



UNIVERSITY OF
LIVERPOOL

Cheminformatics: Quantitative Structure-
Property Relationship Studies on Ames
Mutagenicity and Surfactants' Properties,
and Small Molecule Library Visualisation

Charmaine Siu Man Chu

2022

Thesis submitted in accordance with the requirements of the University of Liverpool
for the degree of Doctor in Philosophy by Charmaine Siu Man Chu

August 2022

PGR Policy on Plagiarism and Dishonest Use of Data
PGR CoP Appendix 4 Annexe 1

PGR DECLARATION OF ACADEMIC HONESTY

NAME (Print)	Charmaine Siu Man Chu
STUDENT NUMBER	200839756
SCHOOL/INSTITUTE	Department of Chemistry
TITLE OF WORK	Cheminfomatics: Quantitative Structure Property Relationship Studies on Ames Mutagenicity and Surfactants Properties, and Small Molecule Library Visualisation

This form should be completed by the student and appended to any piece of work that is submitted for examination. Submission by the student of the form by electronic means constitutes their confirmation of the terms of the declaration.

Students should familiarise themselves with Appendix 4 of the PGR Code of Practice: PGR Policy on Plagiarism and Dishonest Use of Data, which provides the definitions of academic malpractice and the policies and procedures that apply to the investigation of alleged incidents.

Students found to have committed academic malpractice will receive penalties in accordance with the Policy, which in the most severe cases might include termination of studies.

STUDENT DECLARATION

I confirm that:

- I have read and understood the University's PGR Policy on Plagiarism and Dishonest Use of Data.
- I have acted honestly, ethically and professionally in conduct leading to assessment for the programme of study.
- I have not copied material from another source nor committed plagiarism nor fabricated, falsified or embellished data when completing the attached material.
- I have not copied material from another source, nor colluded with any other student in the preparation and production of this material.
- If an allegation of suspected academic malpractice is made, I give permission to the University to use source-matching software to ensure that the submitted material is all my own work.

SIGNATURE.....

DATE.....

Acknowledgement

First, I would like give my deepest thanks to all the people who were involved in the projects carried out within the duration of my PhD. In particular, I would like to thank Prof Neil Berry for giving me the opportunity to work within his group and for his continual guidance throughout.

In preparation for the journal publication related to Chapter 2, I would like to thank Prof Paul O'Neill and Jack Simpson for providing their help and feedback.

I am extremely grateful to Dr Jerry Winter from Unilever for providing essential information and his guidance throughout the projects carried out related to Chapter 3 and 4 of this thesis. I would also like him for providing the network in interacting with industrial research groups.

In addition, I would like to thank Paul Colbon from Liverpool ChiroChem, Ltd. for the initial idea related to Chapter 5 of this thesis. As a result of the work presented in Chapter 5, I was able take two periods of suspension of studies to work in collaboration with Liverpool ChiroChem and Abbvie to further develop ideas from the material presented in the chapter.

Also, special thanks to Chris Swain for providing his insight throughout all works related to Chapter 5 within and out of my PhD, as well as at the final stages of thesis writing.

Lastly, I would like to give a big thanks to my family and friends for their love and support, especially during the difficult times in the midst of pandemic. Thanks to them I was able to maintain my social human interaction even though it was difficult to meet them in person.

Declaration

All work presented within this thesis was research performed during my PhD. Two periods of suspension of studies occurred (April 2018 – November 2018, June 2019 – December 2019) to develop ideas arisen from material presented in Chapter 5, funded via EPSRC Impact Acceleration Account working with Liverpool ChiroChem and Abbvie. All material presented in chapter 5 was generated during my PhD research.

Abstract

Over the past decades, chemistry has increasingly moved towards using theoretical prediction using mathematical methods saving resources and time. As a result of this, a range of tools for calculating 3D conformation of molecules and molecular descriptors has been developed. In combination with workflow platforms which allows flexible data manipulation and statistical analysis, tackling chemistry problems computationally has becoming increasingly accessible.

Cheminformatics is a field within the discipline of computational chemistry which focuses on the application of chemistry theory with information science techniques to problems related to chemistry. A common application of such techniques is quantitative structure-activity/property relationships (QSARs/QSPRs), where the relationship between experimentally observed activity or properties and the molecular structures investigated via computational modelling. Within this thesis, three applications of QSARs/QSPRs have been explored: the relationship between the molecular properties of small molecule and Ames mutagenicity; the relationship between molecular properties of surfactant-like polymers in detergent formulations and their cleaning ability; and the relationship between surfactant molecular properties and critical micelle concentration (CMC).

In relation to QSARs/QSPRs, the field of cheminformatics also involves the development of molecular property calculation suited to describe certain properties of a given molecule. Amphiphilicity, the measure of the extent of hydrophobicity and hydrophilicity of a molecule, is a key description of the amphiphilic surfactant molecules. However, a quantitative measure of such property for a surfactant using calculated or measured molecular properties is not yet established. Therefore, an automated computational workflow was developed which found the boundary between hydrophobic and hydrophilic sections of a surfactant and calculated descriptors related to amphiphilicity. The ability of the resulting descriptors to relate structural properties with the CMC was then compared to predictors of existing QSPR models.

Aside from the numerical outputs, cheminformatics also includes in the visualisation of chemical libraries and molecules. This area is particularly important when results are displayed, whether in form of publications or presentations, as 3D graphics often aids the understanding of concepts and comparisons of molecules. The concept of pharmacophores is an important area in drug discovery and, therefore, it is advantageous to analyse and compare pharmacophores of molecules within libraries to search for possible candidates which match certain criteria. Current tools lack the ability to visualise query pharmacophores of the whole, or a selected part of the, chemical library in 3D, which can be important when trying to identify areas chemical space the library can or cannot explore. Thereupon, workflows were developed to enable such visualisation.

Publications to date

Machine learning – Predicting Ames mutagenicity of small molecules

Charmaine S.M. Chu, Jack D. Simpson, Paul M. O'Neill, Neil G. Berry, Journal of Molecular Graphics and Modelling, Volume 109, 2021, 108011, ISSN 1093-3263, <https://doi.org/10.1016/j.jmgm.2021.108011>.

(<https://www.sciencedirect.com/science/article/pii/S1093326321001820>)

List of Key Abbreviations and Acronyms

Acc	Accuracy
AUROC	Area under the Receiver Operating Characteristic curve
avNNet	Model averaged neural networks
BalAcc	Balanced accuracy
BF	Base formulation
C5	C5.0 classification model
CIC	complementary information content
CMC	Critical micelle concentration
DNN	Deep neural network
ECFPs	Extended-connectivity fingerprints
ESA	Electronegativity surface area
FCFPs	Functional-class fingerprints
FNSA	Fractional partial negative surface area
FPSA	Fractional partial positive surface area
GaP	Gridding and partitioning
GBM	Stochastic gradient boosting
IC	Average information content
InChI	International Chemical Identifier
IUPAC	The International Unions of Pure and Applied Chemistry
k	Gradient of regression line
KNN	K-nearest neighbours
LCC	Liverpool ChiroChem Ltd.
LDA	Linear discriminant analysis
LM	Least square linear regression
logP	Logarithm of the octanol/water partition coefficient
MCC	Matthews correlation coefficient
MDA	Mixture discriminant analysis
Mn	Number average molecular weight
MOMI	Moment of inertia
MP	Molecular properties
MW	Molecular weight
nnet	Neural network

OECD	The Organisation for Economic Co-operation and Development
PCA	Principle component analysis
PHI	Kier flexibility index
PLS	Partial least square
PLSDA	Partial least squares discriminant analysis
PSL	Polymer surfactant library
QSARs	Quantitative Structure-Activity Relationships
QSPRs	Quantitative Structure-Property Relationships
R²	Squared correlation coefficient
RF	Random forest
RiD	Rings in Drugs
RMSE	Root mean square error
RNCS	Relative negative charge surface area
ROC	Receiver Operating Characteristic curve
RPC	Red pottery clay end point
RPCS	Relative positive charge surface area
SDF	Structure data file
Sens	Sensitivity
SMARTS	SMiles ARbitrary Target Specification
SMILES	Simplified Molecular Input Line Entry Specification
SP	Kier & Hall connectivity index
Spec	Specificity
SVM	Support vector machine
TEST	Toxicity Estimation Software Tool
Vis	Viscosity endpoint
XGB	Extreme gradient boosting
YPC	Yellow pottery clay end point

Table of Contents

Acknowledgement.....	ii
Declaration.....	ii
Abstract.....	iii
Publications	iv
List of Key Abbreviations and Acronyms	v
Table of Content	vii
Chapter 1: Introduction	1
1.1. Computational Chemistry and Cheminformatics.....	2
1.2. Quantitative Structure-Activity Relationships (QSARs), Quantitative Structure-Property Relationships (QSPRs) and Predictive Models.....	3
1.2.1. Overview	3
1.2.2. Chemical library selection.....	4
1.2.3. Library Curation.....	5
1.2.4. Inputting data.....	5
1.2.5. Training and test set split.....	13
1.2.6. Modelling algorithms	13
1.2.7. Performance analysis	27
1.3. Visualisation of Chemical Libraries.....	30
1.4. Thesis Overview	33
1.5. References	34
Chapter 2: Predicting Ames Mutagenicity.....	40
2.1. Ames Mutagenicity and its Importance.....	41
2.2. Current State of Art	41
2.3. Our Approach	42
2.4. Process of Library Curation.....	44
2.5. Construction of Ames Mutagenicity Predictive Models.....	44
2.5.1. Training/test set splitting.....	44
2.5.2. Descriptor generation	44
2.5.3. Data pre-processing.....	45
2.5.4. Algorithm selection	45
2.5.5. Model building and performance assessment	45
2.5.6. Model validation via y-randomisation and model robustness.....	47
2.5.7. Data set generation	47
2.5.8. Cross-validation models.....	48
2.5.9. Test set performance check.....	48
2.5.10. Validation via y-randomisation	49
2.5.11. Variable importance of validated models	51

2.5.12.	Comparison of different predictor sets used in model building	52
2.5.13.	Comparison of different algorithms used for predicting mutagenicity.....	55
2.6.	Comparison with Present Work	57
2.6.1.	Performance comparison of top performing models with commercial product	58
2.7.	Conclusion	59
2.8.	References.....	61
Chapter 3: Understanding Polymer Detergent Properties via QSPR method		64
3.1.	Detergent Properties and Polymers as Surfactants	65
3.2.	Polymer Data.....	67
3.2.1.	Data used	67
3.3.	Constructing of Detergent Properties QSPR Models	69
3.3.1.	Data compilation.....	69
3.3.2.	Data pre-processing.....	70
3.3.3.	Initial algorithm selection	71
3.3.4.	Model building and performance assessment	71
3.3.5.	Predictor – observation relationship interpretation	73
3.4.	Constructed models for BF1 data set	74
3.4.1.	Predictor importance across different models.....	77
3.5.	Constructed models for BF2 data sets	83
3.5.1.	Initial constructed models	83
3.5.2.	Models with transformed observation values.....	84
3.5.3.	Models with altered predictors	84
3.6.	Conclusion	90
3.7.	References.....	91
Chapter 4: Novel Surfactant Descriptor – Potential Amphiphilicity.....		93
4.1.	Amphiphilicity and Surfactants	94
4.1.1.	What is potential amphiphilicity.....	95
4.2.	Calculating Potential Amphiphilicity.....	97
4.2.1.	Input molecules	99
4.2.2.	Terminal count.....	100
4.2.3.	Identifying possible cleavage points.....	102
4.2.4.	Enumerate fragmentation pattern	103
4.2.5.	Label fragments as hydrophobic/hydrophilic.....	103
4.2.6.	Hydrophobic/hydrophilic AlogP difference calculation.....	105
4.2.7.	Test to Increasing number of fragments	105
4.2.8.	Export result.....	106

4.3. Critical Micelle Concentration (CMC), Amphiphilicity and Potential Amphiphilicity	
.....	108
4.3.1. CMC QSPR	108
4.4. Utility of Potential Amphiphilicity via QSPR.....	109
4.4.1. Surfactant library generation.....	109
4.4.2. Construction of surfactant CMC QSPR models Part 1	109
4.4.3. Construction of surfactant CMC QSPR models Part 2	121
4.4.4. Comparison with present CMC QSPR models	140
4.5. Conclusion.....	141
4.6. Future Work.....	141
4.7. References	142
Chapter 5: Visualisation of Chemical Functionality for a Chemical Library.....	145
5.1. Chemical functionalities and pharmacophore.....	146
5.2. Visualisation of Chemical Functionality for a Molecule	146
5.3. Process to Visualise Chemical Functionality for a Chemical Library.....	153
5.3.1. Import Library	154
5.3.2. Align library	156
5.3.3. Identify query chemical functionalities	163
5.3.4. Export visualisation	165
5.4. Example of Visualised Chemical Library	172
5.3.1. Settings within the protocol	173
5.3.2. Options within each visualisation outputs	178
5.4. Further Work	181
5.5. References	182
List of Supporting Information.....	184
Chapter 2: Predicting Ames Mutagenicity.....	184
Chapter 3: Understanding Polymer Detergent Properties via QSPR method	184
Chapter 4: Novel Surfactant Descriptor – Potential Amphiphilicity	184
Chapter 5: Visualisation of Chemical Functionality for a Chemical Library.....	184
Pipeline Pilot Protocols	185

Chapter 1: Introduction

1.1. Computational Chemistry and Cheminformatics

Historically, chemistry had been an experimental subject. However, experimental chemistry requires resources which can be costly and can be time consuming. As experimental chemistry has elements of trial and error, there is the possibility an experiment is not successful and the resources and time spent go to waste. Therefore, in the past four decades, movement towards theoretical prediction using mathematical methods over experimental chemistry has increased exponentially [1]. A major reason for this movement is due to the potential reduction in time and money in a research program. As information technology advances, the power of the hardware and software for theoretical predictions has increased and the cost of computing decreased [1]. Using well developed software, one can often cut time required and cost needed compared to the resources required for experimental chemistry. This discipline of using mathematical methods for calculation of molecular properties or simulation of molecular behaviour is called computational chemistry [2].

As experience accumulates over decades, there is a vast range of computational codes developed with increasing capability and accuracy in the predictions they are designed for. This includes software such as alvaDesc [3] for calculating molecular descriptors (properties) and DataWarrior [4] which can calculate 3-dimensional (3D) conformations of molecules and perform substructure searching and structure-based similarity analysis. In addition, nowadays there are many aids available on the internet for learning programming languages and there are various platforms which allow one to write computer scripts or construct protocols or workflows for manipulating data which fits one's need, such as Pipeline Pilot [5], KNIME [6] and Jupyter Notebooks [7]. Utilising such tools and chemical knowledge, it becomes increasingly more convenient to tackle chemistry problems with the aid of computational methods.

Within the discipline of computational chemistry, cheminformatics is the field which focuses on the application of chemistry theory with information science techniques to descriptive and prescriptive problems related to chemistry, such as drug discovery [8]. Such application includes chemical graph theory, quantitative structure-activity/property relationships and 3D pharmacophores [9], very briefly introduced below and relevant to future chapters.

Chemical graph theory is a branch of mathematical chemistry which forms the basis of many 2-dimensional based molecular representation and description solution [9-13]. In this theory, a molecule can be represented by a graph, where the atoms are considered as vertices and molecular bonds as edges (**Figure 1.1**) [10]. It is noted that the hydrogen atoms are often omitted. The theory then proceeds to reduce this topological structure of the molecule to single numbers which characterises its structural properties such as molecular branching [10, 14].

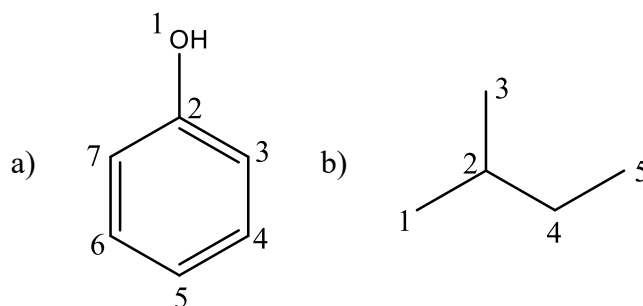


Figure 1.1. An example of chemical graph of a) phenol and b) isopentane.

Quantitative structure-activity/property relationships modelling is a branch of computational modelling which relates experimentally observed activities or properties to molecular structure [9, 15]. The attempts at constructing quantitative relationships to explain the relation between structure and biological activity date back to the beginning of the previous century [15], although it is generally accepted that the quantitative structure-activity relationships (QSARs) Hansch and Fujita constructed in the 1960s are the first QSARs explaining the biological activity of a series of structurally related molecules [9, 15]. In their studies, they made regression analysis on the relationship between the octanol/water partition coefficients (logP) and biological activities for structural series such as benzoic acids on mosquito larvae, phenyl ethyl phosphate insecticides on houseflies, and carcinogenic compounds on mice [16].

A 3D pharmacophore is defined as the specific spatial arrangement required for the specific interactions between the chemical functionalities, such as hydrogen-bond donors and charged groups of a molecules, and the biological target for its activities [2, 9, 17, 18]. By studying the 3D pharmacophores of structure with known activities and comparing them to the pharmacophores of different structures in chemical libraries, it is possible to identify structures with similar activities in unexplored chemical space [9, 19].

Within this thesis, various areas of the computational chemistry and cheminformatics have been being explored. In *Chapter 2*, curation of an existing Ames mutagenicity library and construction of QSAR predictive models using the curated library and comparison of the model performance against existing commercial and open source models will be presented. In *Chapter 3*, we move away from small molecules and look into modelling polymer containing surfactant formulation against their viscosity and cleaning ability, and identifying structural properties that affects these end points. In *Chapter 4*, the novel process of automating the search for the hydrophobic and hydrophilic section boundary of surfactant, calculation of related descriptors, and comparison of their ability in relating structural properties with the critical micelle concentration to predictors of existing models will be detailed. Finally, in *Chapter 5*, the novel approach taken to visualise query pharmacophores of the whole, or a selected part of a chemical library will be described. Within this chapter, an introduction of the key ideas for *Chapter 2, 3* and *4* will be presented in *1.2. Quantitative Structure-Activity Relationships (QSARs), Quantitative Structure-Property Relationships (QSPRs) and Predictive Models*, and key ideas of *Chapter 5* will be presented in *1.3 Visualisation of Chemical Libraries*.

1.2. Quantitative Structure-Activity Relationships (QSARs), Quantitative Structure-Property Relationships (QSPRs) and Predictive Models

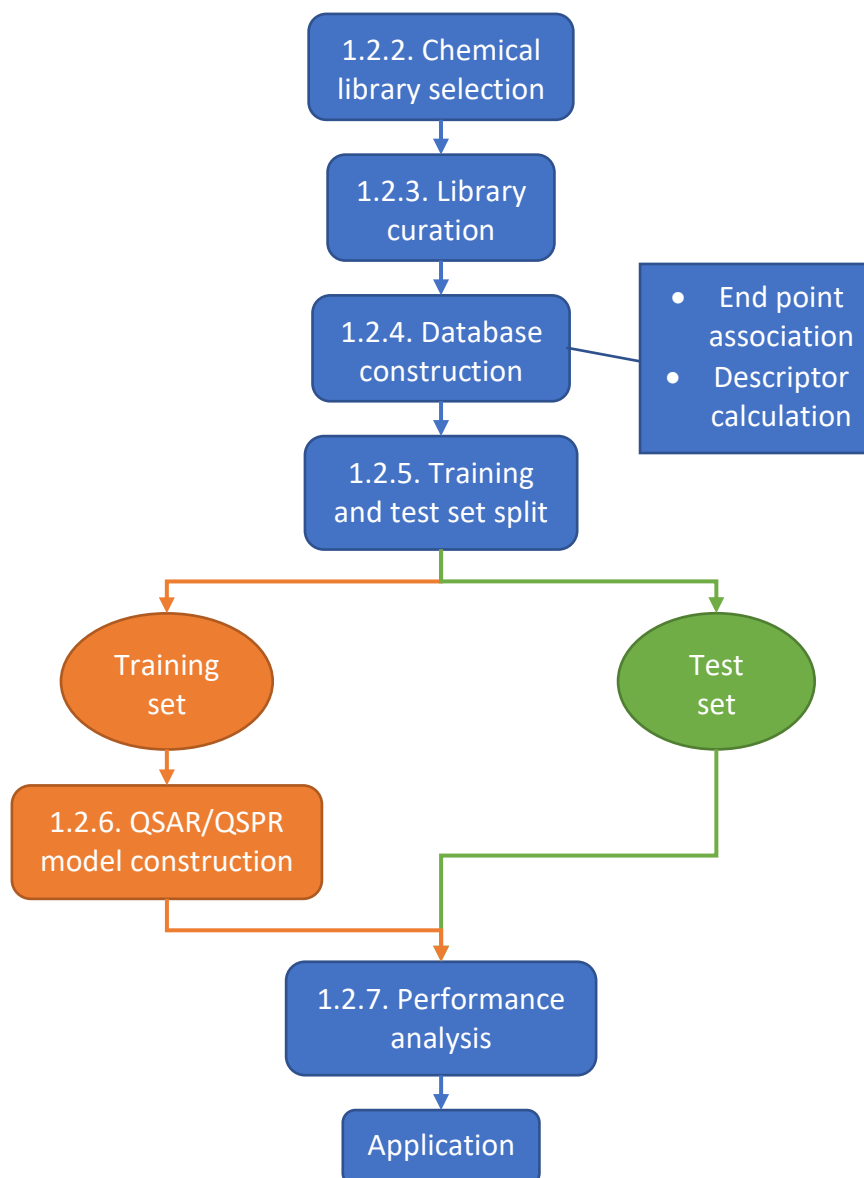
1.2.1. Overview

Quantitative structure-activity relationships (QSARs) or Quantitative structure-property relationships (QSPRs) are the computational modelling technique used for revealing relationships between molecular properties of chemical compounds and biological activities or physical properties [9, 15, 20]. In general, QSAR and QSPR models utilises data analysis, mathematical and statistical methods in order to find the empirical relationships (QSAR and QSPR models) between the property of interest P (biological activity or physical property) and structural properties D_1, D_2, \dots, D_n in form of

$$P = k(D_1, D_2, \dots, D_n) \quad \text{(Equation 1.1)},$$

where k is the mathematical transformations empirically established which when applied to the descriptors (D_1, D_2, \dots, D_n) should calculate the property value (P) for all molecules [21]. Using such a relationship, P of a chemical compound can be predicted using the compound's structural properties. In order to find the empirical relationships, a curated database and a

selection of modelling algorithms are required. **Scheme 1.1** shows the generic workflow of developing QSAR and QSPR [12, 21].



Scheme 1.1. The generic workflow of developing QSAR and QSPR models for predictive modelling.

1.2.2. Chemical library selection

When developing QSAR or QSPR models, it is important to understand the purpose of the models. As the relationships developed are dependent on the structures, and the choice of the molecules in the chemical library influence the quality and the type of information the models can capture [12]. For example, when constructing QSPR models for critical micelle concentration, the concentration where surfactant molecule aggregate to form micelles [22], it is desired to use a library containing molecules which demonstrate surfactant-like properties. Once the chemical library investigating is decided on, it is important to curate the library.

1.2.3. Library Curation

During the preparation of the database for QSAR/QSPR modelling, it is important to curate the chemical library of interest. Molecular descriptors are calculated based on molecular structure provided and, therefore, if there is any error in the molecular structure, this would directly translate into either inability to calculate descriptors for such records or erroneous descriptors [21]. As the models can only be as good as the data available, these errors would then in turn cause any of the models constructed to be inaccurate [21]. The process of curation ensures the structures are chemically correct, standardised and in a format which can be used for descriptor calculation without error. Fourches and Tropsha have described the major steps of data curation workflow that should be taken for the QSAR/QSPR [21, 23], which includes removal of mixtures and inorganic chemical compounds, removal of salts, normalisation of chemotypes and treatment of tautomeric forms using SMILES, a line notation for molecules (details see 1.2.4.2. *Digital storage of molecular structures*). Of course, what steps of curation are taken need to be considered with the type of database used and the purpose of the QSAR/QSPR. The steps of curation taken for each project within this thesis are described in the relevant chapters.

1.2.4. Inputting data

With the chemical library of interest curated, it is essential to construct a database containing the property of interest and the structural properties of the molecules. For each entry of the database, it is necessary to have an associated measured property of interest – the defined endpoint. This is in line with the Organisation for Economic Co-operation and Development (OECD) QSAR guidelines [24]. In most cases, the endpoint is measured experimentally, and hence is common for the endpoint values to be referred to as the observed values. However, as experimental measurements are generally affected by the conditions when the measurement is taken, it is crucial to state the experimental system the QSAR/QSPR is being identified for. In some cases, endpoints are calculated from experimental measurements taken under different conditions using an already established relationship in order to have concurring conditions for the endpoints [25]. Depending on the structure of the endpoint, the type of QSAR/QSPR model that can be constructed differs. If the endpoint is binary or categorical, classification QSAR/QSPR models can be constructed. On the other hand, when the endpoint is continuous, the resulting QSAR/QSPR models are regression models [26, 27].

Apart from the endpoint, the database must also contain molecular properties which describe the entry based on its molecular structure, such as molecular fingerprints, number of rotatable bonds, and moment of inertia for each of the entries.

1.2.4.1. *Digital storage of molecular structures*

Before going into the details of the different properties calculatable for a molecular structure, it is necessary to have a brief understanding how a molecular structure can be stored digitally. Different formats can store different amount of data and within this thesis, four types of molecular structure formats are involved: SMILES, SMARTS, InChI and SDF.

SMILES stands for Simplified Molecular Input Line Entry Specification. It is a line notation for molecules based on the principles of molecular graph theory [13], where the atoms are represented as nodes and bonds are represented as edges (**Figure 1.1**). A SMILES string stores the heavy atoms of the molecule using their element symbols, with their connectivity implied by their position in the SMILES string (**Table 1.1**). It is a fast and compact way of representing molecular structure which is human and machine-readable [13]. It is noted that this method of storing molecular structure does not include the 2D or 3D coordinates of the molecules. The SMILES string of a molecule can be stored with its endpoint and/or properties using delimited

text file formats. Multiple molecules can be stored in the same delimited text file by inputting them on separate lines.

Table 1.1. The difference between SMILES, SMARTS and InChI notation methods

Notation method	String	Matching Structures
SMILES	CCO	
High specification SMARTS	[CH3][CH2][O&-1]	
Low specification SMARTS	CCO	
InChI	InChI=1S/C2H6O/c1-2-3/h3H,2H2,1H3	

SMARTS, standing for SMiles ARbitrary Target Specification, is a language extended from SMILES which describes molecular patterns and properties [28, 29]. It allows the specification of atomic properties, such as charge and number of valences, of SMILES strings at varying specificity. High specificity SMARTS allows specific structures to be described without match with other similar structures that may match using SMILES. On the other hand, low specificity SMARTS allows generic structures to be described which can match with more structures than the corresponding SMARTS string (**Table 1.1**).

InChI is the International Chemical Identifier developed under the sponsorship of the International Unions of Pure and Applied Chemistry (IUPAC). It is a unique chemical identifier which describes a molecule with several layers of information where applicable: the atoms and their bond connectivities, tautomeric information, isotope information, stereochemistry, and electronic charge information (**Table 1.1**) [30]. The InChI of a molecule is a string of characters, converted from inputted structural information to give a unique InChI identifier through normalisation of the structure, canonicalisation to generate unique number label for each atom and serialisation [30]. As InChI normalise and canonicalise the molecular structure, it can be used to identify duplication of molecules in chemical libraries.

SDF stands for structure-data file, which is able to store multiple molecules along with associated data in one file. For each molecule within a SDF, the molecular structure is stored in a Connection table, where the atom information with their 2D or 3D coordinates are stored, followed by their bond connections and types (**Figure 1.2**) [31]. Stereochemistry regarding each atom and bond are also specified in the Connection table where available. As stereochemistry can affect the action of a molecule, which is crucial especially in biology related fields such as pharmacology, it is often the preferred method of storing molecular structures for descriptor calculation.

```

M0001
Spartan 12192118183D

  9  8  0  0  0  0  0  0  0  0  0999 V2000
    0.0768 -0.1925 -0.8970 C  0  0  0  0  0  0  0  0  0  0  0  0
    0.0714 -0.2478  0.6259 C  0  0  0  0  0  0  0  0  0  0  0  0
    0.1279  1.0361  1.2170 O  0  0  0  0  0  0  0  0  0  0  0  0
    0.9587  0.3494 -1.2510 H  0  0  0  0  0  0  0  0  0  0  0  0
   -0.8162  0.3215 -1.2744 H  0  0  0  0  0  0  0  0  0  0  0  0
    0.0870 -1.1996 -1.3274 H  0  0  0  0  0  0  0  0  0  0  0  0
    0.9589 -0.7749  0.9881 H  0  0  0  0  0  0  0  0  0  0  0  0
   -0.8060 -0.8111  0.9816 H  0  0  0  0  0  0  0  0  0  0  0  0
   -0.6585  1.5188  0.9371 H  0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
  2  3  1  0  0  0  0
  1  4  1  0  0  0  0
  1  5  1  0  0  0  0
  1  6  1  0  0  0  0
  2  7  1  0  0  0  0
  2  8  1  0  0  0  0
  3  9  1  0  0  0  0
M  END
> <Formula>
C2H6O

> <HBA Count>
1

> <HBD Count>
1

> <Molecular Wt. (amu)>
46.069

> <Name>
ethanol

$$$$

```

Figure 1.2. The SDF of an ethanol molecule generated using Spartan'18 [32], with associated properties: chemical formula (Formula), hydrogen-bond acceptor count (HBA count), hydrogen-bond donor count (HBD count), molecular weight (Molecular Wt. (amu)) and name of molecule (Name).

1.2.4.2. *Calculatable descriptors*

With the molecular structure files in hand, they can then be imported into computer programs such as Pipeline Pilot [5], KNIME [6], CDK Descriptor Calculator [33] and alvaDesc [3] for calculation of molecular descriptors. For some molecular descriptors, different calculation tools may use a different underlying algorithm for its calculation, and therefore may yield different values. Hence, it is important to ensure all descriptors for the same QSAR/QSPR

project are calculated using the same tool with its version noted. Molecular descriptors can be classified by the level of complexity they describe in 0, 1, 2, 3 or 4-dimensional (**Figure 1.3**) [11, 12].

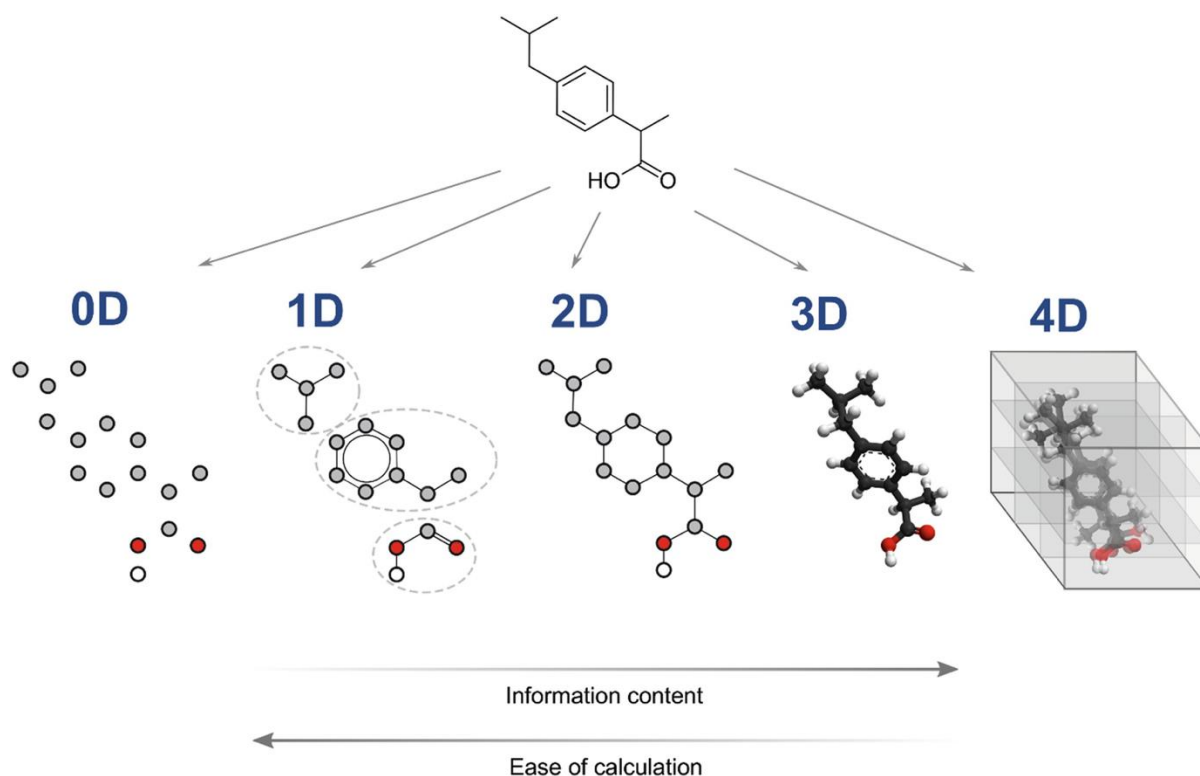


Figure 1.3. Graphical example of different molecular representations of the same structure (ibuprofen). Taken from [12]. The relationship between chosen dimensionality and information content/ease of calculation of the related descriptors is also illustrated.

0-dimensional (0D) descriptors are the simplest molecular descriptors based on the chemical formula, the specification of the chemical elements and their occurrence in a molecule [12]. As 0D descriptors do not consider any connectivity information, they are low in information content and highly degenerate, such that different molecules may have the same values (**Table 1.2**) [11, 12]. Common 0D descriptors includes number of atoms, atom counts (e.g. number of carbon atoms, number of oxygen atoms) and molecular weight.

Table 1.2. Example 0D descriptors for ethanol and methoxymethane

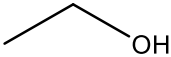
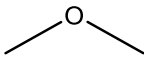
Name	ethanol	methoxymethane
Structure		
Chemical formula	C_2H_6O	C_2H_6O
Number of atoms (hydrogen excluded)	3	3
Number of carbon atoms	2	2
Number of oxygen atoms	1	1
Molecular weight	46.07	46.07

1-dimensional (1D) descriptors derived from the substructure list representation of a molecule, which consist of a list of structural fragments of a molecule [11]. Structural fragments of a

CHAPTER 1: INTRODUCTION

molecule can be any atom-centred fragments, functional groups or substituent of interest present in the molecule [11, 12]. The substructure list representation does not require the complete knowledge of the molecular structure and therefore multiple molecules, such as isomers, can have the same values [11, 12]. 1D descriptors are usually presented as a binary code encoding the presence or absence of the given structural fragment or in terms of their occurrence frequency (**Table 1.3**) [12]. Examples of 1D descriptors includes number of hydrogen bond donors and acceptors, and they are typically used in substructure analysis, similarity/diversity analysis, and virtual screening and design of chemical libraries [11].

