result for each transformation performed on each environment is important. This in turn allows us to observe outliers but prevent misleading effects on general distributions. We also made the decision to remove the few extreme outliers where values are  $\pm 4$  standard deviations for a specific transformation performed on a specific environment from the core analysis; however, these instances have been examined to understand the chemistry involved in changing a property significantly ( $\pm 4$  standard deviations). The reason for this approach is because such extreme outliers are likely to be due to known structural modifications, such as change in ion class of a compound, which will allow us to identify 'true' outliers.

### 4.3 Results and Discussion

#### 4.3.1 Analysis of the physicochemical properties and assay result change over the course of a project

In the first instance we analysed several physicochemical properties, such as the number of hydrogen bond donors and acceptors, number of rotatable bonds and the number of rings as a way of identifying how the chemistry of the compounds change over the course of a project (Figure 21). The median number of each property for both the starting and ending compound at different stages on the project (of which has been split into 10 parts and normalised between 0 and 1; 10 segments) is observed.

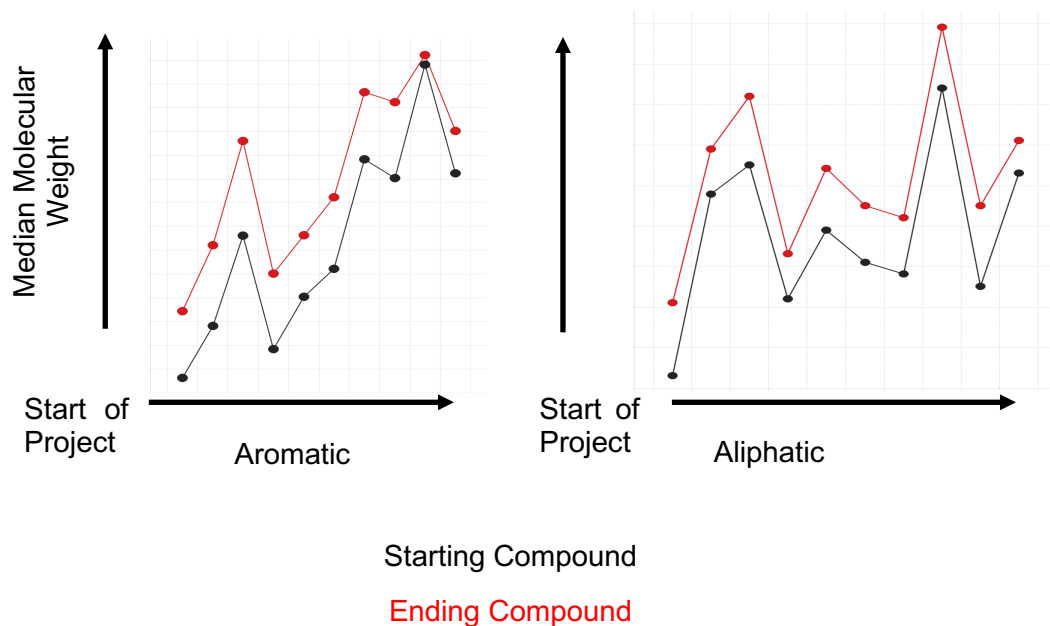


Figure 21: At each stage of a project (10 stages) the median molecular weight of the starting (compound transformed from) and ending (compound transformed to) compound is shown. Median molecular weight increases throughout the course of the project, more so for aromatic systems than aliphatic systems.

Figure 21 shows the median molecular weight throughout the course of a project for both the starting compound and the end compound, and the analysis highlights and supports the findings of Lipinski *et al*<sup>124</sup> that compounds grow through the course of a project in terms of molecular weight. However, at points in the evolution of the project, the median physicochemical properties of the starting compound do not differ from those of the ending compound (Table 9). It is also observed that the penultimate step before the end of a project (9<sup>th</sup> segment) the properties differ the most compared to the rest of the project, likely as a result of compounds becoming too large and the chemists attempt to maintain the favourable properties but reduce the size and lipophilicity of the compound at the very end of the project. The most notable differences are observed in the 9<sup>th</sup> segment (the penultimate step before the end of the project), for aromatic systems where the compounds show an increase in the properties before dropping again. This phenomenon is not observed for aliphatic systems where the increase in molecular weight is not as noticeable as that for aromatic systems, and is likely due to aliphatic systems being more sensitive to the types of transformations that can be performed on them in comparison to aromatic systems. It is noteworthy that the median molecular weight for the ending compound is still always larger than the first compound indicating that the molecular weight increases during the transformation process.

Table 9: For each system and for both the starting and ending compounds throughout the course of the project, the median number of rotatable bonds, hydrogen bond donors and acceptors as well as the number of rings. The



heatmap colour represents the frequency of each property as does the bar chart allowing for direct comparisons to be made between starting and ending compound and project.

| Aromatic Systems  |   |   |  |  |   |   |                                   |                                 |
|---|---|---|--|--|---|---|-----------------------------------|---------------------------------|
| Difference Between Compound 1 and Earliest Project Date (10 bins) | Number of Rotatable Bonds Starting Compound | Number of Rotatable Bonds Ending Compound | Number of Hydrogen Bond Donors Starting Compound | Number of Hydrogen Bond Donors Ending Compound | Number of Hydrogen Bond Acceptors Starting Compound | Number of Hydrogen Bond Acceptors Ending Compound | Number of Rings Starting Compound | Number of Rings Ending Compound |
| $x < 0.10$  | 5   | 5   | 1  | 2  | 5   | 5   | 4                                 | 4                               |
| $0.10 \leq x < 0.20$  | 5   | 5   | 2  | 2  | 4   | 5   | 4                                 | 4                               |
| $0.20 \leq x < 0.30$  | 5   | 5   | 1  | 2  | 5   | 6   | 4                                 | 4                               |
| $0.30 \leq x < 0.40$  | 4.5   | 5   | 2  | 2  | 5   | 6   | 4                                 | 4                               |
| $0.40 \leq x < 0.50$  | 5   | 5   | 2  | 2  | 6   | 6   | 4                                 | 4                               |
| $0.50 \leq x < 0.60$  | 5   | 5   | 2  | 2  | 5   | 6   | 4                                 | 4                               |
| $0.60 \leq x < 0.70$  | 5   | 5   | 2  | 2  | 6   | 6   | 4                                 | 4                               |
| $0.70 \leq x < 0.80$  | 6   | 6   | 2  | 2  | 6   | 6   | 4                                 | 4                               |
| $0.80 \leq x < 0.90$  | 10  | 10  | 3  | 3  | 8   | 8   | 2                                 | 2                               |
| $0.90 \leq x$   | 5   | 5   | 2  | 2  | 6   | 6   | 4                                 | 4                               |
| Aliphatic Systems   |   |   |  |  |   |   |                                   |                                 |
| Difference Between Compound 1 and Earliest Project Date (10 bins) | Number of Rotatable Bonds Starting Compound | Number of Rotatable Bonds Ending Compound | Number of Hydrogen Bond Donors Starting Compound | Number of Hydrogen Bond Donors Ending Compound | Number of Hydrogen Bond Acceptors Starting Compound | Number of Hydrogen Bond Acceptors Ending Compound | Number of Rings Starting Compound | Number of Rings Ending Compound |
| $x < 0.10$  | 5   | 6   | 2  | 2  | 5   | 6   | 4                                 | 4                               |
| $0.10 \leq x < 0.20$  | 6   | 6   | 2  | 2  | 6   | 6   | 4                                 | 4                               |
| $0.20 \leq x < 0.30$  | 6   | 7   | 2  | 2  | 6   | 7   | 4                                 | 4                               |
| $0.30 \leq x < 0.40$  | 5   | 6   | 1  | 2  | 6   | 6   | 4                                 | 4                               |
| $0.40 \leq x < 0.50$  | 6   | 7   | 2  | 2  | 6   | 6   | 4                                 | 4                               |
| $0.50 \leq x < 0.60$  | 6   | 6   | 2  | 2  | 6   | 6   | 4                                 | 4                               |
| $0.60 \leq x < 0.70$  | 5   | 6   | 2  | 2  | 6   | 6   | 4                                 | 4                               |
| $0.70 \leq x < 0.80$  | 6   | 7   | 2  | 2  | 6   | 6   | 4                                 | 4                               |
| $0.80 \leq x < 0.90$  | 6   | 7   | 2  | 2  | 6   | 7   | 4                                 | 4                               |
| $0.90 \leq x$   | 7   | 7   | 2  | 2  | 7   | 7   | 4                                 | 4                               |

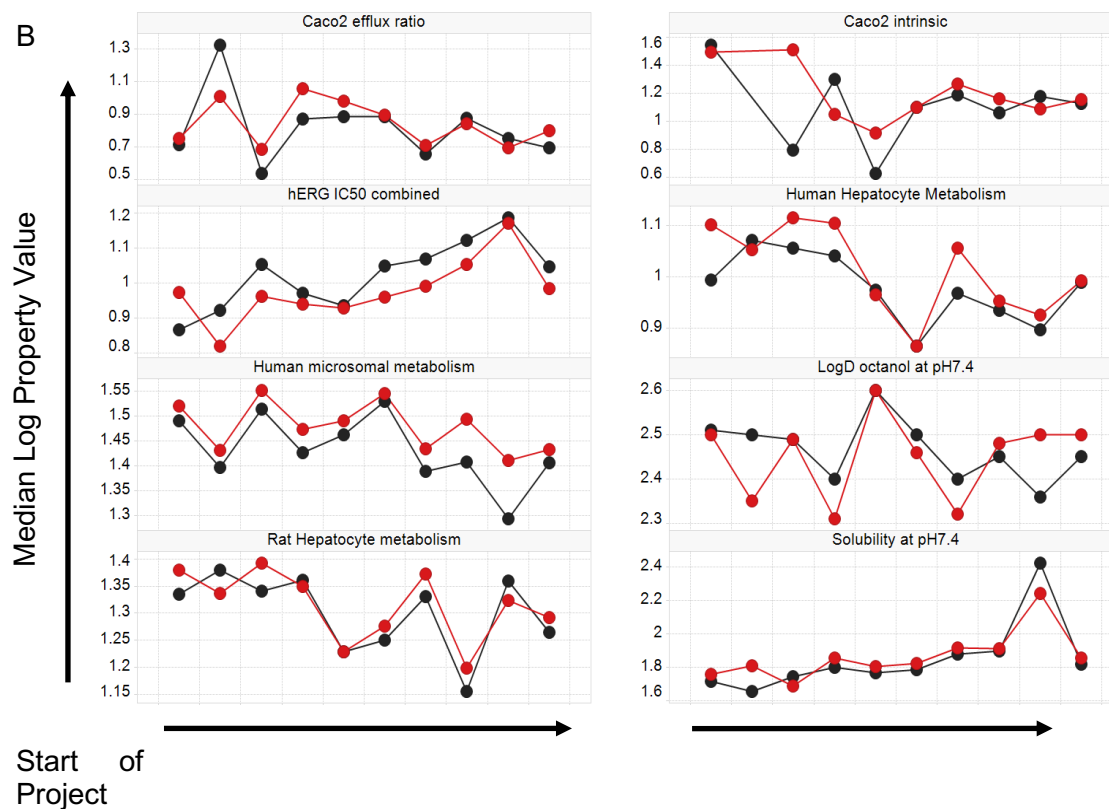
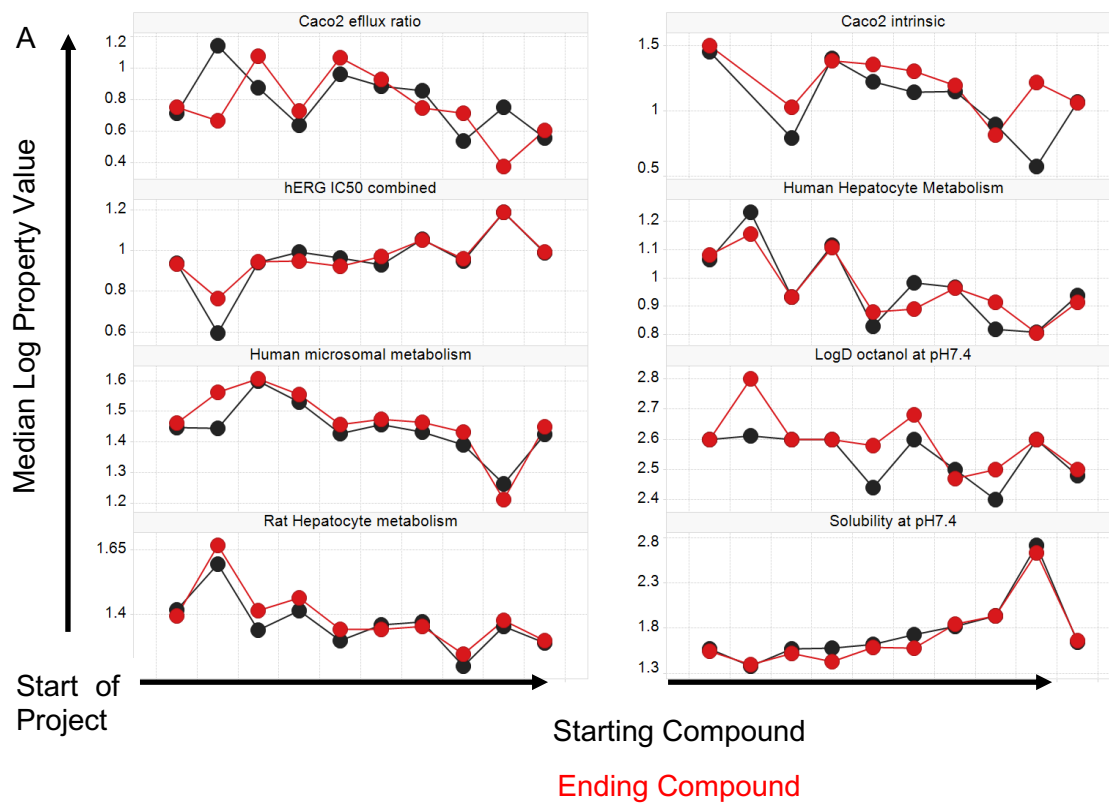


Figure 22: Median log property value of the starting compound (compounds transformed from) and the ending (compound transformed to) compound throughout the course of the project for each test. (A) Aromatic Systems, (B) Aliphatic systems.

Figure 22 shows how the assay property values change throughout the course of the project for both aromatic and aliphatic system. Regardless of the system, generally the trends are the same, highlighting the fact that irrespective of what system the transformation is being performed on, the design goal of the compound (such as increasing solubility or decreasing logD) ultimately leads to the same desired property changes, even if the route to these changes differs.

Comparing Table 9, Figure 21 and Figure 22 offers an overview of how the compounds within a project generally evolve throughout the course of a project and the effects of changing the chemistry of the compound on the identified properties. It is generally observed that the trends across the whole project between aromatic and aliphatic systems do not notably differ. A notable example is for the median logD and solubility that both have an increased log property value at the penultimate stage of the project (segment 9), (log value increasing from the previous segment for the ending compound). At the same point in a project the number of rotatable bonds also increases.

One notable difference between systems is that of the median logD, on aliphatic systems, differ, between the starting and ending compound at the penultimate stage of the project (9<sup>th</sup> segment). No such difference is identified in aromatic systems. Despite this, for both systems, this 9<sup>th</sup> segment of the project sees the highest median log solubility result of compounds across all stages of the project.

Comparisons of compounds and their properties have been previously published, and it has been shown that 10 or fewer rotatable bonds in a compound are likely to have good oral bioavailability in rats, whereas increased number of rotatable bonds is bad for the permeation rate<sup>179</sup>. In Figure 22 and Table 9 it is observed that permeability (Caco-2 intrinsic) decreases through the course of the project, and although the number of rotatable bonds does not drastic change throughout the course of the project, in both aromatic and aliphatic systems the largest median of rotatable bonds are observed towards the end of the project thus supporting the authors' findings<sup>179</sup>. Furthermore, Caco-2 efflux ratios are observed to increase, regardless of the system in the median log property value early in the project (especially in the 2<sup>nd</sup> segment of the project, out of the 10 segments that the project has been split into). The median log property value for Caco-2 efflux for compound 1 increases from 0.71 to 1.32 log units, and it then proceeds to decrease to 0.54 while never increasing above 0.89 log units. It has been previously shown, that the permeability of marketed drugs does not correlate with other properties well such as logD. The reason for this observation is that different drugs interact with different transporters, but it is not clear which drugs/transporters are responsible or how much they influence the permeability of Caco-2<sup>180</sup>.

Analysing the trends of the median log property difference for each assay test for either aromatic or aliphatic systems shows that the trends observed do not always correspond with the chemists' expectations. Generally, the log median Caco-2 efflux ratio decreases throughout the project for both aromatic and aliphatic systems. The median log value for human microsomal metabolism increases first before decreasing, and typically ends up with the same median log value as it did at the start of the project for aromatic systems. As mentioned previously, a lower human microsomal metabolism is preferred and for aliphatic systems the trend does show a decrease in the median log property value, albeit a minor change from the start of the project. The median log value of human hepatocyte decreases as expected throughout the project, while for rat hepatocyte metabolism the median log property values for aliphatic values decrease slightly through a project. The median log property value for solubility increases throughout the project but decreases towards the end for aromatic systems. Despite this, the solubility ends up about the same value as what was originally observed at the beginning of the project. To add to this, generally, the ending compound solubility for aromatic systems falls just below the value for the starting compound (generally the solubility decreases) whereas it increases fractionally on aliphatic systems. For Caco-2 intrinsic, the median log value increases for both aromatic and aliphatic systems before decreasing in both cases. Finally, for hERG the log property value increases throughout the project for aliphatic systems (with an initial decrease at the beginning of the project) while for aromatic systems the median log property decreases before increasing and decreasing again at the end of the project.

In a published report the introduction of groups that contained hydrogen bond acceptors or donors resulted in a larger decrease in solubility than their effect on lipophilicity would suggest<sup>162</sup>. However, in our analysis the median number of hydrogen bond acceptors or donors for starting compound appears to remain neutral throughout the project, as does the median log solubility value. To test this, we calculated the correlation coefficient. The results are shown in Figure 23 where a spearman rank test was used to calculate the correlation coefficient (A) and correlations (B) with a p-value > 0.01 being considered insignificant. The figure shows that the values are positively correlated, however, this is not a significant finding based on a p-value of > 0.01. Increasing the p-value to > 0.05 would result in two of the values being significant. Hydrogen bond donors and solubility would be significantly positively correlated if the p-value significance level was raised to >0.05.