Table 1.3. Example 1D descriptors for ethanol and methoxymethane

Name	ethanol	methoxymethane
Structure		
Number of hydrogen bond donors	1	0
Number of hydrogen bond acceptors	1	1

2-dimensional (2D) descriptors also include topological descriptors. These descriptors are derived from algorithms applied to the molecular graph of the molecule, which defines the connectivity of the atoms in a molecule in terms of the presence and nature of chemical bonds (**Figure 1.1**) [11, 12]. This type of descriptor is usually sensitive to structural features such as size, shape, symmetry, branching and cyclicality [11]. In addition to the structural features, specific chemical properties, such as mass and polarisability, and presence or absence of hydrogen bond donors and acceptors, can also be considered in combination [12]. Examples of 2D descriptors includes Kier and Hall's connectivity index and information content. Kier and Hall's connectivity index includes information from atom and molecular size to branch point and environment [14]. Information content indices measure molecular symmetry, where diversity of the elements is accounted for (**Figure 1.4**) [34-36].

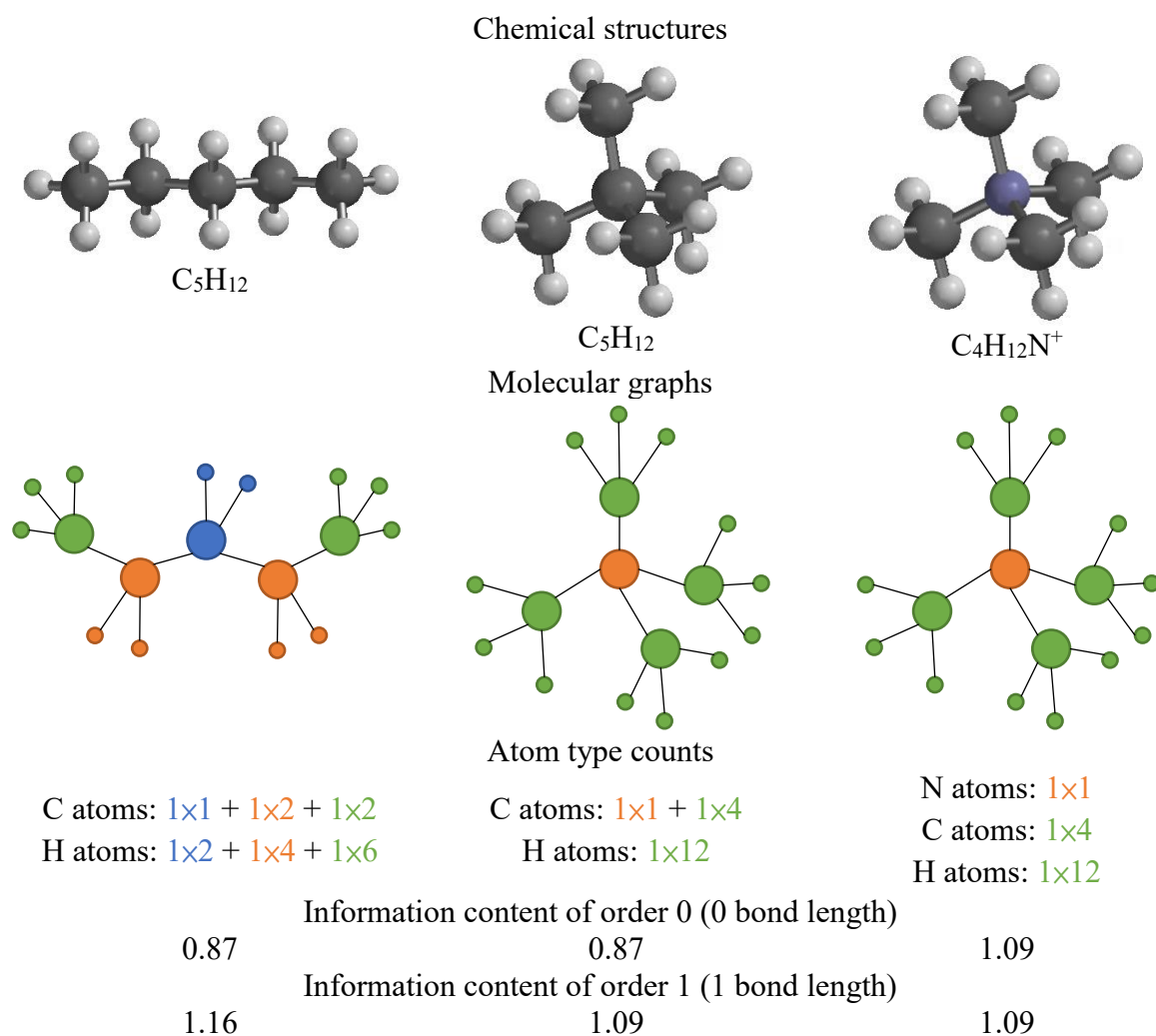


Figure 1.4. Illustration displaying the difference in information content of order 0 and 1 for two C_5H_{12} isomers, pentane (left) and neopentane (middle), and tetramethylammonium (right).

3-dimensional (3D) descriptors are geometric descriptors derived from the 3D representation of a molecule. A 3D representation of a molecule defines the molecule in terms of atom types and x, y, z -coordinates, perceives the molecule as a rigid geometrical object in space, allowing the representation of the overall spatial configuration of the atoms in the molecule as well as the nature and connectivity of the atoms (**Figure 1.5**) [11, 12]. 3D descriptors are high in their information content [37] and can measure the steric and size properties of a molecule [11], however, there are precautions that need to be kept in mind in calculation and usage in relation to the geometric optimisation of the molecule [12]. First, the coordinates of the atoms in the 3D molecular representation are influenced by the geometric optimisation method used [38], and the optimisation is often conducted in the gas phase. Second, for a highly flexible molecule, i.e. a molecule with a high number of rotatable bonds, there is more than one similar energy conformation available, and in reality, these conformers would exist in mixed abundance. Third, the optimised geometry may not be the geometry where the molecule is active for its action, the bioactive conformation. This is in particular observed for molecules expressing their pharmaceutical and biological action [39]. Examples of 3D descriptors includes charged partial surface area and moment of inertia. Charged partial surface area describes the charged surface area in relation to its total solvent accessible surface area [35]. The moment of inertia determines the torque required for the molecule to spin along a specified axis, e.g. x -axis (**Figure 1.6**), which gives an indication of the shape of the molecule [2, 35].

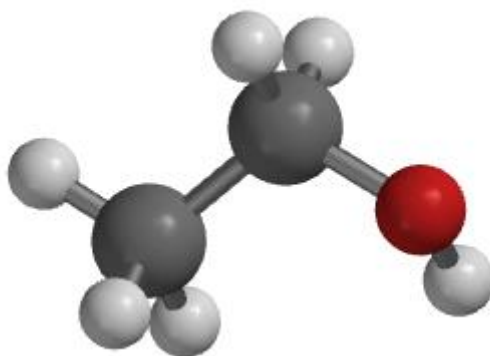


Figure 1.5. 3D representation of ethanol, generated using Spartan'18 [32].

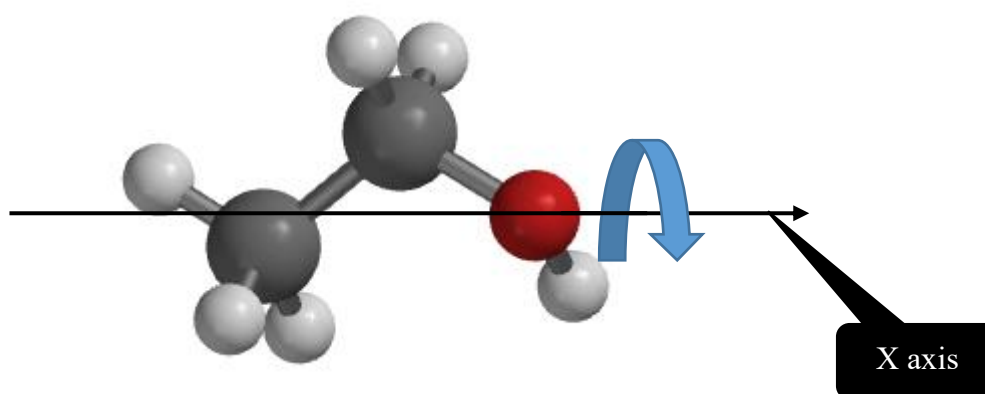


Figure 1.6. Illustration of moment of inertia along the X axis for ethanol.

4-dimensional (4D) descriptors are descriptors with additional information to the molecular geometry. One type of 4D descriptor involves embedding the molecule in a 3D grid of thousands of evenly spaced grid points and mapping the interaction of a probe, such as steric and electron distribution, to the surface of the molecule [11, 12]. Examples of this type of descriptors includes Comparative Molecular Field Analysis [40] and Comparative Molecular Similarity Indices Analysis [41] descriptors. Another type of 4D descriptors are ensemble-based descriptors, such that they describe the properties based on an ensemble of molecules, conformation, protonation states and/or orientations [42-45]. This type of descriptor incorporates conformational and alignment freedom to the descriptors, which 3D descriptors are not able to [12].

In addition to molecular descriptors, molecular fingerprints can also be calculated from the molecular structure. Molecular fingerprints can be based on the 2D or 3D structure of the molecule. Three common types of 2D molecular fingerprints are structural keys, path-based fingerprints and circular fingerprints. Structural key fingerprints encode the structure of a molecule in a binary bit string where each bit corresponds to the presence or absence of a pre-defined structural feature, and Molecular ACCess System (MACCS) key is a commonly used structural key [46, 47]. Path-based fingerprints, such as Daylight fingerprint [48], generate a binary bit string of a fixed length encoding the substructure of molecule, where the substructure is defined by following a path up to a defined number of bonds [47]. A circular fingerprint encodes the molecular structure by circular atom neighbourhoods, defined by the fingerprint diameter or radius [46, 49]. Example of circular fingerprints includes extended-connectivity

fingerprints (ECFPs) and functional-class fingerprints (FCFPs), where the latter is a variation of the former which generalises the atoms of a molecule by their functional classes [49]. Using molecular fingerprints, the similarity of molecular structures can be compared using similarity coefficients such as the Tanimoto coefficient. The Tanimoto coefficient quantifies the similarity between two structures by normalising the number of fingerprint bits commonly present in both structures using the number of bits which is only present in one of the structures [50]. Tanimoto coefficients range from 0 to 1, where pairs of structures with a Tanimoto coefficient of 1 means the structures are identical by fingerprint.

Depending on the type of molecules the investigating chemical library contains, the descriptors one might want to calculate for QSAR/QSPR modelling can differ as outlined below.

1.2.4.3.1. Small molecules

Small molecules are low molecular weight molecules (< 900 Da) such as monosaccharides and drug molecules [51, 52]. As they are small, the complete structure of this type of molecule can easily be stored using SMILES and SDF. Therefore, they can easily be imported into molecular descriptor calculators such as CDK Descriptor Calculator [33] and alvaDesc [3] to calculate their 0-3D descriptors.

1.2.4.3.2. Surfactants

Surfactants are molecules which are able to sit at the interface between different phases to lower surface tension. A surfactant molecule is typically amphiphilic and comprises of hydrophobic (water repelling) and hydrophilic (water attracting) sections [2]. Similar to small molecules, surfactant molecules are of a size which can be stored using SMILES and SDF, and therefore their 0-3D descriptors can easily be calculated. However, in addition to the descriptors for the whole surfactant molecule, descriptors for the hydrophobic and hydrophilic sections of a surfactant have been shown to be important in describing the critical micelle concentration [25]. Although this can be calculated by importing the hydrophobic and hydrophilic fragments of the surfactant molecule into molecular descriptor calculators, it is necessary to bear in mind that the 2-3D descriptors calculated may not be an accurate description of the fragment in a picture of the whole molecule as the hydrophobic and hydrophilic fragments are disconnected from each other. Complications can also arise with the consideration of the relative position of the hydrophobic and hydrophilic sections, which is an important factor to the properties of surfactant exhibited (further details described in *Chapter 4: Novel Surfactant Descriptor – Potential Amphiphilicity*) [53].

1.2.4.3.3. Polymer

In comparison to small molecules and surfactants, polymers are more challenging in calculating their descriptors. Polymers vary in size as they are constructed of repeating units where the number of repeat units for a single polymer can vary [54, 55]. As a polymer exists as a mixture of different lengths, it is difficult to truly capture the molecular structure in a SMILES string or SDF. Due to the varying size, it is common to store the polymer structure as the average structure. Another common way for polymer data to be stored is in some format which denotes the repeat units the polymer contains and how the repeat units are joined together. In recent years, method such as Hierarchical Editing Language for Macromolecules (HELM) [56, 57] and BigSMILES [58] had been developed as a line notation for storing polymer structure based on the repeat units and their attachment points, although their direct compatibility with molecular descriptor calculators is unknown. Nonetheless, using the average structure or the repeat units, it is possible to calculate 0-2D descriptors with molecular descriptor calculators. However, similar to the hydrophobic and hydrophilic fragments of surfactant molecules, it is

necessary to bear in mind that the 2D descriptors calculated using the repeat units of a polymer might not be an accurate description of the repeat units in a picture of the whole molecule due to the disconnection. Calculation of 3D descriptors for polymers can be difficult as the 3D conformation of a polymer molecule is dependent on environmental factors, such as solvent used and the concentration when in solution [59].

In addition to the above, it is important that the structural properties for each QSAR/QSPR study are calculated using the same molecular descriptor calculator where possible. This is because for some properties, such as neighbourhood information content, different molecular descriptor calculators can have a slightly different algorithm for the calculation, which would lead to a different result for the same chemical compound.

In order to avoid confusion, for the rest of this thesis, any structural properties calculated for an entry in the databases used in QSAR/ QSPR modelling are referred to as descriptors and any of these descriptors contributing to the relationships established (D_1, D_2, \dots, D_n in **Equation 1.1**) are referred to as predictors.

1.2.5. Training and test set split

When constructing QSAR/QSPR models for prediction, it is important to split the database into training and test sets. In QSAR/ QSPR, the training set data are used to establish the relationship between the property of interest and descriptors. Once a relationship is established, the accuracy of the relationship is clarified by applying the relationship to the test set and assessing the performance (detail see *1.2.6. Performance analysis*). This is a measure to assess the predictivity of the constructed model on unknown data and reducing overfitting, where the model is only accurate in predicting the data it has seen during construction [26]. As the training set data define the domain of applicability of the constructed models, when carrying out the training and test set split, it is important to ensure the training and test set are balanced in their structure and/or endpoint to maximise the generality and the applicability of the models [60]. For example, for binary classification problems, it would be optimal for the test set to contain an equal number of samples for each class. For regression problem, the continuous endpoint would be partitioned into equal size bins for the endpoint and an equal number of samples from each bin would be selected for the test set.

Once the training and test sets are split, the descriptors of the both the training and test set need to be examined. First, the descriptors in the training set need to be inspected for near zero variance and cross-correlations. Near zero variance, as the words suggest, mean there is close to no variance within the descriptor value range. Such descriptors lack distinctive information to differentiate between different entries and can be identified by the ratio of the most common value to the second most common value [26]. Cross-correlated descriptors mean there is a high correlation between the descriptors and therefore the information such descriptors can provide the model is very similar [26]. As inclusion of such descriptors often adds more complexity to the constructed models in comparison to the information they can provide, it is often a good measure to only retain one of the cross-correlated descriptors for model construction. In addition to the examination of descriptors within the training set, it is also necessary to ensure the value of the descriptors for the test set is within the same range of values in the training set, which is another method of defining the applicability domain of the model construction [60].

1.2.6. Modelling algorithms

Modelling algorithms can be separated into two categories based on the style of endpoint: classification and regression. When the endpoint is binary or categorical, classification model algorithms are used to construct a predictive model for the data set [12, 15, 26, 27]. On the

other hand, when the endpoint is continuous, regression model algorithms are used [15, 26, 27]. Within both categories of classification and regression model algorithms, there are three sub-groups identifying the type of algorithm used: linear, non-linear and decision tree or rule-based [26, 61].

Linear models refer to models which are linear in the parameters such that for regression problems, the model can be written directly or indirectly in forms of

$$P = a_0 + a_1D_{i1} + a_2D_{i2} + \dots + a_nD_{in} + e \quad (\text{Equation 1.2}),$$

where a_0 represents the estimated intercept, a_j represents the estimated coefficients for the j^{th} predictor, D_{in} represents the value of the n^{th} predictor for the i^{th} sample, and e represents the random error that cannot be explained by the model [26, 27]. These models cannot capture the nonlinear relationships between predictors and the endpoint unless by adding non-linear terms manually [26]. For classification problems, the models seek to separate the samples into groups based on the characteristics of predictor variations. For binary classification problems, some models are based on treating the categories as 0s and 1s and use the linear regression model to predict whether the sample is closer to 0 or 1 [26].

Different to linear models, non-linear models do not require one to know explicitly or specify the non-linearity of the relationship prior to model construction [26]. These inherently non-linear models are more frequently seen in QSARs and QSPRs, covering a wide range of frequently used modelling techniques such as support vector machine (SVM, further details described in 1.2.6.4.2. *Support vector machines (SVM)*) and k-nearest neighbours (KNN, further details described in 1.2.6.4.3. *K-nearest neighbours (KNN)*).

Decision tree or rule-based models are also non-linear in nature. Decision tree-based models use one or more nested if-then statements for the predictors to partition the samples [26, 27]. These models are called tree models as the if-then statements act as the branches and the endpoints are like the leaves at the end of the branches. Rule-based models refer to models which have their if-else statements collapsed into independent conditions, which can then be simplified or pruned for better coverage of the data with a smaller number of rules [26]. For both decision tree and rule-based models, they are highly interpretable when the model only consists of a single decision tree or a single set of rules. However, these single tree/rule models are known to be unstable and less-than-optimal in their predictive performance [26]. This is because slight change in the data can affect the tree/rule drastically and if a sample falls out of the predictor space defined by the predictors during the model construction phase, a large prediction error would arise. In order to overcome these weaknesses, the frequently used tree/rule-based models, such as random forest (RF), incorporates multiple ensembles of decision trees. The prediction is then made as by majority vote or averaging of the prediction of the ensembles.

As QSAR and QSPR uses existing data to identify the underlying relationships, it is natural for the identified relationship to best fit the data that is used to construct the relationship. However, in the field of predictive modelling, if the model is highly constrained to the data used in construction, the model would be overfitted for the data and can have difficulty in predicting the endpoint for an unknown sample, especially if predictors of the unknown sample fall “out” of the predictor space of the data used in construction. In order to reduce overfitting, separation of the database into training and test set is a key step allowing the predictive performance of the constructed model(s) to be tested on data which was unknown to the construction process [26]. During the model construction, the training set would be split into k roughly equal size sets, and each set would be held out one at a time and the rest of the training set would be put

through for model construction. Each of the held-out set would then be predicted by the model constructed using the rest of the training set to assess the predictive performance of the model (cross-validation). In addition, for most modelling algorithms, there are a set of kernel parameter which constrict overfitting [26]. These kernel parameters can be tuned by repeating the modelling process with a range of values for the kernel parameters to find the optimal which gives rise to the least overfitting. These above measures for reducing overfitting can easily be carried out within the programming language R.

R is an openly available programming language and software environment for statistical computing and graphics [62, 63]. For the purpose of this thesis, aside from the “base” package, a user-created package “caret” (classification and regression training) was used extensively as it contains a number of very useful functions for:

- Balanced training/test set splitting – training sets used to build the predictive model and test sets to test the model’s action are required for validation of the constructed model predictivity,
- Descriptor pre-processing – processing of the descriptors into a format suitable for the modelling algorithm,
- Model tuning using resampling – training set is resampled and cross-validated for each kernel parameter to find the optimal value(s) [26],
- Variable importance estimation – estimation of the importance of the variables using different methods depending on the modelling algorithm,

and much more [64]. It also allows various model training using unified syntax by calling on the relevant function from other packages and returns the result in a uniform fashion, allowing easy comparison of various models trained with the same data set.

Within this thesis, three linear, six non-linear and five decision tree algorithms were explored across the different projects, each with their own pros and cons as explained later, in search for models with optimal performance, which are interpretable as is, or can be made interpretable with aid of simpler models that have comparable performance.

1.2.6.1. *Linear models*

1.2.6.1.1. *Least square linear regression (LM)*

Least square linear regression (LM) is the simplest regression model algorithm available which computes a vector that contains the coefficient for each predictor to find the plane that minimises the sum-of-squared errors between the observed and the predicted response [20, 26, 65, 66]. With this simplicity, it is easier to analyse and interpret compared with other regression model algorithms [65, 67, 68]. However, it is prone to flaws which can hinder the use of the model or cause it to not be interpretable, and measures are required to avoid these. Such flaws include that the direct interpretability of the coefficients produced by LM is limited to the following conditions: none of the predictors can be determined from a combination of one or more of the other predictors and the number of predictors are less than the number of observations [20, 26, 65, 66]. Only under this circumstance can the coefficients produced be used directly for the analysis of the predictor – observation relationships. In order to fulfil the conditions, the descriptors need to be pre-processed to reduce collinearity and the dimensionality of the predictor space. Another drawback of this algorithm related to the descriptors is that any nonlinear structure within the data cannot be identified [20, 26, 65]. In order to accommodate any nonlinear structure present within the data, addition of modified descriptors is necessary. However, this adds to the number of the descriptors present and

therefore can cause the resulting coefficients to be not easily interpretable [15]. In addition, LM have several assumptions regarding the residual of the predictions. It assumes the residuals are normally distributed and the variance is constant [27, 66, 69]. These assumptions can be checked using diagnostic plots: quantile-quantile (QQ) plot and scale-location plot. QQ plots allow the examination of the distribution of the residuals (**Figure 1.7a**), whereas scale-location plots allow the examination of the variance of the residuals (**Figure 1.7b**) [27].

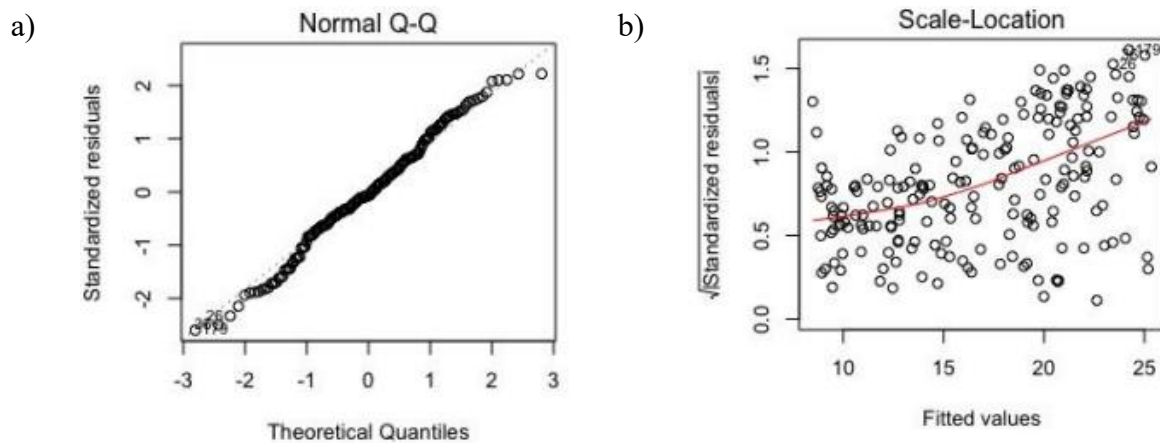


Figure 1.7. Example a) QQ plot and b) scale-location plot for linear regression assumption diagnostics. Taken from [27]. When the residuals are normally distributed, the data points follow the diagonal dashed lined in the QQ plot. Constant variance in residuals can be demonstrated by a horizontal line with equally spread points in a scale-location plot, which is not the case in the example.

1.2.6.1.2. Partial least square (PLS)

Partial least square (PLS) is another linear regression model algorithm which can accommodate correlated predictors. It finds linear combinations of predictors as components that maximally summarise the variation of the predictors while needing those components to have maximum correlation with the response at the same time (**Figure 1.8**) [9, 15, 26, 27, 61]. This algorithm makes a balance between the predictor space dimension reduction, which usually hinders the use of LM, and a predictive relationship with the response, and the number of components retained at the end of the model can be tuned to find the optimum [26]. On the other hand, as the predictors are combined prior to relation establishment, the resulting model can be more difficult to interpret than LM [61].

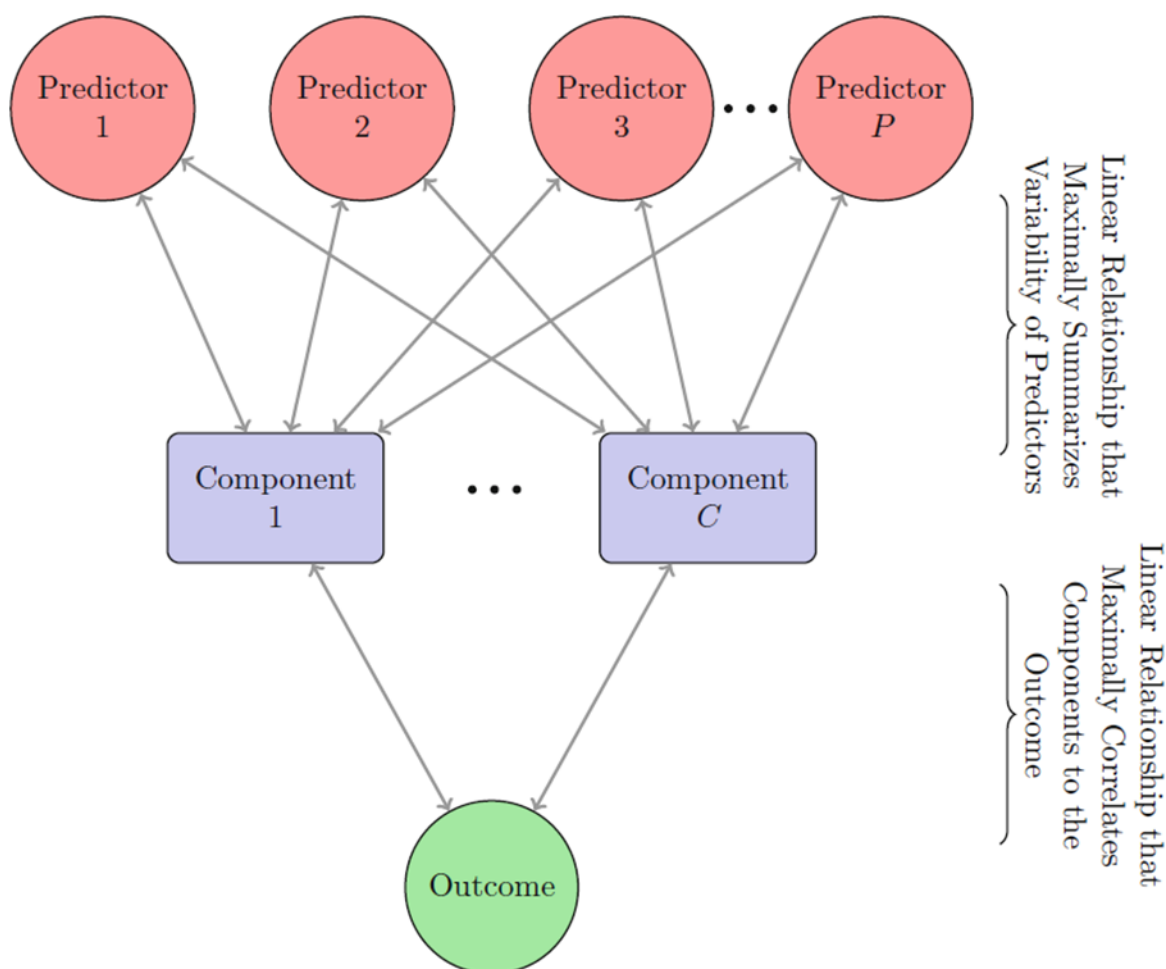


Figure 1.8. A diagram representing the structure of a PLS model, taken from [26]. PLS finds components that simultaneously summarise variation of the predictors while being optimally correlated with the outcome [26].

Within drug discovery, PLS had been successfully applied to a database of 751 known human A_{2A} adenosine receptor antagonists (**Figure 1.9**) in validating the general pharmacophore hypothesis for the receptor by Bacilieri *et al.* [70]. Within the research, conformations of the antagonist at the binding site of the receptor had been calculated and two conformations had been selected for each antagonist: one conformation with the best scoring function for its interaction with the binding site, another conformation which best reproduces the binding interaction observed in the crystallographic pose of the receptor. 3D descriptors of the both sets of conformations were then calculated and PLS models have been constructed on each set of conformations. The performance of the models was then compared to verify the similarity of the conformation in their interaction with the binding site.

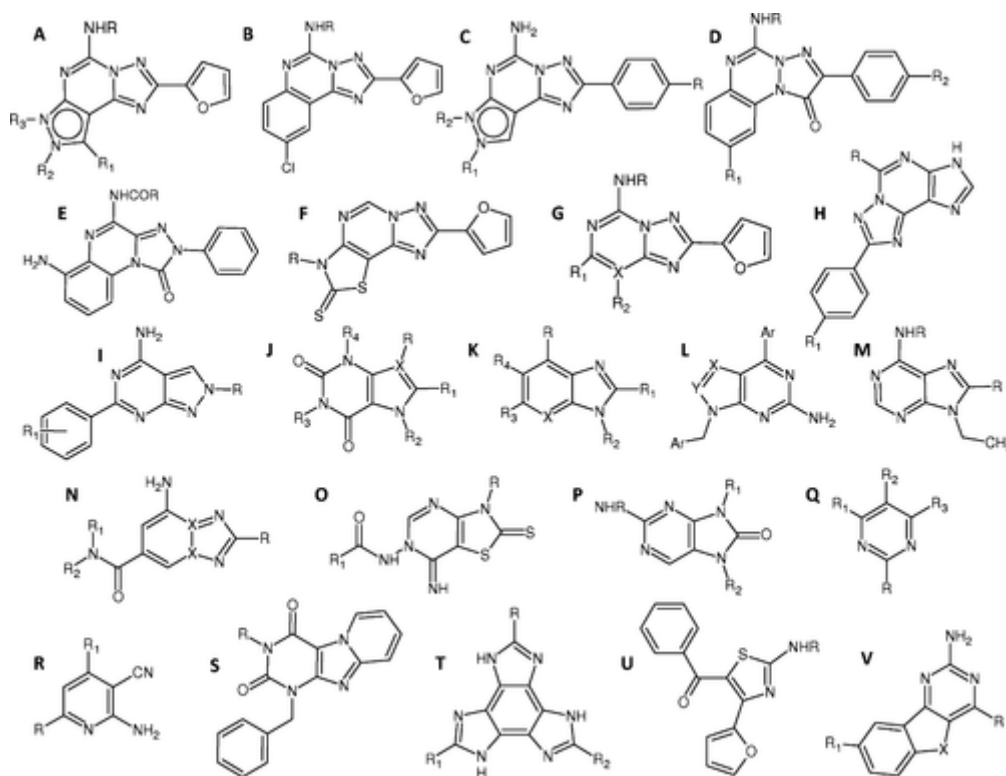


Figure 1.9. Molecular scaffolds of the 751 selected human A2A adenosine receptor antagonist in the validation of the general pharmacophore hypothesis for the receptor by Bacilieri *et al.*, taken from [70].

1.2.6.1.3. Partial least squares discriminant analysis (PLSDA)

Partial least squares discriminant analysis (PLSDA) is an application of the linear regression model partial least squares (PLS) to a classification problem, where it aims to find a straight line which divides the predictor space into two [26, 71, 72]. The algorithm involves the search of latent variables, composed of a defined number of predictor components, which reduce the predictor space dimension and optimises correlation with the categorical response represented as dummy variables (0s and 1s) [26]. New samples are predicted as a number for dummy variables of each class and the class with the largest predicted value is the predicted class [26].

1.2.6.2. Non-linear models

1.2.6.2.1. Mixture discriminant analysis (MDA)

Mixture discriminant analysis (MDA) is a non-linear classification model based on linear discriminant analysis (LDA), where instead of each class coming from a single normal distribution, each data point has a probability of belonging to each class [26, 27]. Within MDA, each class is represented by multiple multivariate normal distributions, and this number can be controlled [26]. A single multivariate normal distribution is then generated from the multiple multivariate normal distribution by creating a per-class mixture and the class of a new sample is then determined by the position of it within the multivariate normal distribution for each class [26].

1.2.6.2.2. Support vector machines (SVM)

Support vector machines (SVM) were originally developed by Vapnik for classification problems which seek to find the optimal hyperplane within the predictor space which can separate the classes [15, 27, 61, 73]. The optimal hyperplane would have the maximum

separation between the classes and the data points defining this separation are called support vectors, which are used to find the class probability for the new samples [26, 61, 73]. For a data set that is not completely separable, SVM utilises different kernel functions to map the input space into a high dimensional feature space, allowing calculation without needing to transform the predictors [26, 27, 61, 73]. When there are data points that lie on the wrong side of the hyperplane or within the margin, the margin is penalised by adding a cost [26, 27].

The SVM algorithm had also been adapted to solve regression problems. In regression problems, SVM is used to minimise the effect of outliers on the regression equations where ϵ controls the width of the ϵ -insensitive zone which the data points contributes linearly to the regression [26, 73]. For data points that are outside of the ϵ -insensitive zone, they are penalised by the cost parameter [26, 73]. As in classification problems, SVM also utilises different kernel functions in regression problems to allow calculation of the distance between a data point and the regression line without transforming the predictors (**Figure 1.10**) [26, 73].