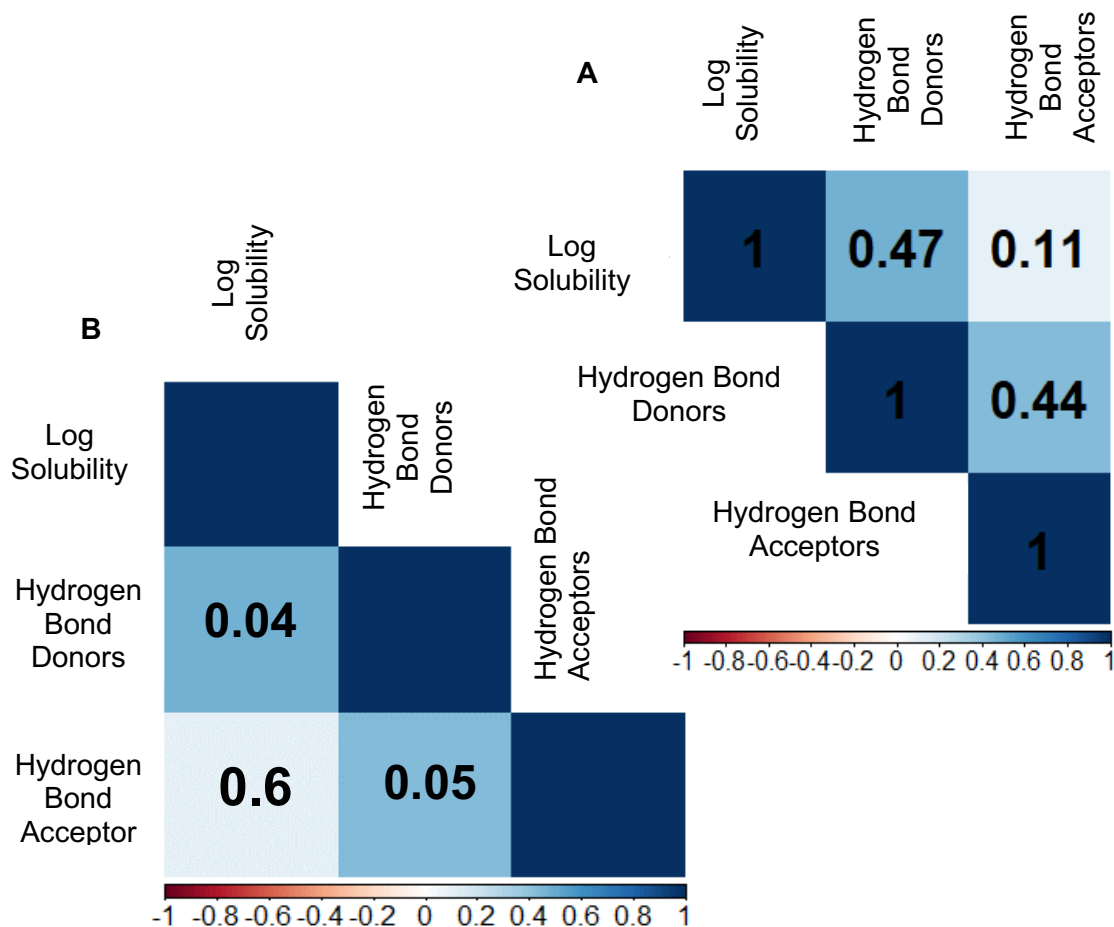


Figure 23: Correlation coefficient and statistical significance of the median log solubility property value for compound 1 against the median number of hydrogen bond donors and hydrogen bond acceptors for compound 1 over the course of the project. (A) represents the correlation coefficient and (B) the statistical significance. The project was split into 20 bins and the median property result and physicochemical properties were calculated for each bin. The darker the blue, the more positively correlated the values are.

#### 4.3.2 Analysis of the most frequently observed MMP transformations and their effect on compound properties

The same properties as we analyse in the previous section are now considered on the level of individual and groups of MMP transformations. Density plots for each of the top 5 transformations are shown below, performed on the top 5 atomic environments for aromatic (Figure 24) and aliphatic (Figure 25) systems.

Despite there being many reasons for an MMP to be investigated in a synthetic project, including feasibility and building block availability, chance matches to previous compounds, library design and potency optimisation will involve many changes to be made in order to improve the properties of a compound. The aim of our analysis was to highlight the matched pairs where a property varies significantly and therefore may signify influence a chemical

change of interest for future implementation. Figures 22 and 23 show the effect of each transformation on the molecular environments on the measured assay properties (logD solubility, metabolism, permeability and hERG).

#### 4.3.2.1 Aromatic Systems

An example with a notable change is the increase in logD when replacing a hydrogen atom with a chlorine atom on an aromatic system, particularly when on the most frequently observed aromatic local atomic environment (the immediate local environment that the transformation is performed on) (Figure 24). Chlorine is more lipophilic than hydrogen and therefore it would be expected that compounds that involve this transformation will show an overall increase in lipophilicity. When logD increases it is expected that solubility decreases, which is demonstrated when hydrogen is replaced by chlorine. Replacing a fluorine atom with a hydrogen atom will change the properties of a compound in different ways depending on the local atomic environment in which the transformation is carried out. For example, human hepatocyte metabolism is shown to increase on the third most frequently occurring aromatic atomic environment (an aromatic ring with a nitrogen attached off the ring). However, as can be seen in Figure 24, the other most frequently occurring aromatic environments do not share the same observed human hepatocyte metabolism increase between the starting compound and the ending compound; only a minimal change (not significant) is observed.

Replacing fluorine by a hydrogen shows an increase in solubility on the fourth most frequently occurring aromatic atomic environment, a heterocyclic ring containing a nitrogen, of which is not observed by the other frequently occurring environments. Whereas, the same transformation, on the same environment, decreases the logD more significantly, than the other most frequently observed transformations of which is as expected given that we see an increase in solubility.

Hydrogen to chlorine increases hERG activity for environment 4 but decreases hERG for environment 3 (Figure 24). Basic compounds are known to induce hERG activity, while  $\pi$ -stacking interactions are important for the interaction of a compound with the hERG protein. For environment 3, with the aniline type nitrogen, hERG activity is decreased which could be due to the decrease in the basicity of the amine due to the presence of the electron withdrawing chlorine group, as well as decreasing the electron density in the aromatic ring, reducing the potential for  $\pi$ -stacking<sup>181</sup>. An interesting example is when hydrogen is replaced by chlorine on a pyridine like environment (4<sup>th</sup> most frequently occurring atomic environment), where we observe an increase in hERG. Generally, when you decrease the basicity of the nitrogen, you would expect a decrease in hERG activity.

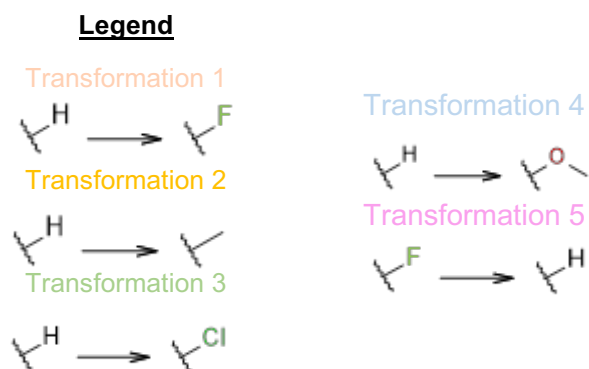
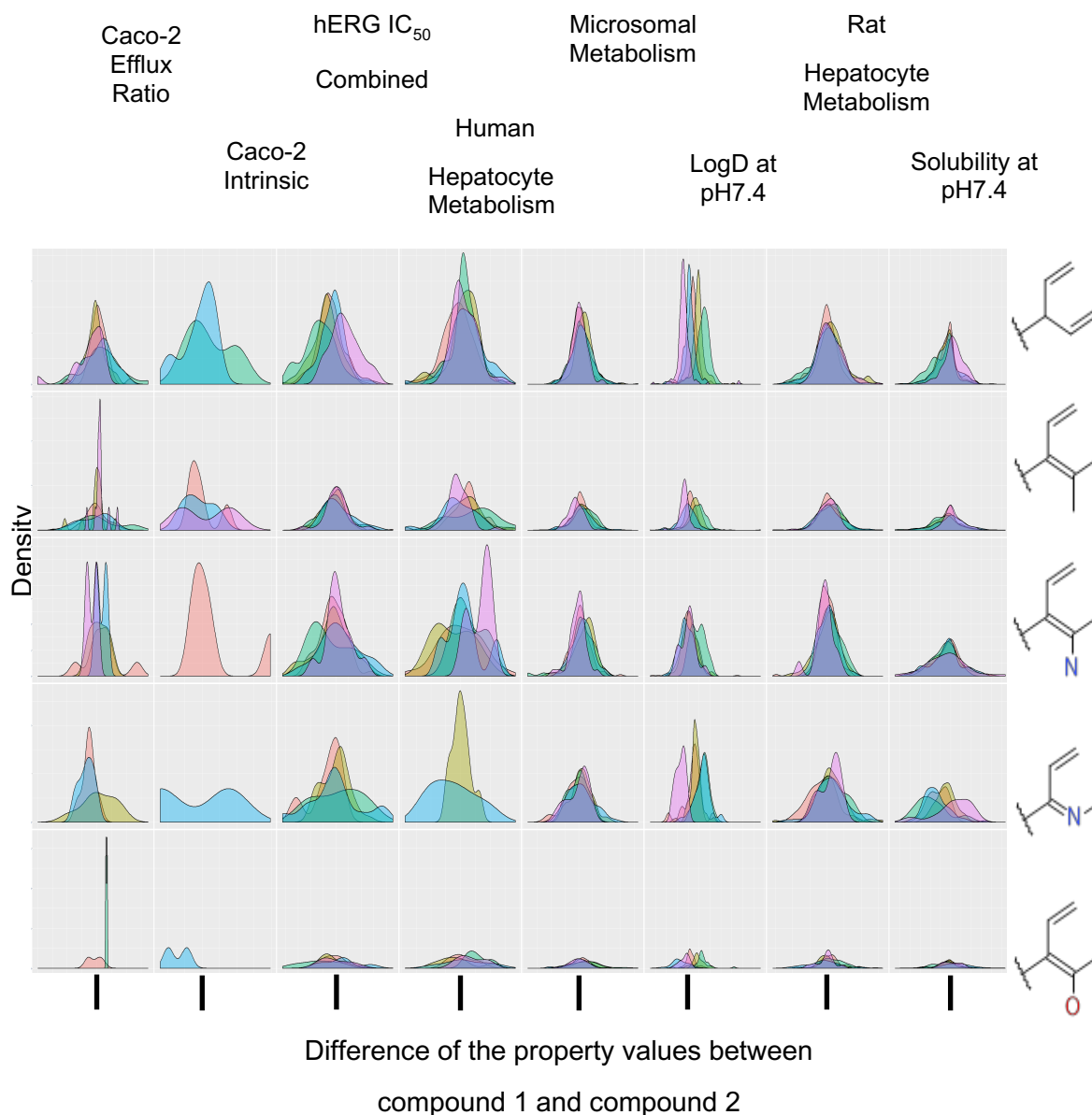


Figure 24: Density plot showing the difference in property values from compound 1 to compound 2 as a function of the top 5 transformations and atomic environments where data was available, for aromatic systems. The transformation and the atomic environments that the transformation is performed on effects the property change.

Values outside  $\pm 4$ SDs were removed from the plots. The difference in property was calculated as the property value of compound 2 – the property value of compound 1.

#### 4.3.2.2 Aliphatic Systems

We have identified some interesting properties for aliphatic systems (Figure 25). The second most frequently occurring aliphatic local atomic environments, in which the transformations are identified on, which are identified as tertiary amine/di-ethyl amine (Figure 25) show the effects on the properties of a compound by replacing a hydrogen atom with a methyl and *vis versa* a methyl group with a hydrogen atom. What we find is that the solubility is not significantly influenced by such transformations.

The 5<sup>th</sup> most frequently occurring transformation (hydrogen  $\gg$  ethyl) is shown to increase the logD regardless of the substitution position (environment) in which they are observed on which they are performed. The solubility however decreases when hydrogen is replaced by ethyl on the fourth most frequently substituted aliphatic position, which tends to be a secondary amine. This same trend is not observed on the other frequently substituted positions.

When analysing the transformations performed on the 4<sup>th</sup> most frequently occurring atomic environment, a secondary amine, the third most frequently occurring transformation of hydrogen to oxygen shows that permeability decreases due to less permeability and therefore less likelihood of crossing the lipid membrane. LogD decreases for the 4<sup>th</sup> most frequently occurring transformations (hydrogen to a methoxy group), when the electron-donating group is feeding electrons into the aromatic ring.



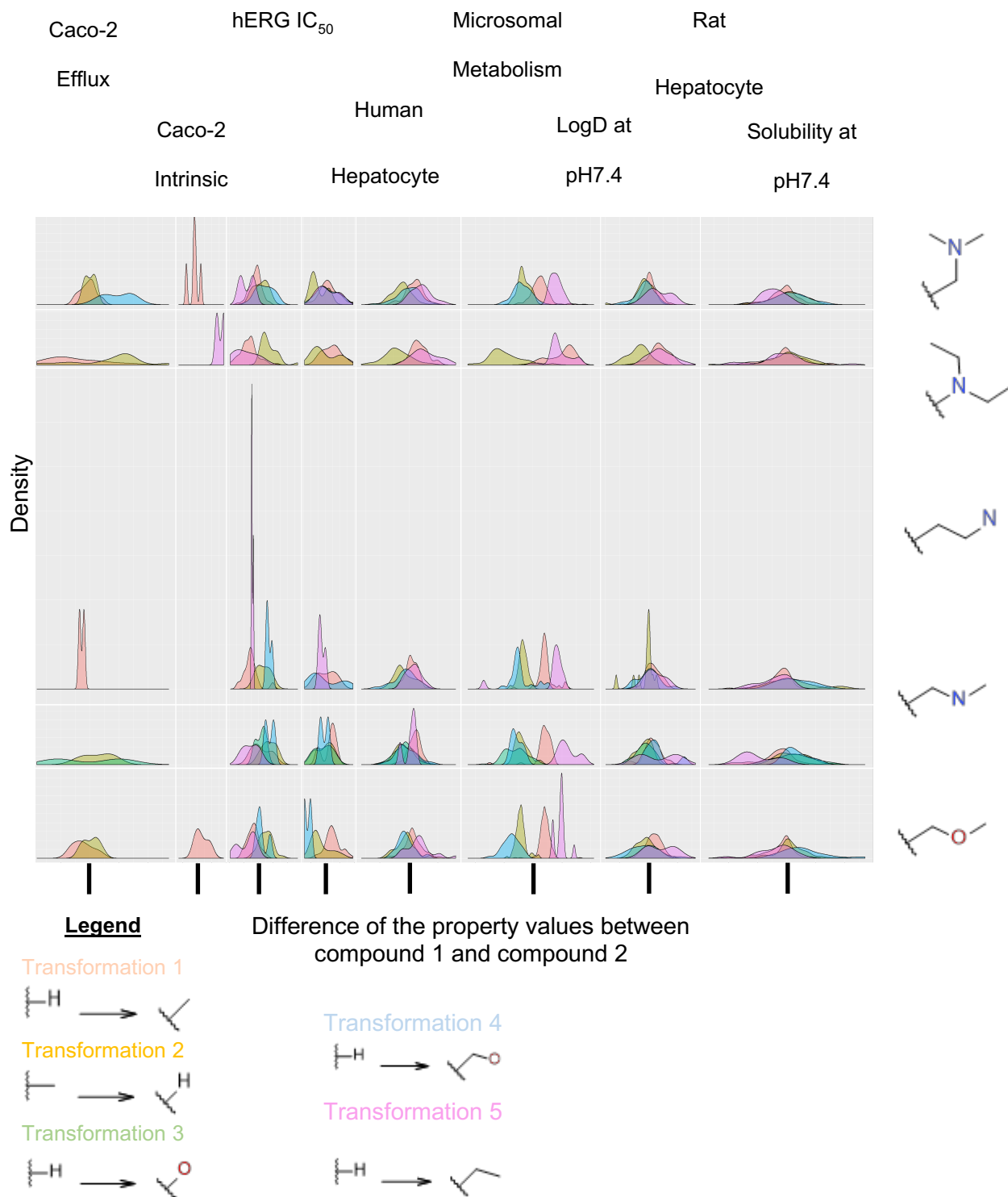


Figure 25: Density plot showing the difference in property values from compound 1 to compound 2 as a function of the top 5 transformations and atomic environments where data was available, for aliphatic systems. The transformation and the atomic environments that the transformation is performed on affects the property change. Values outside  $\pm 4SDs$  were removed from the plots. The difference in property was calculated as the property value of compound 2 – the property value of compound 1.

#### 4.3.3 Analysis of both the proportion of significant property changes as well as the quantitative amount of change, as a function of performing transformations on different atomic environments

The proportion of times that the property increases the measured property value against the times that the measured property value decreased was considered. The aim of this analysis was to observe the effects on properties of interest across the five most frequently occurring environments for each system (aromatic and aliphatic) and we can identify that the type of atomic environment that the transformation is performed on, can drastically affect the result of the measured assay test.

The study was then extended to understand by how much the property changes in each direction (significant increase, significant decrease or minimal change), and the median log value of each measured assay property as a function of the top 5 transformations performed on the top 5 atomic environments for aromatic and aliphatic systems was assessed. The aim was to understand how much each transformation affects the measured assay property and therefore understand which transformation yields a more favourable response on a given environment. In addition, it is also possible to see instances where a transformation has a larger change in the median property value compared to other transformations performed on the same atomic environment.

In a previous study<sup>81</sup> the authors analysed each set of property value data (hERG, solubility and lipophilicity) and the most frequent transformations. Following this, each of the most frequently occurring transformations were split into unfavourable, zero and favourable changes. In the case of hERG results<sup>81</sup>, unfavourable denotes an increase in the binding affinity, and favourable relates to a decrease in binding affinity. The frequently occurring transformations that the authors identify correlate well with those identified in our study. The authors show that many the transformations (40%) have no effect on the hERG activity and this correlates well with our own findings (59% on aliphatic systems and 63% on aromatic systems) (Table 10).