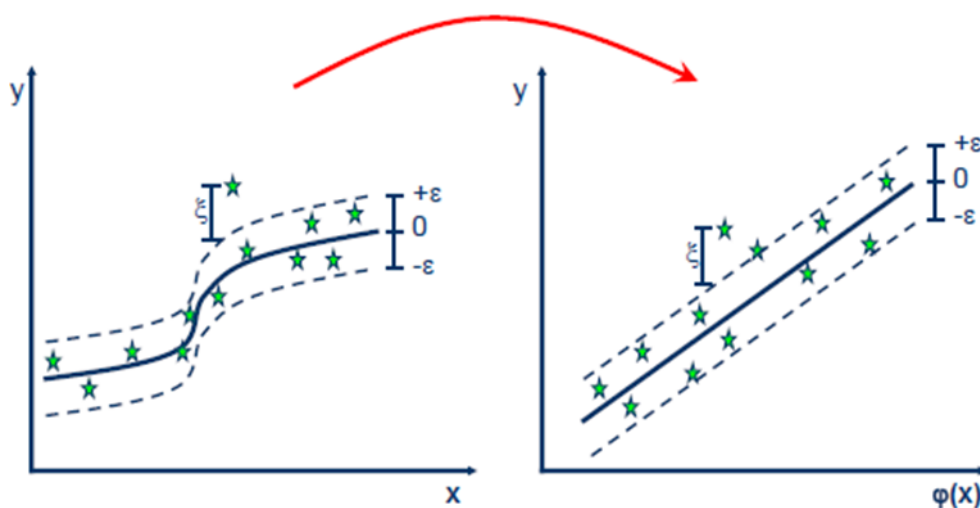


Figure 1.10. An example of SVM for regression, taken from [74]. Left: SVM before kernel function transformation; Right; SVM after kernel function transformation for non-linear data.

SVM has had many success within the pharmaceutical industry, including virtual screening [75, 76], predicting the likeness of a molecular compound to be a kinase inhibitor [77], drug-induced ototoxicity prediction [78], milk / plasma drug concentration prediction [79] and cytochrome P450 inhibitor / non-inhibitor classification [80]. In the study of drug-induced ototoxicity prediction by Zhou *et al.* [78], a database of 572 reported ototoxic small molecules (positive) and 347 reported non-ototoxic small molecules (negative) was used. A test set of 11 positives and 63 negatives was extracted from the database and three subsets of the database were selected based on the risk or strength in ototoxicity. 0-2D descriptors were then calculated for these molecules and predictive models were constructed using a SVM based method, where their performance was compared with other models developed within the study using the test set.

1.2.6.2.3. K-nearest neighbours (KNN)

K-nearest neighbours (KNN) is another widely used non-linear model algorithm that simply predicts a new sample using the K-closest samples from the training set [12, 15, 26, 27, 61]. Due to this nature, KNN construction is solely based on the observations from the training set and therefore cannot be summarised by a model clearly [26]. In regression, the predicted value

of the new sample is the mean of the K-closest samples' observation value and the value of K can be determined by resampling [26, 27, 61]. In classification, the class probability of the new sample is the proportion of K-closest samples' observed in each class [12, 15, 26, 27].

KNN has found some applications such as cytochrome P450 inhibitor / non-inhibitor classification [80] and predicting the isoform 2 of human cyclooxygenase (COX-2) inhibition within the pharmaceutical field [81]. In the prediction of COX-2 inhibition by Baurin *et al.* [81], 354 COX-2 inhibitors were investigated (**Figure 1.11**). 2-3D descriptors were calculated and the log value of inhibitor concentration for inhibiting 50% of the enzymatic activity (pIC_{50}) of COX-2 was modelled with KNN method. The model was then analysed in a classification style method against other models developed within the research by inducing a threshold of $pIC_{50} = 7.5$ to classify the activity of the inhibitors (i.e. active: $pIC_{50} \geq 7.5$, inactive: $pIC_{50} < 7.5$).

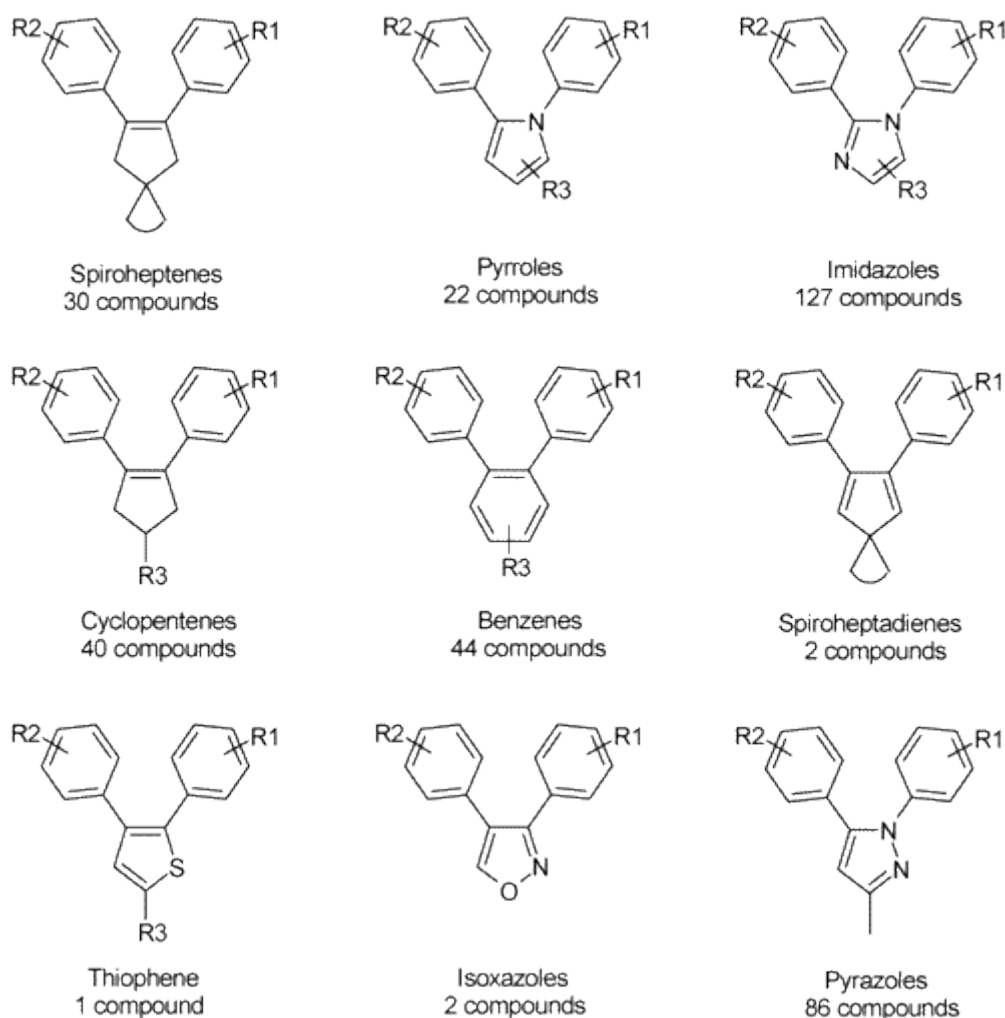


Figure 1.11. Molecular scaffold of the 354 COX-2 inhibitors were investigated in the prediction of COX-2 inhibition by Baurin *et al.*, taken from [81].

1.2.6.2.4. Neural networks (nnet) and their derivatives

Neural networks (nnet) are non-linear regression model algorithms that were inspired by brain function [26, 61]. Within this type of algorithm, linear combinations of predictors are transformed by a sigmoidal function to form hidden units, where predicted values are then computed by a linear combination of all the hidden units (**Figure 1.12**) [15, 26, 61, 73]. The

number of hidden units is one of the tuning parameters of nnets, with weight decay as the other, which penalise and regularise the model to reduce overfitting. However, a nnet is an algorithm where the generation of hidden units relies on random values initially, which means the model generated is usually not the global solution, but a local solution that is dependent on the initial random values, inducing model instability [26]. As a solution to this, model averaged neural networks (avNNet) are often used where several models are created with different initial random values and averaged to produce a more stable model [26]. In relation to nnet, deep neural network (DNN) is one of the algorithms amongst the deep learning algorithms, a more recently expanding yet fast-growing field of statistical algorithms. In essence, DNN is a variation of nnet where the nnet only has one layer of hidden units while DNN have multiple layers of hidden units, and the number of hidden units within each layer can be tuned to find the optimum [82].

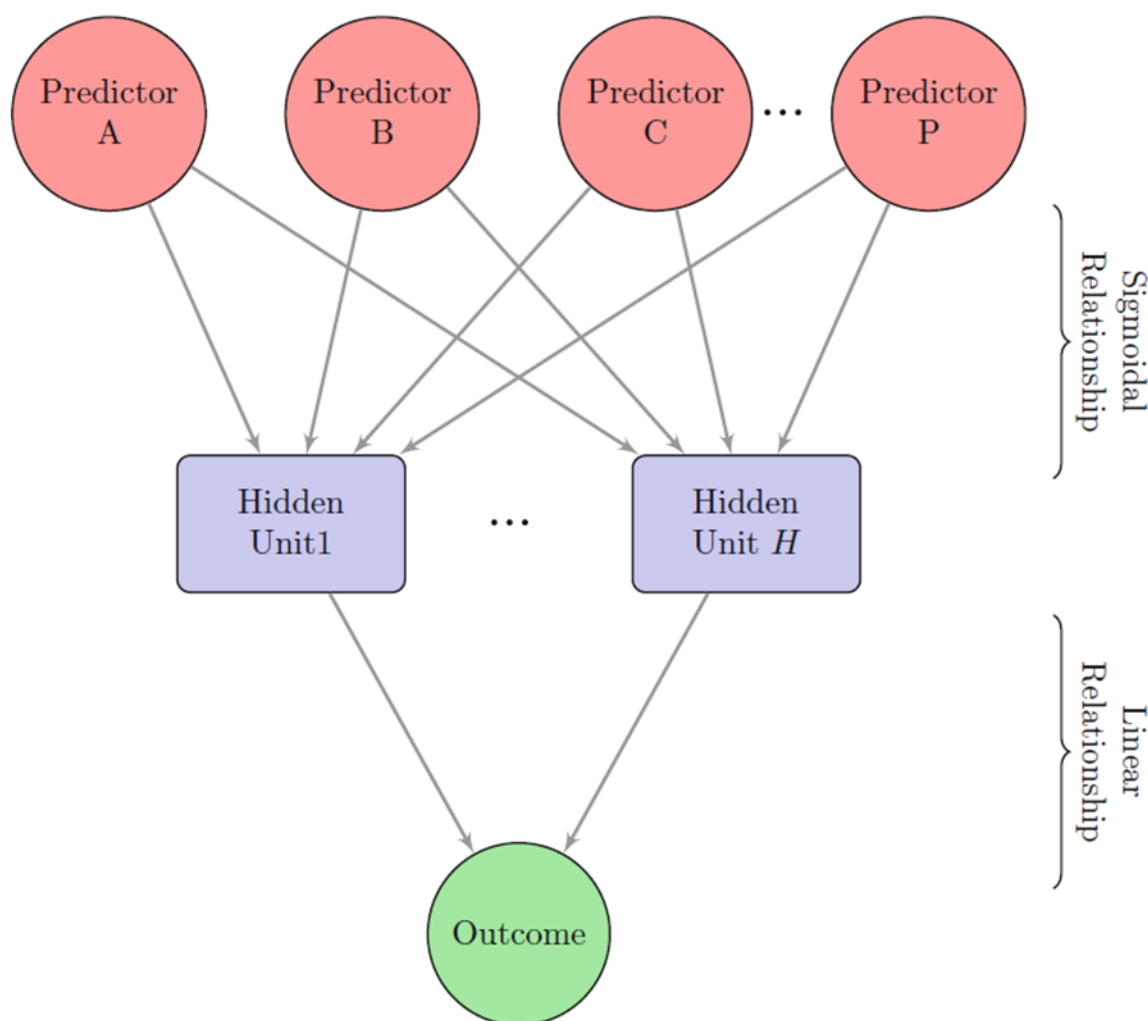


Figure 1.12. A diagram of a neural network with a single hidden layer, taken from [26].

Application of neural network derivatives include estimation of partition coefficients for drug discovery [83], predicting the likeness of a molecular compound to be a kinase inhibitor [77] and predicting the isoform 2 of human cyclooxygenase (COX-2) inhibition within the pharmaceutical field [81]. In the prediction of the “kinase inhibitor-likeness” of molecules by Briem *et al.* [77], a database of 565 active and 7194 inactive compounds was used, where the activity is classed by the inhibitor concentration for inhibiting 50% of the enzymatic activity

CHAPTER 1: INTRODUCTION

(IC₅₀). Due to the unbalanced dataset, an ensemble-based sampling approach was taken where the inactive compounds were split into 13 subsets and 13 training sets were formed by combining each of the 13 inactive compound subsets with the 565 active compounds. Fragment-based descriptors were then calculated for the database and nnet models were constructed for the 13 training sets. The models were then used to predict the activity of 10 kinase inhibitors not present in the original database (**Figure 1.13**). The final class prediction for the 10 kinase inhibitors was derived from the majority voting of the predictions made by the 13 models. The performance of the nnet model ensemble was finally compared with other model ensembles constructed in the study.

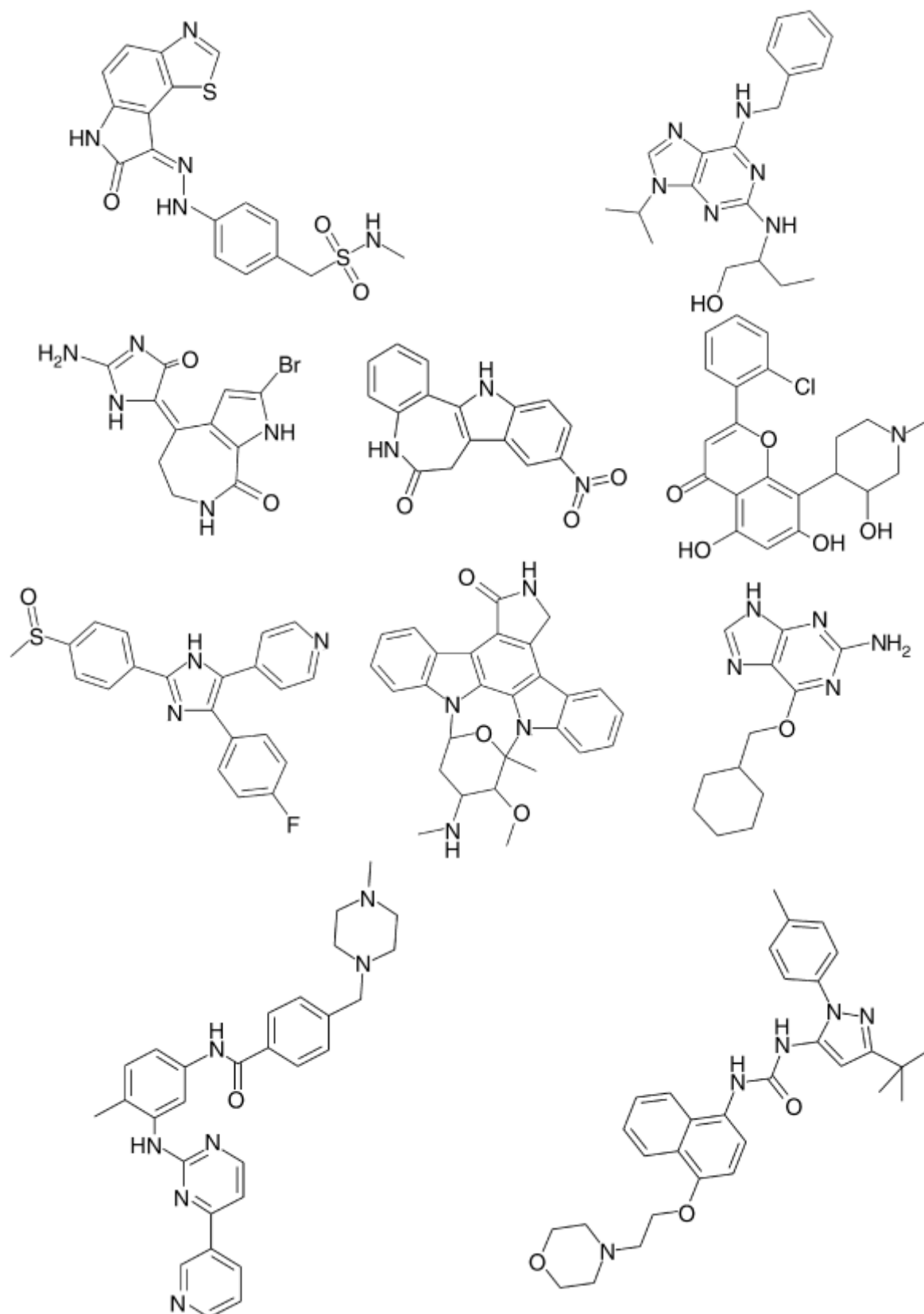


Figure 1.13. Structures of the test set used in the prediction of the “kinase inhibitor-likeness” of molecules by Briem *et al.*, taken from [77].

1.2.6.3. *Tree/Rule-based models*

1.2.6.3.1. C5.0

C5.0 (C5) is an advanced version of Quinlan's C4.5 classification model [26]. The algorithm of C5 proceeds by splitting the training set based on the predictors which provide the maximum information gain, such that the minimum number of predictors is used for the split. The subsets defined by the first split are then each split again, this time based on a different set of predictors. The process is then repeated until the subsets cannot be split any further. At this stage, the conditions of the final splits are examined, and the ones which do not contribute significantly to the accuracy of the model are removed [26, 84]. Although there is little literature on the improvements C5 contain, the author of the "caret" package unravelled the program source code and found that the improvements within C5 lead to generation of smaller and simpler trees than C4.5, including carrying out a final global pruning procedure that attempts to remove the sub-trees until the error rate exceeds that when there was no pruning [26]. C5 is the only model used that does not contain any parameters for tuning.

1.2.6.3.2. Random forest (RF)

Random forest (RF) is a robust tree modelling algorithm used widely in statistical analysis and predictive modelling, that has the capability of producing one of the most accurate models without the need of pre-processing the input data [27, 85]. RF is essentially an ensemble of multiple classification trees, where each tree is built using bootstrap samples of the input data and a random subset of the top k predictors at each split of the tree (**Figure 1.14**), specified by m_{try} [26, 27, 61, 85].

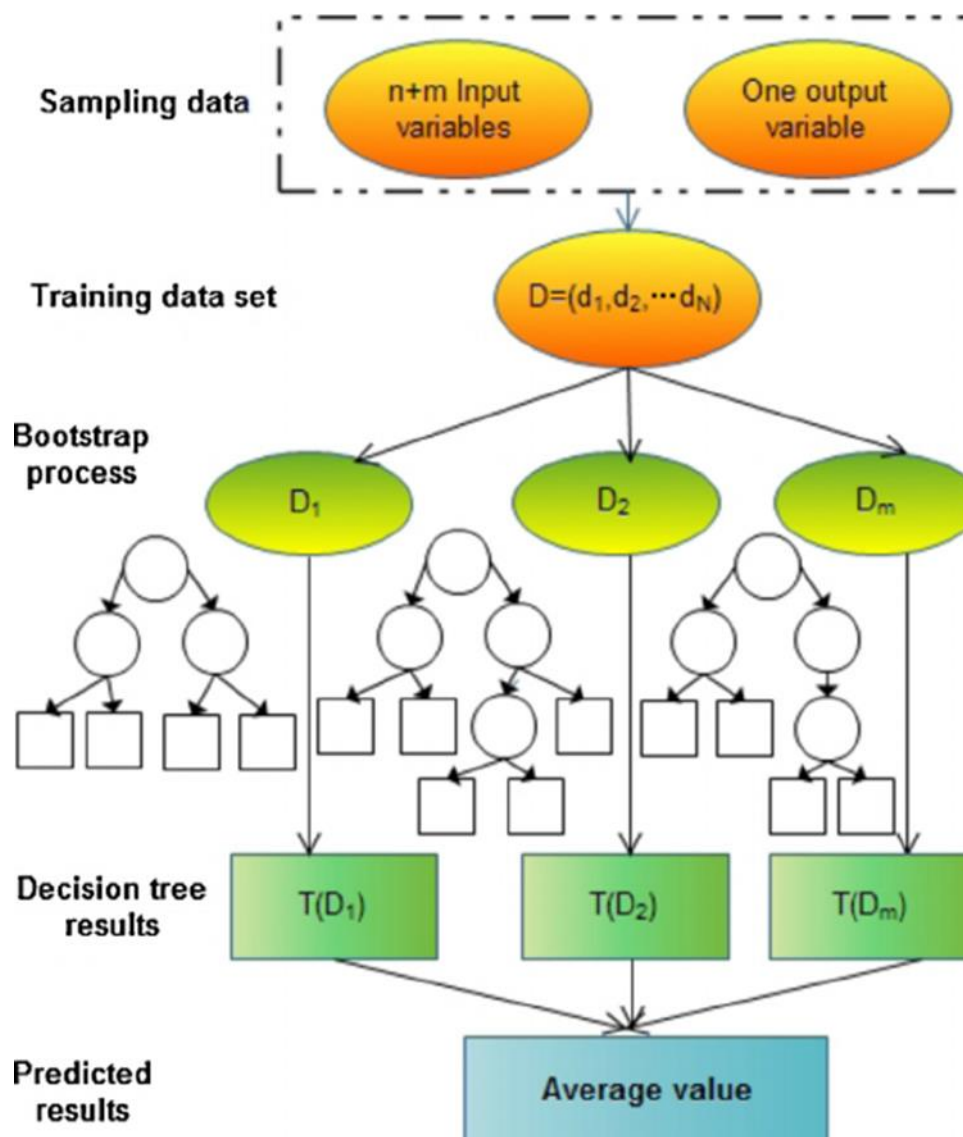


Figure 1.14. A flowchart of a random forest modelling algorithm taken from [86].

In classification problems, each of the classification trees produces a prediction for the class of a new sample as a vote, and the proportion of votes in each class across all the trees within the ensemble is the predicted probability for each class [26, 85]. In regression problems, the predicted value of a sample is in return the mean of the predicted values of each of the trees [26, 85]. Due to this ensemble nature of RF, it is not easy to gain an understanding of the relationships between the predictors and the observations, however, it is possible to quantify the impact of the predictors within the ensemble on prediction [26, 27, 85].

Successful uses of RF include cytochrome P450 inhibitor / non-inhibitor classification [80] and prediction of drug aqueous solubility [87]. In the prediction of drug aqueous solubility by Palmer *et al.* [87], a database of 988 organic molecules with experimental aqueous solubility data was used. The database was split into a training set of 658 molecules and test set of 330 molecules by random partition. 3D conformations of the molecules were generated from their SMILES strings and over 200 descriptors, including 126 2D descriptors and 36 3D descriptors were calculated. RF models were then constructed based on the 2D descriptors and a combination of 2D and 3D descriptors. The importance of the 2D and 3D descriptors was

compared by comparing the performance of the models and the predictor importance of the models which used the combination of the 2D and 3D descriptors. The performance of the model was also compared with other models constructed within the studies and in previous studies.

1.2.6.3.3. Stochastic gradient boosting (GBM)

Stochastic gradient boosting (GBM) is a variation of boosted tree model which uses a loss function in combination with a weak learner [26]. A loss function is a function which calculates the distance between the current prediction and the observation [88], such as squared error for regression and area under the Receiver Operating Characteristic curve (AUROC) for classification; decision tree with depth restriction serves as a weak learner [26]. Within the algorithm, it seeks to find an additive model that minimises the loss function with multiple iterations [26, 89]. GBM holds similarity to RF where both are ensembles of tree models, however, the tree models within GBM are built using a randomly selected subset of the input data at each iteration, dependent on previously computed trees. The tree models within GBM also have depth restriction and unequal contributions to the final model, which differs from RF [26]. GBM can have the tendency to over-fit as, although with restriction, regression trees seek to find the optimum model for the given data, and therefore can be low in prediction for new data. As a countermeasure a regularisation parameter, shrinkage, is used to constrain the learning progress, where the coefficients in the decision tree are made to shrink towards 0 [26]. An alternative version of GBM, extreme gradient boosting (XGB) contains more regularisation parameters within the algorithm to aid the control of over-fitting [26, 90], which results in possibly better performance.

These boosting methods have found their use in areas such as organic compound boiling point prediction in structure-property relationship analysis [91] and prediction of biological activity (e.g. as cyclooxygenase-2 inhibitors) of molecules [92]. In the boiling point prediction of organic compounds by Zhang *et al.* [91], a database of 2475 organic compounds with known boiling point was split into a training set of 1856 compounds and a test set of 619 compounds. 3D conformations of the compounds were then calculated and 810 0-3D descriptors were calculated from the 3D conformations. GBM models were constructed based on the 2D and 3D descriptors separately and their performance was compared. A comparison of the GBM models was also made with other models constructed within the study.

1.2.6.3.4. Cubist

Cubist is a rule-based modelling algorithm that was originally only available in a commercial capacity [26]. Compared to other rule-based modelling algorithms, although the model tree construction involved is almost identical to a regression tree model construction, cubist has some significant differences in the techniques used during the smoothing process for reducing the chance of over-fitting, rule creation and pruning for rule reduction/combination [26]. Cubist also has the option to generate model committees to aid bias reduction of the final rule-based model in a boosting-like procedure [26]. Another aspect of cubist which differs from other rule-based modelling algorithm is the ability to adjust the predicted value with the K most similar samples from the training set (neighbours), where the adjustment is accordant to the difference between the predicted value of the new sample and its closest neighbours [26]. However, there are no established techniques for measuring the importance of each predictor within the model [26].

Uses of cubist include the modelling of immediate release tablet formulation database within the pharmaceutical field by Shao *et al.* [93]. Within the study, a database of 205 tablet

formulations containing the ingredient composition, process conditions and tablet properties was used. The eight measured tablet properties were treated as the endpoints. The database was split into a training set of 177 record and a test set of 28 records based on the randomised experimental design of the database origin. Cubist models were then calculated for each endpoint and their performance was compared with other models constructed within the same study.

1.2.7. Performance analysis

Depending on if the model is a regression model or classification model, a different set of metrics is used to assess its performance.

1.2.7.1. Regression performance metrics

When the performance of a regression model is being assessed for its validity, Golbraikh and Tropsha originally suggested the following criteria which need to be all fulfilled for the model to be deemed acceptable [94]:

- Cross-validated R^2 via internal resampling on training set > 0.5 ,
- R^2 on test set > 0.6 ,
- R^2 through origin (R_0^2) close to R^2 ,
- $\frac{R^2 - R_0^2}{R^2} < 0.1$ or $\frac{R^2 - R_0^2}{R^2} < 0.1$, and
- Corresponding $0.85 \leq k \leq 1.15$ or $0.85 \leq k' \leq 1.15$,

where R^2 is the squared correlation coefficient, the correlation coefficient between the predicted and observed value; R_0^2 is the inversed squared correlation coefficient through origin; k is the gradient of predicted values vs. observed values; k' is the gradient of observed values vs. predicted values.

They emphasised the predictive ability of a regression model can only be estimated using an external test set that was not used for building the model, and moreover that both the normal R^2 and the R^2 through the origin must have similar values for the model to have a high predictive ability. However more recently Alexander, Tropsha and Winkler published a paper which emphasised the importance of RMSE and suggested that for a predictive model, the following criteria needs to be fulfilled [95]:

- High R^2 on test set and
- Low RMSE of test set predictions,

where RMSE is the root mean square error, the average distance between the observed and predicted values.

1.2.7.1.1. Examples of regression performance analysis

In the study of predicting drug aqueous solubility by Palmer *et al.* [87], PLS, SVM, nnet and RF were used to construct regression models on 988 organic molecules for the prediction of their solubility. As part of the analysis, the performance metric of the four models during cross validation and on the test set were compared (**Table 1.4**), and RF was found to be better performing then the other three models during both cross validation and prediction of test set.

Table 1.4. The R^2 and RMSE result for the 10-fold cross-validation within the training set and for the prediction of the test set in the study of predicting drug aqueous solubility by Palmer *et al.* [87]

Model	Cross-validated R^2	Cross-validated RMSE	Test set R^2	Test set RMSE
PLS	0.856	0.787	0.859	0.773
nnet	0.864	0.742	0.866	0.751
SVM	0.880	0.726	0.878	0.720
RF	0.896	0.685	0.890	0.690

In the study of organic compound boiling point prediction by Zhang *et al.* [91], GBM models were constructed based on the 2D or 3D descriptors of 2475 organic compounds with observed boiling point. In order to compare the performance of the models, the R^2 and RMSE of the two models were compared, where the 2D descriptor-based models were found to have a better performance (**Table 1.5**).

Table 1.5. The result of the cross-validation within the training set and for the prediction of the test set based on 2D and 3D descriptors in the study of organic compound boiling point prediction by Zhang *et al.* [91]

Descriptors based upon	Cross-validated R^2	Cross-validated RMSE	Test set R^2	Test set RMSE
2D	17.89	0.957	18.19	0.954
3D	19.96	0.946	20.11	0.948

1.2.7.2. Classification performance metrics

The performance of classification models can be assessed using a variety of metrics: Sensitivity (Sens, **Equation 1.3**) and specificity (Spec, **Equation 1.4**) which describes how much of each class are correctly predicted [26], accuracy (Acc, **Equation 1.5**) which describes the overall rate of true predictions for all observations [26], balanced accuracy (BalAcc, **Equation 1.6**) which describes accuracy with data skewness considerations [96], area under the Receiver Operating Characteristic curve (AUROC) which can assess how much better the model prediction is over a random guess [26, 97] and Kappa. Many of these metrics are based on a combination of aspects of a confusion matrix for the model (**Table 1.6**) and for all of these metrics, a higher value indicates better performance.

Table 1.6. An outline of a confusion matrix for a classification model where Class A is chosen as the positive class and Class B is the negative class

		Observation		
		Class A (positive)	Class B (negative)	
Prediction	Class A (positive)	True Positive (TP)	False Positive (FP)	Positive = TP + FP
	Class B (negative)	False Negative (FN)	True Negative (TN)	Negative = FN + TN
		True = TP + FN	False = FP + TN	Total = TP + FP + FN + TN

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (\text{Equation 1.3})$$

$$\text{Specificity} = \frac{TN}{FP + TN} \quad (\text{Equation 1.4})$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad (\text{Equation 1.5})$$

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \quad (\text{Equation 1.6})$$

Kappa (**Equation 1.7**) assesses the accuracy aspect of a model with class distribution considerations which can have a value between -1 to 1. However, normally Kappa of a model is a value between 0 and 1, where it is commonly interpreted as follows:

- < 0.20 = poor agreement
- 0.20 – 0.40 = fair agreement
- 0.40 – 0.60 = moderate agreement
- 0.60 – 0.80 = good agreement
- 0.80 – 1.00 = very good agreement

Models with moderate agreement or above are usually considered to be good. However, Kappa can be prone to error induced by prevalence [98, 99] of the data and therefore as a safety net, Kappa should be considered together with accuracy such that models with high accuracy and Kappa values are ones that are truly in good agreement [98].

$$\text{Kappa} = \frac{(TP + TN) - (True \times Positive + False \times Negative)}{1 + (True \times Positive + False \times Negative)} \quad (\text{Equation 1.7})$$

The Matthews correlation coefficient (MCC) is another metric for analysing the performance of the models based on the confusion metrics. It is a class symmetrical metric, where switching the positive and negatives leads to the same result, calculated by **Equation 1.8** [100]. However, although MCC is considered more informative than Kappa in binary classification problems, it is found that MCC and Kappa generate similar and concordant scores when Kappa is positive [100].

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (\text{Equation 1.8})$$

1.2.7.2.1. Examples of classification performance analysis

In the study of drug-induced ototoxicity prediction by Zhou *et al.* [78], a database of 572 reported ototoxic small molecules (positive) and 347 reported non-ototoxic small molecules (negative) was used. After the extraction of a test set composed of 11 positives and 63 negatives from the database, three SVM based classification models were constructed based on the subset of the remaining database, defined by the risk or strength in ototoxicity of the molecule. The performance of the models was then compared using sensitivity, specificity, accuracy and MCC (**Table 1.7**). As a result, model 2 was found to have better performance in predicting

both the positives and negative across the training set and test set in comparison to the other two models.

Table 1.7. The performance metric for the three SVM based model on the training set and test set in the study of drug-induced ototoxicity prediction by Zhou *et al.* [78]

Model	Training set				Test set			
	Sens	Spec	Acc	MCC	Sens	Spec	Acc	MCC
SVM model 1	0.847	0.785	0.823	0.629	0.636	0.889	0.723	0.500
SVM model 2	0.829	0.901	0.867	0.734	0.818	0.921	0.853	0.707
SVM model 3	0.594	0.986	0.914	0.689	0.405	1.000	0.609	0.435

1.2.7.3. *y*-randomisation and Z score

Validation of models can be carried out by random shuffling of the observations before training, building models in the same way as for the true data (e.g. hyperparameter optimisation) and the best models applied to the test set and performance metrics obtained (*y*-randomisation) [21]. A minimum of three repeats of *y*-randomisation can be used to test the validity of models built of the true data. Model robustness (**Equation 1.9**) can be calculated subsequently [101], where H is the metric chosen for robustness calculation, e.g. H = Kappa for classification models and H = R² for regression models.

$$Z = \frac{H_{original\ training} - Average(H_{y-randomised\ training})}{Standard\ deviation(H_{y-randomised\ training})} \quad (\text{Equation 1.9})$$

If the original model was valid, the overall performance of *y*-randomised models should be greatly reduced in comparison, with an expected measure of performance being close to random. This can be observed by a high Z score, with Z > 3 considered as significant [102].

1.3. Visualisation of Chemical Libraries

Aside from QSAR and QSPRs, computational chemistry can also aid the visualisation of chemical libraries. When investigating a chemical library, there are often aspects of the library one is interested in important to their research, such as the 2D and 3D structure of the molecules, the proportion of the library with a certain number of rings, the number of molecules with charged atoms, the 3-dimensionality of the molecules, etc. These areas can often be visualised by usage of 2D or 3D graphs to present molecular properties of interest regarding the library, or presenting the molecular graph of the individual molecular structures (**Figure 1.15**). Such visualisations are beneficial when trying to present one's work to an audience, in form of publications or presentations, as they help the presenter to present and the audience to grasp the image.