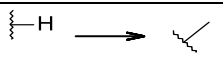
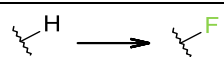
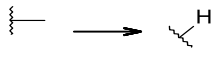
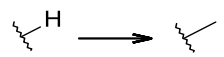
Table 10: Percentage increase, decrease and minimal change observed for hERG, logD and Solubility

| Aliphatic Systems    |      |      |            | Aromatic Systems     |      |      |            |
|----------------------|------|------|------------|----------------------|------|------|------------|
| Change               | hERG | LogD | Solubility | Change               | hERG | LogD | Solubility |
| Minimal Change       | 59%  | 31%  | 39%        | Minimal Change       | 63%  | 38%  | 38%        |
| Significant Decrease | 21%  | 34%  | 30%        | Significant Decrease | 19%  | 29%  | 33%        |
| Significant Increase | 20%  | 34%  | 31%        | Significant Increase | 18%  | 33%  | 29%        |

The study analyses the change in the top 30 transformations, regardless of specific atomic environment. There is agreement in the most frequently occurring transformations between the authors' study and this work (Table 11), with exception of specific rank positions in terms of occurrence of transformation. This is unsurprising given a different dataset.

The highest proportion of instances that correspond to significantly decreasing the hERG shows that for aliphatic systems, there are no transformations in the top 20 most frequently occurring transformations that significantly decrease hERG over 50% of the time. For aliphatic systems there are five instances where hERG decreases significantly over 50% of the time. In the authors study, there is only one instance where the transformation decreases hERG over 50% of the time and is the 28<sup>th</sup> most frequently observed transformation of their study. Replacing either a hydrogen atom or a methyl group with a benzene group yields a high proportion of occurrences where hERG is reported to decrease significantly. This is surprising because of the increase created in lipophilicity.


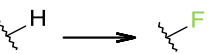
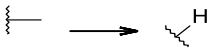

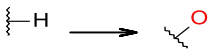
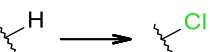
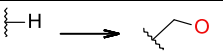
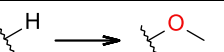
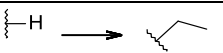
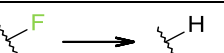
Table 11: Percentage of occurrences identified in our study of hERG significant increasing, decreasing or having minimal effect because of the transformation performed

| Aliphatic Systems   |          |         |         | Aromatic Systems   |          |         |         |
|---|----------|---------|---------|--|----------|---------|---------|
| Transformation  | Min Diff | Sig Dec | Sig Inc | Transformation   | Min Diff | Sig Dec | Sig Inc |
|  | 70%      | 22%     | 8%      |  | 78%      | 15%     | 7%      |
|  | 65%      | 7%      | 28%     |  | 71%      | 20%     | 8%      |

|  |     |            |     |  |     |     |     |
|--|-----|------------|-----|--|-----|-----|-----|
|  | 40% | 4%         | 56% |  | 48% | 45% | 7%  |
|  | 53% | 1%         | 46% |  | 72% | 13% | 15% |
|  | 50% | 50%        |     |  | 70% | 8%  | 23% |
|  | 35% | <b>59%</b> | 6%  |  | 66% | 15% | 20% |
|  | 66% | 9%         | 25% |  | 52% | 4%  | 44% |
|  | 31% |            | 69% |  | 61% | 5%  | 34% |
|  | 13% | 25%        | 63% |  | 60% | 31% | 9%  |
|  | 54% | 44%        | 3%  |  | 60% | 36% | 4%  |
|  | 63% | 21%        | 16% |  | 69% | 23% | 9%  |
|  | 33% | <b>67%</b> |     |  | 58% | 2%  | 41% |
|  | 30% | <b>63%</b> | 7%  |  | 61% | 30% | 9%  |
|  | 67% | 17%        | 17% |  | 62% | 31% | 7%  |
|  | 30% | <b>65%</b> | 4%  |  | 67% | 10% | 22% |
|  | 45% | <b>51%</b> | 4%  |  | 73% | 12% | 15% |
|  | 64% | 9%         | 27% |  | 76% | 8%  | 15% |
|  | 59% | 26%        | 15% |  | 73% | 9%  | 18% |
|  | 55% | 40%        | 5%  |  | 76% | 20% | 3%  |
|  | 58% |            | 42% |  | 54% | 9%  | 37% |

The first is that of when a methyl group is replaced by a methoxy, which when in our studies was identified on aliphatic systems (Table 12) a significant decrease in the logD was reported<sup>81</sup> and n shows agreement with the published study. However, in our study when performed on aromatic systems, the transformation does not reach the 50% threshold, but does come very close. The replacement of chlorine to methoxy is another transformation that we have identified of decreasing logD and is supported by the findings of the previous study<sup>81</sup>. Next, the authors of the report show that a hydrogen being replaced by an oxygen moves the logD in a favourable direction in over 83% of the occurrences; Table 12 shows that when the transformation is carried out on aliphatic systems there is a strong correlation with the authors' findings. Replacing a hydrogen atom with an alcohol functional group is again consistent with the previously reported findings when performed on aliphatic systems (Table 12). The last two transformations that the authors report as greater than 50% occurrence of logD moving in a favourable direction is that of chlorine replaced by a nitrile group and a nitrogen replaced by an oxygen. Neither of these is observed in the top 20 frequently occurring transformations on either aromatic or aliphatic systems in our study. However, replacing hydrogen with a nitrile group does show that 50% of the occurrences significantly decrease the logD. It has been shown<sup>182</sup> that replacing a hydrogen with a nitrile group can significantly reduce the logD of a compound, however, replacing a halogen or a methyl group by a nitrile will decrease the logD even more significantly, presumably due to the steric effects of a nitrile group.

Table 12: Percentage of occurrences in our study of logD significant increasing, decreasing or having minimal effects because of the transformation performed

| Aliphatic Systems   |          |            |         | Aromatic Systems   |          |         |         |
|---|----------|------------|---------|--|----------|---------|---------|
| Transformation  | Min Diff | Sig Dec    | Sig Inc | Transformation   | Min Diff | Sig Dec | Sig Inc |
|  | 39%      | 5%         | 57%     |  | 67%      | 6%      | 27%     |
|  | 34%      | <b>62%</b> | 4%      |  | 40%      | 5%      | 55%     |
|  | 18%      | <b>80%</b> | 2%      |  | 15%      | 2%      | 82%     |
|  | 22%      | <b>74%</b> | 4%      |  | 63%      | 15%     | 22%     |
|  | 10%      | 3%         | 88%     |  | 70%      | 27%     | 4%      |

|  |     |            |     |  |     |            |     |
|--|-----|------------|-----|--|-----|------------|-----|
|  | 2%  | 2%         | 96% |  | 42% | <b>52%</b> | 6%  |
|  | 48% | 42%        | 10% |  | 16% | <b>80%</b> | 4%  |
|  | 11% | <b>88%</b> | 1%  |  | 18% | <b>79%</b> | 2%  |
|  | 5%  | <b>95%</b> |     |  | 50% | 42%        | 8%  |
|  | 14% | 2%         | 84% |  | 22% | 1%         | 77% |
|  | 68% | 13%        | 19% |  | 62% | 24%        | 14% |
|  | 8%  | 1%         | 91% |  | 48% | 45%        | 7%  |
|  | 15% | 2%         | 83% |  | 48% | 3%         | 50% |
|  | 34% | 19%        | 47% |  | 7%  | 1%         | 92% |
|  | 3%  | 2%         | 95% |  | 63% | 26%        | 11% |
|  | 11% | 2%         | 87% |  | 70% | 20%        | 10% |
|  | 5%  | <b>58%</b> | 37% |  | 41% | 46%        | 13% |
|  | 19% | <b>52%</b> | 29% |  | 65% | 4%         | 31% |
|  | 53% | 14%        | 34% |  | 61% | 12%        | 26% |
|  | 6%  | <b>92%</b> | 1%  |  | 20% | <b>75%</b> | 5%  |

Finally, the authors study<sup>81</sup>, considers how solubility is affected by the transformation in terms of a significant increase, decrease or a minimal change in the solubility. The authors do not identify any instances where the solubility is affected favourably in over 50% of the cases, however we show, Table 13 it is shown that there are four instances where the solubility

increases over 50% of the time, including the removal of a boc-protecting group, replacing a chlorine and replacing a benzene.

Table 13: Percentage of occurrences identified in our study of solubility significant increasing, decreasing or having minimal effects a result of the transformation performed

| Aliphatic Systems |          |         |            | Aromatic Systems |          |         |            |
|-------------------|----------|---------|------------|------------------|----------|---------|------------|
| Transformation    | Min Diff | Sig Dec | Sig Inc    | Transformation   | Min Diff | Sig Dec | Sig Inc    |
|                   | 48%      | 30%     | 22%        |                  | 49%      | 37%     | 15%        |
|                   | 43%      | 24%     | 33%        |                  | 43%      | 40%     | 18%        |
|                   | 37%      | 15%     | 48%        |                  | 38%      | 53%     | 8%         |
|                   | 38%      | 14%     | 48%        |                  | 39%      | 38%     | 23%        |
|                   | 38%      | 44%     | 18%        |                  | 50%      | 18%     | 32%        |
|                   | 24%      | 69%     | 7%         |                  | 39%      | 23%     | 38%        |
|                   | 36%      | 17%     | 47%        |                  | 40%      | 6%      | <b>55%</b> |
|                   | 45%      | 16%     | 39%        |                  | 45%      | 10%     | 45%        |
|                   | 11%      | 4%      | <b>85%</b> |                  | 36%      | 46%     | 18%        |
|                   | 37%      | 49%     | 14%        |                  | 48%      | 43%     | 9%         |
|                   | 42%      | 23%     | 35%        |                  | 41%      | 25%     | 34%        |
|                   | 22%      | 73%     | 5%         |                  | 44%      | 10%     | 46%        |

|  |     |     |            |  |     |     |            |
|--|-----|-----|------------|--|-----|-----|------------|
|  | 49% | 39% | 13%        |  | 38% | 51% | 11%        |
|  | 28% | 52% | 20%        |  | 29% | 64% | 7%         |
|  | 20% | 66% | 14%        |  | 47% | 25% | 28%        |
|  | 41% | 47% | 12%        |  | 42% | 34% | 23%        |
|  | 27% | 42% | 32%        |  | 51% | 25% | 23%        |
|  | 41% | 24% | 36%        |  | 46% | 27% | 27%        |
|  | 44% | 39% | 17%        |  | 50% | 26% | 23%        |
|  | 20% | 8%  | <b>72%</b> |  | 39% | 10% | <b>52%</b> |

Our analysis of the data contributes to the previous reports; this section extends this work by explicitly comparing the effects on property values for different frequently observed atomic environments in aromatic or aliphatic systems. Notably, we observe instances where our findings are in concurrence with the previous study, however, when delving deeper and splitting the most frequently occurring transformations by the system they are identified on (aromatic or aliphatic), it shows that the concurrence is not always true and the system needs to be very carefully considered, playing a major part in the “success” of the transformation to move a desired property in a favourable way.

#### 4.3.3.1 Aromatic Systems

Regardless of the atomic environment, many of the transformations in the top 5 most frequently occurring, result in a significant increase in the logD between the starting compound and the ending compound suggesting that some changes are not as sensitive to the atomic environment.



An interesting example is observed when hydrogen is replaced by a chlorine on the third (aromatic ring structure with a nitrogen coming off the ring system) and the fourth (heterocyclic environment with a nitrogen in the ring (pyridine like)) most frequently occurring aromatic atomic environments (where the transformation took place) do not observe any instances of hydrogen being replaced by chlorine significantly decreasing the human hepatocyte metabolism. The instances on the third aromatic atomic environment (nitrogen coming off the aromatic ring) does not have all instances described as significant unlike the case when the nitrogen is within the ring (meaning there are instances of minimal change being observed). When the nitrogen is in the ring, nucleophilic substitution reactions can occur with the chlorine<sup>183</sup> which is readily substituted due to its strong leaving group abilities<sup>184</sup>. In a related literature example<sup>185</sup> a nucleophilic aromatic substitution reaction was catalysed by rat liver microsomes (specifically, Glutathione S-Transferase 1) and the authors report that the nucleophilic aromatic substitution of 2-chloropyridine derivatives was affected by the position of the substituents as well as the strength of the electron-withdrawing properties of the substituents. Replacing hydrogen by chlorine on an aromatic ring bearing a nitrogen functionality results in human hepatocyte metabolism significantly increasing at a median quantity of just under 0.5 log units (Figure 26). Most of the instances when this transformation occurs on this environment (nitrogen coming off the aromatic ring (3<sup>rd</sup> most frequently occurring)), result in a minimal change of the human hepatocyte metabolism, however where there is a significant increase would be worth exploring.

This replacement of hydrogen to chlorine generally shows the large instances of significant decreases in hERG, which corresponds to what the chemists might expect to see. However, on the second and fourth most commonly occurring atomic environments, on which the transformation takes place, a phenyl group with a carbon attached to the ring and the ring system with the nitrogen atom inside the do not show a majority proportion of instances where the hERG value is decreased. In the instance of hydrogen to chlorine being carried out on the second most commonly occurring aromatic atomic environments (phenyl group with a carbon attached off the ring) shows a roughly proportionate significant increase or decrease in hERG value. Although, in instances where it does significantly decrease, it does so on par with the significant increases observed on the other environments (between 0.4 and 0.5 log units, Figure 27). However, most instances for this example are minimal changes and are therefore not significant. The instances on the fourth most frequently occurring atomic environment (ring system with a nitrogen in it) suggest the hERG property increases, therefore, the replacement of a hydrogen with a chlorine on a pyridine like system is unlikely to yield favourable hERG activity. Having said this, hERG does not decrease significantly for the other frequently

occurring transformations. Such cases have been discussed<sup>186</sup> including the largest observed change involved replacing an imidazole with a methyl tetrazole and also replacing a hydrogen to a methoxy in aromatic rings reduces hERG binding. Our study shows significant decreases of hERG when this transformation (hydrogen to methoxy) is performed on the observed aromatic atomic environments, however, the proportion of these instances is minimal in comparison to occurrences of significant increases and changes that were deemed minimal, again supporting the need to fully understand the chemistry that the transformation is being performed on.

Another finding in this thesis that supports those found in the previous study<sup>162</sup> discussed earlier, where the addition of a bromine to an aromatic system, shown that the solubility decreased in 98% of occurrences. The 104 instances out of a total of 153 (68% of instances) show a solubility decrease, for which 37% of the instances were considered a minimal change (failed to reach the 0.3 log property value difference) and only 7% were considered a significant increase in the log solubility property value. Therefore, as the previous study noted, if solubility is causing a problem in the compound, a bromine should not be added and if present in the structure should be removed<sup>162</sup>.

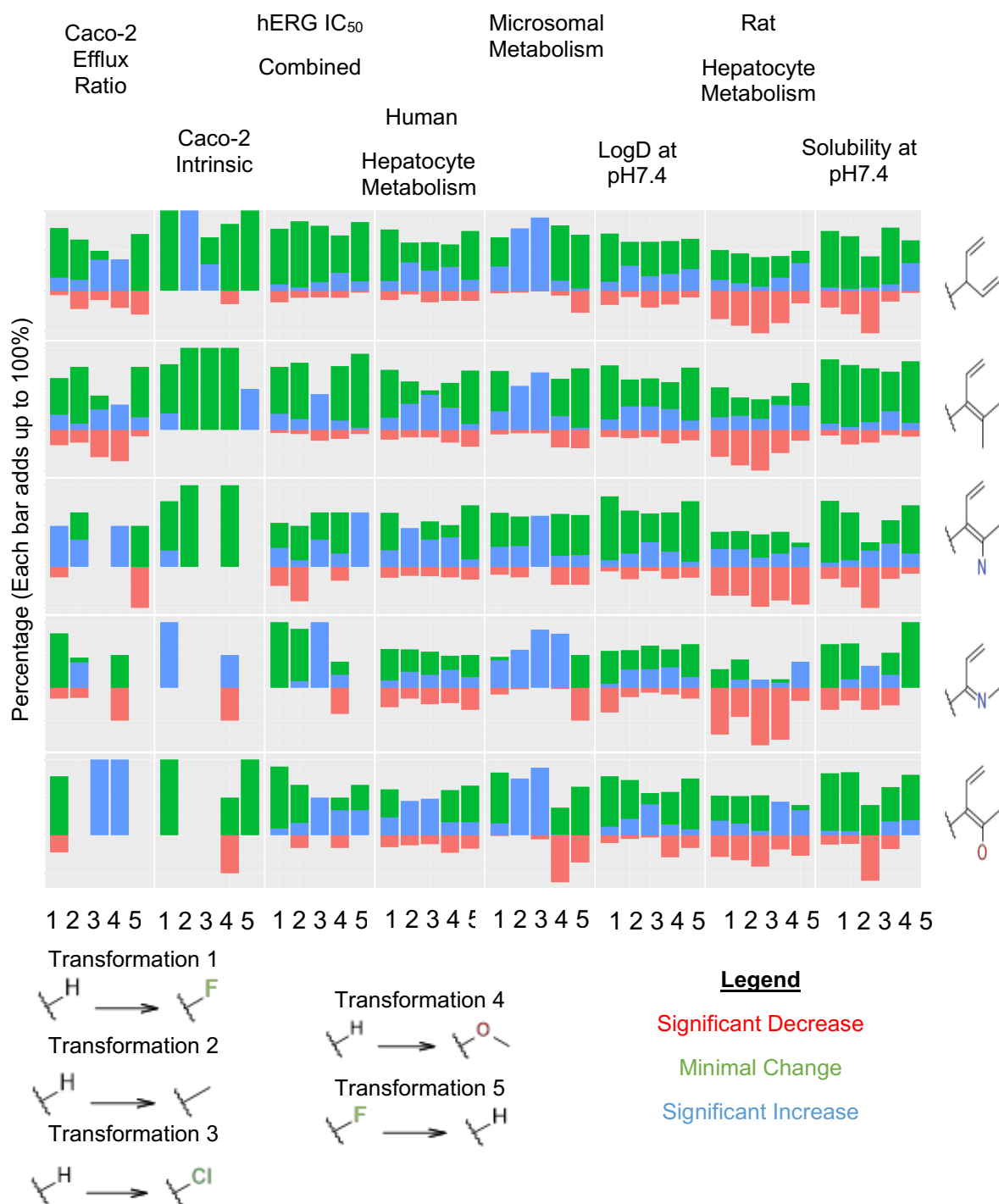


Figure 26: Bar chart showing the percentage of changes of *properties* for each of the top 5 most frequently occurring transformation against the top 5 most frequently occurring atomic environments on aromatic systems. It shows that in many cases – a minimal log change is made to the measured properties. It also shows that that the same transformation on the same environment can result in different property changes. The same transformation is likely to have *different effects on the property* depending on the atomic environment the transformation is performed on. Additionally, some transformations lean more towards an effect on the *property* regardless of the atomic environment.