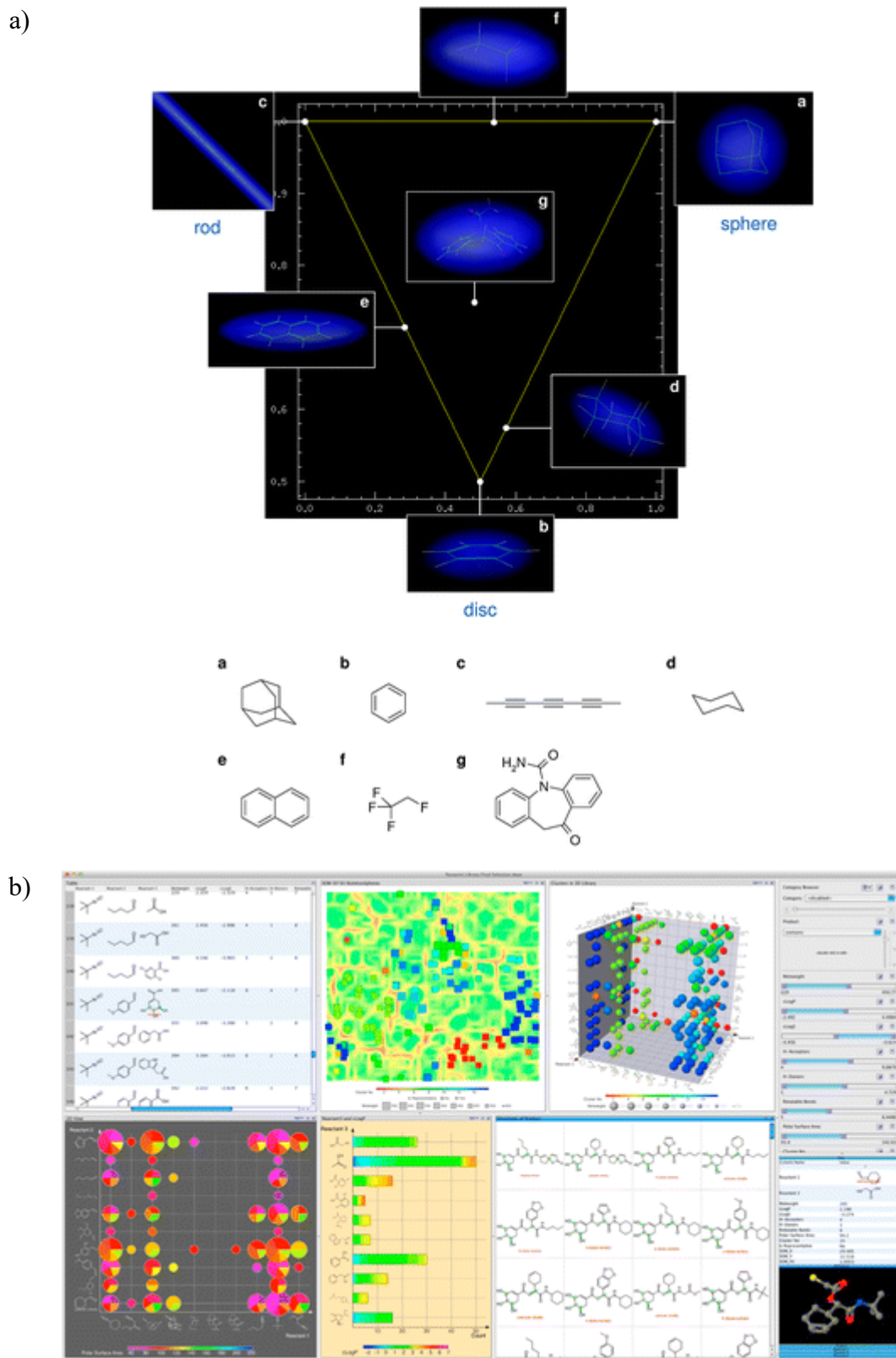


Figure 1.15. Examples of visualisation of chemical libraries a) using normalized PMI ratios as shape descriptors for molecule a to g [103], and b) using DataWarrior [4].

In medicinal chemistry, important images in relation to molecular structures could include their shape, their 3-dimensionality, the location of their pharmacophores (e.g. the hydrogen bond

donors and acceptors on the structure, **Figure 5.16**). Approaches to capture the shape and 3-dimensionality include the use of normalised principal of inertia to present the chemical library as to how rod, disc or sphere-like the molecular structures are [103], and using the plane of best fit and the distance of the heavy atoms from the plane of best fit of the molecular structure to determine how three-dimensional it is [104]. On the other hand, approaches to identify and visualise the pharmacophores of molecular structures within chemical libraries, and measure their diversity includes Pharma [105], HookSpace [106] and gridding and partitioning [107].

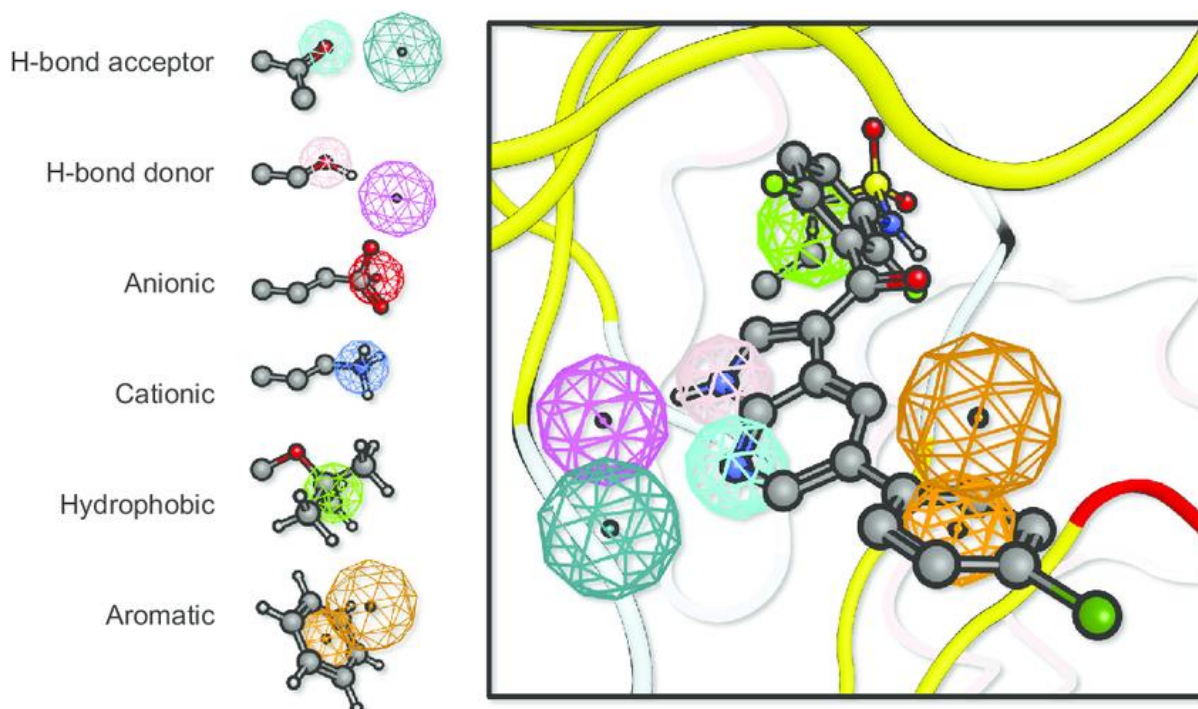


Figure 5.16. Example structure with pharmacophoric features highlighted, taken from [108].

With description of each of the methods detailed in *Chapter 5: Visualisation of Chemical Functionality for a Chemical Library*, Pharma is a search tool built to identify and calculate pharmacophore features [105], HookSpace assesses the diversity of functional groups for a chemical library using the spatial relationship between functional group pairs [106], and gridding and partitioning which tracks the position of the pharmacophores in relation to a predefined core [107]. Application of these approaches is mainly in the analytical area, for example Pharma was used in the benchmark exercise of unpublished data from pharmaceutical companies by Carlson *et al.* for matching small molecule ligands to pharmacophore models developed by them [109]; Gridding and partitioning formed the base of the shape comparison algorithm in the Quantum Isostere Database, a web-based tool for predicting bioisosteric replacements [110]. However, although these approaches can analyse and compare pharmacophores of molecules within a library, they do not hold the ability to visualise query pharmacophores of the whole, or a selected part of the chemical library in 3D, which can be important when trying to identify areas of chemical space the library can or cannot explore [19].

CHAPTER 1: INTRODUCTION

1.4. Thesis Overview

Chapter 2 describes the development of robust QSARs for predicting Ames mutagenicity of molecules and compare the models vs. commercial and open source alternatives. QSPRs for polymer cleaning properties are developed in *Chapter 3* and key structural molecular properties are identified.

A novel protocol for automated search of hydrophobic and hydrophilic section boundary of surfactant is explained in *Chapter 4* with surfactant QSPRs explored incorporating the descriptors generated from the protocol.

Chapter 5 described the protocol for visualising pharmacophore of a chemical library.

1.5. References

1. Grant, G. H.; Graham Richards, W. *Computational Chemistry*; Oxford University Press, 1995.
2. McNaught, A. D.; Wilkinson, A. *IUPAC. Compendium of Chemical Terminology, 2nd ed. (the "Gold Book")*. Blackwell Scientific Publications, Oxford (1997), (accessed 14/07/2021).
3. Mauri, A. alvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints. In *Ecotoxicological QSARs*, Roy, K. Ed.; Springer US, 2020; pp 801-820.
4. Sander, T.; Freyss, J.; von Korff, M.; Rufener, C. DataWarrior: an open-source program for chemistry aware data visualization and analysis. *Journal of Chemical Information and Modeling* **2015**, *55* (2), 460-473. DOI: 10.1021/ci500588j From NLM.
5. *Chemistry Collection: Basic Chemistry User Guide, Pipeline Pilot Release 16.5.0.143*; Accelrys Software Inc.: San Diego, 2016. (accessed 2018).
6. Berthold, M. R.; Cebon, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In *Data Analysis, Machine Learning and Applications*, Berlin, Heidelberg, 2008; Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R., Eds.; Springer Berlin Heidelberg: pp 319-326.
7. *Jupyter Notebooks*. <https://jupyter.org> (accessed 28/01/2022).
8. Brown, F. K. Chapter 35 - Chemoinformatics: What is it and How does it Impact Drug Discovery. In *Annual Reports in Medicinal Chemistry*, Bristol, J. A. Ed.; Vol. 33; Academic Press, 1998; pp 375-384.
9. Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*; Springer, 2007. DOI: 10.1007/978-1-4020-6291-9.
10. Estrada, E.; Bonchev, D. Chemical Graph Theory. 2013; pp 1538-1558.
11. Consonni, V.; Todeschini, R. Molecular Descriptors. In *Recent Advances in QSAR Studies: Methods and Applications*, Puzyn, T., Leszczynski, J., Cronin, M. T. Eds.; Springer Netherlands, 2010; pp 29-102.
12. Grisoni, F.; Ballabio, D.; Todeschini, R.; Consonni, V. Molecular Descriptors for Structure-Activity Applications: A Hands-On Approach. *Methods Mol Biol* **2018**, *1800*, 3-53. DOI: 10.1007/978-1-4939-7899-1_1 From NLM.
13. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* **1988**, *28* (1), 31-36. DOI: 10.1021/ci00057a005.
14. Mozrzymas, A. Molecular connectivity indices for modeling the critical micelle concentration of cationic (chloride) Gemini surfactants. *Colloid Polym Sci* **2017**, *295* (1), 75-87. DOI: 10.1007/s00396-016-3979-3 PubMed.
15. Van de Waterbeemd, H.; Rose, S. Chapter 23. Quantitative Approaches to Structure-Activity Relationships. 2003; pp 351-369.
16. Hansch, C.; Fujita, T. ρ - σ - π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *Journal of the American Chemical Society* **1964**, *86* (8), 1616-1626. DOI: 10.1021/ja01062a035.
17. Harrold, M. W.; Zavod, R. M. *Basic Concepts in Medicinal Chemistry*; American Society of Health-System Pharmacists, 2013.
18. Roy, K.; Kar, S.; Das, R. N. Chapter 10 - Other Related Techniques. In *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*, Roy, K., Kar, S., Das, R. N. Eds.; Academic Press, 2015; pp 357-425.
19. Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discovery Today* **2002**, *7* (17), 903-911. DOI: [https://doi.org/10.1016/S1359-6446\(02\)02411-X](https://doi.org/10.1016/S1359-6446(02)02411-X).

20. Flom, P. *The Disadvantages of Linear Regression*. 2017. <http://sciencing.com/disadvantages-linear-regression-8562780.html> (accessed 04/07/2017).
21. Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics* **2010**, 29 (6-7), 476-488, <https://doi.org/10.1002/minf.201000061>. DOI: <https://doi.org/10.1002/minf.201000061> (accessed 2021/07/14).
22. Helenius, A.; McCaslin, D. R.; Fries, E.; Tanford, C. [63] Properties of detergents. In *Methods in Enzymology*, Vol. 56; Academic Press, 1979; pp 734-749.
23. Fourches, D.; Muratov, E.; Tropsha, A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *Journal of chemical information and modeling* **2010**, 50 (7), 1189-1204. DOI: 10.1021/ci100176x PubMed.
24. OECD. *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models*; 2014. DOI: <https://doi.org/10.1787/9789264085442-en>.
25. Katritzky, A. R.; Pacureanu, L.; Dobchev, D.; Karelson, M. QSPR Study of Critical Micelle Concentration of Anionic Surfactants Using Computational Molecular Descriptors. *Journal of Chemical Information and Modeling* **2007**, 47 (3), 782-793. DOI: 10.1021/ci600462d.
26. Kuhn, M.; Johnson, K. *Applied predictive modeling*; Springer New York, 2013. DOI: <https://doi.org/10.1007/978-1-4614-6849-3>.
27. Kassambara, A. *Machine Learning Essentials: Practical Guide in R*; shtda, 2018.
28. Sayle, R. *1st-class SMARTS patterns*. Daylight, 1998. <https://www.daylight.com/meetings/summerschool98/course/basics/ref/sayle/> (accessed 19/12/2021).
29. Daylight Chemical Information Systems, I. *SMARTS Tutorial*. https://www.daylight.com/dayhtml_tutorials/languages/smarts/ (accessed 19/01/2021).
30. Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *Journal of Cheminformatics* **2015**, 7 (1), 23. DOI: 10.1186/s13321-015-0068-4.
31. Biovia. *CTfile formats*. 2020. https://discover.3ds.com/sites/default/files/2020-08/biovia_ctfileformats_2020.pdf (accessed 17/12/2021).
32. *Spartan'18*; Wavefunction, Inc.: (accessed 2017).
33. *CDK Descriptor Calculator*; <http://www.rguha.net/code/java/cdkdesc.html> (accessed 30/09/2020).
34. Sabirov, D. S.; Shepelevich, I. S. Information Entropy in Chemistry: An Overview. *Entropy* **2021**, 23 (10), 1240.
35. Todeschini, R.; Consonni, V.; Mannhold, R.; Kubinyi, H.; Folkers, G. *Molecular Descriptors for Chemoinformatics: Volume I: Alphabetical Listing / Volume II: Appendices, References*; Wiley, 2009.
36. *Molecular Descriptors Guide*; U.S. Environmental Protection Agency, 2020.
37. Kubinyi, H. *3D QSAR in drug design: volume 1: theory methods and applications*; Springer Science & Business Media, 1993.
38. Rybinska, A.; Sosnowska, A.; Barycki, M.; Puzyn, T. Geometry optimization method versus predictive ability in QSPR modeling for ionic liquids. *Journal of Computer-Aided Molecular Design* **2016**, 30 (2), 165-176. DOI: 10.1007/s10822-016-9894-3.
39. Nicklaus, M. C.; Wang, S.; Driscoll, J. S.; Milne, G. W. Conformational changes of small molecules binding to proteins. *Bioorganic & Medicinal Chemistry* **1995**, 3 (4), 411-428. DOI: 10.1016/0968-0896(95)00031-b From NLM.

40. Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society* **1988**, *110* (18), 5959-5967. DOI: 10.1021/ja00226a005 From NLM.
41. Klebe, G.; Abraham, U.; Mietzner, T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *Journal of Medicinal Chemistry* **1994**, *37* (24), 4130-4146. DOI: 10.1021/jm00050a010 From NLM.
42. Andrade, C. H.; Pasqualoto, K. F.; Ferreira, E. I.; Hopfinger, A. J. 4D-QSAR: perspectives in drug design. *Molecules* **2010**, *15* (5), 3281-3294.
43. Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *Journal of the American Chemical Society* **1997**, *119* (43), 10509-10524. DOI: 10.1021/ja9718937.
44. Vedani, A.; Briem, H.; Dobler, M.; Dollinger, H.; McMasters, D. R. Multiple-conformation and protonation-state representation in 4D-QSAR: the neurokinin-1 receptor system. *Journal of Medicinal Chemistry* **2000**, *43* (23), 4416-4427. DOI: 10.1021/jm000986n From NLM.
45. Vedani, A.; McMasters, D. R.; Dobler, M. Multi-conformational ligand representation in 4D-QSAR: Reducing the bias associated with ligand alignment. *Quantitative Structure-Activity Relationships* **2000**, *19* (2), 149-161.
46. O'Boyle, N. M.; Sayle, R. A. Comparing structural fingerprints using a literature-based similarity benchmark. *Journal of Cheminformatics* **2016**, *8* (1), 36. DOI: 10.1186/s13321-016-0148-0.
47. Zagidullin, B.; Wang, Z.; Guan, Y.; Pitkänen, E.; Tang, J. Comparative analysis of molecular fingerprints in prediction of drug combination effects. *Briefings in Bioinformatics* **2021**, *22* (6), bbab291. DOI: 10.1093/bib/bbab291 (accessed 12/18/2021).
48. Daylight Chemical Information Systems, I. *Fingerprints - Daylight Theory*. <https://www.daylight.com/dayhtml/doc/theory/theory.finger.html> (accessed 18/12/2021).
49. Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50* (5), 742-754. DOI: 10.1021/ci100050t.
50. Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics* **2015**, *7* (1), 20. DOI: 10.1186/s13321-015-0069-3.
51. *Small molecule*. Nature, <https://www.nature.com/subjects/small-molecules> (accessed 18/12/2021).
52. Macielag, M. J. Chemical Properties of Antimicrobials and Their Uniqueness. In *Antibiotic Discovery and Development*, Dougherty, T. J., Pucci, M. J. Eds.; Springer US, 2012; pp 793-820.
53. Georgiou, G.; Lin, S.-C.; Sharma, M. M. Surface-Active Compounds from Microorganisms. *Bio/Technology* **1992**, *10* (1), 60-65. DOI: 10.1038/nbt0192-60.
54. Mayo, F. R.; Lewis, F. M. Copolymerization. I. A Basis for Comparing the Behavior of Monomers in Copolymerization; The Copolymerization of Styrene and Methyl Methacrylate. *Journal of the American Chemical Society* **1944**, *66* (9), 1594-1601. DOI: 10.1021/ja01237a052.
55. Walling, C. Copolymerization. XIII.1 Over-all Rates in Copolymerization. Polar Effects in Chain Initiation and Termination. *Journal of the American Chemical Society* **1949**, *71* (6), 1930-1935. DOI: 10.1021/ja01174a009.
56. Zhang, T.; Li, H.; Xi, H.; Stanton, R. V.; Rotstein, S. H. HELM: A Hierarchical Notation Language for Complex Biomolecule Structure Representation. *Journal of Chemical Information and Modeling* **2012**, *52* (10), 2796-2806. DOI: 10.1021/ci3001925.

CHAPTER 1: INTRODUCTION

57. Milton, J.; Zhang, T.; Bellamy, C.; Swayze, E.; Hart, C.; Weisser, M.; Hecht, S.; Rotstein, S. HELM Software for Biopolymers. *Journal of Chemical Information and Modeling* **2017**, *57* (6), 1233-1239. DOI: 10.1021/acs.jcim.6b00442.
58. Lin, T.-S.; Coley, C. W.; Mochigase, H.; Beech, H. K.; Wang, W.; Wang, Z.; Woods, E.; Craig, S. L.; Johnson, J. A.; Kalow, J. A.; et al. BigSMILES: A Structurally-Based Line Notation for Describing Macromolecules. *ACS Central Science* **2019**, *5* (9), 1523-1531. DOI: 10.1021/acscentsci.9b00476.
59. Su, W.-F. Polymer Size and Polymer Solutions. In *Principles of Polymer Design and Synthesis*, Su, W.-F. Ed.; Springer Berlin Heidelberg, 2013; pp 9-26.
60. Kar, S.; Roy, K.; Leszczynski, J. Applicability Domain: A Step Toward Confident Predictions and Decidability for QSAR Modeling. In *Computational Toxicology: Methods and Protocols*, Nicolotti, O. Ed.; Springer New York, 2018; pp 141-169.
61. Lewis, R. A.; Wood, D. Modern 2D QSAR for drug discovery. *WIREs Computational Molecular Science* **2014**, *4* (6), 505-522, <https://doi.org/10.1002/wcms.1187>. DOI: <https://doi.org/10.1002/wcms.1187> (accessed 2021/12/20).
62. R: What is R? <https://www.r-project.org/about.html> (accessed 05/07/2021).
63. Kabacoff, R. I. *R in action: Data analysis and graphics with R*; Manning Publications, 2011.
64. Kuhn, M. *The caret package*. <http://topepo.github.io/caret/index.html> (accessed 08/11/2016).
65. ClockBackward. *Ordinary Least Squares Linear Regression: Flaws, Problems and Pitfalls*. 2009. <http://www.clockbackward.com/2009/06/18/ordinary-least-squares-linear-regression-flaws-problems-and-pitfalls/> (accessed 04/07/2017).
66. Peña, E. A.; Slate, E. H. Global Validation of Linear Model Assumptions. *J Am Stat Assoc* **2006**, *101* (473), 341. DOI: 10.1198/016214505000000637 PubMed.
67. NIST/SEMATECH. *Linear Least Square Regression*. <http://www.itl.nist.gov/div898/handbook/pmd/section1/pmd141.htm> (accessed 04/07/2017).
68. Weisstein, E. W. *Least Squares Fitting*. <http://mathworld.wolfram.com/LeastSquaresFitting.html> (accessed 04/07/2017).
69. Kassambara, A. *Linear Regression Assumptions and Diagnostics in R: Essentials*. 2018. <http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/> (accessed).
70. Bacilieri, M.; Ciancetta, A.; Paoletta, S.; Federico, S.; Cosconati, S.; Cacciari, B.; Taliani, S.; Da Settimo, F.; Novellino, E.; Klotz, K. N.; et al. Revisiting a Receptor-Based Pharmacophore Hypothesis for Human A2A Adenosine Receptor Antagonists. *Journal of Chemical Information and Modeling* **2013**, *53* (7), 1620-1637. DOI: 10.1021/ci300615u.
71. Brereton, R. G.; Lloyd, G. R. Partial least squares discriminant analysis: taking the magic away. *J. Chemometr.* **2014**, *28* (4), 213-225, <https://doi.org/10.1002/cem.2609>. DOI: <https://doi.org/10.1002/cem.2609> (accessed 2021/12/20).
72. Barker, M.; Rayens, W. Partial least squares for discrimination. *J. Chemometr.* **2003**, *17* (3), 166-173, <https://doi.org/10.1002/cem.785>. DOI: <https://doi.org/10.1002/cem.785> (accessed 2021/12/20).
73. Pirhadi, S.; Shiri, F.; Ghasemi, J. B. Multivariate statistical analysis methods in QSAR. *RSC Advances* **2015**, *5* (127), 104635-104665, 10.1039/C5RA10729F. DOI: 10.1039/C5RA10729F.
74. Sayad, S. *Support Vector Machine - Regression (SVR)*. http://www.sadaysad.com/support_vector_machine_reg.htm (accessed 11/07/2017).

75. Jorissen, R. N.; Gilson, M. K. Virtual Screening of Molecular Databases Using a Support Vector Machine. *Journal of Chemical Information and Modeling* **2005**, *45* (3), 549-561. DOI: 10.1021/ci049641u.
76. Sato, T.; Yuki, H.; Takaya, D.; Sasaki, S.; Tanaka, A.; Honma, T. Application of Support Vector Machine to Three-Dimensional Shape-Based Virtual Screening Using Comprehensive Three-Dimensional Molecular Shape Overlay with Known Inhibitors. *Journal of Chemical Information and Modeling* **2012**, *52* (4), 1015-1026. DOI: 10.1021/ci200562p.
77. Briem, H.; Gunther, J. Classifying "kinase inhibitor-likeness" by using machine-learning methods. *Chembiochem : a European journal of chemical biology* **2005**, *6* (3), 558-566. DOI: 10.1002/cbic.200400109 From NLM.
78. Zhou, S.; Li, G.-B.; Huang, L.-Y.; Xie, H.-Z.; Zhao, Y.-L.; Chen, Y.-Z.; Li, L.-L.; Yang, S.-Y. A prediction model of drug-induced ototoxicity developed by an optimal support vector machine (SVM) method. *Computers in Biology and Medicine* **2014**, *51*, 122-127. DOI: <https://doi.org/10.1016/j.compbimed.2014.05.005>.
79. Zhao, C.; Zhang, H.; Zhang, X.; Zhang, R.; Luan, F.; Liu, M.; Hu, Z.; Fan, B. Prediction of Milk/Plasma Drug Concentration (M/P) Ratio Using Support Vector Machine (SVM) Method. *Pharmaceutical Research* **2006**, *23* (1), 41-48. DOI: 10.1007/s11095-005-8716-4.
80. Vasanathan, P.; Taboureau, O.; Oostenbrink, C.; Vermeulen, N. P. E.; Olsen, L.; Jørgensen, F. S. Classification of Cytochrome P450 1A2 Inhibitors and Noninhibitors by Machine Learning Techniques. *Drug Metabolism and Disposition* **2009**, *37* (3), 658. DOI: 10.1124/dmd.108.023507.
81. Baurin, N.; Mozziconacci, J. C.; Arnoult, E.; Chavatte, P.; Marot, C.; Morin-Allory, L. 2D QSAR consensus prediction for high-throughput virtual screening. An application to COX-2 inhibition modeling and screening of the NCI database. *Journal of Chemical Information and Modeling* **2004**, *44* (1), 276-285. DOI: 10.1021/ci0341565 From NLM.
82. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks* **2015**, *61*, 85-117. DOI: <https://doi.org/10.1016/j.neunet.2014.09.003>.
83. Devillers, J. EVA/PLS versus autocorrelation/neural network estimation of partition coefficients. *Perspect. Drug Discov. Design* **2000**, *19* (1), 117-131, Review. DOI: 10.1023/a:1008771606841.
84. Local, I. W. S. *C5.0 node*. <https://www.ibm.com/docs/en/watson-studio-local/2.1.0?topic=modeling-c50-node> (accessed 21/12/2021).
85. Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of chemical information and computer sciences* **2003**, *43* (6), 1947-1958. DOI: 10.1021/ci034160g.
86. Liu, M.; Liu, X.; Li, J.; Ding, C.; Jiang, J. Evaluating total inorganic nitrogen in coastal waters through fusion of multi-temporal RADARSAT-2 and optical imagery using random forest algorithm. *International Journal of Applied Earth Observation and Geoinformation* **2014**, *33*, 192-202. DOI: <https://doi.org/10.1016/j.jag.2014.05.009>.
87. Palmer, D. S.; O'Boyle, N. M.; Glen, R. C.; Mitchell, J. B. O. Random Forest Models To Predict Aqueous Solubility. *Journal of Chemical Information and Modeling* **2007**, *47* (1), 150-158. DOI: 10.1021/ci060164k.
88. Chollet, F.; Allaire, J. J. *Deep Learning with R*; Manning Publications, 2018.
89. Friedman, J. H. Stochastic gradient boosting. *Computational Statistics & Data Analysis* **2002**, *38* (4), 367-378. DOI: [https://doi.org/10.1016/S0167-9473\(01\)00065-2](https://doi.org/10.1016/S0167-9473(01)00065-2).
90. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA; 2016.

91. Zhang, J. H.; Liu, Z. M.; Liu, W. R. QSPR study for prediction of boiling points of 2475 organic compounds using stochastic gradient boosting. *J. Chemometr.* **2014**, *28* (3), 161-167, Article. DOI: 10.1002/cem.2587.
92. Mustapha, I. B.; Saeed, F. Bioactive Molecule Prediction Using Extreme Gradient Boosting. *Molecules* **2016**, *21* (8), 11, Article. DOI: 10.3390/molecules21080983.
93. Shao, Q.; Rowe, R. C.; York, P. Investigation of an artificial intelligence technology--Model trees. Novel applications for an immediate release tablet formulation database. *European Journal of Pharmaceutical Sciences* **2007**, *31* (2), 137-144. DOI: 10.1016/j.ejps.2007.03.004 From NLM.
94. Golbraikh, A.; Tropsha, A. Beware of q²! *Journal of Molecular Graphics and Modelling* **2002**, *20* (4), 269-276. DOI: [http://dx.doi.org/10.1016/S1093-3263\(01\)00123-1](http://dx.doi.org/10.1016/S1093-3263(01)00123-1).
95. Alexander, D. L. J.; Tropsha, A.; Winkler, D. A. Beware of R²: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *Journal of Chemical Information and Modeling* **2015**, *55* (7), 1316-1322. DOI: 10.1021/acs.jcim.5b00206.
96. Brodersen, K. H.; Ong, C. S.; Stephan, K. E.; Buhmann, J. M. The Balanced Accuracy and Its Posterior Distribution. In *2010 20th International Conference on Pattern Recognition*, 23-26 Aug. 2010, 2010; pp 3121-3124. DOI: 10.1109/ICPR.2010.764.
97. Hallinan, J. *Assessing and Comparing Classifier Performance with ROC Curves*. 2014. <http://machinelearningmastery.com/assessing-comparing-classifier-performance-roc-curves-2/> (accessed 07/07/2017).
98. Carpenter, B. *High Kappa Values are not Necessary for High Quality Corpora*. 2012. <https://lingpipe-blog.com/2012/10/02/high-kappa-not-necessary-high-quality-corpora/> (accessed 07/07/2017).
99. Eugenio, B. D.; Glass, M. The Kappa Statistic: A Second Look. *Computational Linguistics* **2004**, *30* (1), 95-101. DOI: 10.1162/089120104773633402.
100. Chicco, D.; Warrens, M. J.; Jurman, G. The Matthews Correlation Coefficient (MCC) is More Informative Than Cohen's Kappa and Brier Score in Binary Classification Assessment. *IEEE Access* **2021**, *9*, 78368-78381. DOI: 10.1109/ACCESS.2021.3084050.
101. Luo, M.; Wang, X. S.; Roth, B. L.; Golbraikh, A.; Tropsha, A. Application of Quantitative Structure-Activity Relationship Models of 5-HT_{1A} Receptor Binding to Virtual Screening Identifies Novel and Potent 5-HT_{1A} Ligands. *Journal of Chemical Information and Modeling* **2014**, *54* (2), 634-647. DOI: 10.1021/ci400460q.
102. Rodgers, A. D.; Zhu, H.; Fourches, D.; Rusyn, I.; Tropsha, A. Modeling Liver-Related Adverse Effects of Drugs Using kNearest Neighbor Quantitative Structure-Activity Relationship Method. *Chemical Research in Toxicology* **2010**, *23* (4), 724-732. DOI: 10.1021/tx900451r.
103. Sauer, W. H. B.; Schwarz, M. K. Molecular Shape Diversity of Combinatorial Libraries: A Prerequisite for Broad Bioactivity. *Journal of chemical information and computer sciences* **2003**, *43* (3), 987-1003. DOI: 10.1021/ci025599w.
104. Firth, N. C.; Brown, N.; Blagg, J. Plane of Best Fit: A Novel Method to Characterize the Three-Dimensionality of Molecules. *Journal of Chemical Information and Modeling* **2012**, *52* (10), 2516-2525. DOI: 10.1021/ci300293f.
105. Koes, D. R.; Camacho, C. J. Pharmer: Efficient and Exact Pharmacophore Search. *Journal of Chemical Information and Modeling* **2011**, *51* (6), 1307-1314. DOI: 10.1021/ci200097m.
106. Boyd, S. M.; Beverley, M.; Norskov, L.; Hubbard, R. E. Characterising the geometric diversity of functional groups in chemical databases. *Journal of Computer-Aided Molecular Design* **1995**, *9* (5), 417-424. DOI: 10.1007/BF00123999.

CHAPTER 1: INTRODUCTION

107. Leach, A. R.; Green, D. V.; Hann, M. M.; Judd, D. B.; Good, A. C. Where are the GaPs? A rational approach to monomer acquisition and selection. *Journal of Chemical Information and Modeling* **2000**, *40* (5), 1262-1269. DOI: 10.1021/ci0003855 From NLM.
108. Qing, X.-Y.; Lee, X. Y.; De Raeymaecker, J.; Tame, J.; Zhang, K.; De Maeyer, M.; Voet, A. Pharmacophore modeling: Advances, Limitations, And current utility in drug discovery. *Journal of Receptor, Ligand and Channel Research* **2014**, *7*, 81-92. DOI: 10.2147/JRLCR.S46843.
109. Carlson, H. A.; Smith, R. D.; Damm-Ganamet, K. L.; Stuckey, J. A.; Ahmed, A.; Convery, M. A.; Somers, D. O.; Kranz, M.; Elkins, P. A.; Cui, G.; et al. CSAR 2014: A Benchmark Exercise Using Unpublished Data from Pharma. *Journal of Chemical Information and Modeling* **2016**, *56* (6), 1063-1077. DOI: 10.1021/acs.jcim.5b00523.
110. Devereux, M.; Popelier, P. L. A.; McLay, I. M. Quantum Isostere Database: A Web-Based Tool Using Quantum Chemical Topology To Predict Bioisosteric Replacements for Drug Design. *Journal of Chemical Information and Modeling* **2009**, *49* (6), 1497-1513. DOI: 10.1021/ci900085d.

Chapter 2: Predicting Ames Mutagenicity

2.1. Ames Mutagenicity and its Importance

Within the modern drug discovery field, the mutagenicity of a compound is a crucial property that can restrict the development of a particular compound series at all stages of drug development due to its close relationship with carcinogenicity [1, 2]. In order to assist the early identification of potential mutagenic compounds and hence reduce the time and expense associated with hit to lead optimisation, *in silico* prediction of compound mutagenicity has attracted much attention from several research groups [3-5]. One of the most widely used assays for testing the mutagenicity of a compound is the Ames test, invented by Professor Bruce Ames in the early 1970s [6-9]. The Ames test contains a bacterial revertant mutation assay with a simulation of mammalian metabolism, which is highly sensitive for chemicals which can induce genetic damage and frameshift mutation in the environment [10]. However, there are limitations to mutagenicity detection using this approach, including identification of false-positives and false-negatives amongst the outputs [11], and the interlaboratory reproducibility rate is not 100% [12-14]. Nonetheless, it still serves well as a quick and cheap alternative to the standard carcinogen assays on test animals.

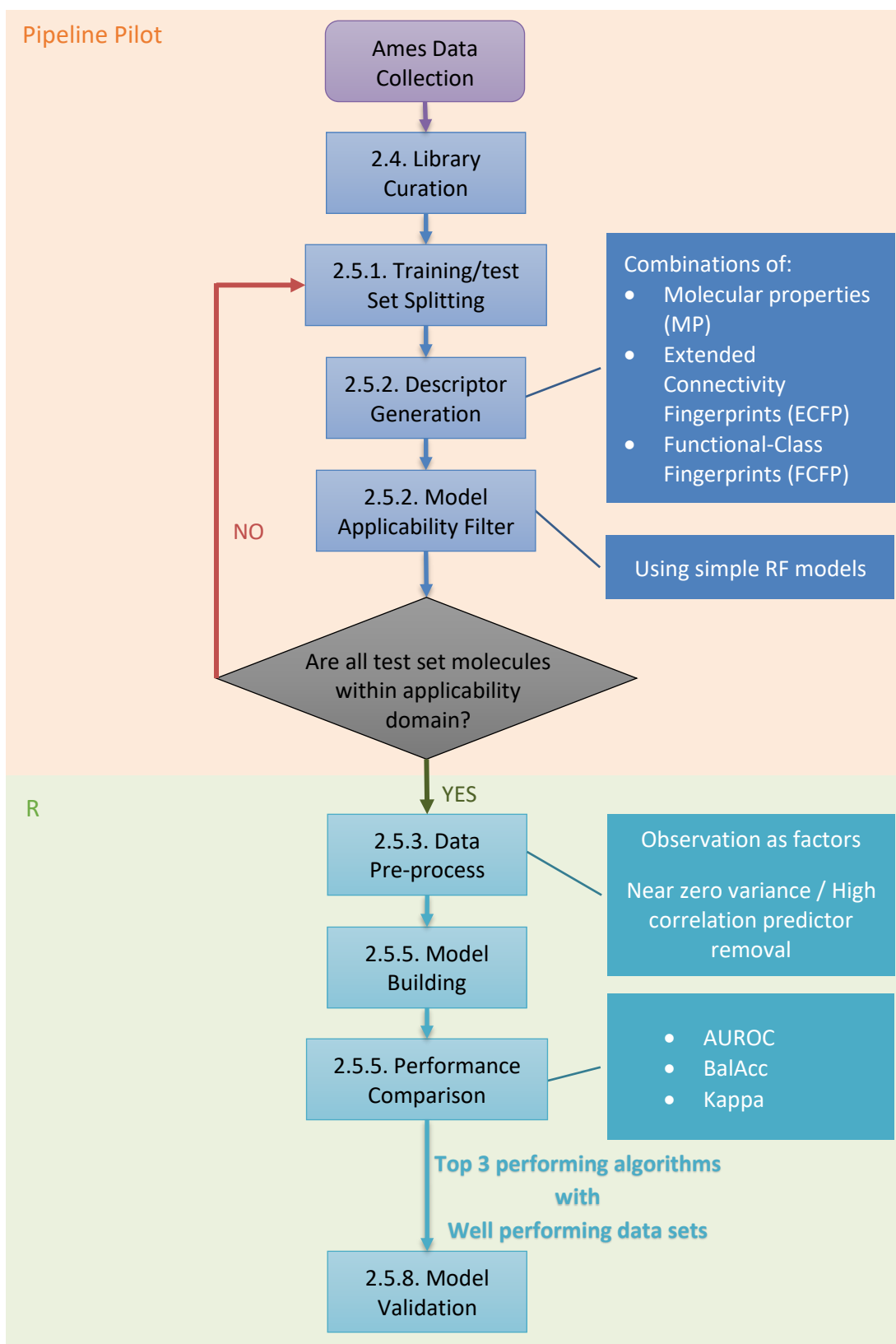
2.2. Current State of Art

As the Ames test can take considerable time and material to carry out especially where very large compound libraries need to be analysed, the availability of a reliable *in silico* model would be advantageous. By using robust predictive *in silico* models, the number of Ames tests needed to be carried out can be reduced, thus reducing the time and resources needed. In addition, an *in silico* model could be used to evaluate compounds prior to synthesis. Over the past several decades, there have been many statistical models [15, 16] and structural alert based models [3, 17, 18] published in literature alongside various commercial [4, 19] and open source software packages [19] which attempt to address Ames mutagenicity prediction. In one of the works, the performance of commercial programmes for Ames prediction (DEREK, MultiCASE, and an off-the-shelf Bayesian machine learner in Pipeline Pilot) have been compared with statistical models [15]. It was found that the statistical models constructed within the work outperform the commercial programs when analysing the corresponding Receiver Operating Characteristic curve (ROC). However, the predictive accuracy and robustness of these models are not yet satisfactory as their application domain is bound by the database which the models were constructed from and none of them have over 95% predictive accuracy [20-22]. Although for theoretical reasons, it is impossible to reach 100% accuracy with statistical models and with the imperfect reproducibility of the experimental Ames test itself, it is difficult to achieve models with accuracy over 95%, it is still important to investigate this method to overcome this hurdle.

In order to try and overcome the above difficulties, Xu *et al.* constructed a large database based on five different sources containing more than 8300 compounds with experimentally derived mutagenicity [5]. Using this, the Xu group reported some predictive models using molecular fingerprints, a type of molecular descriptor widely used in similarity searching [23], virtual screening [24] and classification [25, 26]. In addition, recently an Ames/QSAR (Quantitative Structure Activity Relationship) International Challenge Project has been reported where 12 QSAR vendors across the world have worked in collaboration to test and improve their Ames QSAR tools using a database of over 12,000 molecules established by the Division of Genetics and Mutagenesis, National Institute of Health Sciences of Japan [14].

2.3. Our Approach

We have taken the openly available mutagenicity database published by Xu *et al.* and produced a different range of models in order to try to identify improved models to those from Xu *et al.* [5]. By combining a range of physicochemical molecular properties and molecular fingerprints as compound descriptors and an alternative selection of modelling algorithms, we aimed to identify models that have excellent predictive ability and that can provide molecular insights to uncover aspects of the molecules that cause mutagenicity. **Scheme 2.1** displays the overall approach, with further details and discussion presented below.



Scheme 2.1. A flow chart of the process of searching for the best performing model and data set. Rounded rectangle: raw data; rectangle: processes; coloured background: Software used. Each step is further described in 2.4. *Process of Library Curation*, 2.5.1. *Training/test set splitting*, 2.5.2. *Descriptor generation*, 2.5.3. *Data pre-processing*, 2.5.5. *Model building and performance assessment* and 2.5.8. *Validation via y-randomisation*.

2.4. Process of Library Curation

The raw data used in this study was Xu's Ames data collection provided within 'In silico Prediction of Chemical Ames mutagenicity' as **Supporting Information 2.1a – c** [5]. The data collection consists of a training set (7617 molecules), an external validation set (731 molecules) and a balanced external set (234 molecules). The balanced external set is a subset of the external validation set in which the number of mutagens and non-mutagens are balanced. We discovered that the library *Xu et al.* presented in their work contained over 2000 duplicates, and therefore we curated the library.

The curation was achieved using Pipeline Pilot 2017 [27] via the following steps, as suggested by Tropsha [28] while taking into account the steps *Xu et al.* [5] followed. Firstly, any inorganic molecules, defined as those without carbon atoms within the structure, were removed. Secondly, each molecule was analysed and any molecules with unspecified stereochemistry were removed. Thirdly, the molecules were standardised using the InChI key [29]. Fourthly, any salt fragments, defined with the built-in salt fragment list in Pipeline Pilot, were removed to leave the ionic drug molecule. Finally, any duplicates across the data collection were identified and removed using the InChI key. For any instances where the Ames mutagenicity differs between the duplication of structures, both instances are removed.

2.5. Construction of Ames Mutagenicity Predictive Models

2.5.1. Training/test set splitting

Continuing in Pipeline Pilot 2017 [27], the curated data collection was split into training and test sets by the following steps. Firstly, the data collection was clustered into 1000 clusters using extended connectivity fingerprints [30] (ECFP, diameter = 6) with a maximal dissimilarity partitioning relocation method. Secondly, where possible, a mutagenic and a non-mutagenic representative closest to the cluster centre were taken from each cluster as candidates for the test set. Thirdly, the representative molecules from any cluster with only one representative were put into the test set. Fourthly, the mutagenic representatives from the clusters with two representatives were put into the test set until the threshold of 500 molecules was reached. Fifthly, the non-mutagenic representatives from the rest of the clusters were put into the test set. Finally, any molecules not in the test set were put into the training set.

2.5.2. Descriptor generation

Next, remaining in Pipeline Pilot 2017, molecular properties (MP) (AlogP, molecular weight, number of atoms, number of hydrogen acceptors, number of hydrogen donors, number of rotatable bonds, number of rings, number of aromatic rings, molecular surface area, molecular polar surface area, molecular polar solvent-accessible surface area, molecular solubility, logD), extended connectivity fingerprints [30] (ECFP, diameter = 4, 2048 bits) and functional-class fingerprints [30] (FCFP, diameter = 4, 2048 bits) for both the training and test set were calculated. Various combinations of these generated predictors sets were exported as separate .csv files for each of the training and test sets, resulting in seven predictor data sets:

1. ECFP: 2048 bits of extended connectivity fingerprints only,
2. FCFP: 2048 bits of functional-class fingerprints only,
3. MP: 13 molecular properties,
4. ECFP+FCFP: combination of 2048 bits of extended connectivity fingerprints and 2048 bits of functional-class fingerprints,

5. MP+ECFP: combination of 13 molecular properties and 2048 bits of extended connectivity fingerprints,
6. MP+FCFP: combination of 13 molecular properties and 2048 bits of functional-class fingerprints,
7. MP+ECFP+FCFP: combination of 13 molecular properties, 2048 bits of extended connectivity fingerprints and 2048 bits of functional-class fingerprints.

Using the training set of each predictor data set, an applicability domain was defined by tracking the property range and analysing the optimum prediction space [31]. The molecules from the test set were then filtered for model applicability by analysing them against the defined applicability domain. Any molecules which did not pass the model applicability filter were swapped for the next molecule closest to the cluster centre within the same cluster with the same mutagenicity label from the training set. The applicability domain was then defined again via the same process until no more swaps could be made. Any molecules which did not pass the model applicability filter were then removed from the test set and placed back into the training set. The training and test sets were then exported as .sdf files (**Supporting Information 2.2 – 2.3**) and .csv files for further calculation.

2.5.3. Data pre-processing

For each of the seven predictor data sets (see 2.5.2. *Descriptor generation*), the training and test set were imported into R and the column containing the mutagenicity label was converted into factors, the categorical data type within R.

Using the training set, near zero variances predictors, predictors with low frequency ratio for the most common value over the second most common value were removed. As a result, two variations of each predictor set are selected using the following rules:

- Variation 1: the full set of filtered predictors where predictors with frequency ratio above $\#observation/10$ were removed
- Variation 2: the reduced set of filtered predictors where predictors with frequency ratio above $\#observation/100$ were removed

Predictors with pair-wise absolute correlations over 0.9 were identified with `caret::findCorrelation`. For each predictor pair, the average correlation with the rest of the predictors was calculated and the predictor with the higher average correlation was removed to give the final version of the two variations of each predictor set.

2.5.4. Algorithm selection

One linear (PLSDA), three non-linear (MDA, SVM and KNN) and four tree/rule-based (C5, RF, GBM and XGB) model algorithms were chosen to investigate their performance on the seven data sets, covering the range of simpler and more interpretable models to potentially more robust but more complex models for identifying the optimally performing methods.

2.5.5. Model building and performance assessment

Model construction using `caret::train()` with the default tuning parameters outlined above and 10-fold cross validation on the training set for each of the predictor data sets (see 2.5.2. *Descriptor generation*) proceeded. For PLSDA, SVM and KNN, the predictors were centred and scaled within `caret::train()` using the `preProcess` option. The constructed models were then tested using the test set.

The performance of classification models was assessed using a variety of metrics: Sensitivity (Sens, **Equation 2.1**) and specificity (Spec, **Equation 2.2**) which describes how much of each class are correctly predicted [32], accuracy (Acc, **Equation 2.3**) which describes the overall rate of true predictions for all observations [32], balanced accuracy (BalAcc, **Equation 2.4**) which describes accuracy with data skewness considerations [33], area under the Receiver Operating Characteristic curve (AUROC) which can assess how much better the model prediction is over a random guess [32, 34] and Kappa. Many of these metrics are based on a combination of aspects of a confusion matrix for the model (**Table 2.1**) and for all of these metrics, a higher value indicates better performance.

Table 2.1. An outline of a confusion matrix for a classification model where Class A is chosen as the positive class and Class B is the negative class

		Observation		
		Class A (positive)	Class B (negative)	
Prediction	Class A (positive)	True Positive (TP)	False Positive (FP)	Positive = TP + FP
	Class B (negative)	False Negative (FN)	True Negative (TN)	Negative = FN + TN
		True = TP + FN	False = FP + FN	Total = TP + FP + FN + TN

$$Sensitivity = \frac{TP}{TP + FN} \quad (\text{Equation 2.1})$$

$$Specificity = \frac{TN}{FP + TN} \quad (\text{Equation 2.2})$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (\text{Equation 2.3})$$

$$Balanced Accuracy = \frac{Sensitivity + Specificity}{2} \quad (\text{Equation 2.4})$$

Kappa (**Equation 2.5**) assesses the accuracy aspect of a model with class distribution considerations which can have a value between -1 to 1. However, normally Kappa of a model is a value between 0 and 1, where it is commonly interpreted as follows:

- < 0.20 = poor agreement
- 0.20 – 0.40 = fair agreement
- 0.40 – 0.60 = moderate agreement
- 0.60 – 0.80 = good agreement
- 0.80 – 1.00 = very good agreement

Models with moderate agreement or above are usually considered to be good. However, Kappa is prone to error induced by prevalence [35, 36] of the data and therefore as a safety net, Kappa should be considered together with accuracy such that models with high accuracy and Kappa values are ones that are truly in good agreement [35].

$$Kappa = \frac{(TP + TN) - (True \times Positive + False \times Negative)}{1 + (True \times Positive + False \times Negative)} \quad (\text{Equation 2.5})$$

The Matthews correlation coefficient (MCC) is another metric considered for analysing the performance of the models based on the confusion metrics. However, although MCC is considered more informative than Kappa in binary classification problems [37], as our predictive models only consisted of $Kappa > 0$ (see 2.5.8. *Cross-validation models*, 2.5.9. *Test set performance check*, **Supporting Information 2.4** and **2.5**), MCC and Kappa generate similar and concordant scores above this threshold, MCC was not added to the assessment.

Across all models constructed, the number of times a modelling algorithm produced a model with $Kappa > 0.45$ on the test set were counted and the top three (SVM, RF and XGB) were selected for validation. The threshold was chosen to be 0.45 as this can still be considered to fall within the moderately accurate classification. Details of the results are described within 2.5.10. *Validation via y-randomisation*.

2.5.6. Model validation via y-randomisation and model robustness

Validation of the SVM, RF and XGB models using all the predictor data sets apart from the MP data set (for reason explained within 2.5.10. *Validation via y-randomisation*) was carried out by random shuffling of the observations before training using `base::sample()`. This was repeated three times for SVM, RF and XGB to test the validity of original models. The Model robustness, Z (**Equation 2.6**), is calculated subsequently [38].

$$Z = \frac{Kappa_{original\ training} - Average(Kappa_{y-randomised\ training})}{Standard\ deviation(Kappa_{y-randomised\ training})} \quad (\text{Equation 2.6})$$

If the original model was valid, the overall performance of y-randomised models should be greatly reduced in comparison, with an expected measure of performance being close to random. This can be observed by a high Z score, with $Z > 3$ considered as significant [39]. The variable importance of the predictors of the original models were also calculated using `caret::varImp`.

2.5.7. Data set generation

From Xu's Ames Data collection [5], a total of 5395 unique molecules were identified. These 5395 molecules were split into a training set of 4402 molecules (2549 mutagens and 1853 non-mutagens) and a test set of 993 molecules (498 mutagens and 495 non-mutagens). A total of 4109 predictors (13 molecular properties and 2×2048 bits fingerprint from ECFP and FCFP) were calculated for each molecule (**Table 2.2**). Using the training set, near zero variance and highly correlated predictors were removed to give the final number of predictors for each data set as shown in **Table 2.2**.

Table 2.2. The number of predictors generated within Pipeline pilot and after data pre-processing

Data set	Original number of predictors	Variation 1 [#]	Variation 2 [#]
ECFP	2048	1425	196
FCFP	2048	788	138
MP	13	8	8
ECFP+FCFP	4096	2200	337
MP+ECFP	2061	1433	204
MP+FCFP	2061	796	146
MP+ECFP+FCFP	4109	2208	345

[#] See 2.4. Data pre-processing for detail

2.5.8. Cross-validation models

Within this study, a total of 112 binary classification models were constructed using the combination of eight algorithms with the two variations of predictor set for each of the seven predictor data sets (see 2.5.2. Descriptor generation). All the models found showed good performance of AUROC > 0.7, Spec > 0.5, Sens > 0.7, balanced accuracy > 0.6 and accuracy > 0.6. 45 out of the 112 constructed models also have Kappa > 0.7 and accuracy > 0.85, which shows they are in good agreement. On the other hand, most models constructed with the variation 1 of the predictor sets are seen to have a slightly better performance than the corresponding model built with the variation 2 of the predictor sets (differences: AUROC +0.022 ± 0.033, Spec +0.025 ± 0.045, Sens +0.025 ± 0.044, balanced accuracy +0.025 ± 0.041, accuracy +0.025 ± 0.042, Kappa +0.051 ± 0.084). The detailed performances of these models are given in **Supporting Information 2.4**.

2.5.9. Test set performance check

All of the 112 constructed models were assessed using the test set, with the detailed performance of these models given in **Supporting Information 2.5**. In comparison to the training set, the performance on the test set was seen to decrease for each model, especially in Kappa (-0.285 ± 0.116), Sens (-0.221 ± 0.052) and AUROC (-0.149 ± 0.043). Although the Kappa metric for these models dropped notably in general, 44 of the 112 models still had Kappa > 0.45 and accuracy > 0.7 on the test set (**Table 2.3** and **Supporting Information 2.5**). In particular, most models constructed with the MP descriptors had very poor performance where specificity was less than 0.5, sensitivity less than 0.5 and balanced accuracy between 0.5 – 0.7. Again, the models constructed with the variation 1 of the predictor sets are seen to have approximately the same performance generally (AUROC +0.002 ± 0.027, Spec -0.005 ± 0.046, Sens +0.010 ± 0.038, balanced accuracy +0.002 ± 0.024, accuracy +0.002 ± 0.024, Kappa +0.004 ± 0.048) as the corresponding model built with the variation 2 of the predictor sets. With the decrease in performance on the test set in comparison to the training set, concerns of possible overfitting were considered. However, this possibility was reduced via two approaches taken throughout the model construction process. First, the training and test split was carried out in a fashion which allows the test set to cover all the chemical structural domain the entire database covers. Secondly, during the 10-fold cross validation at the model construction phase, the integrated model tuning identifies the model kernel parameters with the best predictive performance and least overfitting.

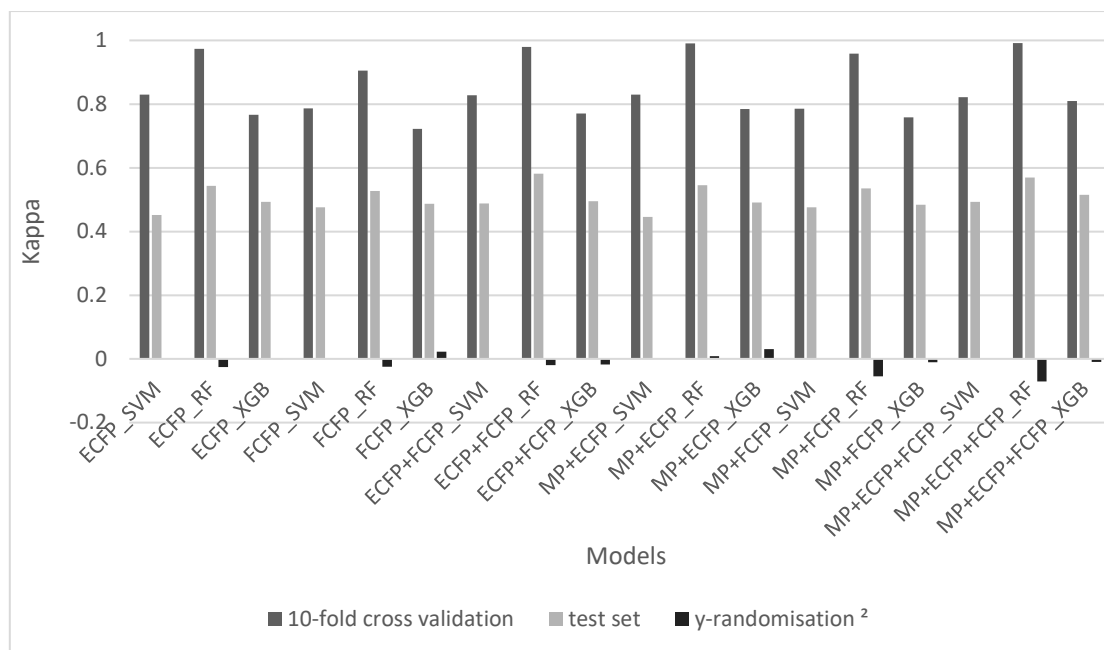
Table 2.3. The number of times a modelling algorithm produced a model with Kappa > 0.45 on the test set

Modelling algorithm	Number of models with Kappa > 0.45 on the test set
PLSDA	3
MDA	2
SVM	11
KNN	0
RF	11
C5	0
GBM	6
XGB	11

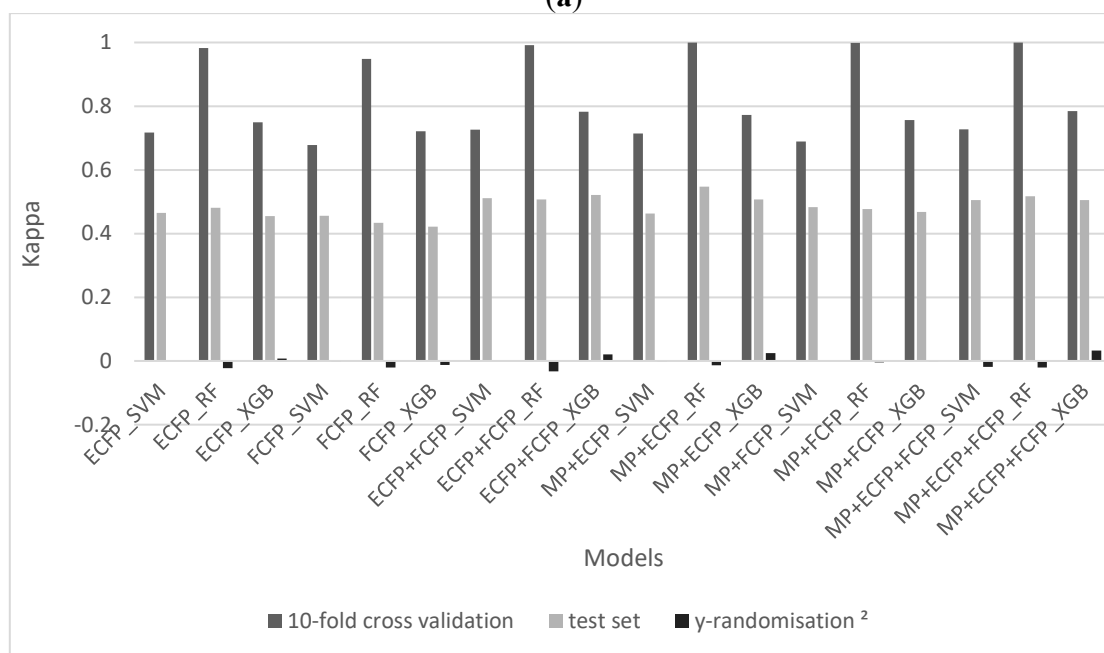
2.5.10. Validation via y-randomisation

Y-randomisation was performed on the models with Kappa > 0.45 on the test set (**Table 2.3.**; SVM, RF and XGB). Kappa was selected as the focusing analysis metric as the accuracy of the models with Kappa > 0.45 also had moderate to good performance when looking at the other metrics (**Supporting Information 2.4** and **2.5**). As the MP data set has been identified to have poor performance when used to construct models alone, y-randomisation was not performed for the models which were constructed using only the MP descriptors. As expected, the performance of the y-randomised models for SVM, RF and XGB was reduced greatly for both the 10-fold cross validation and test set, verifying that the performance of the models is much better than random (**Figure 2.1**, **Supporting Information 2.6**). This is supported by the high Z scores of the models (**Table 2.4**). It is to note that some SVM models have Z score of infinity due to all predictions of the y-randomised models during training were in the positive class, leading to the denominator of Z score, standard deviation of the training y-randomised Kappa, being zero.

CHAPTER 2: PREDICTING AMES MUTAGENICITY



(a)



(b)

Figure 2.1. Kappa of the selected classification models¹ using (a) the variation 1 of the predictor sets; (b) variation 2 of the predictor sets.¹ ECFP extended connectivity fingerprint, FCFP functional class fingerprint, MP molecular properties, RF random forest, SVM support vector machine, XGB extreme gradient boosting.² The average test set Kappa of the repeats.

Table 2.4. Z scores of the selected classification models

Predictor data set	Predictor variation #	Mode 1	Z Score	Predictor data set	Predictor variation #	Mode 1	Z Score
ECFP	1	SVM	Inf	ECFP	2	SVM	Inf
ECFP	1	RF	54.96	ECFP	2	RF	93.75
ECFP	1	XGB	69.88	ECFP	2	XGB	23.08
FCFP	1	SVM	Inf	FCFP	2	SVM	144.64
FCFP	1	RF	124.43	FCFP	2	RF	43.67
FCFP	1	XGB	42.22	FCFP	2	XGB	25.62
ECFP+FCFP	1	SVM	Inf	ECFP+FCFP	2	SVM	Inf
ECFP+FCFP	1	RF	54.24	ECFP+FCFP	2	RF	39.90
ECFP+FCFP	1	XGB	11.29	ECFP+FCFP	2	XGB	71.20
MP+ECFP	1	SVM	Inf	MP+ECFP	2	SVM	104.18
MP+ECFP	1	RF	142.39	MP+ECFP	2	RF	176.00
MP+ECFP	1	XGB	22.12	MP+ECFP	2	XGB	93.11
MP+FCFP	1	SVM	Inf	MP+FCFP	2	SVM	209.77
MP+FCFP	1	RF	148.82	MP+FCFP	2	RF	148.34
MP+FCFP	1	XGB	40.25	MP+FCFP	2	XGB	38.50
MP+ECFP+FCFP	1	SVM	Inf	MP+ECFP+FCFP	2	SVM	22.78
MP+ECFP+FCFP	1	RF	50.91	MP+ECFP+FCFP	2	RF	124.69
MP+ECFP+FCFP	1	XGB	101.95	MP+ECFP+FCFP	2	XGB	140.75

[#] See 2.4. Data pre-processing for detail

Inf - infinity

2.5.11. Variable importance of validated models

The variable importance of the models that successfully passed the y-randomisation validation assessment was calculated and is summarised in **Table 2.5**. From this, we can conclude MP predictors are often important in contributing toward constructing a good predictive model, followed by ECFP and closely FCFP. However, we must bear in mind that the MP predictors were never used alone in the models analysed as MP predictors alone produce poor predictive models. This suggests the MP predictor requires additional structural information from the fingerprint predictors to provide vital information for a good predictive model. A detailed summary of the average variable importance per predictor is given in **Supporting Information 2.7**.

Table 2.5. The overall importance of each predictor type within the SVM, RF and XGB models

Predictor type	Number of models present in	Overall importance ¹ ± SD
MP ²	6	51.8 ± 12.0
ECFP	8	21.6 ± 1.7
FCFP	8	21.8 ± 2.6

¹ sum of average importance (**Table S2.3**) divided by total number of predictors within each predictor type

²individual predictors are shown in **Table 2.6**

2.5.12. Comparison of different predictor sets used in model building

From the results of the 10-fold cross validation and external validations (**Figure 2.1**, **Supporting Information 2.4** and **2.5**), we can conclude that the MP predictors are not sufficient alone to construct a good model using most of the algorithms. This is expected as the number of MP predictors is very small and the distribution of all 8 molecular properties, including Lipinski's Rule of Five parameters [40, 41] such as number of hydrogen bond donors and LogD, overlap greatly (**Figure 2.2**), so a simple set of descriptors would not be able to distinguish mutagen from non-mutagen in general. On the other hand, when the MP predictors are used in combination with fingerprint predictors, they can have a relatively high variable importance (**Table 2.5** and **Supporting Information 2.7**). In particular, the molecular solubility and molecular surface area have high average variable importance of 69.6 ± 31.2 and 62.11 ± 25.2 respectively whereas the lowest average importance of the MP predictors is the number of hydrogen bond donors (30.8 ± 26.6) (**Table 2.6**). This suggest that although the MP predictors alone are not sufficient, they can still contribute towards the construction of a good model.

CHAPTER 2: PREDICTING AMES MUTAGENICITY

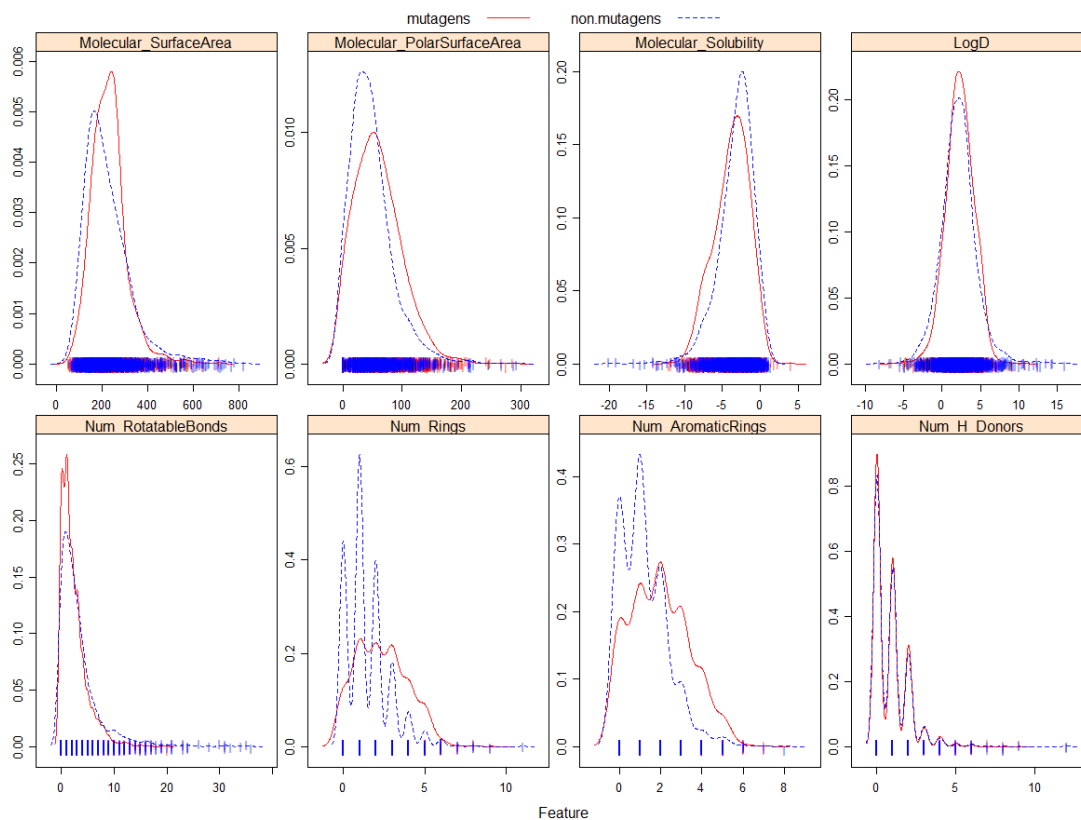


Figure 2.2. Distribution of the 8 retained predictors on the training set from the MP predictor data set.

Table 2.6. The overall importance the MP predictors within the SVM, RF and XGB models, correlations displayed in **Figure 2.3**.

Predictors	Average importance	SD
LogD	59.8	20.8
Molecular polar solvent accessible surface area	52.4	25.1
Molecular solubility	69.6	31.2
Molecular surface area	62.1	25.2
Number of aromatic rings	37.5	32.2
Number of hydrogen bond donors	30.8	26.6
Number of rings	54.4	44.3
Number of rotatable bonds	47.8	32.4

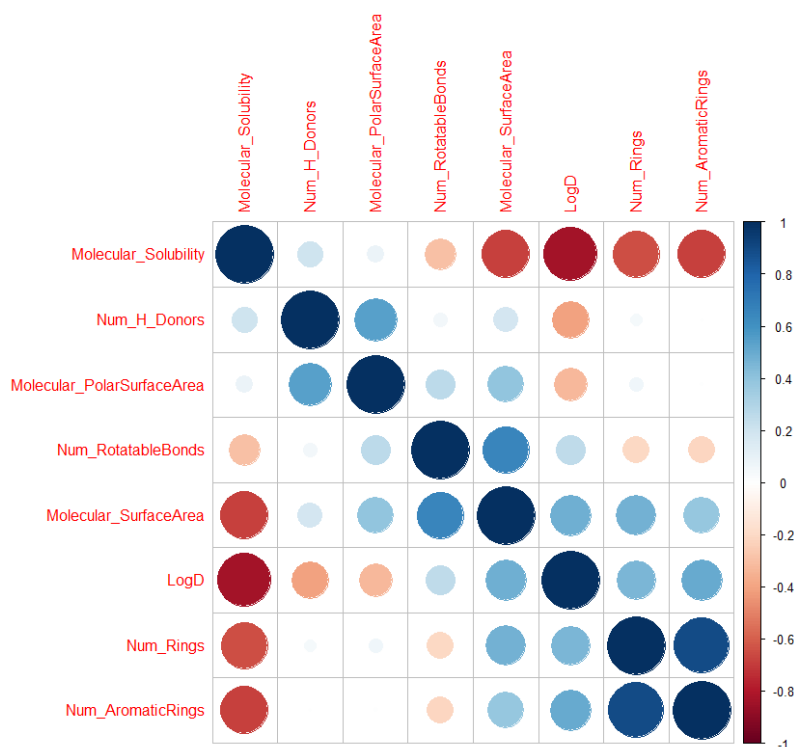


Figure 2.3. Correlation plot of the MP predictors. All correlations are between -0.9 and 0.9.