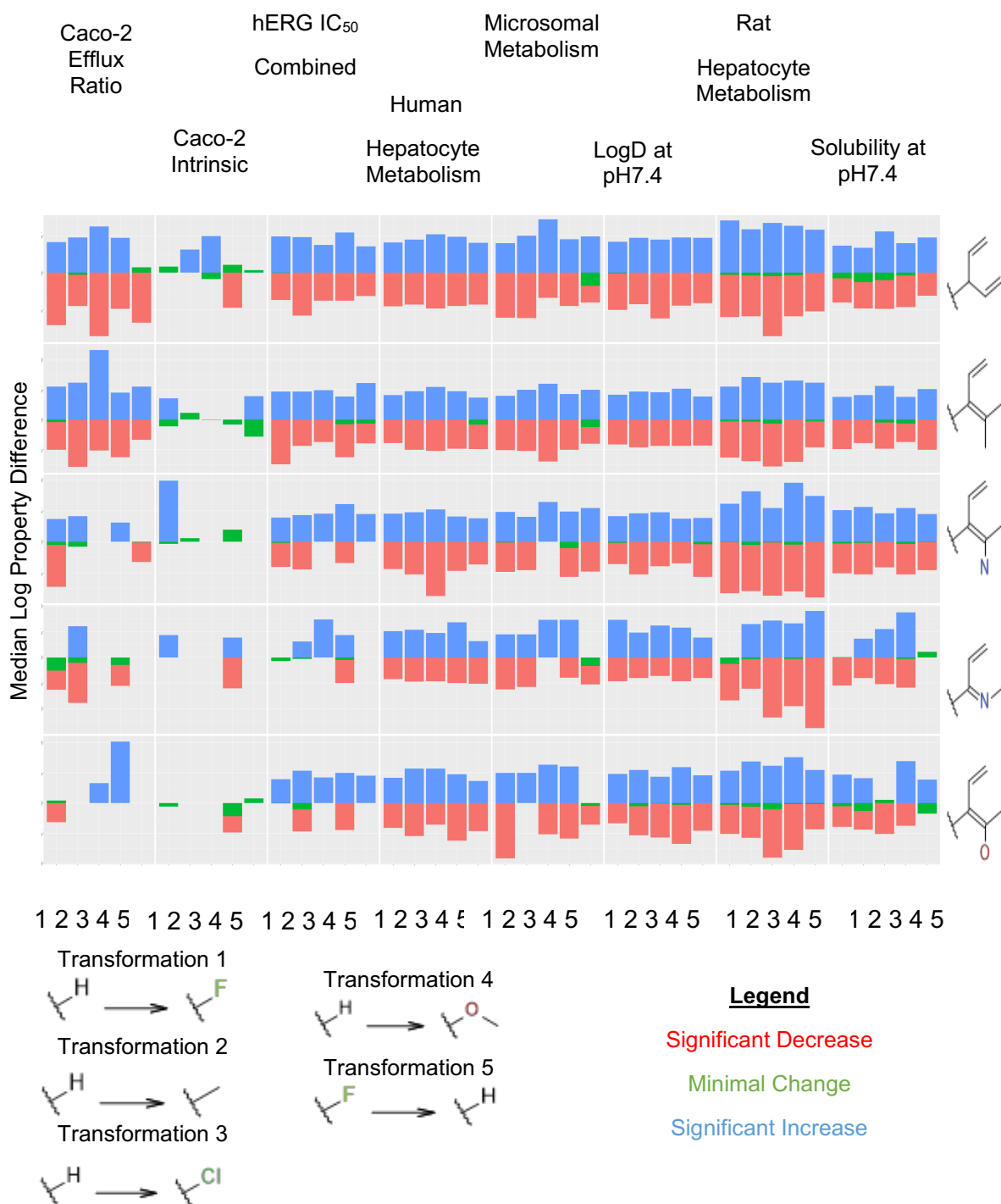


Figure 27: Bar chart showing the median log property change of *properties* for each of the top 5 most frequently occurring transformation against the top 5 most frequently occurring atomic environments on aromatic systems. It shows that in many cases – a minimal log change is made to the measured properties. It also shows that that the same transformation on the same environment can result in different property changes. The same transformation is likely to have *different effects on the property* depending on the atomic environment the transformation is performed on. Additionally, some transformations lean more towards an effect on the *property* regardless of the atomic environment.

#### 4.3.3.2 Aliphatic Systems

Hydrogen replaced by a methyl group or an ethyl group is shown to generally increase the logD regardless of the atomic environment (Figure 28). In instances where it does decrease the logD, it does so significantly with no minimal changes in the logD observed. The greatest median significant decrease observed in these two examples is approximately one log unit difference in logD when hydrogen is replaced by a methyl group on a secondary amine and nearly 1.7 log units decrease when hydrogen is replaced by an ethyl group on an ether (Figure 29), whereas methyl replaced by a hydrogen or hydrogen being replaced by an oxygen or alcohol, where data is available, is shown to decrease the logD (Figure 28). An exception to this is hydrogen being replaced by an oxygen on a primary amine, likely a result of aliphatic amines being oxidised by a radical-chain mechanism, specifically when primary aliphatic amines undergo oxidation one of the molecules of the organic inhibitor participates in chain termination<sup>187</sup>.

Another interesting example is that of the effects of a methyl replaced by a hydrogen and hydrogen replaced by an oxygen; the second and third most frequently occurring transformations on aliphatic systems, respectively, on the human hepatocyte metabolism. Notably when these transformations are identified on a tertiary amine/ di-ethyl amine human hepatocyte metabolism is shown to decrease in the majority of cases by approximately 0.4 log units when a methyl is replaced by a hydrogen, whereas for a hydrogen being replaced by an oxygen the human hepatocyte metabolism is significantly increased in all observed instances (Figure 28) and increases by approximately 0.3 log units.

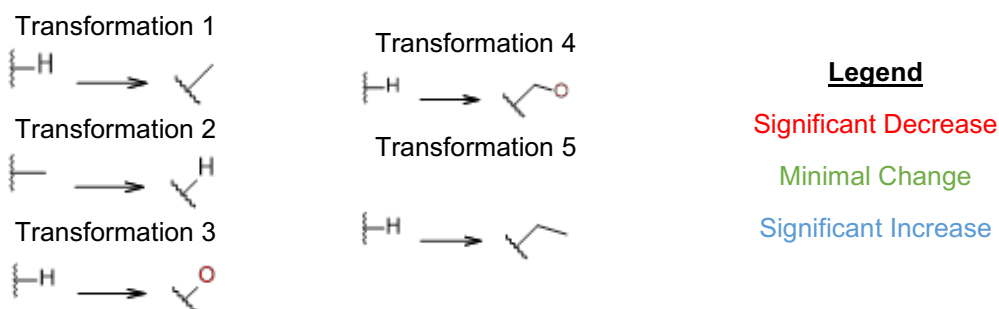
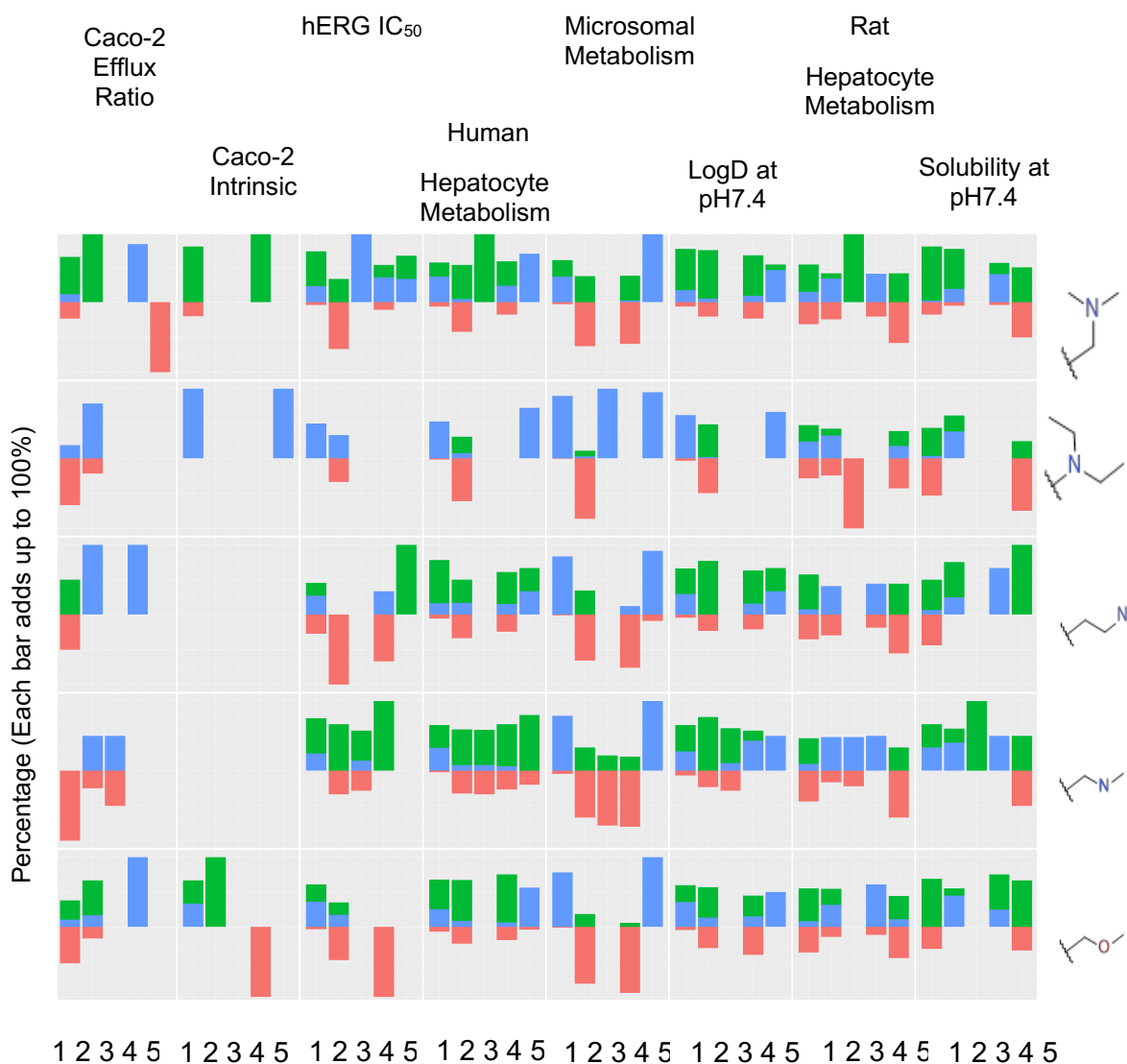


Figure 28: Bar chart showing the percentage of changes of properties for each of the top 5 most frequently occurring transformation against the top 5 most frequently occurring atomic environments on aliphatic systems. It shows that in many cases – a minimal log change is made to the measured properties. It also shows that that the same transformation on the same environment can result in different property changes. The same transformation is likely to have different effects on the property depending on the atomic environment the transformation is performed on. Additionally, some transformations lean more towards an effect on the property regardless of the atomic environment.

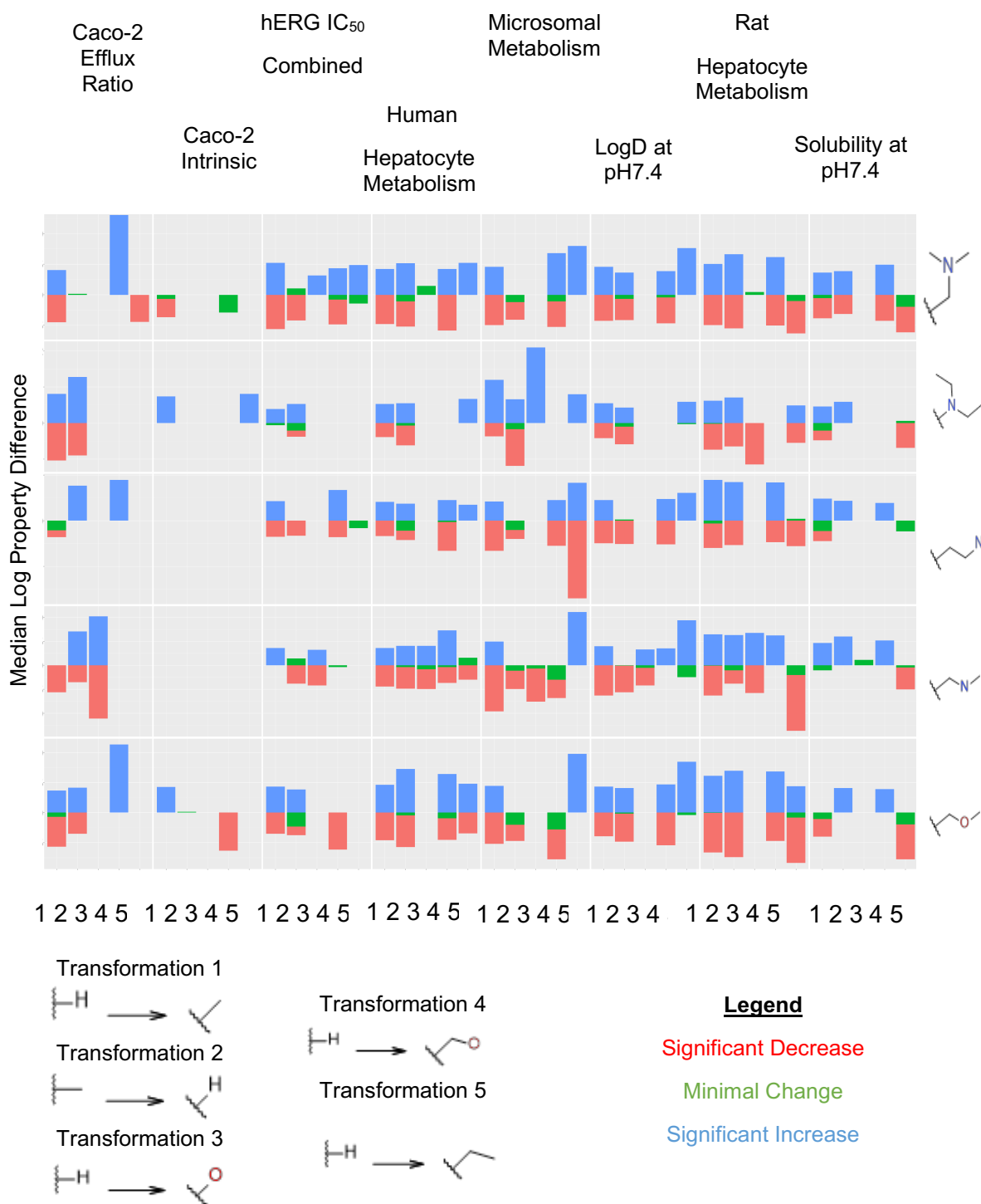


Figure 29: Bar chart showing the Median log property value of *properties* for each of the top 5 most frequently occurring transformation against the top 5 most frequently occurring atomic environments on *aliphatic systems*. It shows that in many cases – a minimal log change is made to the measured properties. It also shows that that the same transformation on the same environment can result in different property changes. The same transformation is likely to have *different effects on the property* depending on the atomic environment the transformation is performed on. Additionally, some transformations lean more towards an effect on the *property* regardless of the atomic environment.

#### 4.3.4 Analysis of extreme outliers ( $\pm 4$ Standard Deviations)

In our study we investigated the chemistry of those transformations performed on atomic environments where regardless of what the property was, the properties measured value change between the starting compound and the ending compound was considered an extreme change ( $\pm 4$  standard deviations). This analysis was performed for both aromatic and aliphatic systems and confirms that these extreme outliers are investigated for very specific reasons, such as changing the acidity of the compound, which has large effects on the property values. Overall, all observed transformations that result in a large change in a property value are the result of a change in the ion class, or the addition or removal of a large fragment.

##### 4.3.4.1 Aromatic Systems

On aromatic systems, the most frequently occurring atomic environment (the local environment) that the transformation is performed on is the most predominant environment that these extreme changes occur on (Figure 30 and Table 14). In total 8 atomic environments have transformations performed on them that result in an extreme property change. In most instances there is only one occurrence of such an event occurring, however, there are 3 transformation and-atomic environment combinations that occur more than once. These are the 157<sup>th</sup>, 159<sup>th</sup> and 295<sup>th</sup> most frequently occurring transformations when performed on the most frequently occurring atomic environment. In all three instances, the transformation involves replacing a molecular fragment with a carboxylic acid. The addition of a carboxylic acid, will likely be a conscious decision made by the chemists, either to drastically change the compounds properties, or even to alter binding and/ or potency of the compound.

There are 4 instances in which the transformation performed does not replace a molecular fragment with a carboxylic acid and only two transformations that do not involve a carboxylic acid at all (Table 14). The 226<sup>th</sup> and 380<sup>th</sup> most frequently occurring transformation on the most frequently occurring atomic environment, the 357<sup>th</sup> and 368<sup>th</sup> most frequently occurring transformations on the 9<sup>th</sup> and 26<sup>th</sup> most frequently occurring atomic environments, respectively.