During the 10-fold cross validation, the ECFP predictors are seen to have a slightly better performance ($+0.01 \pm 0.13$) than the FCFP predictors in four out of six performance metrics on average when using the same algorithm, whether alone or in combination with the MP predictors. ECFP is a circular fingerprint which represent molecular structures by circular atom neighbourhoods, defined by the fingerprint diameter, while FCFP is a variation of ECFP which generalises the atoms of a molecule by their functional classes [30]. As the fingerprints are generated using the circular atom neighbourhoods, it is able to capture any novel substructures.

For 16 out of 32 models, FCFP is seen to have a slightly better performance ($+0.01 \pm 0.15$) than ECFC in four out of six performance metrics on average when using the same algorithm, whether alone or in combination with the MP predictors.

From the model performance (**Supporting Information 2.4** and **2.5**), variable importance (**Table 2.5**) and the above analysis, we can rate the predictor set as $MP < ECFP \leq FCFP$ for the amount of information they each contain that is crucial to mutagenicity prediction (**Table 2.7**). It was also noted that by combining the predictor sets, better performing models were constructed. Together with the fact that the variation 1 of the predictor sets results in models with slightly better performance, we can conclude that increased numbers of predictors allow the generation of models which perform better.

Table 2.7. Summary of predictor type performance

Predictor type	Number of models with Kappa > 0.7 in 10-fold cross validation	Number of models with Kappa > 0.45 on either or both validation sets	Overall importance ¹ ± SD
MP	24	21	51.8 ± 12.0
ECFP	30	31	21.6 ± 1.7
FCFP	28	32	21.8 ± 2.6

Colour code: green highest, orange middle, red lowest. ¹ sum of average importance (**Supporting Information 2.7**) divided by total number of predictors within each predictor type on models selected for y-randomisation validation

2.5.13. Comparison of different algorithms used for predicting mutagenicity

Out of the eight algorithms (PLSDA, MDA, SVM, KNN, C5, RF, GBM and XGB) used within this study, RF was seen to have the best performance with the test set across all data sets (ECFP, FCFP, MP, ECFP+FCFP, MP+ECFP, MP+FCFP, MP+ECFP+FCFP) when using the variation 1 of the predictor sets and in four out of seven data sets when using the variation 2 of the predictor sets (AUROC > 0.70, Sens > 0.65, Spec > 0.65, balanced accuracy > 0.65, Kappa > 0.45).

RF is a robust tree modelling algorithm used widely in statistical analysis and predictive modelling. RF is an ensemble of multiple classification trees, and due to this nature, it can be difficult to gain an understanding of the relationships between the predictors and the observations. However, it is possible to quantify the impact of the predictors within the ensemble on prediction using the improvement criteria aggregated across the ensemble [32]. As mentioned previously, due to the poor performance of the MP predictors alone, the models with MP predictors only were excluded from the analysis. Upon inspection of the variable importance of the nine best performing models (**Table 2.8** and **Supporting Information 2.7**), we can identify that one to five predictors with importance of over 70. It is clear that when MP predictors are provided for model construction (**Model 2.5 – 2.9**), molecular surface area, number of rings and logD are seen to have high importance in all cases, closely followed by molecular solubility. It is to note that in these models, the molecular fingerprints generally do not have importance over 70. Although the molecular fingerprints do not have high importance, as mentioned previously, the MP predictors alone do not give good models. Therefore, even though the importance is not high, the molecular fingerprints must hold some importance in the models.

Table 2.8. The variable importance of the selected best performing RF models

Model No.	Predictor data set	Predictor variation [#]	Important predictors
2.1	ECFP	1	ECFP bit: 925
2.2	ECFP	2	ECFP bit: 1069
2.3	FCFP	1	FCFP bits: 466, 870 and 1966
2.4	ECFP+FCFP	1	ECFP bit: 925
2.5	MP+ECFP	1	ECFP bit: 925, molecular solubility, molecular surface area, molecular polar surface area, number of rings and logD
2.6	MP+ECFP	2	Molecular solubility, molecular surface area, number of rings and logD
2.7	MP+FCFP	1	Molecular solubility, molecular surface area, molecular polar surface area, number of rings and logD
2.8	MP+ECFP+FCFP	1	Molecular solubility, molecular surface area, molecular polar surface area, number of rings and logD
2.9	MP+ECFP+FCFP	2	Molecular surface area, number of rings and logD

[#] See 2.5.3. *Data pre-processing* for detail

Aside from the best performing RF models, the performance of the other algorithms was looked at in an attempt to aid the understanding from the RF models. XGB was found to have the best performance with the test set across all data sets (ECFP, FCFP, ECFP+FCFP, MP+ECFP, MP+FCFP, MP+ECFP+FCFP) when using the variation 1 of the predictor sets and in three out of seven data sets when using the variation 2 of the predictor sets (AUROC > 0.65, Sens > 0.65, Spec > 0.60, balanced accuracy > 0.65, Kappa > 0.45). On the other hand, SVM was found to have the best performance in four of seven data sets when using the variation 2 of the predictor sets (AUROC > 0.75, Sens > 0.65, Spec > 0.75, balanced accuracy > 0.70, Kappa > 0.45). Although the performance of these models does not look as good as the 10-fold cross validation training data and are not to the desired level (e.g. AUROC > 0.9), they are still good models as all the metrics are closely matching and the Kappa signifies fair to moderate agreement.

SVM is an algorithm widely used for predictive modelling as it has shown to have great capability of fitting non-linear relationships within the pharmaceutical industry [42-44]. However, as the predictors within SVM models are transformed by the radial basis kernel function, it is the type of model where the analysis of predictor – observation relationship can only be achieved via the number of times a predictor appears in the model. Therefore, the use of SVM models to identify the direct predictor – observation relationship poses some difficulty. Nonetheless, upon close inspection of the variable importance of the four top performing SVM models (**Table 2.9** and **Supporting Information 2.7**), we can identify that 7 and 65 predictors have importance of over 70. In particular, number of rotatable bonds and molecular solubility have high importance in both of the models which used MP predictors (**Model 2.13** and **2.14**), whereas only >20% of the used fingerprint predictors in such models have an importance of over 70.

Table 2.9. The variable importance of the selected best performing SVM models

Model No.	Predictor data set	Predictor variation [#]	Important predictors
2.10	ECFP	2	7 ECFP bits
2.11	FCFP	2	10 FCFP bits
2.12	MP+FCFP	2	27 FCFP bits + molecular solubility + number of rotatable bonds
2.13	MP+ECFP+FCFP	2	36 ECFP bits + 27 FCFP bits + molecular solubility + number of rotatable bonds

[#] See 2.5.3. *Data pre-processing* for detail

XGB is a relatively new modelling algorithms which holds great potential in tackling machine learning problems [45]. As a tree-based algorithm, it would be possible to derive some rules for the underlying predictor – observation relationship; however, this is not a straightforward task. Upon inspection of the variable importance of the eight top performing XGB models (**Table 2.10** and **Supporting Information 2.7**), we can identify that number of rings and number of aromatic rings both have importance of 100 in two out of the 10 models (**Model 2.20** and **2.21**), whereas for molecular fingerprints, ECFP bit: 925 and FCFP bit: 870 have both made their appearance in four and three of the analysing models respectively, with importance over 70.

Table 2.10. The variable importance of the selected best performing XGB models

Model No.	Predictor data set	Predictor variation [#]	Important predictors
2.14	ECFP	1	ECFP bit: 925
2.15	FCFP	1	FCFP bit: 870, 1226 and 1966
2.16	ECFP+FCFP	1	ECFP bit: 925
2.17	ECFP+FCFP	2	FCFP bit: 870 and 2007
2.18	MP+ECFP	1	ECFP bit: 925 and number of rings
2.19	MP+ECFP	2	Number of rings
2.20	MP+FCFP	1	FCFP bit: 870 and number of aromatic rings
2.21	MP+ECFP+FCFP	1	ECFP bit: 925 and number of aromatic rings

[#] See 2.5.3. *Data pre-processing* for detail

Nonetheless, as the performance of the constructed XGB models do not match the desired values, especially when looking at Kappa ($Kappa > 0.6$), we decided not to attempt to examine further the predictor – observation relationship.

2.6. Comparison with Present Work

In addition to the model validation process, predictions of the Ames mutagenicity of the curated data was calculated using the commercial mutagenicity categorical model in StarDrop [46] and open source Toxicity Estimation Software Tool (TEST) [47] for comparison. The StarDrop mutagenicity categorical model is based on a range of decision tree algorithms which employs the C4.5 algorithm introduced by Quinlan [46, 48]. TEST predicts Ames mutagenicity via the weighted average of predictions from several different

cluster models and the estimation based on the three nearest chemical neighbours in the training set of TEST to the predicting molecule [47]. Both of these models were constructed using the same database [15]. A confusion matrix was created for the training set and test set individually.

In order to compare the performance of the top performing models, the StarDrop Ames mutagenicity category model and TEST, mutagenic predictions were made for the Class A mutagenic chemicals from the Ames/QSAR International Challenge Project [14]. The SMILES of the Class A mutagenic chemicals were extracted from the relevant PDF files, and any incomplete SMILES were removed as were molecules without a provided CAS number. The molecules were then processed and compiled as described in 2.1. *Data preparation* and 2.2. *Data compilation*. The mutagenicity of the molecules was predicted using the top performing models, the StarDrop Ames mutagenicity categorical model and TEST. As only mutagenic chemicals are available in this dataset (non-mutagens were not published), the performance of these models was compared using sensitivity only.

2.6.1. Performance comparison of top performing models with commercial product

When comparing the average Kappa of the selected classification models verified via y-randomisation and Z score with the Kappa from StarDrop and TEST (**Table 2.11**), it is clear that there is a clear drop in performance for the selected classification models between the training set and test set. The decrease in performance also existed for TEST, however this is not as prominent as the constructed classification models. On the other hand, StarDrop shows comparable performance across the training and test sets (Kappa difference = 0.07). The classification models are constructed using the training set and assessed using the test set, while 4657 of the 6512 molecules involved in the overall construction process of the StarDrop model and TEST were included in the training and test set. One note needed to be taken is that across the training and test sets, 202 and 102 molecules did not have a prediction (predict result = N/A).

Table 2.11. Kappa of the selected classification models and StarDrop Ames mutagenicity category model on the training and test set

Model	Training set	Test set
SVM ¹	0.76 ± 0.06	0.48 ± 0.02
RF ¹	0.98 ± 0.03	0.52 ± 0.04
XGB ¹	0.76 ± 0.02	0.49 ± 0.03
StarDrop	0.69	0.62
TEST	0.70 (0.62) ²	0.57 (0.41) ²

¹average Kappa (**Supporting Information 2.4** and **2.5**) of the selected classification models and their standard deviation

²Kappa within the bracket calculated for the unpredicted molecules counted as falsely predicted (i.e. unpredicted mutagens as false negative and unpredicted non-mutagens as false positive)

Therefore, the Class A mutagenic chemicals from the Ames/QSAR International Challenge Project [14] was used as a fairer comparison between the selected classification models, the StarDrop model and TEST. The Class A mutagenic chemicals were published much more recently and therefore are more unlikely to be involved in the construction process of the StarDrop model and TEST. After data curation and cross-checking against the molecules involved in model construction of our models, StarDrop's model and TEST, 508

mutagens were extracted from Ames/QSAR International Challenge Project [14] and used for comparison. From **Table 2.12**, we can note that our selected classification models are marginally better performing than both the StarDrop model and TEST. Here again, TEST failed to predict 71 molecules of the extracted Class A mutagenic chemical. However, we have to bear in mind that this only captures the ability of the models to predict mutagens correctly as no non-mutagens are involved in this analysis.

Table 2.12. Sensitivity of the selected classification models and StarDrop Ames mutagenicity category model on the Class A mutagenic chemicals from the Ames/QSAR International Challenge Project [14]

Model	Sens
SVM ¹	0.62 ± 0.01
RF ¹	0.62 ± 0.02
XGB ¹	0.61 ± 0.02
StarDrop	0.56
TEST	0.57 (0.49) ²

¹average sensitivity (**Supporting Information 2.4** and **2.5**) of the selected classification models and their standard deviation

²Sensitivity within the bracket calculated for the unpredicted molecules counted as falsely predicted (i.e. false negatives)

From **Table 2.11**, we can see that our constructed models have a better chance in correctly predicting an Ames mutagenic compound than the StarDrop model and TEST. This then brought our interest to if our models can correctly predict the mutagenicity of the Class A mutagenic chemical which one or zero models within the Ames/QSAR International Challenge Project predicted correctly [14]. Out of the 36 curated Class A mutagenic chemicals which were correctly predicted by one or fewer models within the Ames/QSAR International Challenge Project, our models can correctly predict the mutagenicity of 28 of them (**Supporting Information 2.8**). In comparison, the StarDrop model and TEST can only each correctly predict six compounds. Again, here TEST fails to produce prediction for six of the molecules. This further shows the capability of our models to correctly predict mutagens, possibly due to the different chemical space the training set covers in comparison to the StarDrop model and models within the Ames/QSAR International Challenge Project. However, as the specificity of models are also important due to the possibility of predicting a strong drug candidate incorrectly as mutagen, similar analysis using non-mutagens will be necessary.

2.7. Conclusion

Within this work, we attempted to build models with comparable performance to the models *Xu et al.* presented using a different range of descriptors and modelling algorithms. During the process, we discovered that the library *Xu et al.* presented in their work contained duplicates, and therefore we curated the library identifying 5395 unique molecules. As this resulting library is different to the library *Xu et al.* used for their work, comparison of model performance against their work was not carried out. After constructing 112 models using eight different algorithms, we discovered SVM, RF and XGB to have the best performance. The RF models had the best performance across most data sets during the 10-fold cross validation training (AUROC > 0.95, Sens > 0.90, Spec > 0.95, balanced accuracy > 0.95, Kappa > 0.90) and on the test set (AUROC > 0.75, Sens > 0.75, Spec > 0.65, balanced accuracy > 0.70, Kappa > 0.40); the SVM and XGB models

CHAPTER 2: PREDICTING AMES MUTAGENICITY

had good performance during the 10-fold cross validation (AUROC > 0.90, Sens > 0.85, Spec > 0.75, balanced accuracy > 0.80, Kappa > 0.65) and on the test set (AUROC > 0.65, Sens > 0.65, Spec > 0.60, balanced accuracy > 0.65, Kappa > 0.30).

With the RF, SVM and XGB models, we discovered that the MP descriptors showed the highest importance when used in combination with molecular fingerprints. Such descriptors are logD, molecular solubility, molecular surface area, number of aromatic rings, number of rings and number of rotatable bonds, with their importance differing with different modelling algorithms.

By comparing the performance of our top performing models against the StarDrop Ames mutagenicity prediction model and TEST using the Class A mutagenic compounds from the Ames/QSAR International Challenge Project, we found our models to be better performing in predicting mutagens correctly. We also discovered that our models were able to predict some of the Class A mutagenic compounds where one or less models were able to predict correctly in the Ames/QSAR International Challenge Project. Attempts to improve the performance and robustness of the models would involve searching for more experimental Ames data and extend the variation of modelling algorithms used, as well as the range of molecular descriptors calculated.

2.8. References

1. Custer, L. L.; Sweder, K. S. The Role of Genetic Toxicology in Drug Discovery and Optimization. *Curr. Drug Metab.* **2008**, *9* (9), 978-985, Review. DOI: 10.2174/138920008786485191.
2. Kramer, J. A.; Sagartz, J. E.; Morris, D. L. The application of discovery toxicology and pathology towards the design of safer pharmaceutical lead candidates. *Nature Reviews Drug Discovery* **2007**, *6* (8), 636-649. DOI: 10.1038/nrd2378 From NLM.
3. Kazius, J.; McGuire, R.; Bursi, R. Derivation and validation of toxicophores for mutagenicity prediction. *Journal of medicinal chemistry* **2005**, *48* (1), 312-320. DOI: 10.1021/jm040835a From NLM.
4. White, A. C.; Mueller, R. A.; Gallavan, R. H.; Aaron, S.; Wilson, A. G. E. A multiple in silico program approach for the prediction of mutagenicity from chemical structure. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis* **2003**, *539* (1-2), 77-89, Article. DOI: 10.1016/s1383-5718(03)00135-9.
5. Xu, C.; Cheng, F.; Chen, L.; Du, Z.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. In silico Prediction of Chemical Ames Mutagenicity. *Journal of chemical information and modeling* **2012**, *52* (11), 2840-2847. DOI: 10.1021/ci300400a.
6. Ames, B. N.; Durston, W. E.; Yamasaki, E.; Lee, F. D. Carcinogens are Mutagens: A Simple Test System Combining Liver Homogenates for Activation and Bacteria for Detection. *Proceedings of the National Academy of Sciences of the United States of America* **1973**, *70* (8), 2281-2285. PMC.
7. Ames, B. N.; Gurney, E. G.; Miller, J. A.; Bartsch, H. Carcinogens as Frameshift Mutagens: Metabolites and Derivatives of 2-Acetylaminofluorene and Other Aromatic Amine Carcinogens. *Proceedings of the National Academy of Sciences* **1972**, *69* (11), 3128-3132.
8. Ames, B. N.; Lee, F. D.; Durston, W. E. An Improved Bacterial Test System for the Detection and Classification of Mutagens and Carcinogens. *Proceedings of the National Academy of Sciences of the United States of America* **1973**, *70* (3), 782-786. PMC.
9. McCann, J.; Spingarn, N. E.; Kobori, J.; Ames, B. N. Detection of carcinogens as mutagens: bacterial tester strains with R factor plasmids. *Proceedings of the National Academy of Sciences of the United States of America* **1975**, *72* (3), 979-983. PMC.
10. Mortelmans, K.; Zeiger, E. The Ames Salmonella/microsome mutagenicity assay. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **2000**, *455* (1-2), 29-60. From NLM.
11. Walmsley, R. M.; Billinton, N. How accurate is in vitro prediction of carcinogenicity? *British Journal of Pharmacology* **2011**, *162* (6), 1250-1258. DOI: 10.1111/j.1476-5381.2010.01131.x PMC.
12. Benigni, R.; Giuliani, A. Computer-assisted analysis of interlaboratory Ames test variability. *Journal of Toxicology and Environmental Health* **1988**, *25* (1), 135-148. DOI: 10.1080/15287398809531194.
13. Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Kovalishyn, V. V.; Prokopenko, V. V.; Tetko, I. V. Applicability domain for in silico models to achieve accuracy of experimental measurements. *Journal of Chemometrics* **2010**, *24* (3-4), 202-208. DOI: 10.1002/cem.1296.
14. Honma, M.; Kitazawa, A.; Cayley, A.; Williams, R. V.; Barber, C.; Hanser, T.; Saiakhov, R.; Chakravarti, S.; Myatt, G. J.; Cross, K. P.; et al. Improvement of quantitative structure-activity relationship (QSAR) tools for predicting Ames mutagenicity: outcomes of the Ames/QSAR International Challenge Project. *Mutagenesis* **2019**, *34* (1), 3-16. DOI: 10.1093/mutage/gey031 PubMed.

15. Hansen, K.; Mika, S.; Schroeter, T.; Sutter, A.; ter Laak, A.; Steger-Hartmann, T.; Heinrich, N.; Müller, K.-R. Benchmark Data Set for in Silico Prediction of Ames Mutagenicity. *Journal of chemical information and modeling* **2009**, *49* (9), 2077-2081. DOI: 10.1021/ci900161g.
16. Votano, J. R.; Parham, M.; Hall, L. H.; Kier, L. B.; Oloff, S.; Tropsha, A.; Xie, Q.; Tong, W. Three new consensus QSAR models for the prediction of Ames genotoxicity. *Mutagenesis* **2004**, *19* (5), 365-377. DOI: 10.1093/mutage/geh043 From NLM.
17. Ferrari, T.; Gini, G. An open source multistep model to predict mutagenicity from statistical analysis and relevant structural alerts. *Chemistry Central journal* **2010**, *4 Suppl 1*, S2. DOI: 10.1186/1752-153x-4-s1-s2 From NLM.
18. Zheng, M.; Liu, Z.; Xue, C.; Zhu, W.; Chen, K.; Luo, X.; Jiang, H. Mutagenic probability estimation of chemical compounds by a novel molecular electrophilicity vector and support vector machine. *Bioinformatics* **2006**, *22* (17), 2099-2106. DOI: 10.1093/bioinformatics/btl352 From NLM.
19. Hillebrecht, A.; Muster, W.; Brigo, A.; Kansy, M.; Weiser, T.; Singer, T. Comparative evaluation of in silico systems for ames test mutagenicity prediction: scope and limitations. *Chemical Research in Toxicology* **2011**, *24* (6), 843-854. DOI: 10.1021/tx2000398 From NLM.
20. Kazius, J.; McGuire, R.; Bursi, R. Derivation and Validation of Toxicophores for Mutagenicity Prediction. *Journal of Medicinal Chemistry* **2005**, *48* (1), 312-320. DOI: 10.1021/jm040835a.
21. Ferrari, T.; Gini, G. An open source multistep model to predict mutagenicity from statistical analysis and relevant structural alerts. *Chemistry Central Journal* **2010**, *4* (1), S2, journal article. DOI: 10.1186/1752-153x-4-s1-s2.
22. Zheng, M.; Liu, Z.; Xue, C.; Zhu, W.; Chen, K.; Luo, X.; Jiang, H. Mutagenic probability estimation of chemical compounds by a novel molecular electrophilicity vector and support vector machine. *Bioinformatics* **2006**, *22* (17), 2099-2106. DOI: 10.1093/bioinformatics/btl352.
23. Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *Journal of Chemical Information and Computer Sciences* **1998**, *38* (6), 983-996. DOI: 10.1021/ci9800211.
24. Ewing, T.; Baber, J. C.; Feher, M. Novel 2D fingerprints for ligand-based virtual screening. *Journal of chemical information and modeling* **2006**, *46* (6), 2423-2431. DOI: 10.1021/ci060155b From NLM.
25. Chen, L.; Li, Y.; Zhao, Q.; Peng, H.; Hou, T. ADME Evaluation in Drug Discovery. 10. Predictions of P-Glycoprotein Inhibitors Using Recursive Partitioning and Naive Bayesian Classification Techniques. *Molecular Pharmaceutics* **2011**, *8* (3), 889-900. DOI: 10.1021/mp100465q.
26. Wang, S.; Li, Y.; Wang, J.; Chen, L.; Zhang, L.; Yu, H.; Hou, T. ADMET Evaluation in Drug Discovery. 12. Development of Binary Classification Models for Prediction of hERG Potassium Channel Blockage. *Molecular Pharmaceutics* **2012**, *9* (4), 996-1010. DOI: 10.1021/mp300023x.
27. *Chemistry Collection: Basic Chemistry User Guide, Pipeline Pilot Release 16.5.0.143*; Accelrys Software Inc.: San Diego, 2016. (accessed 2018).
28. Tropsha, A. *Best Practices for QSAR Model Development, Validation, and Exploitation*; 2010. DOI: 10.1002/minf.201000061.
29. Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *Journal of Cheminformatics* **2015**, *7* (1), 23, journal article. DOI: 10.1186/s13321-015-0068-4.

30. Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Journal of chemical information and modeling* **2010**, *50* (5), 742-754, Article. DOI: 10.1021/ci100050t.
31. *Pipeline Pilot Statistics and Data Modelling Collection 2016, Pipeline Pilot Release 16.5.0.143*; Accelrys Software Inc.: San Diego, 2016. (accessed 2018).
32. Kuhn, M.; Johnson, K. *Applied predictive modeling*; Springer, 2013.
33. Brodersen, K. H.; Ong, C. S.; Stephan, K. E.; Buhmann, J. M. The Balanced Accuracy and Its Posterior Distribution. In *2010 20th International Conference on Pattern Recognition*, 23-26 Aug. 2010, 2010; pp 3121-3124. DOI: 10.1109/ICPR.2010.764.
34. Hallinan, J. *Assessing and Comparing Classifier Performance with ROC Curves*. 2014. <http://machinelearningmastery.com/assessing-comparing-classifier-performance-roc-curves-2/> (accessed 07/07/2017).
35. Carpenter, B. *High Kappa Values are not Necessary for High Quality Corpora*. 2012. <https://lingpipe-blog.com/2012/10/02/high-kappa-not-necessary-high-quality-corpora/> (accessed 07/07/2017).
36. Eugenio, B. D.; Glass, M. The Kappa Statistic: A Second Look. *Computational Linguistics* **2004**, *30* (1), 95-101. DOI: 10.1162/089120104773633402.
37. Chicco, D.; Warrens, M. J.; Jurman, G. The Matthews Correlation Coefficient (MCC) is More Informative Than Cohen's Kappa and Brier Score in Binary Classification Assessment. *IEEE Access* **2021**, *9*, 78368-78381. DOI: 10.1109/ACCESS.2021.3084050.
38. Luo, M.; Wang, X. S.; Roth, B. L.; Golbraikh, A.; Tropsha, A. Application of Quantitative Structure–Activity Relationship Models of 5-HT_{1A} Receptor Binding to Virtual Screening Identifies Novel and Potent 5-HT_{1A} Ligands. *Journal of chemical information and modeling* **2014**, *54* (2), 634-647. DOI: 10.1021/ci400460q.
39. Rodgers, A. D.; Zhu, H.; Fourches, D.; Rusyn, I.; Tropsha, A. Modeling Liver-Related Adverse Effects of Drugs Using kNearest Neighbor Quantitative Structure–Activity Relationship Method. *Chemical Research in Toxicology* **2010**, *23* (4), 724-732. DOI: 10.1021/tx900451r.
40. 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 IPII of original article: S0169-409X(96)00423-1. The article was originally published in *Advanced Drug Delivery Reviews* *23* (1997) 3–25.1. *Advanced Drug Delivery Reviews* **2001**, *46* (1), 3-26. DOI: [https://doi.org/10.1016/S0169-409X\(00\)00129-0](https://doi.org/10.1016/S0169-409X(00)00129-0).
41. Lipinski, C. A. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies* **2004**, *1* (4), 337-341. DOI: <https://doi.org/10.1016/j.ddtec.2004.11.007>.
42. Vasanthanathan, P.; Taboureau, O.; Oostenbrink, C.; Vermeulen, N. P. E.; Olsen, L.; Jorgensen, F. S. Classification of Cytochrome P450 1A2 Inhibitors and Noninhibitors by Machine Learning Techniques. *Drug Metab. Dispos.* **2009**, *37* (3), 658-664, Article. DOI: 10.1124/dmd.108.023507.
43. Zhao, C. Y.; Zhang, H. X.; Zhang, X. Y.; Zhang, R. S.; Luan, F.; Liu, M. C.; Hu, Z. D.; Fan, B. T. Prediction of milk/plasma drug concentration (MIP) ratio using support vector machine (SVM) method. *Pharm. Res.* **2006**, *23* (1), 41-48, Article. DOI: 10.1007/s11095-005-8716-4.
44. Zhou, S.; Li, G. B.; Huang, L. Y.; Xie, H. Z.; Zhao, Y. L.; Chen, Y. Z.; Li, L. L.; Yang, S. Y. A prediction model of drug-induced ototoxicity developed by an optimal support vector machine (SVM) method. *Comput. Biol. Med.* **2014**, *51*, 122-127, Article. DOI: 10.1016/j.combiomed.2014.05.005.

CHAPTER 2: PREDICTING AMES MUTAGENICITY

45. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA; 2016.
46. *StarDrop*; Optibrium Ltd: United Kingdom, 2011. (accessed 01/10/2021).
47. *User's Guide for T.E.S.T. (Toxicity Estimation Software Tool)*; U.S. Environmental Protection Agency: United States, 2020. (accessed 01/10/2021).
48. *Toxicity Prediction in StarDrop*; Optibrium Ltd: United Kingdom, 2011. (accessed 01/10/2021).

Chapter 3:
Understanding Polymer Detergent Properties via
QSPR method

3.1. Detergent Properties and Polymers as Surfactants

A detergent is a surfactant or a mixture of surfactants with cleaning properties in dilute aqueous solutions [1]. One of the keys to the cleaning properties of detergents is their ability to lower the interfacial tension between different phases, such that oil and grease can be removed from surfaces [2]. The ability of a detergent to lower the interfacial tension is often partly gauged by its critical micelle concentration (CMC). The critical micelle concentration describes the concentration where the surfactant molecules aggregate to form clusters called micelles which have their hydrophobic groups clustering towards the oil phase and the hydrophilic groups pointing outwards towards the aqueous phase [3]. When the phase changes from monomeric surfactant solution to micellar solution, this invokes a sharp change of physical properties [4]. This sharp change can be measured via various different method (e.g. surface tension, conductivity), and the suitable method of measurement is dependent on the nature of surfactant investigated [4]. The CMC is dependent on the experimental conditions, such as temperature and the equipment used, as these can affect the physical properties of the solution containing the surfactant molecules [5]. In addition, although the discontinuity of the physical properties is sharp, the phase change from monomeric surfactant solution to micellar solution occurs over a small range of concentration, and therefore is difficult to obtain a single precise value.

Although CMC is the key scientific parameter scientists look at when trying to understand the cleaning properties of detergents, for commercial detergent products, developers would often look at other properties, such as the viscosity and result of cleaning tests, when developing new liquid detergent formulas. This is because consumers would often pay attention to how viscous the product is when they select them from the shelf and how well they can remove stains. Therefore, construction of quantitative structure-property relationships (QSPRs) to aid understanding of the properties of the detergent formulas that affects viscosity and cleaning results is important for developers.

Traditionally, surfactants commonly refer to molecules with an ionic hydrophilic head and a carbon chain hydrophobic tail. This is the structure of surfactant that formed bar soap in the early days (**Figure 3.1**) [1]. However, as research advances, surfactants also refers to any molecule with hydrophobic and hydrophilic section which can sit at the interface between difference phases to lower surface tension [1]. This includes non-ionic molecules with sections distinctively more hydrophilic then the rest of the molecule which can arrange themselves in solution to form micelles.

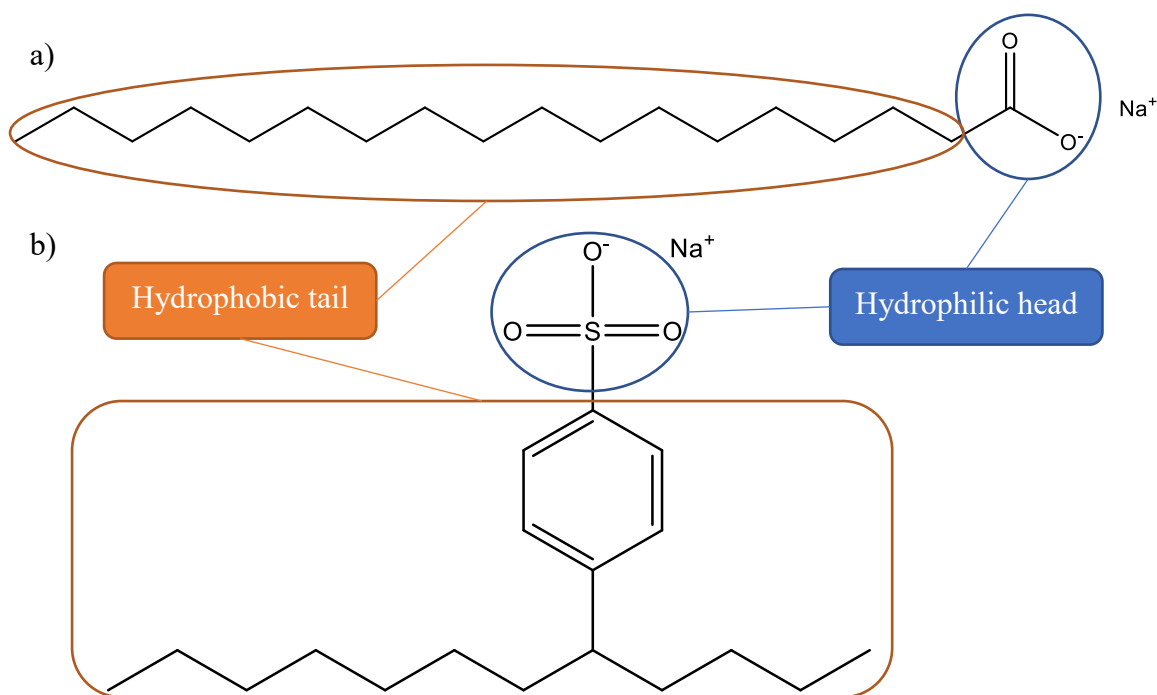


Figure 3.1. Example of common surfactant a) sodium stearate and b) 4-(5-dodecyl) benzenesulfonate with their hydrophilic head and hydrophobic tail labelled.

Polymers are macromolecules which are composed of small repeated units, monomers, and copolymers are polymers formed of more than one species of repeated unit [1]. Block copolymers are copolymers with each species of the repeated units in distinct blocks [1] (**Figure 3.2a**). When these blocks have distinctively different hydrophilicity and their relative size and position allows them to form micelles at its CMC, the polymer can be defined as a surfactant [6, 7] (**Figure 3.2b**). Theoretically the relative hydrophilicity and size of the junction blocks would determine the CMC of the polymer and therefore its ability to remove stains; the total length of the polymer chains would increase the viscosity of the detergent formula it is present in as they can tangle with each other.

CHAPTER 3: UNDERSTANDING POLYMER DETERGENT PROPERTIES VIA QSPR METHOD

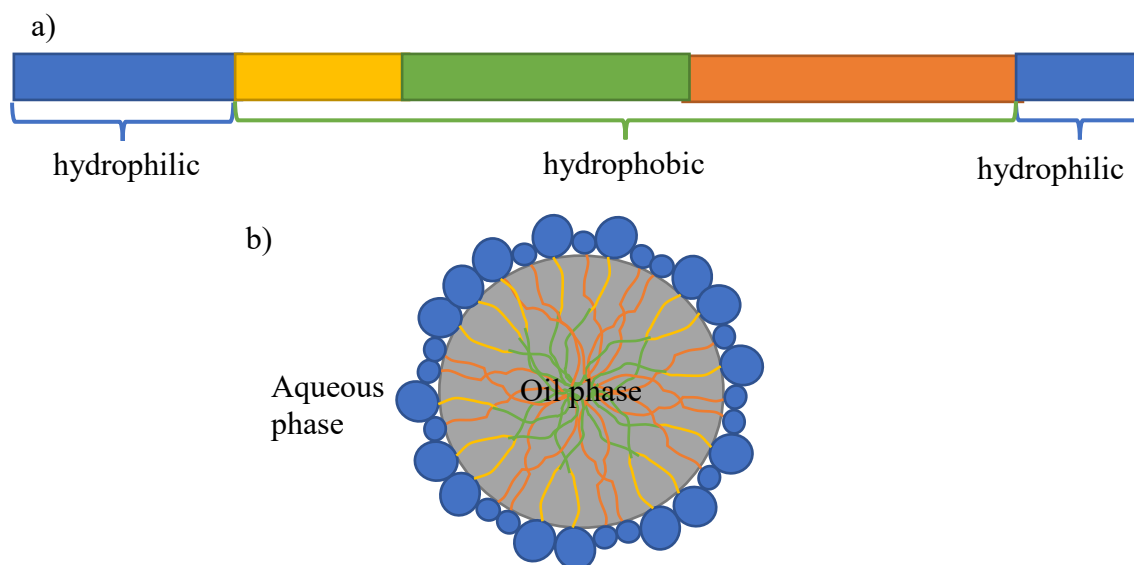


Figure 3.2. Example of a) block polymer and b) how block polymer can form micelles.

3.2. Polymer Data

Although it is possible to include polymers with surfactant properties in detergent formulas, QSPR with polymer data is an ongoing challenge. First, the composition of the polymer is not always certain [8, 9]. With block copolymers, they are usually formed via living polymerisation techniques, such that the block of the polymer would grow as long as the conditions allow. However, as multiple blocks would form at the same time, it is not possible to fully control all forming polymer chains' length, but instead indicate the range of length of the blocks, i.e. there is an unknown distribution of polymer lengths typically. Another challenging point of polymer data is the molecular structures. Currently companies would each store their own polymer data with their inhouse database using their own methods and there is no one universal structure to how one should structure their database. Therefore, different databases would have different construction to each other and may contain different properties. This makes usage and analysis of polymer data from different sources require extra steps to unify and curate the data. This may not be such a problem when dealing with a single database, however, if one was to apply previously constructed QSPR models derived from one database on another, difficulties may arise.

3.2.1. Data used

The data collection used in this project is a polymer surfactant library (PSL) provided by Dr Jerry Winter, Unilever. The members of the PSL are all star-block polymers based on chemically similar cores and various chemically similar arms of different sizes, tipped with different chemical functionalities (**Figure 3.3**). This data collection is anonymised as it is part of a proprietary technology development programme. Therefore, that no attempt is made in this chapter to explain the structure property relationships in terms that could be used to infer the proprietary chemistries involved, but it is the purpose to show that useful models can be created using descriptors derived from the chemical structure of a real industrially relevant data set. A limited set of chemical characteristics and properties of the PSL members have been derived from the structures and are listed in the first column in **Supporting Information 3.1**. The observation of this data collection is measured on formulations in which each member of a PSL is added to one of two "base formulations" (BF1 and BF2) containing other typical cleaning product ingredients in water. The two base

CHAPTER 3: UNDERSTANDING POLYMER DETERGENT PROPERTIES VIA QSPR METHOD

formulations vary in composition, but both contain a 10% water hole for adding water and PSL sample. Each of the formulations were evaluated for their properties as via one or more of the three possible tests, each providing a different endpoint:

- The viscosity (measured in mPa.s) at the shear rate of about 21 s^{-1} experienced by the fluid while being poured (Vis)
- A reflectance change (measured in ΔE) measure of the amount of red pottery clay removed from a “white” polyester fabric in a test wash (RPC)
- A reflectance change (measured in ΔE) measure of the amount of yellow pottery clay removed from a “white” polyester fabric in a test wash (YPC)

Within the data collection, various calculated properties of the PSL members were provided as descriptors for the formulations (**Supporting Information 3.1**).

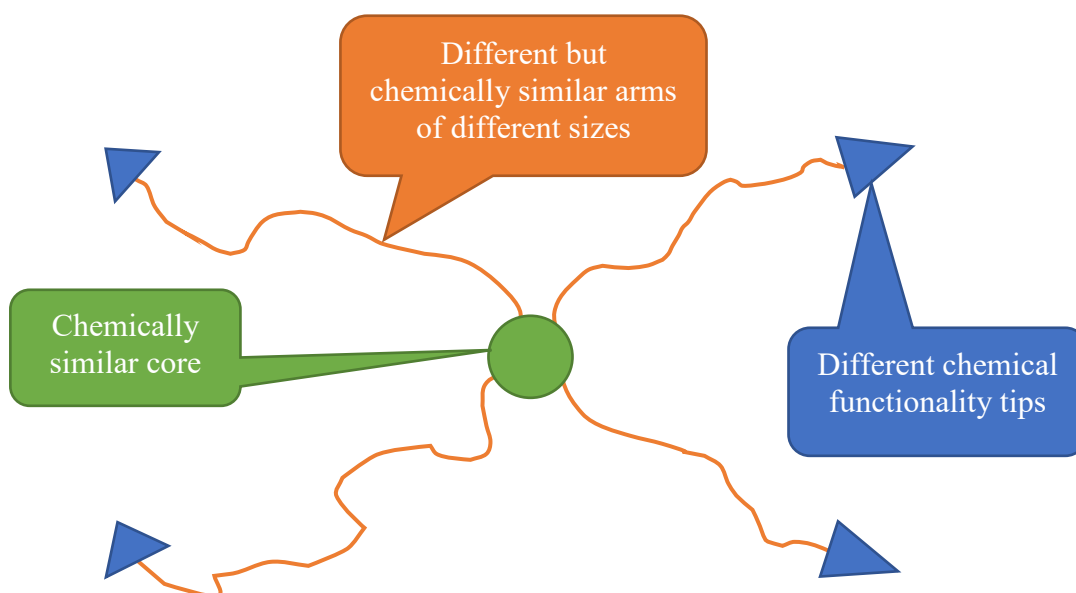
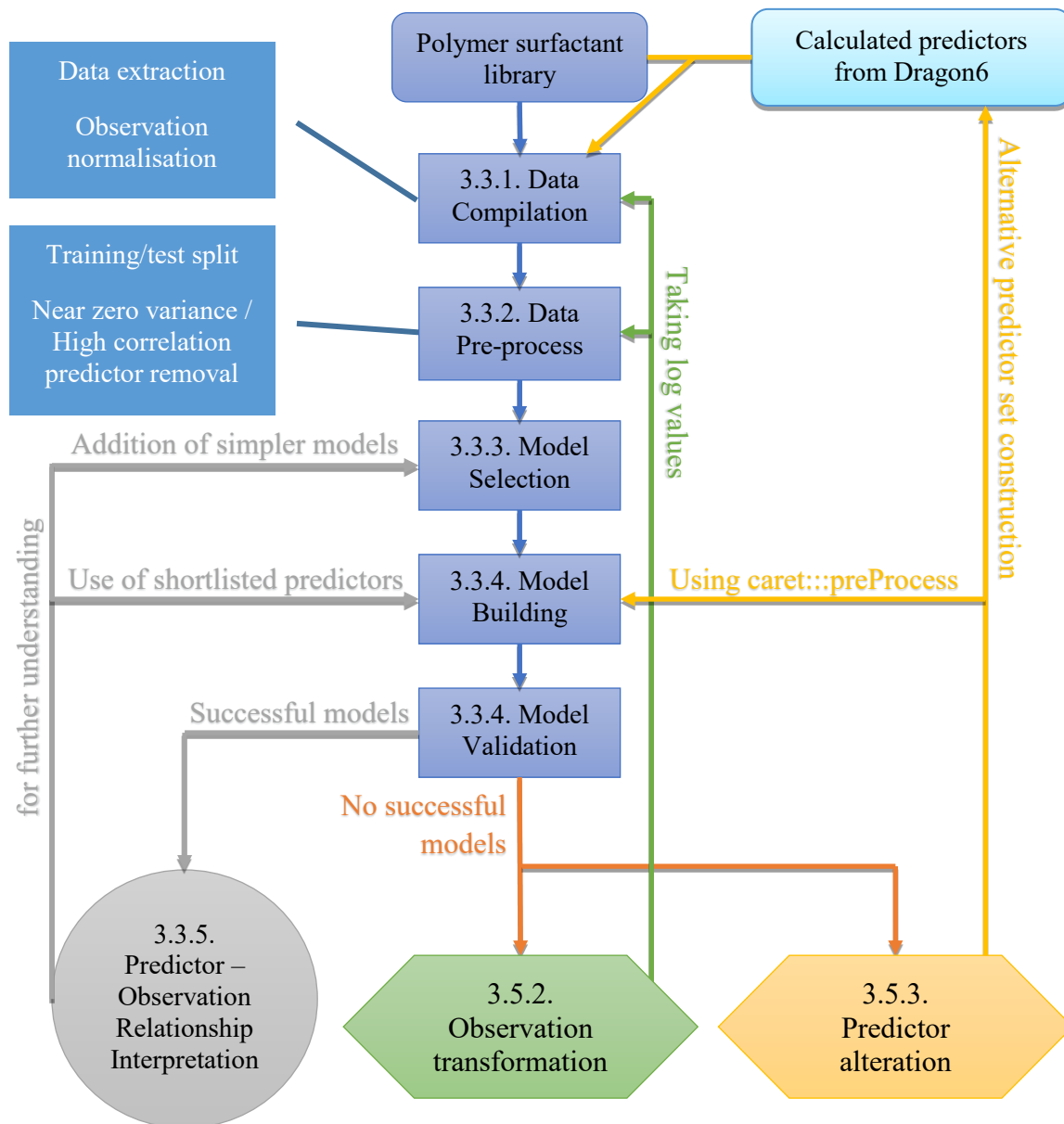


Figure 3.3. Illustration displaying the generic structures of the members of the PSL, including the points of variation for the structures.

CHAPTER 3: UNDERSTANDING POLYMER DETERGENT PROPERTIES VIA QSPR METHOD

3.3. Constructing of Detergent Properties QSPR Models

Scheme 3.1 displays the overall workflow for developing and testing QSPR models.



Scheme 3.1. A flow chart of the process of achieving a successful model. Rounded rectangles: raw data; rectangles: processes; circle: interpretation; hexagon: solutions ideas. Each step is further described in 3.3.1. *Data compilation*, 3.3.2. *Data pre-processing*, 3.1.1. *Initial algorithm selection*, 3.3.4. *Model building and performance assessment*, 3.1.1. *Predictor – observation relationship interpretation*, 3.5.2. *Models with transformed observation values* and 3.5.3. *Models with altered predictors*.

3.3.1. Data compilation

Each of the spreadsheets of the PSL excel file containing the descriptors for the PSL members, formulation compositions and test results were exported as .csv files, which were then imported into R [10] as individual data frames.

CHAPTER 3: UNDERSTANDING POLYMER DETERGENT PROPERTIES VIA QSPR METHOD

For the data frame from the test result spreadsheet, the observations were normalised against their respective controls unique to the particular batch of experiments. The batch control normalised observations were then normalised again against their respective controls for the specific series of formulation and end point. For all data frames, the rows with no observations were removed and only columns with essential information for data compilation and model building are retained.

Next, data regarding to the formulation involving each base formulation (BF1 and BF2) were extracted from the formulation composition spreadsheets. Information regarding the percentage composition of the base formulation and water within the formulation composition data frame was removed and reformatted to accommodate an extra column for the presence of a second polymer ingredient. The descriptors for the polymer used in each formulation was extracted from the PSL spreadsheet.

The data frames for each combination of base formulation and endpoint (Vis, RPC and YPC) were then merged to give one data frame using the formulation ID. The descriptors of the polymer involved in the relevant formulation was then merged with the newly formed data frames. Once the data frames are composed, all columns other than the observations and the descriptors listed in the first column of **Supporting Information 3.1** were removed to leave only the essential information required for model construction. At this stage, a data frame for each of the data sets, BF1_Vis, BF2_Vis, BF2_RPC and BF2_YPC, was formed with 83, 198, 170 and 76 observations in them respectively (**Table 3.1**).

Table 3.1. The number of observations for each data set

End Point [#]	Base formulation	
	BF1	BF2
Vis	83	198
RPC	-	170
YPC	-	76

[#] see 3.2.1. Data sets for detail

3.3.2. Data pre-processing

For each data set, a training/test split was carried out with a 9:1 ratio using `caret::createDataPartition` to ensure the resulting training set and test set contain the correct ratio of observation values from each percentile based section of the overall observation values [11].

Near zero variance predictors were removed to reduce the likelihood of error induced by the lack of information of these predictors by the following rules:

- Variation 1: the full set of filtered predictors where predictors with frequency ratio above $\#observation/5$ were removed (**Supporting Information 3.1**).
- Variation 2: the reduced set of filtered predictors where predictors were removed following the below criteria
 - Predictors with frequency ratio above $\#observation/25$ for total $\#observation < 100$ or
 - Predictors with frequency ratio above $\#observation/50$ for total $\#observation$ between 100-1000 (**Supporting Information 3.1**).

CHAPTER 3: UNDERSTANDING POLYMER DETERGENT PROPERTIES VIA QSPR METHOD

Correlation plots of the predictors were also examined. However, due to the nature of the predictors, if the highly correlated predictors were removed, only one predictor would be left. Therefore, no procedures were carried out to reduce the correlation between predictors.

Any factor predictors were manually transformed into integers to avoid errors arising during model building.

3.3.3. Initial algorithm selection

Initially four non-linear model algorithms (SVM [12, 13], KNN [13, 14], nnet and avNNet [13]) and three decision tree/rule-based model algorithms (RF [13, 15], cubist [13] and XGB [16]) were selected to investigate their performance on the four data sets, covering the range of simple yet interpretable (e.g. KNN) to robust but black-box like models (e.g. cubist) for discovering the optimal performing model that can be interpreted to uncover some of the predictor – observation relationships. Each of the algorithms contain different tuning parameters as described in **Table 3.2**.

Table 3.2. Tuning parameters for the selected algorithms

Algorithm	Tuning parameter
SVM	ϵ cost
KNN	K
Nnet	Decay Size
avNNet	Decay Size Bag
RF	m_{try}
Cubist	Committees Neighbours
XGB	N rounds Max depth Eta Gamma Col sample by tree Minimum child weight

3.3.4. Model building and performance assessment

Model construction using `caret::train()` with the default tuning parameters and 10-fold cross validation on the training set for each data set (see 3.3.1. *Data compilation*) proceeded. For PLSDA, SVM and KNN, the predictors were centred and scaled within `caret::train()` using the `preProcess` option. The constructed models were then tested using the test set.

CHAPTER 3: UNDERSTANDING POLYMER DETERGENT PROPERTIES VIA QSPR METHOD

Commonly, when assessing the performance of a regression model, the following metrics are assessed:

- Correlation coefficient (R^2): measures how well the predictions match the observations
- Root mean square error (RMSE): standard deviation of the prediction error
- Gradient of regression line (k)

Normally, these metrics are calculated by prediction vs. observation. Variant of the metrics includes the inverse metrics (denoted by dash) calculated by observation vs prediction and through the origin metrics (denoted by subscript 0).

When the performance of a model is being assessed for its validity, Golbraikh and Tropsha originally suggested the following criteria which need to be all fulfilled for a regression model to be deemed acceptable [17]:

- Cross-validated R^2 via internal resampling on training set > 0.5
- R^2 on test set > 0.6
- R^2 through origin (R_0^2) close to R^2
- $\frac{R^2 - R_0^2}{R^2} < 0.1$ or $\frac{R^2 - R_0'^2}{R^2} < 0.1$
- And the corresponding $0.85 \leq k \leq 1.15$ or $0.85 \leq k' \leq 1.15$ (See 1.2.7.1. Regression performance metrics for full definition)

They emphasised the predictive ability of a model can only be estimated using an external test set that was not used for building the model, and more over that both the normal R^2 and the R_0^2 must have similar values for the model to have a high predictive ability. However more recently Alexander, Tropsha and Winkler published a paper which emphasised the importance of RMSE and suggested that for a predictive model, the following criteria needs to be fulfilled [18]:

- High R^2 on test set
- Low RMSE of test set predictions

Generally, R^2 on training set is higher than R^2 on test set as the data is seen by the model during construction. However, in the cases where R^2 on training set is calculated based on the model during the cross-validation stage, it is possible that the R^2 on training set is lower than the R^2 on test set as the hold-out data during the cross-validation stage is not seen by the model during construction.

In addition, the Z score had been proposed by a number of QSPR papers [19, 20] as a measurement of robustness of models. The Z score compares the performance of the models to multiple repeats of models with their observation randomly shuffled before training. In this project, ten repeats of the y-randomised models were carried out and the Z score was calculated following Equation 3.1 [21].

$$Z = \frac{R^2_{original\ training} - Average(R^2_{y-randomised\ training})}{Standard\ deviation(R^2_{y-randomised\ training})} \quad (\text{Equation 3.1})$$

CHAPTER 3: UNDERSTANDING POLYMER DETERGENT PROPERTIES VIA QSPR METHOD

If the original model was valid, the overall performance of y-randomised models should be greatly reduced in comparison, resulting in a high Z score. Models with Z scores of over 3 are considered as significant [21].

CHAPTER 3: UNDERSTANDING POLYMER DETERGENT PROPERTIES VIA QSPR METHOD

Taking in account of both literatures and Z score, the follow modified criteria for validity was used to assess the successes of the tuned models in terms of their predictability and understanding the trend within the data:

- On training set
 - Cross-validated $R^2 > 0.5$
 - $RMSE < 0.5$
- On test set
 - (Adjusted) $R^2 > 0.6$
 - (Adjusted) R_0^2 close to R^2
 - $\left| \frac{R^2 - R_0^2}{R^2} \right| < 0.1$
 - $RMSE < 0.35$
 - Slope of R_0^2 regression line: $0.85 \geq k \geq 1.15$
 - Z score > 3

After analysis of the models constructed, models fulfilling all criteria were tuned with an expanded set of tuning parameters as shown in **Table 3.3** to refine the models.

Table 3.3. The difference between the defaults set and expanded set of tuning parameters used

Algorithm	Tuning parameter	Default set	Expanded set
Nnet	Decay	0, 0.0001, 0.1	0, 0.00001, 0.0001, 0.001, 0.01, 0.1
	Size	1, 3, 5	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
avNNet	Decay	0, 0.0001, 0.1	0, 0.00001, 0.0001, 0.001, 0.01, 0.1
	Size	1, 3, 5	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
	Bag	FALSE	FALSE
Cubist	Committees	1, 10, 20	1, 5, 10, 20, 50, 100
	Neighbours	0, 5, 9	0, 1, 3, 5, 7, 9

3.3.5. Predictor – observation relationship interpretation

As all the selected algorithms are non-linear or tree like models, it is difficult to interpret the correlation between the predictors and the observation, i.e. does the predictor have a positive or negative effect on the observation. Therefore, two linear model algorithms (LM and PLS) were used as a simple tool on data sets for which successful models had been found after tuning, in an attempt to aid the understanding of the predictor – observation relationships by looking for trends of a linear relationship. In addition, if these linear models do not perform well, it shows the necessity to use the more complex model algorithms as the observation cannot be explained by simple models.

Both LM and PLS were used with the in the same fashion as the other model algorithms in 3.3.3. *Initial algorithm selection.* For LM, the predictors were centred, scaled and principle components generated `caret::train()` using the `preProcess` option due to the high correlation between the predictors. The predictors for PLS were centred and scaled within `caret::train()`

CHAPTER 3: UNDERSTANDING POLYMER DETERGENT PROPERTIES VIA QSPR METHOD

using the `preProcess` option and the number of components to retain is tuned by extending the maximum number of components to retain to the number of predictors available to find the optimum model.

As the linear models did not perform well to provide enough insight to fully discover the predictor – observation relationship, refinement of the models by shortlisting the predictors was carried out in attempt to refine any potential relationships already noted. Predictors were shortlisted using the best performing model's variable importance by only including predictors that took part in the model, i.e. by removing predictors with zero importance. The shortlisted predictors were then subject to model building and validation process all model algorithms used previously, i.e. LM, PLS, SVM, KNN, RF, cubist and XGB.

3.4. Constructed models for BF1 data set

14 models were constructed using SVM, KNN, `nnet`, `avNNet`, RF, cubist and XGB initially for the BF1 dataset containing 83 observations with viscosity as the endpoint, where the observed values were normalised against the controls. Out of the 85 descriptors provided with the database for each observation, 60 and 48 descriptors remained respectively in the variation 1 and 2 predictor set after the removal of the near zero variance predictors as described in 3.3.2. *Data pre-processing*. On analysis of the models trained using the default parameters, all 14 models have $R^2 > 0.5$ on both the training set and the test set. Therefore, in an attempt to narrow down the selection of models for tuning, the models were assessed against the modified criteria for validity (**Table 3.4**). Upon validity assessment, `nnet` and `avNNet` fulfilled all criteria with the variation 1 predictor set (**Model 3.5** and **3.7**) while cubist fulfilled all criteria with both variation 1 and 2 predictor sets (**Model 3.11 – 3.12**). Consequently, these three models were tuned over an extended set of tuning parameters (**Table 3.3**) in attempt to refine the models. However, while cubist maintained its performance and fulfilled all criteria after tuning, both `nnet` and `avNNet` failed to fulfil one or more criteria after tuning (**Table 3.5**). This phenomenon of `nnet` and `avNNet` losing their validity can be explained by the random value dependency initially within their algorithms. Both of these algorithms involve calculation of a set of hidden units from the inputted predictors initially, which are then used to calculate the prediction value. In the calculation of the hidden units, the coefficients which relate the predictors to the hidden units are initialised to random values, which is then optimised via their specialised algorithms. Although the algorithm of `avNNet` aims to reduce this model instability, the results illustrates that `avNNet` still has the possibility of not finding the global optimum.

CHAPTER 3: UNDERSTANDING POLYMER DETERGENT PROPERTIES VIA QSPR METHOD

Table 3.4. Performance of the original models against the validity criteria

Model No.	Modelling Algorithm	Predictor variation	Training R ² *	Training RMSE*	Test R ²	R ² - R ₀ ²	$\frac{ R^2 - R_0^2 }{R^2}$	Test RMSE	k ₀	Z Score	Criteria fulfilled
3.1	SVM	1	0.776	0.121	0.802	0.108	0.134	0.218	0.720	19.198	5
3.2	SVM	2	0.782	0.118	0.857	0.073	0.085	0.208	0.717	19.525	7
3.3	KNN	1	0.740	0.125	0.739	0.175	0.236	0.191	0.801	27.755	5
3.4	KNN	2	0.713	0.135	0.798	0.127	0.160	0.187	0.786	27.730	5
3.5	nnet	1	0.749	0.125	0.912	0.052	0.057	0.118	0.916	27.991	8
3.6	nnet	2	0.715	0.126	0.860	0.065	0.076	0.193	0.768	29.194	7
3.7	avNNNet	1	0.747	0.124	0.909	0.054	0.059	0.119	0.920	24.679	8
3.8	avNNNet	2	0.700	0.129	0.859	0.066	0.077	0.193	0.765	22.976	7
3.9	RF	1	0.788	0.117	0.816	0.122	0.150	0.158	0.859	12.002	6
3.10	RF	2	0.769	0.118	0.801	0.132	0.165	0.161	0.864	10.218	5
3.11	Cubist	1	0.761	0.122	0.895	0.056	0.063	0.141	0.873	4.011	8
3.12	Cubist	2	0.764	0.125	0.901	0.054	0.059	0.137	0.876	4.168	8
3.13	XGM	1	0.784	0.120	0.761	0.165	0.216	0.167	0.877	24.941	6
3.14	XGM	2	0.770	0.118	0.788	0.142	0.180	0.164	0.870	24.046	6

*Training R² and Training RMSE extracted from average of 10-fold cross validation

Bold: values that meets the criteria (detail see 3.3.4. *Model building and performance assessment*)

CHAPTER 3: UNDERSTANDING POLYMER DETERGENT PROPERTIES VIA QSPR METHOD

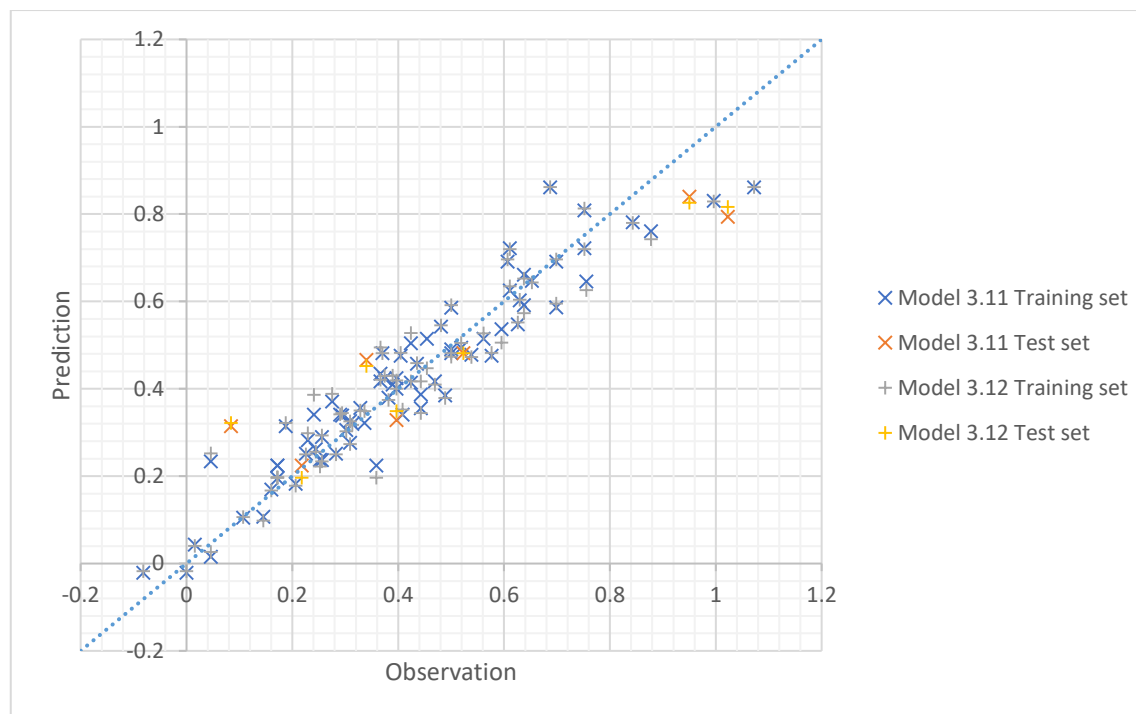


Figure 3.4. Plot of observation against prediction of the training set and test set for **Model 3.11** and **3.12**.

Table 3.5. Performance of the refined models, tuned over an expended set of tuning parameters, against the validity criteria

Model No.	Modelling Algorithm	Predictor variation	Training R^2 *	Training RMSE*	Test R^2	$R^2 - R_0^2$	$\left \frac{R^2 - R_0^2}{R^2} \right $	Test RMSE	k_0	Z Score	Criteria fulfilled
3.5T	nnet	1	0.802	0.114	0.823	0.104	0.127	0.177	0.818	18.968	5
3.7T	avNNet	1	0.795	0.114	0.902	0.050	0.056	0.164	0.794	18.457	7
3.11T	Cubist	1	0.758	0.125	0.887	0.060	0.068	0.150	0.855	6.601	8
3.12T	Cubist	2	0.757	0.126	0.902	0.053	0.059	0.136	0.878	10.603	8

*Training R^2 and Training RMSE extracted from average of 10-fold cross validation

Bold: values that meets the criteria (detail see 3.3.4. *Model building and performance assessment*)

3.4.1. Predictor importance across different models

Having identified some very good models it is important to assess the predictor importance in order to help understand what and how the viscosity of the formulations is influenced, even at a semi-quantitative molecular level. Due to the complexity of the cubist model algorithm, there is currently no established method to measure the predictor importance of a cubist model [14]. The only statistic related to the predictor importance available is the usage of predictors as a percentage of times each predictor was used in a condition and/or a linear model within the final model using the `varImp()` function within the `caret` package [22]. However, it is important to note that this is a form of usage count and therefore does not give any in depth information to how the predictors were used and hence their importance.

3.4.1.1. Predictor importance of the cubist models

There are some similarities and differences between the predictor importance for the models built with the two predictor set variants (**Model 3.11T – 3.12T**). For both predictor set variations, fewer than half, of the predictors available (**Supporting Information 3.2**) were used in the model. All the predictors used within the model with the variation 2 of the predictor set are also used for the model with the variation 1 predictor set, where the repeated use of those predictors indicates the high possibility of them in containing some useful information. On the other hand, there are some predictors that are not used within either of the model with variation 1 or 2 predictor set. Together with the high correlation between the groups of predictors indicated earlier on (3.3.2. *Data pre-processing*), this implies that some of the information included within those unused predictors has already been provided by those which have been used. For example, the `ChargeContourDensity` and `ChargeDensity` predictors are highly correlated and as both of them are some form of mole fraction of total charge measure and are effectively the same for the datasets investigated. Out of all predictors, mole fraction of total charge at pH7 (`ChargeDensity_pH7`), Abraham's parameter for hydrogen bond accepting at pH7 (`ABS_HBondAcceptor_pH7`) and number average molecular weight (`Mn`) are some of the top most used predictors with their usage between 70% and 100% for both models (**Supporting information 3.3**), therefore it is reasonable to consider that they contain highly important information. For **Model 3.12T**, number average molecular weight (`PSL_MW`) have the second highest usable of 82.2%. However, using the cubist models alone is not possible to define the effect each of the predictor has as the usage only accounts for the number of times a predictor was used in a linear model or a split within the algorithm and ignores the neighbour-based correction the algorithm also performs during prediction.

In general, the predictors that are used for both of the models should be considered to contain some possibly significant information, followed by the predictors that are only used in the model with the variation 1 predictor set, whereas the predictors that are not used in both of the models should be considered as predictors containing information that is already covered by other predictors or insignificant information.

3.4.1.2. Predictor importance of the neural network models

As the cubist model cannot give a true indication of the variable importance, the `nnet` and `avNNet` models using the variation 2 predictor set (**Model 3.5 and 3.7**) before tuning were used to aid the understanding as they both have comparable performance to cubist models. For neural networks, `caret` determined the variable importance by partitioning the connection weights between the input and hidden layer [22].

The variable importance from from the `nnet` and `avNNet` are somewhat contradictory; while `Mn` have the score of 66.53 (maximum score possible: 100) for the `nnet` model, its importance

score in the avNNet model is only 24.93 where on the other hand ABS_HBondAcceptor_pH7 has the score of 100 for the avNNet model but only score 5.17 for its importance within the avNNet model. The predictors with 70% - 100% usage in either of the cubist models generally have a reverse order of relative importance in the nnet and avNNet models. However, this could be explained by the fact that the avNNet model is an average of several nnet models. Due to the reported higher stability of the avNNet algorithm in comparison to the nnet algorithm, its variable importance should be deemed more reliable than those from the nnet, and hence takes the focus of the subsequent analysis.

The predictors with 70% - 100% usage in either of the cubist models also have a similar trend in the avNNet model. It is to note that ABS_HBondAcceptor_pH7, which has 78% and 88% usage in the cubist models with variation 1 and 2 predictor sets respectively, has 100% importance in the avNNet model, suggesting that it is one of the most important predictors. ChargeDensity_pH7, which was also highlighted for its high usage in the cubist models previously (85% and 100% for **Model 3.11** and **3.12** respectively), also has a relatively high variable importance of 87% in the avNNet model to support its possibility of high overall importance. On the other hand, the other highlighted predictor for the cubist models, Mn (usage: 73% and 76% for **Model 3.11** and **3.12** respectively), has a low variable importance of 25% within the avNNet model in comparison, suggesting that it might not actually contain much useful information. Overall, Abraham's parameter for hydrogen bond accepting at specified pH (ABS_HBondAcceptor), Abraham's parameter for hydrogen bond accepting divided by the molecular volume at specified pH (ABS_HBondAcceptor_d), mole fraction of total charge (ChargeDensity) and Abraham's parameter for hydrogen bond donation divided by the molecular volume (ABS_HBondDonor_d) are the predictors that are used within the cubist models which have variable importance over 85%, suggesting they should be the factors of the polymers within the formulations influencing the viscosity. However, as the structure of the neural network algorithms transforms the predictors using a sigmoidal function prior to combining them to calculate the hidden units, and the prediction is calculated as a combination of the hidden units, it is not possible to define which predictors have positive or negative effects on the predicted values.

Investigation of any trends within the variable importance for the nnet and avNNet models was also carried out. It is important to note that for the predictors noted with _pH# (**Supporting Information 3.1**, first column) can be grouped together by what property of the PSL they are describing. The difference between them is the pH of the environment when the measurement/calculation was carried out, e.g. ChargeDensity_pH7, ChargeDensity_pH8 and ChargeDensity_pH9 represent the mole fraction of total charge of the PSL at pH 7, 8 and 9 respectively. However, upon investigation the effect of the pH on the variable importance was not clear.

3.4.1.3. *Predictor – observation relationship interpretation using LM and PLS*

As the analysis of the variable importance of the cubist and the neural network models cannot establish the relationships between the predictors and the observations due to the nature of the algorithm, two approaches were taken in attempt to contribute to the understanding of any relationship existing. One of which is the use of LM and PLS. Although as expected these linear models do not fulfil all of the modified criteria for validity, the performance of PLS models (**Model 3.17** and **3.18**) and LM model with predictor variation 1 (**Model 3.15**) are acceptable as their values are not too far from the margin set for the criteria (within ± 0.05 of the margin, **Table 3.6**).

CHAPTER 3: UNDERSTANDING POLYMER DETERGENT PROPERTIES VIA QSPR METHOD

Table 3.6. Performance the test set of each of the linear models with the variation 1 prediction set against the validity criteria.

Model No.	Modelling Algorithm	Predictor variation	Training R ² *	Training RMSE*	Test R ²	$R^2 - R_0^2$	$\frac{ R^2 - R_0^2 }{R^2}$	Test RMSE	k ₀	Z Score	Criteria fulfilled
3.15	LM	1	0.687	0.139	0.835	0.099	0.119	0.167	0.835	30.640	6
3.16	LM	2	0.626	0.150	0.858	0.090	0.105	0.169	0.790	14.852	6
3.17	PLS	1	0.675	0.138	0.858	0.082	0.096	0.163	0.828	21.898	7
3.18	PLS	2	0.657	0.144	0.859	0.088	0.103	0.164	0.806	18.539	6

*Training R² and Training RMSE extracted from average of 10-fold cross validation

Bold: values that meets the criteria (detail see 3.3.4. *Model building and performance assessment*)

Another attempt to increase the understanding of the predictor – observation relationships was to build models with the list of predictors that took part in the cubist models. However, only the cubist and nnet models with the shortlisted variation 1 predictor set fulfil all the validity criteria (**Table 3.7**). We had already established that these models are not suitable for the analysis of the correlation between predictor and observation for the following reasons; for cubist, it is only possible to obtain the usage of the predictors within the algorithm for a linear model or a split, without the neighbour-based correction; for neural network models, the relation between predictors and the outputting prediction involves overlapping sigmoidal and linear relationships. Therefore, further investigation on the variable importance of these models were not investigated. On the other hand, the LM and PLS models here again have values close to the margins of the validity criteria, therefore it was considered suitable to further investigate these models (**Model 3.19 – 3.22**) in combination with the ones with the original sets of predictors to identify potential predictor – observation relationships. It is to note that the RF model using variation 1 predictor set is also close to the margin, however, due to its complex tree ensemble nature, it is again not suitable for analysis of the correlation between predictor and observation.