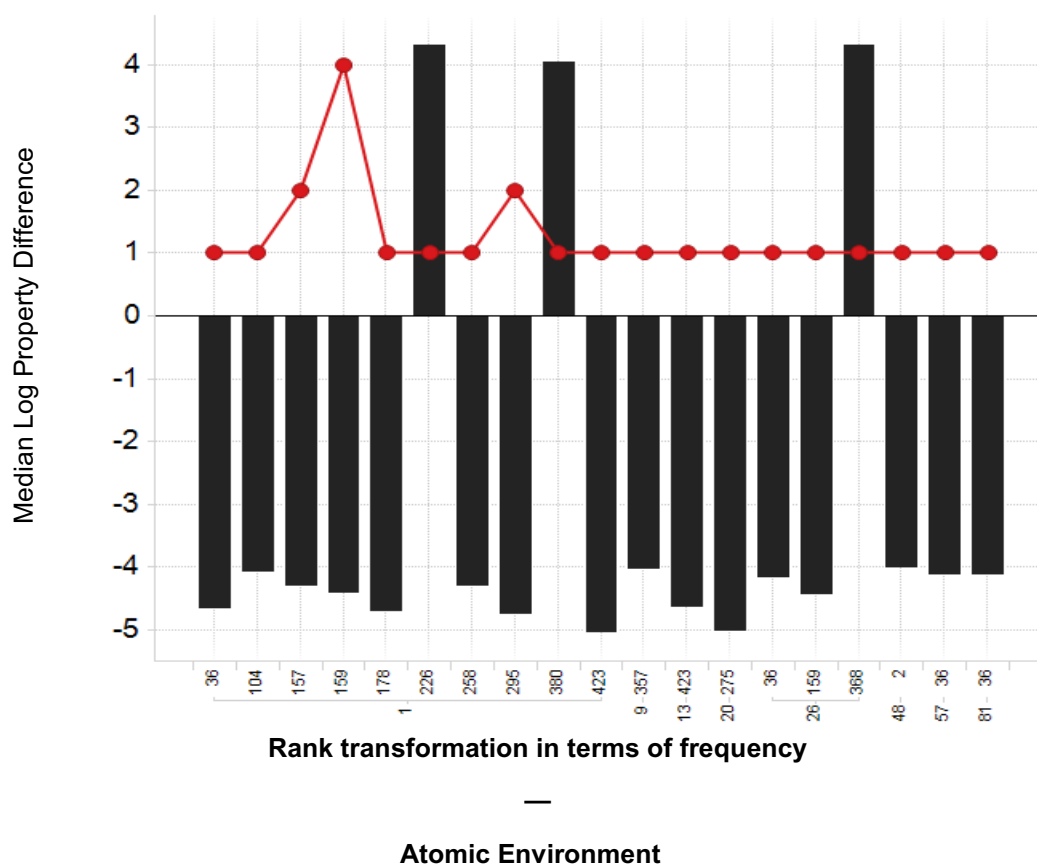
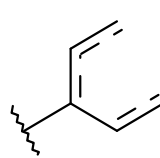
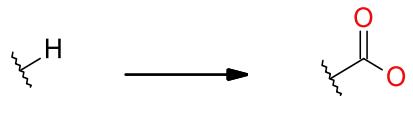
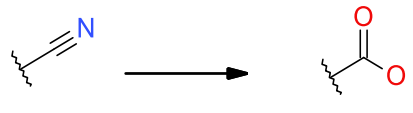
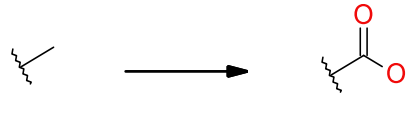
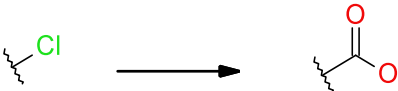
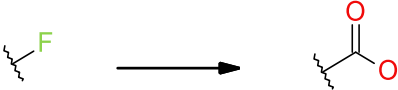
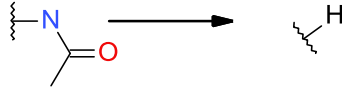
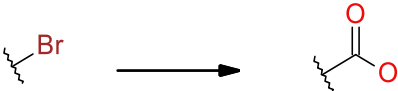
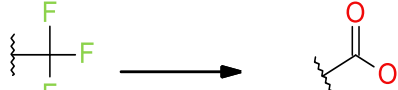
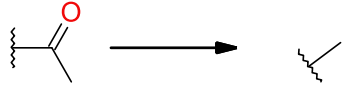

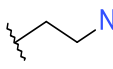
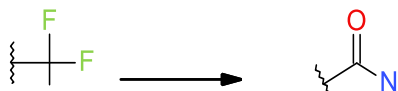
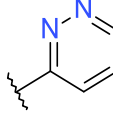
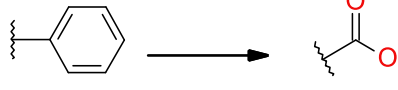
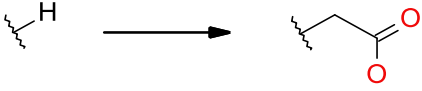
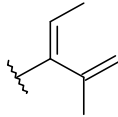
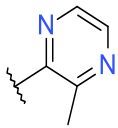
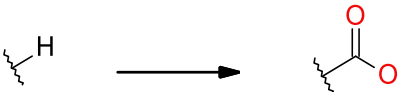
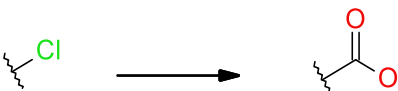
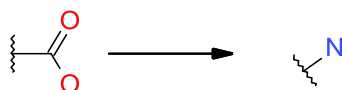
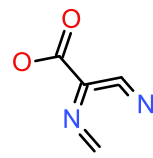
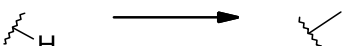
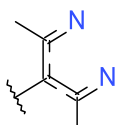

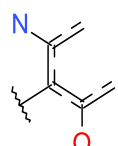



Figure 30: A combination plot showing for each aromatic atomic environment and each transformation the median property value difference (unspecified) (black bars) and the occurrences of this atomic environment and transformation that had extreme outliers. The most frequently occurring atomic environment on aromatic systems has the greatest number of transformations performed on it that result in atomic environments.

Table 14: Atomic environments and transformations performed on them for aromatic systems that resulted in a property change (unspecified) up  $\pm 4$  standard deviations.

| Rank Atomic Environment | Atomic Environment  | Rank Transformation Aromatic | Transformation   |
|-------------------------|---|------------------------------|--|
| 1                       |  | 36                           |  |
|                         |   | 104                          |  |
|                         |   | 157                          |  |

|    |   |     |  |
|----|---|-----|--|
|    |   | 159 |    |
|    |   | 178 |    |
|    |   | 226 |    |
|    |   | 258 |    |
|    |   | 295 |    |
|    |   | 380 |   |
|    |   | 423 |  |
| 9  |  | 357 |  |
| 13 |  | 423 |  |
| 20 |   | 275 |  |

|    |   |     |  |
|----|---|-----|--|
|    |    |     |  |
| 26 |    | 36  |    |
|    |   | 159 |    |
|    |   | 368 |    |
| 48 |   | 2   |    |
| 57 |  | 36  |  |
| 81 |  | 36  |  |

#### 4.3.4.2 Aliphatic Systems

With regards to those transformations performed on aliphatic systems and the extreme outliers, there are more transformation-atomic environment pairs than those on aromatic systems (Figure 31 and Table 15). We identified 19 different atomic environments being involved; however, a high percentage of transformation-atomic environments pairs occur more than once (~53%). Again, the transformations that involve carboxylic acids feature heavily in these extreme property changes ( $\pm 4$  standard deviations), however, there is more variety in that there are more transformations that do not involve a carboxylic acid in comparison to aromatic analogues. These changes can be rationalised on chemical grounds; for example, incorporating an ester affects the acidity or the compound. It is therefore important to

acknowledge extreme outliers but not allow them to hide other outliers that may be unexpected and of great interest. The 15<sup>th</sup> most frequently occurring aliphatic atomic environment (the local environment that the transformation is identified on) has the greatest number of unique transformations performed on it with regards to observed extreme value changes. Unlike the transformations performed on aromatic systems, only one transformation is performed on the most frequently occurring aliphatic atomic environment that results in an extreme property change ( $\pm 4$  standard deviations).

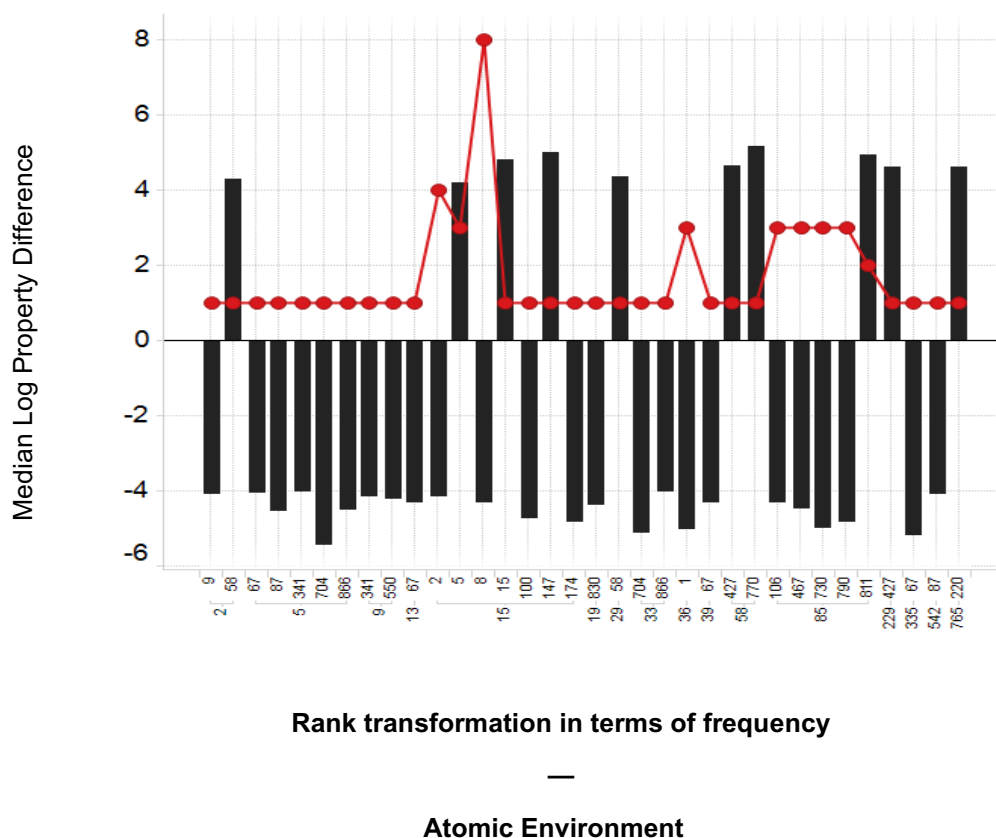
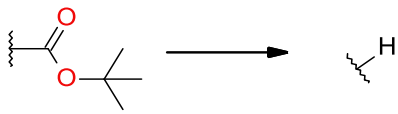
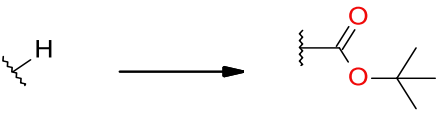
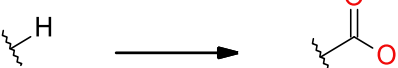
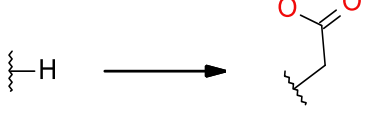

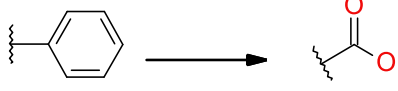
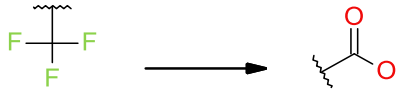
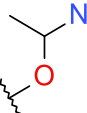
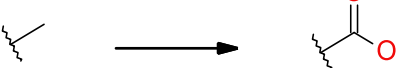
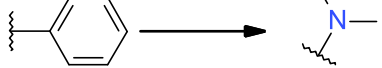
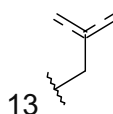
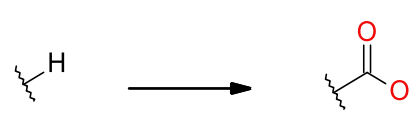
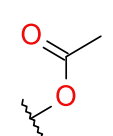
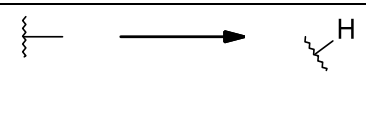
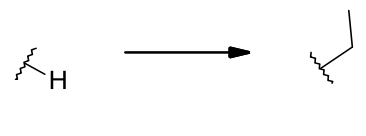
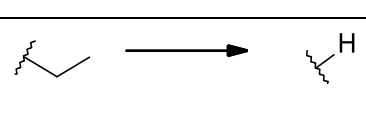
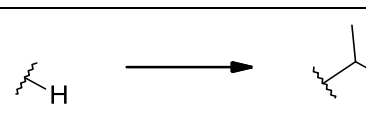
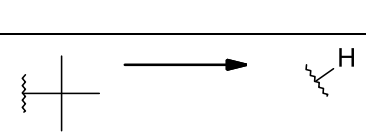
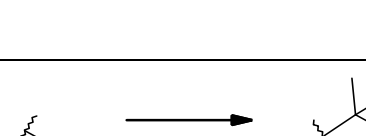
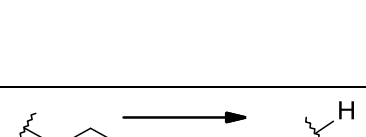
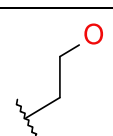
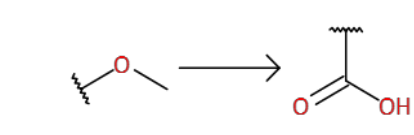
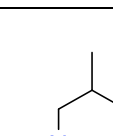
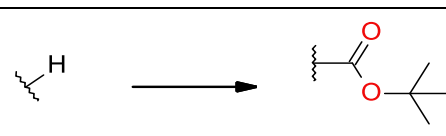
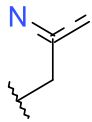
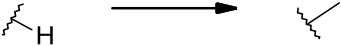
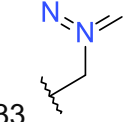
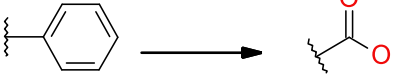
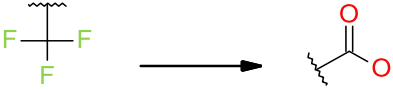
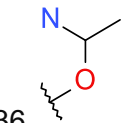
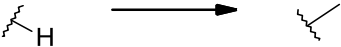
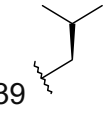
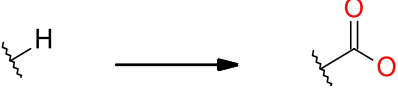

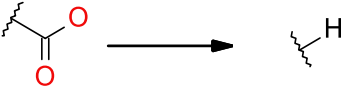
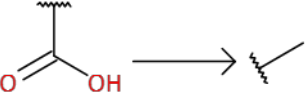
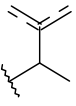
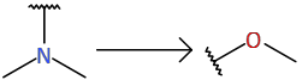
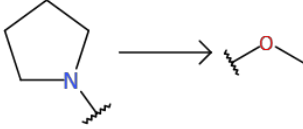
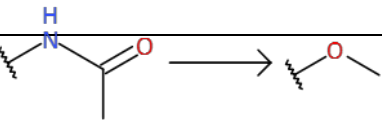


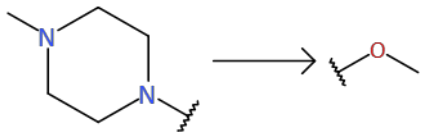
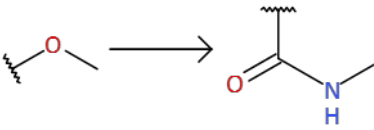
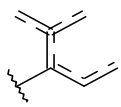
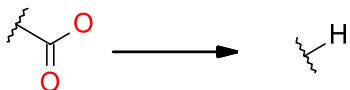
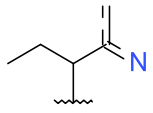
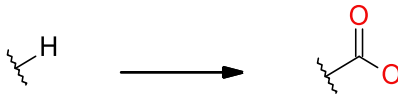
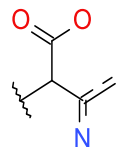
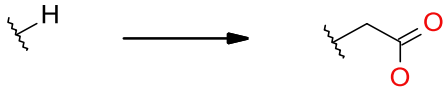
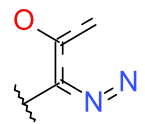
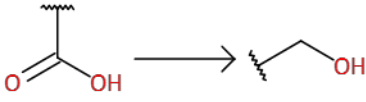
Figure 31 A combination plot showing for each aliphatic atomic environment and each transformation the median property value difference (unspecified) (black bars) and the occurrences of this atomic environment and transformation that had an extreme outlier. There are many atomic environments and transformations that have resulted in an extreme property change.

Table 15: Atomic environments and transformations performed on them for aliphatic systems that resulted in a property change (unspecified) up  $\pm 4$  standard deviations.

| Rank Atomic Environment | Atomic Environment | Rank Transformation Aliphatic | Transformation   |
|-------------------------|--------------------|-------------------------------|--|
| 2                       |                    | 9                             |    |
|                         |                    | 58                            |    |
| 5                       |                    | 67                            |    |
|                         |                    | 87                            |   |
|                         |                    | 341                           |  |
|                         |                    | 704                           |  |
|                         |                    | 866                           |  |
|                         |                    |                               |   |
| 9                       |                    | 341                           |  |
|                         |                    | 550                           |  |

|    |   |     |  |
|----|---|-----|--|
| 13 |    | 67  |    |
| 15 |  | 2   |    |
|    |   | 5   |    |
|    |   | 8   |    |
|    |   | 15  |    |
|    |   | 100 |  |
|    |   | 147 |  |
|    |   | 174 |  |
| 19 |  | 830 |  |
| 29 |  | 58  |  |

|    |   |     |   |
|----|---|-----|---|
| 32 |    | 1   |     |
| 33 |    | 704 |     |
|    |   | 866 |     |
| 36 |    | 1   |     |
| 39 |    | 67  |     |
| 58 |  | 427 |   |
|    |   | 770 |  |
| 85 |  | 106 |  |
|    |   | 467 |  |
|    |   | 730 |   |
|    |   |     |   |

|     |   |     |  |
|-----|---|-----|--|
|     |   |     |  |
|     |   | 790 |    |
|     |   | 811 |    |
| 229 |   | 427 |    |
| 335 |  | 67  |  |
| 542 |  | 87  |  |
| 765 |  | 220 |  |

#### 4.4 Case Studies of Transformations Performed on Atomic Environments of Which Affected Property Changes Unexpectedly when Increasing the LogD

We next identified some case studies where the increase in logD results in an unexpected effect occurring on one of the other measured assay properties. Generally, it is expected that when logD increases, the endpoints of human microsomal metabolism, human hepatocytes,



rat hepatocytes, and Caco-2 Efflux ratio are expected (and desired) to increase whereas solubility, hERG IC50 and Caco-2 intrinsic permeability are expected to decrease. We identify and consider some examples where a transformation that is performed on an atomic environment results in a significant logD increase as well as a significant unexpected change in other properties.

An analysis extended beyond the top 5 transformations and atomic environments can reveal interesting and unexpected property changes, which are often the result of specific tailoring of a compound. We therefore investigated some of these examples, in particular those where the transformation performed on an atomic environment either: increases the LogD and the solubility in most of the cases; increases the logD but decreases the human microsomal metabolism in most of the cases; or increases the logD but decreases the human hepatocytes in many of the cases. This is the result of the atomic environment beyond three atoms (as analysed here). All transformations performed were attempted in order to alter the measured assay results analysed in this study.

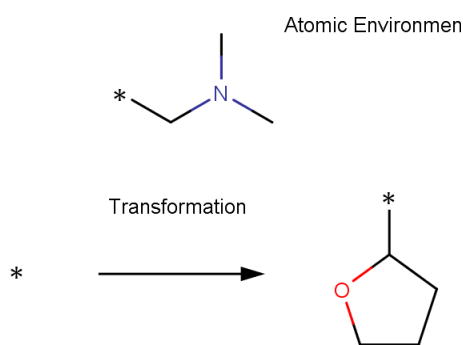
The first case study (Figure 32) was the 235<sup>th</sup> most frequently occurring transformation on the most frequently occurring atomic environment for aliphatic systems. In this case logD increases, as does solubility. The equation used to determine the increase in properties shows that the number of significant increases (+0.3 log units) needs to exceed the number of significant decreases (-0.3) plus the number of minimal changes (a change of between -0.3 and +0.3 log units). Furthermore, the number of occurrences needs to exceed 5 instances: in this example there are 11 instances of the logD increasing significantly with a median significant increase of ~0.7 log units. When this transformation was carried out on this environment for aliphatic systems, there are 6 examples of a significant increase for solubility and a median significant increase of 0.8 log units. There was only one incidence of the solubility decreasing significantly at -0.5 log units. In the remaining incidents of solubility change for this case study designated as a minimal change (3 instances), the median difference shows an increase. When a larger molecular fragment replaces a small molecular fragment ([\*H]), the lipophilicity achieves the expected to increase. When logD increases, it is not expected that solubility will also increase, however, the oxygen atom (which is electronegative and capable of hydrogen bonding) in the transformed molecule fragment will increase the solubility.

**System:**  
 Aliphatic  
**Rank Transformation in terms of frequency:**  
 235  
**Rank Atomic Environment in terms of Frequency:**  
 1

**Majority of the Time When**

**LogD increase**

- Solubility Increases



**Legend**

- Significant Decrease
- Minimal Change
- Significant Increase

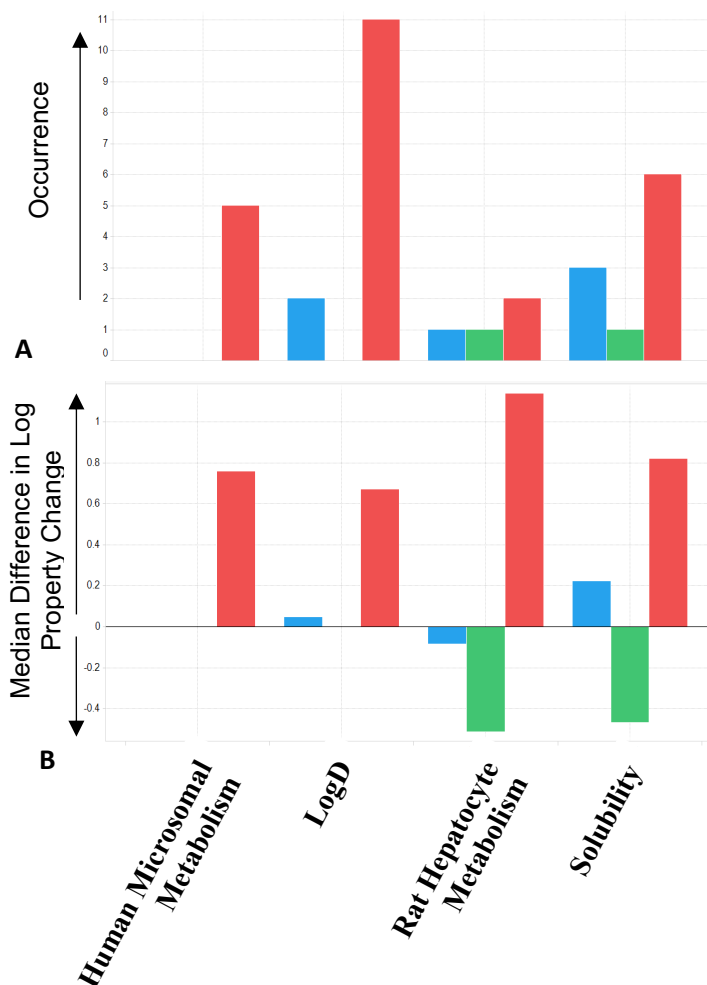


Figure 32: Case study where the transformation that is the 235 most frequently occurring transformation when performed on the most frequently occurring aliphatic atomic environment showed that when LogD significant increases the majority of the time, so does solubility. The occurrence of instances for the minimal change, the significant decrease and significant increase for each test (A) shows the significance of the increases for both logD and solubility. Additionally, (B) shows the median difference of the measured property.

We then considered two case studies involving aromatic systems. The first looks (Figure 33) at the 123<sup>rd</sup> most frequently occurring transformation on the most frequently occurring aromatic atomic environment. This transformation replaces a methyl group with a OCF<sub>3</sub> group, which promotes stability and blocks potential reactivity of the hydrogen atoms. When the logD increases, rat hepatocytes decrease. A previous example showed PhOCF<sub>3</sub> increases logD by ~1 log unit and when replacing a PhOCH<sub>3</sub> exhibits a lower passive permeability despite having

a higher lipophilicity<sup>188</sup>. When a functional group is in a benzylic position, it is generally considered more reactive than when the functional group is on its own<sup>189</sup>.

**System:**

Aromatic

**Rank Transformation in terms of frequency:**

123

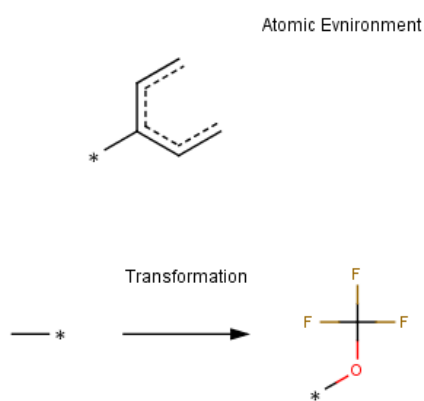
**Rank Atomic Environment in terms of Frequency:**

1

**Majority of the Time When**

**LogD increase**

- Rat hepatocyte decreases



**Legend**

Significant Decrease

Minimal Change

Significant Increase

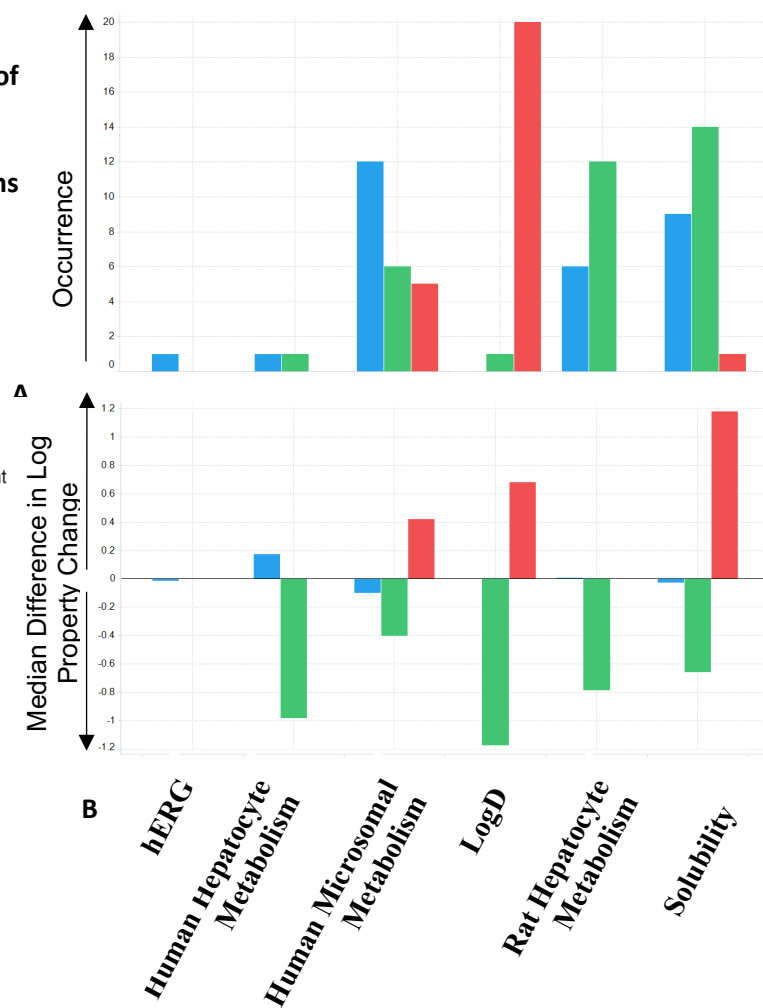


Figure 33: Case study where the transformation that is the 123<sup>rd</sup> most frequently occurring transformation when performed on the most frequently occurring aromatic atomic environment showed that when LogD significant increases the majority of the time, so does rat hepatocyte decreases. The occurrence of instances for the minimal change, the significant decrease and significant increase for each test (A) shows the significance of the increases for both logD and rat hepatocyte. Additionally, (B) shows the median difference of the measured property.

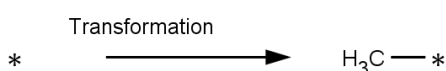
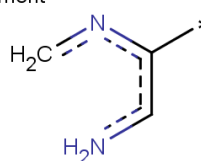
In the final case (Figure 34), the 2<sup>nd</sup> most frequently occurring transformation when performed on the 15<sup>th</sup> most frequently occurring atomic environment on aromatic systems was considered, and we found that as logD increases, human hepatocytes decrease due to electrons being donated into the ring which increases the basicity of the aromatic nitrogen.

**System:**  
 Aromatic  
**Rank Transformation in terms of frequency:**  
 2  
**Rank**                      **Atomic Environment in terms of Frequency:**  
 15

**Majority of the Time When LogD increase**

- Human Hepatocyte decreases

Atomic Environment



**Legend**

Significant Decrease

Minimal Change

Significant Increase

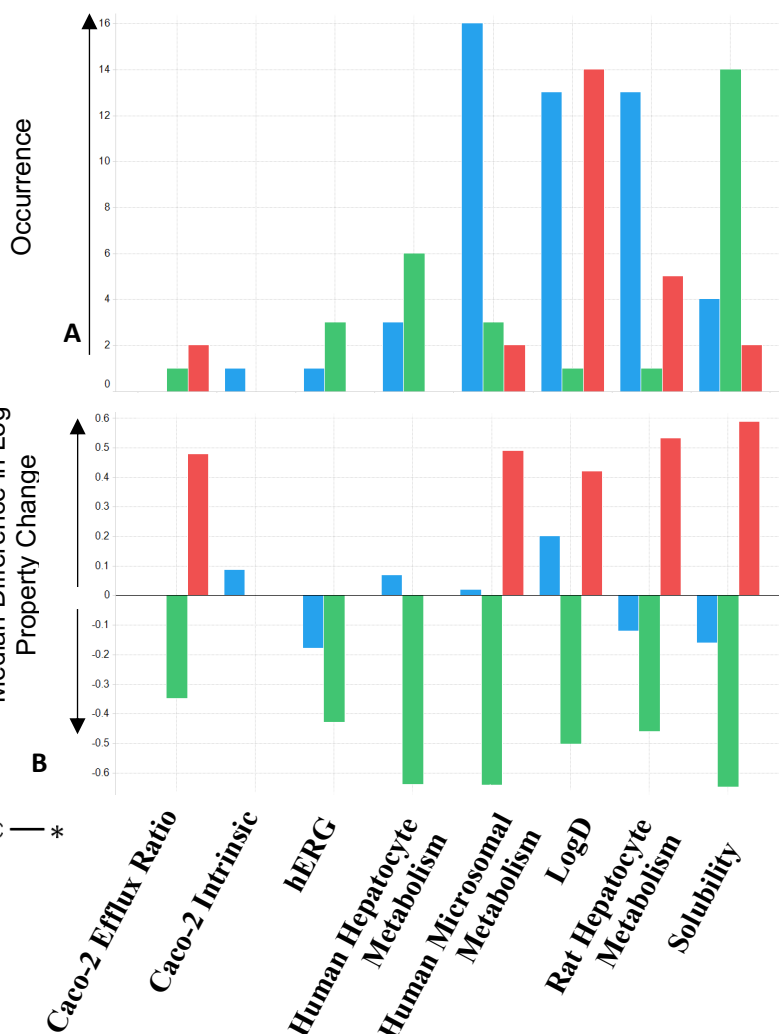


Figure 34: Case study where the transformation that is the 2<sup>nd</sup> most frequently occurring transformation when performed on the 15<sup>th</sup> most frequently occurring aromatic atomic environment showed that when LogD significant increases the majority of the time, so does human hepatocyte decrease. The occurrence of instances for the minimal change, the significant decrease and significant increase for each test (A) shows the significance of the increases for both logD and human hepatocyte. Additionally, (B) shows the median difference of the measured property.

#### 4.5 Chapter overview

Understanding the chemistry of the surrounding atomic environment is crucial when optimising compounds due to the chemical interactions that can occur between the transformed molecular fragment and the local atomic environment. In this study we have shown that even if a particular property change is expected, the atomic environment at which the transformation is performed can result in an unexpected change in the property value.

We have also shown that the proportion of occurrences of a transformation increasing or decreasing a property value on the same atomic environment can still significantly differ. Therefore, extending beyond three levels of atomic environment may be important if the property value is not as expected. However, it is also observed that some transformations prefer a particular property change direction on a given atomic environment.

Following on from this, we show that the atomic environment and the transformation that is performed influence the median log change of the property value. Again, the same transformation, performed on the same environment, rarely has a single preference (i.e. transformation  $X \gg Y$  always increases the solubility on environment A) but other factors, including chemistry beyond level three atomic environment from the transformation, play an important role.

We also show that transformations performed on specific atomic environments where the log property changes by  $\pm 4$  standard deviations are the result of a conscience decision to alter the chemistry of the compound, i.e. by changing the compound to an (carboxylic) acid.

Finally, we give examples where the direction of the property change is not as expected and shows that the surrounding chemistry of the compound is what is driving this unexpected change, highlighting the need to understand how the transformation will interact with the surrounding chemistry.

## Conclusion

The first research chapter investigated when, where and what is published in terms of novel chemistry both on its own and in association with a particular target. There is a need to improve the drug discovery process particularly in terms of lead optimisation because it has been shown many times that even though technology is increasingly becoming much more efficient, we are still not getting an increasing number of FDA approved drugs each year. To aid drug discovery, the use of compounds from several different sources will help during the design process. We found that generally compounds that are associated with a target tends to be published in scientific literature, whereas, novel chemistry tends to be published in patents. However, there are a few reasons why a compound can be published in scientific literature before it is patented, such as in the case of a formulation patents where the compounds in question are patented as part of a bigger objective with other compounds involved in the formulae.

Once a lead compound has been identified the optimisation process often uses matched molecular pairs (MMPs) to improve the compound in various aspects (such as ADMET). The most frequently observed molecular fragments vary drastically between different systems and their effects on property values vary between different atomic environments as well. Chemists should therefore be very much aware of the chemistry of the compound they are trying to perform the transformation on in order to yield the desired effect that they were looking for. Although regulating physical properties and measured assay properties is highly important in the drug discovery process, there are also concerns with areas such as potency and binding that are not considered in this thesis and are very much target specific.

Future work would involve extending this to understand a target basis and target class space to identify the effect that transformations have on particular atomic environments.

Although this is more of a statistical analysis and we will never really know what the chemist was thinking when they register the compound and the design process that was followed. Changes such as replacing fluorine with chlorine and replacing chlorine with fluorine occur frequently and occur in a near equal proportion to each other are potentially the result of chemists testing what they already know. Whereas, specific transformations and their inverse transformation, that occur in a less equal proportion could be a conscious decision by the chemist or even determined by ease of synthesis.

We have also shown that expected trends in property change, for example, increasing the logD you would decrease the solubility, are heavily influenced by the surrounding chemistry of the compound. Many transformations performed on environments alter the direction of the

property change in either direction suggesting that chemistry beyond 3 levels of the atomic environment are also important to consider when not registering the expected results. We have therefore shown that it is important to understand the underlying chemistry that the transformations are performed on, as there are cases where even though the logD increases significantly the solubility also increases significantly, which is not what you would normally expect.

The outcome of this study does not only add valuable information to previously reported studies, but also make a meaningful contribution to the process of using new analytical/processing tools to optimise compounds. Ultimately, this thesis allows for a greater understanding of where novel chemistry is published, and disseminated to the wider community, allowing for a clear direction of where to find relevant information as well as observing how trends have change over the course of history. Finally, this thesis has allowed for a more knowledge-based approach to optimising compounds for lead optimisation processes as well as observing how trends change over the course of a project.

## References

1. The Cost Of Creating A New Drug Now \$5 Billion, Pushing Big Pharma To Change#7896261713c3. Available at: <https://www.forbes.com/sites/matthewherper/2013/08/11/how-the-staggering-cost-of-inventing-new-drugs-is-shaping-the-future-of-medicine/#7896261713c3>. (Accessed: 31st January 2018)
2. Fishman, M. C. & Porter, J. A. Pharmaceuticals: a new grammar for drug discovery. *Nature* **437**, 491–493 (2005).
3. Hughes, J. P., Rees, S. S., Kalindjian, S. B. & Philpott, K. L. Principles of early drug discovery. *Br. J. Pharmacol.* **162**, 1239–1249 (2011).
4. Cragg, G. M. & Newman, D. J. Natural products: A continuing source of novel drug leads. *Biochimica et Biophysica Acta - General Subjects* **1830**, 3670–3695 (2013).
5. Palmer, M. Phenotypic Screening. in *Small Molecule Medicinal Chemistry: Strategies and Technologies* 281–304 (2015). doi:10.1002/9781118771723.ch10
6. Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **432**, 862–865 (2004).
7. Schultes, S. *et al.* Ligand efficiency as a guide in fragment hit selection and optimization. *Drug Discov. Today Technol.* **7**, e157–e162 (2010).
8. Szymański, P., Markowicz, M. & Mikiciuk-Olasik, E. Adaptation of high-throughput screening in drug discovery-toxicological screening tests. *International Journal of Molecular Sciences* **13**, 427–452 (2012).
9. Pereira, D. A. & Williams, J. A. Origin and evolution of high throughput screening. *Br. J. Pharmacol.* **152**, 53–61 (2007).
10. Mayr, L. M. & Fuerst, P. The future of high-throughput screening. *J. Biomol. Screen.* **13**, 443–448 (2008).
11. Schreiber, S. L., Nicolaou, K. C. & Davies, K. Diversity-oriented organic synthesis and proteomics: New frontiers for chemistry & biology. *Chem. Biol.* **9**, 1–2 (2002).
12. Lead Optimisation - Drug Discovery | Sygnature Discovery. Available at: <https://www.sygnaturediscovery.com/drug-discovery/integrated-drug-discovery/lead-optimisation/>. (Accessed: 3rd August 2018)
13. Cheng, K.-C., Korfmacher, W. A., White, R. E. & Njoroge, F. G. Lead Optimization in



Discovery Drug Metabolism and Pharmacokinetics/Case study: The Hepatitis C Virus (HCV) Protease Inhibitor SCH 503034. *Perspect. Medicin. Chem.* **1**, 1–9 (2008).

14. Eddershaw, P., Beresford, A. & Bayliss, M. ADME/PK as part of a rational approach to drug discovery. *Drug Discov. Today* **5**, 409–414 (2000).

15. Patidar, A. K. *et al.* Lead Discovery and Lead Optimization : A Useful Strategy in Molecular Modification of Lead Compound in Analog Design ABSTRACT : *Int. J. Drug Des. Discov.* **2**, 458–463 (2011).

16. Bleicher, K. H., Böhm, H. J., Müller, K. & Alanine, A. I. Hit and lead generation: Beyond high-throughput screening. *Nat. Rev. Drug Discov.* **2**, 369–378 (2003).

17. Transport Across Caco-2 Monolayer : Biological , Pharmaceutical and Analytical Considerations. Available at: <http://pharmaquest.weebly.com/uploads/9/9/4/2/9942916/caco2.pdf>. (Accessed: 21st June 2018)

18. Gao, Y., Gesenberg, C. & Zheng, W. Oral Formulations for preclinical studies: Principle, design, and development considerations. in *Developing Solid Oral Dosage Forms: Pharmaceutical Theory and Practice: Second Edition* 455–495 (Elsevier, 2016). doi:10.1016/B978-0-12-802447-8.00017-0

19. Savjani, K. T., Gajjar, A. K. & Savjani, J. K. Drug Solubility: Importance and Enhancement Techniques. *ISRN Pharm.* **2012**, 1–10 (2012).

20. Knights, K. M., Stresser, D. M., Miners, J. O. & Crespi, C. L. In vitro drug metabolism using liver microsomes. *Curr. Protoc. Pharmacol.* **2016**, 7.8.1-7.8.24 (2016).

21. Sahi, J., Grepper, S. & Smith, C. Hepatocytes as a tool in drug metabolism, transport and safety evaluations in drug discovery. *Curr. Drug Discov. Technol.* **7**, 188–98 (2010).

22. Qt, T., Gene, E. R., Conference, I. & Ic, T. Cardiac toxicity herg. 2–3 (2012).

23. Jackson, C. M., Esnouf, M. P., Winzor, D. J. & Duewer, D. L. Defining and measuring biological activity: Applying the principles of metrology. *Accredit. Qual. Assur.* **12**, 283–294 (2007).

24. Dissociation Constant - Chemistry LibreTexts. Available at: [https://chem.libretexts.org/Textbook\\_Maps/Physical\\_and\\_Theoretical\\_Chemistry\\_Textbook\\_Maps/Supplemental\\_Modules\\_\(Physical\\_and\\_Theoretical\\_Chemistry\)/Equilibria/Chemical\\_Equilibria/Dissociation\\_Constant](https://chem.libretexts.org/Textbook_Maps/Physical_and_Theoretical_Chemistry_Textbook_Maps/Supplemental_Modules_(Physical_and_Theoretical_Chemistry)/Equilibria/Chemical_Equilibria/Dissociation_Constant). (Accessed: 23rd September 2018)

25. Yung-Chi, C. & Prusoff, W. H. Relationship between the inhibition constant (KI) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. *Biochem. Pharmacol.* **22**, 3099–3108 (1973).
26. Gesztelyi, R. *et al.* The Hill equation and the origin of quantitative pharmacology. *Archive for History of Exact Sciences* **66**, 427–438 (2012).
27. Engel, P. C. A hundred years of the Hill equation. *Biochem. J.* **2013**, 1–4 (2013).
28. Weiss, J. N. The Hill equation revisited: uses and misuses. *FASEB J.* **11**, 835–841 (1997).
29. Xie, L., Xie, L. & Bourne, P. E. Structure-based systems biology for analyzing off-target binding. *Current Opinion in Structural Biology* **21**, 189–199 (2011).
30. Jones, S. *et al.* Core Signaling Pathways in Human Pancreatic Cancers Revealed by Global Genomic Analyses. *Science (80-. )*. **321**, 1801–1806 (2008).
31. Jia, J. *et al.* Mechanisms of drug combinations: Interaction and network perspectives. *Nat. Rev. Drug Discov.* **8**, 111–128 (2009).
32. Drug Administration - Drugs - Merck Manuals Consumer Version. Available at: <https://www.msdmanuals.com/en-gb/home/drugs/administration-and-kinetics-of-drugs/drug-absorption>. (Accessed: 10th July 2018)
33. Drug Absorption - Clinical Pharmacology - MSD Manual Professional Edition. *MSD Manual* Available at: <https://www.msdmanuals.com/en-gb/professional/clinical-pharmacology/pharmacokinetics/drug-absorption>. (Accessed: 10th July 2018)
34. Valko, K., Butler, J. & Eddershaw, P. Predictive approaches to increase absorption of compounds during lead optimisation. *Expert Opin. Drug Discov.* **8**, 1225–1238 (2013).
35. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **64**, 4–17 (2012).
36. El-Saadi, M. W., Williams-Hart, T., Salvatore, B. A. & Mahdavian, E. Use of in-silico assays to characterize the ADMET profile and identify potential therapeutic targets of fusarochromanone, a novel anti-cancer agent. *Silico Pharmacol.* **3**, 6 (2015).
37. Drug Administration - Drugs - Merck Manuals Consumer Version. Available at: <https://www.msdmanuals.com/en-gb/home/drugs/administration-and-kinetics-of-drugs/drug->

distribution. (Accessed: 2nd August 2018)

38. Martin, B. K. Potential effect of the plasma on drug distribution. *Nature* **207**, 274–6 (1965).

39. Drug absorption | Pharmacology Education Project. Available at: <https://www.pharmacologyeducation.org/pharmacology/drug-distribution>. (Accessed: 10th July 2018)

40. Wilkinson, G. R. Drug Metabolism and Variability among Patients in Drug Response. *N. Engl. J. Med.* **352**, 2211–2221 (2005).

41. Zanger, U. M. & Schwab, M. Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacology and Therapeutics* **138**, 103–141 (2013).

42. Elimination vs Excretion | Medicinal Chemistry Wiki | FANDOM powered by Wikia. Available at: [http://medicinalchemistry.wikia.com/wiki/Elimination\\_vs\\_Excretion](http://medicinalchemistry.wikia.com/wiki/Elimination_vs_Excretion). (Accessed: 11th July 2018)

43. Le, J. Drug Excretion - Clinical Pharmacology - MSD Manual Professional Edition. *MSD Manual* (2014). Available at: <https://www.msmanuals.com/en-gb/professional/clinical-pharmacology/pharmacokinetics/drug-excretion>. (Accessed: 11th July 2018)

44. Chapter 6. Drug Elimination and Clearance | Applied Biopharmaceutics & Pharmacokinetics, 6e | AccessPharmacy | McGraw-Hill Medical. Available at: <https://accesspharmacy.mhmedical.com/content.aspx?bookid=513&sectionid=41488024#56602262>. (Accessed: 11th July 2018)

45. Bonate, P. L., Reith, K. & Weir, S. Drug interactions at the renal level. *Clin. Pharmacokinet.* **34**, 375–404 (1998).

46. Guengerich, F. P. Mechanisms of Drug Toxicity and Relevance to Pharmaceutical Development. *Drug Metab. Pharmacokinet.* **26**, 3–14 (2011).

47. Rudmann, D. G. On-target and off-target-based toxicologic effects. *Toxicol. Pathol.* **41**, 310–314 (2013).

48. Lipinski, C. A. Avoiding investment in doomer drugs, is poor solubility an industry wide problem? *Curr. Drug Discov.* 17–19 (2001).

49. Sambuy, Y. *et al.* The Caco-2 cell line as a model of the intestinal barrier: influence of

cell and culture-related factors on Caco-2 cell functional characteristics. *Cell Biol. Toxicol.* **21**, 1–26 (2005).

50. Osakwe, O., Rizvi, S. A. A. & Osakwe, O. Preclinical In Vitro Studies: Development and Applicability. in *Social Aspects of Drug Discovery, Development and Commercialization* 129–148 (Elsevier, 2016). doi:10.1016/B978-0-12-802220-7.00006-5

51. Bhal, S. K. Application Note: Lipophilicity Descriptors: Understanding When to Use LogP and LogD. *ACD/Labs PhysChem Softw. Appl. Notes. Available online [http://www.acdlabs.com/resources/knowledgebase/app\\_notes/physchem/](http://www.acdlabs.com/resources/knowledgebase/app_notes/physchem/)* (accessed 09/09/15). 1–4 (2007).

52. Dressman, J. B., Amidon, G. L., Reppas, C. & Shah, V. P. Dissolution testing as a prognostic tool for oral drug absorption: Immediate release dosage forms. *Pharmaceutical Research* **15**, 11–22 (1998).

53. CEREP. Application note: Partition Coefficient ( log D ). **33**, 1–4 (2002).

54. Leon L, Herbert A L, J. L. K. *The Theory and Practice of Industrial Pharmacy*. (Philadelphia : Lea & Febiger, 1990).

55. United States Pharmacopeial Convention. *The United States Pharmacopeia 37 - The National Formulary 32*. (United States Pharmacopeial Convention, 2014).

56. Stationery Office (Great Britain). *British pharmacopoeia 2009*. (Stationery Office, 2008).

57. Darby, F. J., Newnes, W. & Price Evans, D. A. Human liver microsomal drug metabolism. *Biochem. Pharmacol.* **19**, 1514–1517 (1970).

58. Vrbanac, J. & Slaughter, R. ADME in Drug Discovery. in *A Comprehensive Guide to Toxicology in Preclinical Drug Development* 3–30 (Elsevier, 2013). doi:10.1016/B978-0-12-387815-1.00002-2

59. Bowen, R. Hepatic Histology: Hepatocytes. *colorado state university* (1998). Available at: [http://www.vivo.colostate.edu/hbooks/pathphys/digestion/liver/histo\\_hcytes.html](http://www.vivo.colostate.edu/hbooks/pathphys/digestion/liver/histo_hcytes.html). (Accessed: 25th June 2018)

60. Castell, J. V, Jover, R., Martinez-Jimenez, C. P. & Gomez-Lechn, M. J. Hepatocyte cell lines: their use, scope and limitations in drug metabolism studies. *Expert Opin. Drug Metab. Toxicol.* **2**, 183–212 (2006).

61. Sanguinetti, M. C. & Tristani-Firouzi, M. hERG potassium channels and cardiac arrhythmia. *Nature* **440**, 463–469 (2006).
62. Gaulton, A. *et al.* ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–D1107 (2012).
63. Bento, A. P. *et al.* The ChEMBL bioactivity database: An update. *Nucleic Acids Res.* **42**, D1083–D1090 (2014).
64. Index of /pub/databases/chembl/ChEMBLdb/releases/chembl\_21. Available at: [http://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/releases/chembl\\_21/](http://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/releases/chembl_21/). (Accessed: 4th August 2017)
65. GOSTAR: About GOSTAR. Available at: <https://www.gostardb.com/about-gostar.jsp>. (Accessed: 5th April 2018)
66. Jagarlapudi, S. A. R. P. & Kishan, K. V. R. Database Systems for Knowledge-Based Discovery. in *Methods in molecular biology (Clifton, N.J.)* **575**, 159–172 (2009).
67. AstraZeneca. AstraZeneca - Research-Based BioPharmaceutical Company. (2016). Available at: <https://www.astrazeneca.com/>. (Accessed: 5th April 2018)
68. Terrett, N. K., Gardner, M., Gordon, D. W., Kobylecki, R. J. & Steele, J. Combinatorial synthesis - the design of compound libraries and their application to drug discovery. *Tetrahedron* **51**, 8135–8173 (1995).
69. Southan, C., Williams, A. J. & Ekins, S. Challenges and recommendations for obtaining chemical structures of industry-provided repurposing candidates. *Drug Discov. Today* **18**, 58–70 (2013).
70. Southan, C. Expanding opportunities for mining bioactive chemistry from patents. *Drug Discov. Today Technol.* **14**, 3–9 (2015).
71. Formulation Patents—New Formulation of Known Compound - Inventing Patents. Available at: <http://inventingpatents.com/new-formulation-of-a-known-compound/>. (Accessed: 4th August 2017)
72. Kenny, P. W. & Sadowski, J. Structure Modification in Chemical Databases. in *Chemoinformatics in Drug Discovery* **23**, 271–285 (Wiley-VCH Verlag GmbH & Co. KGaA, 2005).
73. Tyrchan, C. & Evertsson, E. Matched Molecular Pair Analysis in Short: Algorithms,

Applications and Limitations. *Comput. Struct. Biotechnol. J.* **15**, 86–90 (2017).

74. Hussain, J. & Rea, C. Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *J. Chem. Inf. Model.* **50**, 339–348 (2010).

75. Gleeson, P., Bravi, G., Modi, S. & Lowe, D. ADMET rules of thumb II: A comparison of the effects of common substituents on a range of ADMET parameters. *Bioorganic Med. Chem.* **17**, 5906–5919 (2009).

76. Hajduk, P. J. & Sauer, D. R. Statistical analysis of the effects of common chemical substituents on ligand potency. *J. Med. Chem.* **51**, 553–564 (2008).

77. Dossetter, A. G. A statistical analysis of in vitro human microsomal metabolic stability of small phenyl group substituents, leading to improved design sets for parallel SAR exploration of a chemical series. *Bioorganic Med. Chem.* **18**, 4405–4414 (2010).

78. Warner, D. J., Griffen, E. J. & St-Gallay, S. A. WizePairZ: A novel algorithm to identify, encode, and exploit matched molecular pairs with unspecified cores in medicinal chemistry. *J. Chem. Inf. Model.* **50**, 1350–1357 (2010).

79. O'Boyle, N. M., Boström, J., Sayle, R. A. & Gill, A. Using matched molecular series as a predictive tool to optimize biological activity. *J. Med. Chem.* **57**, 2704–2713 (2014).

80. Hu, X., Hu, Y., Vogt, M., Stumpfe, D. & Bajorath, J. MMP-cliffs: Systematic identification of activity cliffs on the basis of matched molecular pairs. *J. Chem. Inf. Model.* **52**, 1138–1145 (2012).

81. Papadatos, G. *et al.* Lead optimization using matched molecular pairs: Inclusion of contextual information for enhanced prediction of hERG inhibition, solubility, and lipophilicity. *J. Chem. Inf. Model.* **50**, 1872–1886 (2010).

82. Raymond, J. W., Watson, I. A. & Mahoui, A. Rationalizing lead optimization by associating quantitative relevance with molecular structure modification. *J. Chem. Inf. Model.* **49**, 1952–1962 (2009).

83. Cao, Y., Jiang, T. & Girke, T. A maximum common substructure-based algorithm for searching and predicting drug-like compounds. *Bioinformatics* **24**, i366–i374 (2008).

84. Raymond, J. W. & Willett, P. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *Journal of Computer-Aided Molecular Design* **16**, 521–533 (2002).

85. Lukac, I. *et al.* Turbocharging Matched Molecular Pair Analysis: Optimizing the Identification and Analysis of Pairs. *J. Chem. Inf. Model.* **57**, 2424–2436 (2017).
86. Leeson, P. D. & Springthorpe, B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discov.* **6**, 881–890 (2007).
87. Proudfoot, J. R. The evolution of synthetic oral drug properties. *Bioorganic Med. Chem. Lett.* **15**, 1087–1090 (2005).
88. Senger, S. Assessment of the significance of patent-derived information for the early identification of compound-target interaction hypotheses. *J. Cheminform.* **9**, 26 (2017).
89. ORACLE SQL Developer. Available at: <http://www.oracle.com/technetwork/developer-tools/sql-developer/overview/index.html>. (Accessed: 4th August 2017)
90. Berthold, M. R. *et al.* KNIME - the Konstanz information miner. in *ACM SIGKDD Explorations Newsletter* **11**, 26 (Springer, Berlin, Heidelberg, 2009).
91. Kogej, T. *et al.* Big pharma screening collections: more of the same or unique libraries? The AstraZeneca–Bayer Pharma AG case. *Drug Discov. Today* **18**, 1014–1024 (2013).
92. Ashenden, S. K., Kogej, T., Engkvist, O. & Bender, A. Innovation in Small-Molecule-Druggable Chemical Space: Where are the Initial Modulators of New Targets Published? *J. Chem. Inf. Model.* **57**, 2741–2753 (2017).
93. Arrowsmith, C. H., Bountra, C., Fish, P. V., Lee, K. & Schapira, M. Epigenetic protein families: A new frontier for drug discovery. *Nat. Rev. Drug Discov.* **11**, 384–400 (2012).
94. RDKit. Available at: <http://www.rdkit.org/>. (Accessed: 4th August 2017)
95. Toad for MySQL - Toad World. Available at: <https://www.toadworld.com/products/toad-for-mysql>. (Accessed: 4th August 2017)
96. TIBCO Spotfire. Available at: [https://spotfire.tibco.com/resources/product-trial-cloud/world-simple-place?mkwid=s4c5L5kup&pdv=c&pclid=209299651808&pmt=e&pkw=tibco spotfire&campaign=ggl\\_s\\_uk\\_en\\_spt\\_brand\\_alpha&group=&bt=209299651808&\\_bk=tibco spotfire&\\_bm=e&\\_bn=g&gclid=Cj0KCQjwtpDMBRC4](https://spotfire.tibco.com/resources/product-trial-cloud/world-simple-place?mkwid=s4c5L5kup&pdv=c&pclid=209299651808&pmt=e&pkw=tibco spotfire&campaign=ggl_s_uk_en_spt_brand_alpha&group=&bt=209299651808&_bk=tibco spotfire&_bm=e&_bn=g&gclid=Cj0KCQjwtpDMBRC4). (Accessed: 4th August 2017)
97. Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **39**, 2887–2893 (1996).
98. Tobergte, D. R. & Curtis, S. MOE Molecular Operating Environment. *Journal of*

*Chemical Information and Modeling* **53**, 1689–1699 (2013).

99. Heller, S. R., McNaught, A., Pletnev, I., Stein, S. & Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminform.* **7**, 23 (2015).

100. Test of Equal or Given Proportions. Available at: <http://stat.ethz.ch/R-manual/R-patched/library/stats/html/prop.test.html>. (Accessed: 4th August 2017)

101. R: Pairwise comparisons for proportions. Available at: <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/pairwise.prop.test.html>. (Accessed: 4th August 2017)

102. Rstudio Team. RStudio – Open source and enterprise-ready professional software for R. *RStudio* (2016). Available at: <https://www.rstudio.com/>. (Accessed: 4th August 2017)

103. Southan, C., Varkonyi, P., Boppana, K., Jagarlapudi, S. A. R. P. & Muresan, S. Tracking 20 years of compound-to-target output from literature and patents. *PLoS One* **8**, e77142 (2013).

104. Azoulay, P., Michigan, R. & Sampat, B. N. The anatomy of medical school patenting. *N. Engl. J. Med.* **357**, 2049–56 (2007).

105. Sampat, B. N. Academic patents and access to medicines in developing countries. *Am. J. Public Health* **99**, 9–17 (2009).

106. ChEMBL. Available at: <https://www.ebi.ac.uk/chembl/>. (Accessed: 5th September 2018)

107. Hunter, P. The second coming of epigenetic drugs: A more strategic and broader research framework could boost the development of new drugs to modify epigenetic factors and gene expression. *EMBO Rep.* **16**, 276–279 (2015).

108. Garland, S. L. Are GPCRs still a source of new targets? *J. Biomol. Screen.* **18**, 947–966 (2013).

109. Lagerstrom, M. C. & Schioth, H. B. Structural diversity of G protein-coupled receptors and significance for drug discovery. *Nat Rev Drug Discov* **7**, 339–357 (2008).

110. Schneider, G. Automating drug discovery. *Nat. Rev. Drug Discov.* **17**, 97–113 (2018).

111. Nuti, S. V. *et al.* The use of google trends in health care research: A systematic review. *PLoS ONE* **9**, e109583 (2014).

112. Samaras, L., García-Barriocanal, E. & Sicilia, M.-A. Syndromic surveillance models using Web data: The case of scarlet fever in the UK. *Informatics Heal. Soc. Care* **37**, 106–124



(2012).

113. Karaman, M. W. *et al.* A quantitative analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **26**, 127–132 (2008).

114. Kinase Inhibitors: Global Markets: BIO053B | BCC Research. Available at: <https://www.bccresearch.com/market-research/biotechnology/kinase-inhibitors-markets-bio053b.html>. (Accessed: 4th August 2017)

115. Ghosh, A. K. *et al.* Design of Potent Inhibitors for Human Brain Memapsin 2 ( $\beta$ -Secretase). *J. Am. Chem. Soc.* **122**, 3522–3523 (2000).

116. Christie, G., Hussain, I. & Powell, D. J. Method of screening for inhibitors of Asp2. **2000–2000G**, 34 (2001).

117. Narum, L., Norskov-Lauritsen, L. & Olesen, P. H. Scaffold hopping and optimization towards libraries of glycogen synthase kinase-3 inhibitors. *Bioorganic Med. Chem. Lett.* **12**, 1525–1528 (2002).

118. Albaugh, Pamela, A. *et al.* Pyrrole-2, 5-Dione Derivatives and their use as GSK-3 Inhibitors. (2003).

119. Roder, H. Compositions and Method for the Treatment of Parkinson'S Disease. (2010).

120. Overington, J. P., Al-Lazikani, B. & Hopkins, A. L. How many drug targets are there? *Nat. Rev. Drug Discov.* **5**, 993–996 (2006).

121. Southan, C., Vrkonyi, P. & Muresan, S. Quantitative assessment of the expanding complementarity between public and commercial databases of bioactive compounds. *J. Cheminform.* **1**, 10 (2009).

122. Kojetin, D. J. & Burris, T. P. REV-ERB and ROR nuclear receptors as drug targets. *Nat. Rev. Drug Discov.* **13**, 197–216 (2014).

123. Search - SureChEMBL. Available at: <https://www.surechembl.org/search/>. (Accessed: 4th August 2017)

124. Young, R. J. & Leeson, P. D. Mapping the Efficiency and Physicochemical Trajectories of Successful Optimizations. *J. Med. Chem.* [acs.jmedchem.8b00180](https://doi.org/10.1021/acs.jmedchem.8b00180) (2018). doi:10.1021/acs.jmedchem.8b00180

125. Wenlock, M. C., Austin, R. P., Barton, P., Davis, A. M. & Leeson, P. D. A comparison of physiochemical property profiles of development and marketed oral drugs. *J. Med. Chem.*

46, 1250–1256 (2003).

126. Hann, M. M. *et al.* Molecular obesity, potency and other addictions in drug discovery. *Medchemcomm* **2**, 349 (2011).

127. Turk, S., Merget, B., Rippmann, F. & Fulle, S. Coupling Matched Molecular Pairs with Machine Learning for Virtual Compound Optimization. *J. Chem. Inf. Model.* **57**, 3079–3085 (2017).

128. Kramer, C. *et al.* Learning Medicinal Chemistry Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) Rules from Cross-Company Matched Molecular Pairs Analysis (MMPA). *J. Med. Chem.* **61**, 3277–3292 (2018).

129. Kramer, C., Fuchs, J. E., Whitebread, S., Gedeck, P. & Liedl, K. R. Matched molecular pair analysis: Significance and the impact of experimental uncertainty. *J. Med. Chem.* **57**, 3786–3802 (2014).

130. Roughley, S. D. & Jordan, A. M. The medicinal chemist's toolbox: An analysis of reactions used in the pursuit of drug candidates. *J. Med. Chem.* **54**, 3451–3479 (2011).

131. Gomez, L. Decision Making in Medicinal Chemistry: The Power of Our Intuition. *ACS Med. Chem. Lett.* acsmedchemlett.8b00359 (2018). doi:10.1021/acsmedchemlett.8b00359

132. Kutchukian, P. S. *et al.* Inside the Mind of a Medicinal Chemist: The Role of Human Bias in Compound Prioritization during Drug Discovery. *PLoS One* **7**, e48476 (2012).

133. Lajiness, M. S., Maggiora, G. M. & Shanmugasundaram, V. Assessment of the consistency of medicinal chemists in reviewing sets of compounds. *Journal of Medicinal Chemistry* **47**, 4891–4896 (2004).

134. Wassermann, A. M. & Bajorath, J. Identification of target family directed bioisosteric replacements. *Medchemcomm* **2**, 601–606 (2011).

135. Dossetter, A. G., Griffen, E. J. & Leach, A. G. Matched molecular pair analysis in drug discovery. *Drug Discov. Today* **18**, 724–731 (2013).

136. Pant, S. M. *et al.* Design, Synthesis, and Testing of Potent, Selective Hepsin Inhibitors via Application of an Automated Closed-Loop Optimization Platform. *J. Med. Chem.* **61**, 4335–4347 (2018).

137. Perl. The Perl Programming Language - [www.perl.org](http://www.perl.org). *Dr Dobbs Journal* 64–69 (2011). Available at: <http://www.perl.org/>. (Accessed: 22nd March 2018)

138. KNIME. KNIME | Open for Innovation. (2016). Available at: <https://www.knime.org/>. (Accessed: 22nd March 2018)
139. <https://chemaxon.com/products/marvin>. ChemAxon - Software Solutions and Services for Chemistry & Biology. Available at: <https://chemaxon.com/products/marvin>. (Accessed: 5th March 2018)
140. Anaconda Navigator | Anaconda: Documentation. Available at: <https://docs.anaconda.com/anaconda/navigator>. (Accessed: 23rd February 2018)
141. Mierzejewska, K., Bochtler, M. & Czapinska, H. On the role of steric clashes in methylation control of restriction endonuclease activity. *Nucleic Acids Res.* **44**, 485–495 (2016).
142. Burkner, U. Effects of methyl groups on the geometry and conformational equilibrium of 1,3-dioxanes.
143. Leung, C. S., Leung, S. S. F. F., Tirado-Rives, J. & Jorgensen, W. L. Methyl effects on protein-ligand binding. *J. Med. Chem.* **55**, 4489–4500 (2012).
144. Wermuth, C. G. *The practice of medicinal chemistry*. (Elsevier/Academic Press, 2008).
145. Nassar, A. E. F., Kamel, A. M. & Clarimont, C. Improving the decision-making process in the structural modification of drug candidates: Enhancing metabolic stability. *Drug Discov. Today* **9**, 1020–1028 (2004).
146. Sun, S. & Fu, J. Methyl-Containing Pharmaceuticals: Methylation in Drug Design. *Bioorg. Med. Chem. Lett.* (2018). doi:10.1016/J.BMCL.2018.09.016
147. Hernandez, M., Cavalcanti, S. M., Moreira, D. R., de Azevedo Junior, W. & Leite, A. C. Halogen Atoms in the Modern Medicinal Chemistry: Hints for the Drug Design. *Curr. Drug Targets* **11**, 303–314 (2010).
148. Priimagi, A., Cavallo, G., Metrangolo, P. & Resnati, G. The Halogen Bond in the Design of Functional Supramolecular Materials: Recent Advances. *Acc. Chem. Res.* **46**, 2686–2695 (2013).
149. Gillis, E. P., Eastman, K. J., Hill, M. D., Donnelly, D. J. & Meanwell, N. A. Applications of Fluorine in Medicinal Chemistry. *J. Med. Chem.* **58**, 8315–8359 (2015).
150. Müller, K., Faeh, C. & Diederich, F. Fluorine in pharmaceuticals: Looking beyond intuition. *Science (80-. )*. **317**, 1881–1886 (2007).

151. Hagmann, W. K. The many roles for fluorine in medicinal chemistry. *J. Med. Chem.* **51**, 4359–4369 (2008).
152. Shah, P. & Westwell, A. D. The role of fluorine in medicinal chemistry. doi:10.1080/14756360701425014
153. Kirk, K. L. Fluorine in medicinal chemistry: Recent therapeutic applications of fluorinated small molecules. *J. Fluor. Chem.* **127**, 1013–1029 (2006).
154. Swallow, S. Fluorine in medicinal chemistry. *Prog. Med. Chem.* **54**, 65–133 (2015).
155. Organic Chemistry/Haloalkanes - Wikibooks, open books for an open world. Available at: [https://en.wikibooks.org/wiki/Organic\\_Chemistry/Haloalkanes](https://en.wikibooks.org/wiki/Organic_Chemistry/Haloalkanes). (Accessed: 21st May 2018)
156. Manallack, D. T., Prankerd, R. J., Yuriev, E., Oprea, T. I. & Chalmers, D. K. The significance of acid/base properties in drug discovery. *Chem. Soc. Rev.* **42**, 485–496 (2013).
157. Jagodzinska, M., Huguenot, F., Candiani, G. & Zanda, M. Assessing the bioisosterism of the trifluoromethyl group with a protease probe. *ChemMedChem* **4**, 49–51 (2009).
158. Leroux, F. Atropisomerism, biphenyls, and fluorine: A comparison of rotational barriers and twist angles. *ChemBioChem* **5**, 644–649 (2004).
159. Talele, T. T. The 'cyclopropyl Fragment' is a Versatile Player that Frequently Appears in Preclinical/Clinical Drug Molecules. *J. Med. Chem.* **59**, 8712–8756 (2016).
160. Manly, C. J., Chandrasekhar, J., Ochterski, J. W., Hammer, J. D. & Warfield, B. B. Strategies and tactics for optimizing the Hit-to-Lead process and beyond-A computational chemistry perspective. *Drug Discov. Today* **13**, 99–109 (2008).
161. Brown, D. G. & Boström, J. Analysis of Past and Present Synthetic Methodologies on Medicinal Chemistry: Where Have All the New Reactions Gone? *J. Med. Chem.* **59**, 4443–4458 (2016).
162. Leach, A. G. *et al.* Matched molecular pairs as a guide in the optimization of pharmaceutical properties; a study of aqueous solubility, plasma protein binding and oral exposure. *J. Med. Chem.* **49**, 6672–6682 (2006).
163. Hammett, L. P. Some relations between reaction rates and equilibrium constants. *Chem. Rev.* **17**, 125–136 (1935).
164. Altaf, A. A. *et al.* A Review on the Medicinal Importance of Pyridine Derivatives. [Http://www.sciencepublishinggroup.com](http://www.sciencepublishinggroup.com) **1**, 1 (2015).

165. Sierański, T. Discovering the stacking landscape of a pyridine-pyridine system. *J. Mol. Model.* **23**, (2017).
166. Mignon, P., Loverix, S., De Proft, F. & Geerlings, P. Influence of stacking on hydrogen bonding: Quantum chemical study on pyridine-benzene model complexes. *J. Phys. Chem. A* **108**, 6038–6044 (2004).
167. Administration, U. S. F. and D. Drugs@FDA: FDA approved drug products. *Www.Fda.Gov.Cder/Orange/Default.Htm* (2008). Available at: [www.fda.gov/cder/orange/default.htm](http://www.fda.gov/cder/orange/default.htm). (Accessed: 11th September 2017)
168. Naso-Kaspar, C. K. *et al.* In vitro formation of acetylmorphine from morphine and aspirin in postmortem gastric contents and deionized water. *J. Anal. Toxicol.* **37**, 500–506 (2013).
169. Bhardwaj, V., Gumber, D., Abbot, V., Dhiman, S. & Sharma, P. Pyrrole: A resourceful small molecule in key medicinal hetero-aromatics. *RSC Adv.* **5**, 15233–15266 (2015).
170. Black, D. M., Bakker-Arkema, R. G. & Nawrocki, J. W. An overview of the clinical safety profile of atorvastatin (lipitor), a new HMG-CoA reductase inhibitor. *Arch. Intern. Med.* **158**, 577–84 (1998).
171. Zhang, L., Peng, X.-M., Damu, G. L. V., Geng, R.-X. & Zhou, C.-H. Comprehensive Review in Current Developments of Imidazole-Based Medicinal Chemistry. *Med. Res. Rev.* **34**, 340–437 (2014).
172. Molina, P., Tárraga, A. & Otón, F. Imidazole derivatives: A comprehensive survey of their recognition properties. *Org. Biomol. Chem.* **10**, 1711 (2012).
173. Edition, F. Greene's Protective Groups in Organic Synthesis, Fourth Edition,. *Carbon N. Y.* 1053–1082 (2007). doi:10.1002/0470053488
174. Birch, A. M., Kenny, P. W., Simpson, I. & Whittamore, P. R. O. Matched molecular pair analysis of activity and properties of glycogen phosphorylase inhibitors. *Bioorganic Med. Chem. Lett.* **19**, 850–853 (2009).
175. Hann, M. M. & Keseř, G. M. Finding the sweet spot: The role of nature and nurture in medicinal chemistry. *Nat. Rev. Drug Discov.* **11**, 355–365 (2012).
176. Package 'Hmisc'. 421 (2018). Available at: <https://cran.r-project.org/web/packages/Hmisc/Hmisc.pdf>. (Accessed: 25th June 2018)

177. Wei, T. corrplot: Visualization of a correlation matrix. (2013).
178. CRAN. R: The R Project for Statistical Computing. (2017). Available at: <https://www.r-project.org/>. (Accessed: 27th June 2018)
179. Veber, D. F. *et al.* Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* **45**, 2615–2623 (2002).
180. O'Hagan, S. & Kell, D. B. The apparent permeabilities of Caco-2 cells to marketed drugs: magnitude, and independence from both biophysical properties and endogenite similarities. *PeerJ* **3**, e1405 (2015).
181. Carvalho, J. F. S. *et al.* Strategies to reduce hERG K<sup>+</sup> channel blockade. Exploring heteroaromaticity and rigidity in novel pyridine analogues of dofetilide. *J. Med. Chem.* **56**, 2828–2840 (2013).
182. Fleming, F. F., Yao, L., Ravikumar, P. C., Funk, L. & Shook, B. C. Nitrile-containing pharmaceuticals: Efficacious roles of the nitrile pharmacophore. *J. Med. Chem.* **53**, 7902–7917 (2010).
183. Bomika, Z. A., Andaburskaya, M. B., Pelcher<sup>1</sup>, Y. É. & Dubur, G. Y. Some nucleophilic substitution reactions of 2-chloro-3-cyanopyridines. *Chem. Heterocycl. Compd.* **12**, 896–899 (1976).
184. Westaway, K. C. *et al.* A New Insight into Using Chlorine Leaving Group and Nucleophile Carbon Kinetic Isotope Effects To Determine Substituent Effects on the Structure of S<sub>N</sub>2 Transition States. *J. Phys. Chem. A* **111**, 8110–8120 (2007).
185. Inoue, K. *et al.* Aromatic substitution reaction of 2-chloropyridines catalyzed by microsomal glutathione S-transferase 1. *Drug Metab. Dispos.* **37**, 1797–1800 (2009).
186. Springer, C. & Sokolnicki, K. L. A fingerprint pair analysis of hERG inhibition data. *Chem. Cent. J.* **7**, 167 (2013).
187. Kovtun, G. A. & Aleksandrov, A. L. Oxidation of aliphatic amines by molecular oxygen in the liquid phase. *Bull. Acad. Sci. USSR, Div. Chem. Sci.* **22**, 2156–2158 (1973).
188. Xing, L. *et al.* Fluorine in drug design: A case study with fluoroanisoles. *ChemMedChem* **10**, 715–726 (2015).
189. Ch 11: Benzylic systems. Available at: <http://www.chem.ucalgary.ca/courses/351/Carey5th/Ch11/ch11-9-0.html>. (Accessed: 14th

April 2018)