CHAPTER 3: UNDERSTANDING POLYMER DETERGENT PROPERTIES VIA QSPR METHOD

Table 3.7. Performance of the original models with the variation 1 predictor set shortlisted from the cubist model against the validity criteria

Model No.	Modelling Algorithm	Predictor variation	Training R ² *	Training RMSE*	Test R ²	$R^2 - R_0^2$	$\frac{ R^2 - R_0^2 }{R^2}$	Test RMSE	k ₀	Z Score	Criteria fulfilled
3.19	LM	1	0.617	0.147	0.862	0.086	0.100	0.168	0.791	12.590	6
3.20	LM	2	0.636	0.150	0.881	0.074	0.084	0.162	0.794	13.645	7
3.21	PLS	1	0.628	0.147	0.861	0.087	0.101	0.168	0.792	12.634	6
3.22	PLS	2	0.613	0.150	0.881	0.073	0.083	0.162	0.794	15.283	7
3.23	SVM	1	0.696	0.135	0.804	0.118	0.147	0.229	0.677	10.438	5
3.24	SVM	2	0.708	0.131	0.889	0.057	0.064	0.200	0.717	15.264	7
3.25	KNN	1	0.687	0.137	0.831	0.107	0.129	0.166	0.825	9.292	5
3.26	KNN	2	0.682	0.141	0.825	0.113	0.137	0.169	0.813	12.104	5
3.27	nnet	1	0.707	0.141	0.938	0.038	0.041	0.114	0.867	10.641	8
3.28	nnet	2	0.753	0.124	0.804	0.136	0.168	0.161	0.837	14.428	5
3.29	avNNet	1	0.778	0.120	0.903	0.061	0.068	0.135	0.848	11.807	7
3.30	avNNet	2	0.787	0.117	0.881	0.077	0.087	0.141	0.849	8.957	7
3.31	RF	1	0.785	0.120	0.819	0.120	0.146	0.156	0.862	11.568	6
3.32	RF	2	0.771	0.119	0.810	0.127	0.157	0.158	0.864	10.254	6
3.33	Cubist	1	0.748	0.126	0.899	0.055	0.061	0.132	0.909	6.280	8
3.34	Cubist	2	0.764	0.125	0.859	0.086	0.100	0.144	0.905	9.472	7
3.35	XGM	1	0.780	0.117	0.802	0.130	0.163	0.161	0.872	10.344	6
3.36	XGM	2	0.786	0.119	0.767	0.160	0.208	0.168	0.868	8.946	6

*Training R² and Training RMSE extracted from average of 10-fold cross validation

Bold: values that meets the criteria (detail see 3.3.4. *Model building and performance assessment*)

Caret defines the variable importance of the PLS variable by the weighted sums of the absolute regression coefficients proportionally to the reduction in the sum-of-squared errors [22]. Across all four PLS models (**Model 3.17 – 3.18, 3.21 – 3.22**), there is a big gap in importance between the top two to five most important predictors and the rest of the rest of the predictors. This gap ranges from a drop from 50.23% to 5.75% (**Support information 3.4, Model 3.17**) to the extreme case where the importance drop from 98.76% to 5.86% (**Support information 3.4, Model 3.18**). Within the top two to five most important predictors, Mn and PSL_MW are seen commonly. As previously noted, Mn is one of the top most used predictors in both of the cubist models. Although it did not appear so in avNNet model, the high variable importance of Mn within the four PLS models successfully supported this proposal. PSL_MW is another predictor that is used within both of the cubist models, with a high usage of 82% within the model with the variation 2 predictor set (**Model 3.12T**). However, similar to Mn, it also has a low variable importance within the avNNet model.

With the above analysis, it was reasonable to focus the investigation of the key properties correlated to experimental measurements to the following: ABS_HBondAcceptor_pH7, ABS_HBondAcceptor_d_pH7, ChargeDensity_pH7, ABS_HBondDonor_d_pH7, Mn and PSL_MW. By assessing the distribution of their coefficients across all of the components of the PLS models (**Model 3.17 – 3.18, 3.21 – 3.22, Supporting information 3.5**), it was found that PSL_MW, ABS_HBondAcceptor_pH7 and ABS_HBondAcceptor_d_pH7 contribute positively and ChargeDensity and ABS_HBondDonor_d_pH7 contributes negatively towards the predictor – observation relationship. For Mn, the contribution varies between the components across the models. In a third of the PLS components Mn took part in, it has a negative coefficient, with the remaining two third being positive. However, when considering the overall amplitude of the coefficients, Mn contributes positively towards the predictor – observation relationship in broad terms.

It was found during the analysis of the PLS model coefficients that the change in coefficient in relation to pH for ABS_HBondAcceptor, ABS_HBondAcceptor_d, ABS_HBondDonor_d and ChargeDensity is such that the lower the pH, the larger the coefficient is, i.e. the value of the coefficient gets larger regardless of the sign. On the other hand, on the analysis of the variable importance across the four PLS models, the balance of predictors used within the cubist model with the variation 2 predictor set is tilted towards being in pH ascending order whereas for the predictors used within the cubist model with the variation 1 predictor set is tilted towards being in pH descending order. Combining the trends identified for the neural network models, it is still very difficult to truly establish the relationship between the pH of the predictors and the observations.

Consolidating all the analyses performed regarding the variable importance of the cubist, neural networks and PLS models, the main conclusions drawn were that Mn, PSL_MW, ABS_HBondAcceptor, ABS_HBondAcceptor_d, ChargeDensity and ABS_HBondDonor_d are the descriptors that have the most influence on the observation prediction. Out of which, Mn, ABS_HBondAcceptor and ChargeDensity at pH7 were found to be the most important ones. Mn, PSL_MW, ABS_HBondAcceptor and ABS_HBondAcceptor_d have a positive effect on the outcome whereas ChargeDensity and ABS_HBondDonor_d have a negative effect. It was not possible to draw a clear conclusion regarding the relation between the predictors and the viscosity with the aid of the neural networks and PLS models. Even though the cubist models can be used to predict the viscosity of a new formulation, it was not appropriate for the analysis of the influence of PSL properties on viscosity of current

formulations and therefore is not suitable for aiding the design of new PSL to be included in the formulation.

3.4.1.4. Proposed mechanism of viscosity

Out of the predictors identified to be important, it was found that Mn and PSL_MW increase the viscosity of the formulation. Both of these predictors describe the PSL in terms of its molecular weight where the former describes the number average molecular weight of the PSL which includes the terminal substitution where the latter only describes the molecular weight of the PSL core. According to the Mark-Houwink equation, molecular weight of a polymer is directly related to its contribution to the viscosity of a solution [23], and therefore the relationship between the viscosity of the formulation and the two PSL molecular weight descriptors seems consistent with this. However, as the molecular weight of the PSLs analysed are low for hyperbranched polymers (PSL_MW = 6000 – 215000, where branched polymer MW can vary into 10^6 region [24]), the Mark-Houwink equation alone is not sufficient to explain the phenomenon. Another possible explanation to the correlation between Mn and PSL_MW with viscosity is the interference of the PSL molecules with the surfactant mesomorphic phases (mesophases). Mesophase is the phase of state in between liquid and solid [1] and with surfactants, this contributes to their viscosity. By introducing PSL molecules to a detergent formulation, the PSL molecules would interact with the existing surfactant mesophases [25]. The number of PSL molecules interfering with the surfactant mesophase is correlated to how ordered the mesophases are and therefore the viscosity of the detergent formulation. The lower the molecular weight of the PSL molecules are, the smaller the PSL molecules are and therefore more PSL molecules can interact with the surfactant mesophases.

3.5. Constructed models for BF2 data sets

3.5.1. Initial constructed models

For the BF2 data sets, only SVM successfully produced a model for the BF2_YPC with $R^2 > 0.5$ on the training set and the test set (**Table 3.8, Model 3.41**). However, upon tuning, and assessment against the validity criteria (e.g. $R^2 > 0.6$ on the test set), it was not accepted as a successful model.

Table 3.8. The list of initial models with $R^2 > 0.5$ on either training or test set for the BF2 data sets

Model No.	End point	Modelling Algorithm	Predictor variation	Training set R^2 *	Test set R^2 *
3.37	RPC	Nnet	1	0.51	0.13
3.38	RPC	avNNet	1	0.50	0.13
3.39	Vis	Nnet	1	0.20	0.64
3.40	Vis	avNNet	1	0.24	0.68
3.41	YPC	SVM	1	0.52	0.71
3.42	YPC	Nnet	1	0.36	0.91
3.43	YPC	avNNet	1	0.38	0.92
3.44	YPC	avNNet	2	0.22	0.97

*Training R^2 extracted from average of 10-fold cross validation
 Bold: $R^2 > 0.5$

3.5.2. Models with transformed observation values

As there were no models accepted as a successful model after tuning, observation values were regulated in attempt to find a successful model. Two methods were used for observation value regulation:

- Taking \log_{10} values of the data pre-processed observation values or
- Taking \log_{10} values of the raw data observation values during normalisation within the data compilation process

However, neither method manage to yield any successful predictive models.

3.5.3. Models with altered predictors

As a transformation of the observation values did not produce any successful predictive models, transformations of the predictors were attempted as follows:

- Principle component analysis (PCA) using the `preProcess` option within `caret::train()`
- Yeo Johnson using the `preProcess` option within `caret::train()`
- Limiting the predictors to certain pH by removing the irrelevant pH related predictors
- Using weighted descriptors calculated from the repeat unit descriptors provided within the data collection (62 descriptors) and calculated by Dragon6 [26] (1807 zero to two dimensional descriptors) according to the fraction each repeat unit is present in the polymer

The Yeo Johnson function is a function able to accommodate zero and negative values while eliminating any skewness possibly occurring within the predictors. However, application of the function did not successfully yield any predictive models. Adjustment of the predictors via PCA, a commonly used data reduction method which seeks to find the linear combination of predictors which capture the most possible variance, and limiting the predictors according to their pH by removal of the irrelevant pH related predictors yielded some models with limited success where those models only have $R^2 > 0.5$ on the test set (**Table 3.9**).

Table 3.9. The list of models with $R^2 > 0.5$ on the test set alone after predictor adjustments

End point	Predictor variation	Predictor regulation	Modelling algorithm
Vis	1	PCA	nnet
Vis	1	PCA	avNNet
Vis	1	pH7 predictors only	nnet
Vis	1	pH7 predictors only	avNNet
Vis	1	pH8 predictors only	cubist
Vis	2	pH8 predictors only	
Vis	1	pH8 predictors only	nnet
Vis	1	pH8 predictors only	avNNet
Vis	1	pH9 predictors only	nnet
Vis	1	pH9 predictors only	avNNet
YPC	1	PCA	SVM
YPC	2	PCA	SVM
YPC	1	PCA	nnet
YPC	1	PCA	avNNet
YPC	2	PCA	avNNet
YPC	1	pH7 predictors only	SVM
YPC	1	pH7 predictors only	nnet
YPC	1	pH7 predictors only	avNNet
YPC	2	pH7 predictors only	avNNet
YPC	1	pH8 predictors only	SVM
YPC	2	pH8 predictors only	SVM
YPC	2	pH8 predictors only	nnet
YPC	1	pH8 predictors only	avNNet
YPC	2	pH8 predictors only	avNNet
YPC	1	pH9 predictors only	SVM
YPC	1	pH9 predictors only	nnet
YPC	1	pH9 predictors only	avNNet

As the above methods did not result in any successful models, an alternative set of predictors were calculated by using a combination of the descriptors for the repeat units of the PSLs provided within the data collection and ones calculated by Dragon6 [26]. Using Dragon6, a total of 1807 0-2D descriptors were calculated based on the 2D structure of the repeat units. For each of the PSLs used within the formulations, the predictors were calculated as the weighted descriptors of PSL repeat units (**Example 3.1**). However, this approach following the original data pre-processing procedure also only managed to produce some models with $R^2 > 0.5$ on the test set alone (**Table 3.10**). Predictors with an absolute correlation above 0.95 were then removed to try to improve the performance of the models. Nonetheless, this approach as well only managed to produce some models with $R^2 > 0.5$ on the test set alone, i.e. did not satisfy the criteria for training set (**Table 3.11**).

<i>Repeat units</i>	A	B	C	D	Total
<i>Proportion present within PSL</i>	0.6	0.2	0.1	0.1	1.0
<i>× Mn</i>	90	25	41	12	
<i>= Weighted Mn</i>	54	5	4.1	1.2	64.3

Example 3.1. An example of the calculation of weighted descriptors of PSLs.

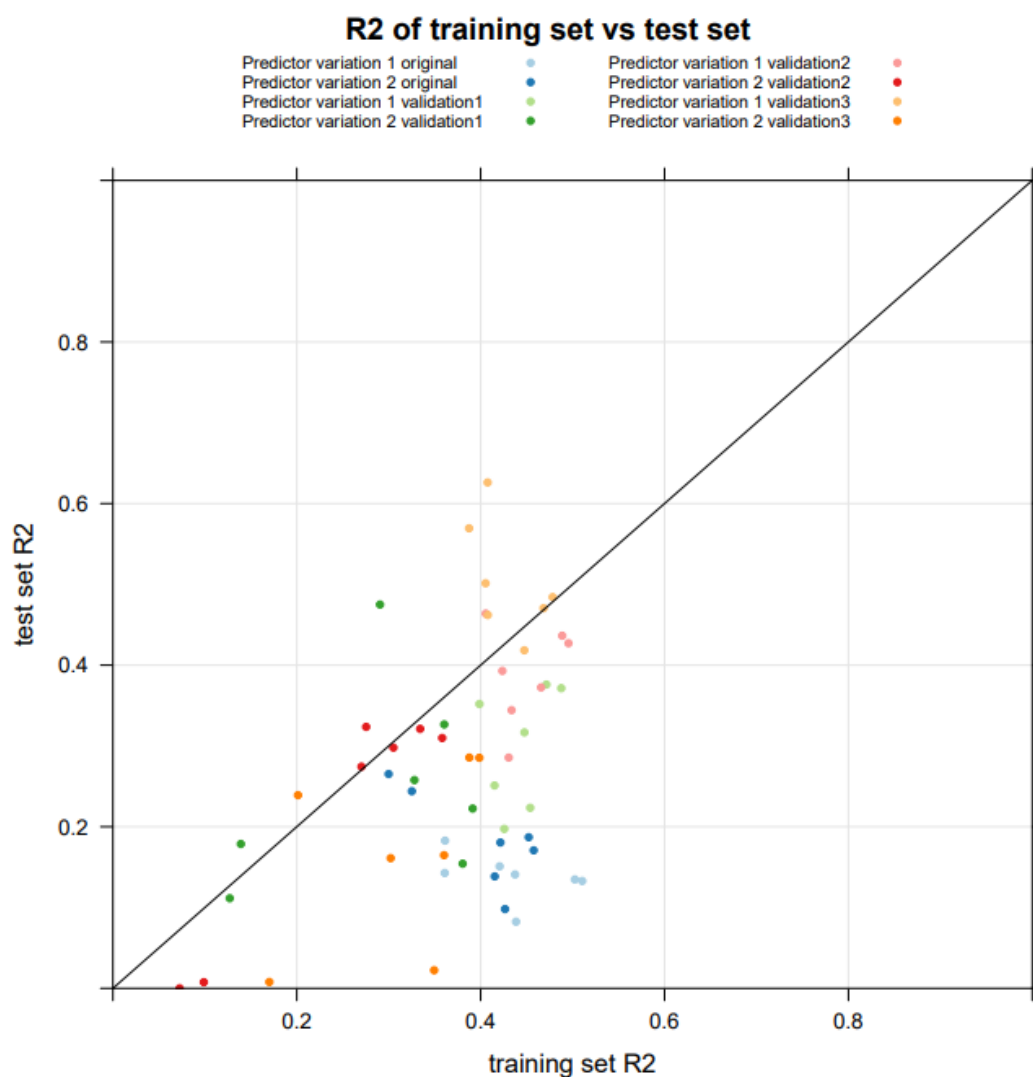
Table 3.10. The list of models with $R^2 > 0.5$ on the test set alone with the weighted descriptors as predictors.

End point	Predictor variation	Modelling algorithm	Test set R^2
RPC	1	KNN	0.53
RPC	2	KNN	0.56
RPC	2	RF	0.50
Vis	1	RF	0.67
Vis	1	Cubist	0.51
YPC	1	SVM	0.57
YPC	1	Cubist	0.92

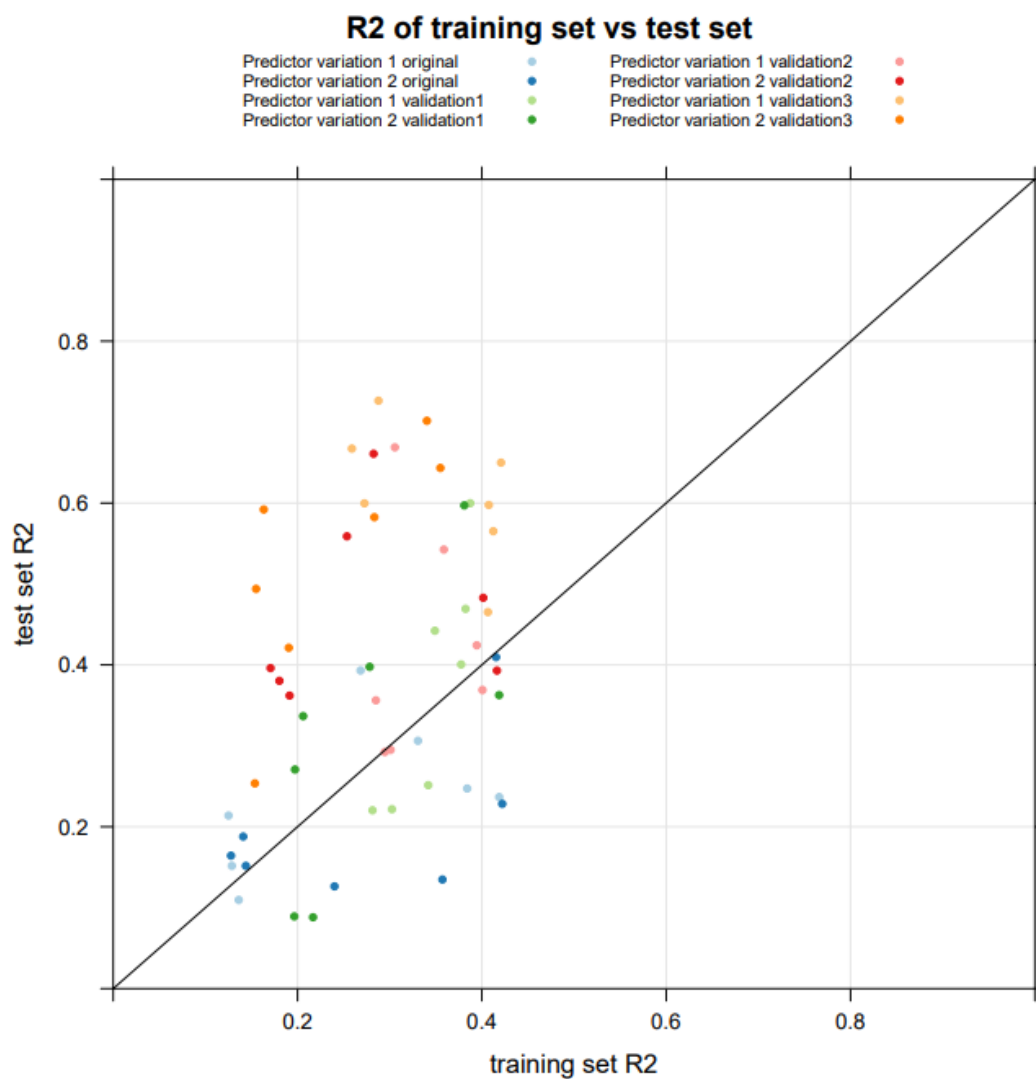
Table 3.11. The list of models with $R^2 > 0.5$ on the test set alone with the weighted descriptors as predictors and highly correlated ones removed.

End point	Predictor variation	Modelling algorithm	Test set R^2
RPC	1	SVM	0.72
RPC	1	KNN	0.50
RPC	1	RF	0.64
RPC	1	Cubist	0.55
RPC	1	XGB	0.61
RPC	2	SVM	0.67
RPC	2	RF	0.54
RPC	2	Cubist	0.68
YPC	1	RF	0.68
YPC	1	Cubist	0.79
YPC	1	XGB	0.78
YPC	2	RF	0.70
YPC	2	Cubist	0.68
YPC	2	XGB	0.65

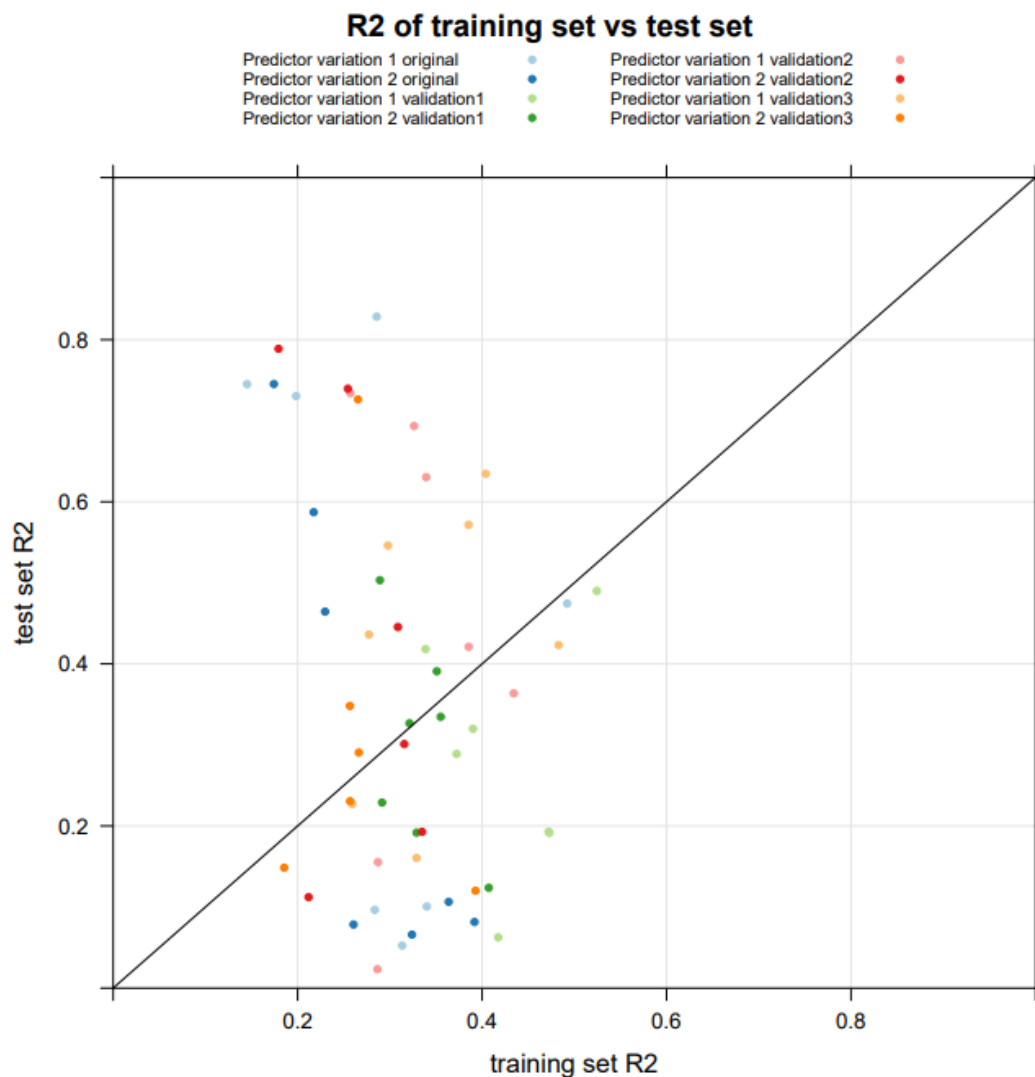
It was noted that with the BF2 data sets, most of them produce models which have a $R^2 > 0.5$ on the test set but not for the training set. This phenomenon was thought to occur due to the training/test split of the data, i.e. there is always a possibility that molecules in the test set are predicted well just as a consequence of which compounds are in the test set. This hypothesis was tested by repeating the data pre-process and model building process with the original predictor sets three times. Each time the data pre-processing is carried out, a different training/test split is performed using a different random seed number. The R^2 values of the all models on the test set was then plotted against the values on the training set according to the data set. If the performance does not cluster in the same area as the original models, it proves that the variation in the training/test split contributes to the difference in performance between the training and test set. As seen in **Figure 3.5**, the test set of models with the BF2_RPC data set usually underperform in comparison to the training set, the BF2_Vis data set tends to outperform and the performance for BF2_YPC is fairly equal in comparison to the training set.



a)



b)



c)

Figure 3.5. A graph of R2 values of test set against training set for a) BF2_RPC, b) BF2_Vis and c) BF2_YPC.

Although none of the methods above were successful in building a predictive model for any of the BF2 data sets, possibly important predictors were defined as those which had 50% variable importance in at least 10 out of the 28 models, including the three sets of models built for the test set performance validation, for each data set as concluded in **Table 3.12**.

Table 3.12. List of predictors above 50% variable importance (description see **Support information 3.1**) across at least 10 models

End point	Predictor variation 1	Predictor variation 2
RPC	mf_FG2 mf_FG9 ABS_LogP_pH9 Hydrophobe_Length	ABS_LogP_pH8
Vis	Mn Molecules_vs_Current mf_FG2 mf_FG9 mf_FG5Hion_pH8 ChargeContourDensity_pH7 IonContourDensity_pH7 ABS_HBondDonor_pH9 ABS_HBondDonor_d_pH9 Hydrophobe_Length	Mn Molecules_vs_Current mf_FG5Hion_pH9 IonContourDensity_pH8 ABS_HBondDonor_pH7 ABS_HBondDonor_pH9
YPC	mf_FG5Hion_pH9 ChargeContourDensity_pH7 ABS_HBondAcceptor_d_pH7	ChargeDensity_pH9 ABS_HBondAcceptor_pH7 ABS_HBondDonor_pH8 ABS_HBondDonor_d_pH9

3.6. Conclusion

Out of all the data sets investigated, only BF1_Vis data set has yielded a predictive model using the cubist algorithm. Within the model, Mn, PSL_MW, ABS_HBondAcceptor, ABS_HBondAcceptor_d, ChargeDensity and ABS_HBondDonor_d are the factors considered to have the most influence on the observation prediction, where Mn, PSL_MW, ABS_HBondAcceptor and ABS_HBondAcceptor_d is thought to have a positive effect on the prediction and ChargeDensity and ABS_HBondDonor_d have a negative effect.

Although no predictive models were successfully built from the BF2 data sets using various approaches, possibly important predictors for each of the data set were marked by the number of times they were used between all of the models.

In order to build a successful model for the BF2 data sets and increase the performance of the simpler models, such as PLS, for BASD1_Vis, predictors that are more closely related to the observations, such as amphiphilicity, will need to be obtained by experiment or calculation for the PSLs – calculation of amphiphilicity related descriptors is described in *Chapter 4: Novel Surfactant Descriptor – Potential Amphiphilicity*.

3.7. References

1. McNaught, A. D.; Wilkinson, A. *IUPAC. Compendium of Chemical Terminology, 2nd ed. (the "Gold Book")*. Blackwell Scientific Publications, Oxford (1997), (accessed 08/06/2021).
2. Cheng, K. C.; Khoo, Z. S.; Lo, N. W.; Tan, W. J.; Chemmangattuvalappil, N. G. Design and performance optimisation of detergent product containing binary mixture of anionic-nonionic surfactants. *Heliyon* **2020**, *6* (5), e03861. DOI: <https://doi.org/10.1016/j.heliyon.2020.e03861>.
3. Helenius, A.; McCaslin, D. R.; Fries, E.; Tanford, C. [63] Properties of detergents. In *Methods in Enzymology*, Vol. 56; Academic Press, 1979; pp 734-749.
4. Chakraborty, T.; Chakraborty, I.; Ghosh, S. The methods of determination of critical micellar concentrations of the amphiphilic systems in aqueous medium. *Arabian Journal of Chemistry* **2011**, *4* (3), 265-270. DOI: <https://doi.org/10.1016/j.arabjc.2010.06.045>.
5. Rosen, M. J.; Kunjappu, J. T. *Surfactants and interfacial phenomena*; John Wiley & Sons, 2012.
6. Mortensen, K. PEO-related block copolymer surfactants. *Colloids and Surfaces A: Physicochemical and Engineering Aspects* **2001**, *183-185*, 277-292. DOI: [https://doi.org/10.1016/S0927-7757\(01\)00546-5](https://doi.org/10.1016/S0927-7757(01)00546-5).
7. Atta, D. Y.; Negash, B. M.; Yekeen, N.; Habte, A. D. A state-of-the-art review on the application of natural surfactants in enhanced oil recovery. *Journal of Molecular Liquids* **2021**, *321*, 114888. DOI: <https://doi.org/10.1016/j.molliq.2020.114888>.
8. Walling, C. Copolymerization. XIII.1 Over-all Rates in Copolymerization. Polar Effects in Chain Initiation and Termination. *Journal of the American Chemical Society* **1949**, *71* (6), 1930-1935. DOI: 10.1021/ja01174a009.
9. Mayo, F. R.; Lewis, F. M. Copolymerization. I. A Basis for Comparing the Behavior of Monomers in Copolymerization; The Copolymerization of Styrene and Methyl Methacrylate. *Journal of the American Chemical Society* **1944**, *66* (9), 1594-1601. DOI: 10.1021/ja01237a052.
10. *R: What is R?* <https://www.r-project.org/about.html> (accessed 2017 07/10/2021).
11. Kuhn, M. *The caret package*. <http://topepo.github.io/caret/index.html> (accessed 07/10/2021).
12. Breerton, R. G.; Lloyd, G. R. Support Vector Machines for classification and regression. *Analyst* **2010**, *135* (2), 230-267, 10.1039/B918972F. DOI: 10.1039/B918972F.
13. Fernández-Delgado, M.; Sirsat, M. S.; Cernadas, E.; Alawadi, S.; Barro, S.; Febrero-Bande, M. An extensive experimental survey of regression methods. *Neural Networks* **2019**, *111*, 11-34. DOI: <https://doi.org/10.1016/j.neunet.2018.12.010>.
14. Kuhn, M.; Johnson, K. *Applied predictive modeling*; Springer New York, 2013. DOI: <https://doi.org/10.1007/978-1-4614-6849-3>.
15. Couronné, R.; Probst, P.; Boulesteix, A.-L. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics* **2018**, *19* (1), 270. DOI: 10.1186/s12859-018-2264-5.
16. Mohammadi, M.-R.; Hadavimoghaddam, F.; Pourmahdi, M.; Atashrouz, S.; Munir, M. T.; Hemmati-Sarapardeh, A.; Mosavi, A. H.; Mohaddespour, A. Modeling hydrogen solubility in hydrocarbons using extreme gradient boosting and equations of state. *Scientific Reports* **2021**, *11* (1), 17911. DOI: 10.1038/s41598-021-97131-8.
17. Golbraikh, A.; Tropsha, A. Beware of q²! *Journal of Molecular Graphics and Modelling* **2002**, *20* (4), 269-276. DOI: [http://dx.doi.org/10.1016/S1093-3263\(01\)00123-1](http://dx.doi.org/10.1016/S1093-3263(01)00123-1).

18. Alexander, D. L. J.; Tropsha, A.; Winkler, D. A. Beware of R²: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *Journal of Chemical Information and Modeling* **2015**, *55* (7), 1316-1322. DOI: 10.1021/acs.jcim.5b00206.
19. Shen, M.; LeTiran, A.; Xiao, Y.; Golbraikh, A.; Kohn, H.; Tropsha, A. Quantitative structure-activity relationship analysis of functionalized amino acid anticonvulsant agents using k nearest neighbor and simulated annealing PLS methods. *Journal of Medicinal Chemistry* **2002**, *45* (13), 2811-2823. DOI: 10.1021/jm010488u From NLM.
20. Khazaei, A.; Sarmasti, N.; Seyf, J. Y.; Rostami, Z.; Zolfigol, M. A. QSAR study of the non-peptidic inhibitors of procollagen C-proteinase based on Multiple linear regression, principle component regression, and partial least squares. *Arabian Journal of Chemistry* **2017**, *10* (6), 801-810. DOI: <https://doi.org/10.1016/j.arabjc.2015.02.016>.
21. Shen, M.; LeTiran, A.; Xiao, Y.; Golbraikh, A.; Kohn, H.; Tropsha, A. Quantitative Structure-Activity Relationship Analysis of Functionalized Amino Acid Anticonvulsant Agents Using k Nearest Neighbor and Simulated Annealing PLS Methods. *Journal of Medicinal Chemistry* **2002**, *45* (13), 2811-2823. DOI: 10.1021/jm010488u.
22. Williams, C. K.; Engelhardt, A.; Cooper, T.; Mayer, Z.; Ziem, A.; Scrucca, L.; Tang, Y.; Candan, C.; Hunt, T.; Kuhn, M. M. Package 'caret'. **2017**.
23. McNaught, A. D. *Compendium of chemical terminology*; Blackwell Science Oxford, 1997.
24. Park, I. I. H.; Choi, E. J. Characterization of branched polyethyleneimine by laser light scattering and viscometry. *Polymer* **1996**, *37* (2), 313-319. DOI: [https://doi.org/10.1016/0032-3861\(96\)81104-9](https://doi.org/10.1016/0032-3861(96)81104-9).
25. Kumaraswamy, G.; Sharma, K. P. Chapter Seven - Polymer and Colloidal Inclusions in Lyotropic Lamellar and Hexagonal Surfactant Mesophases. In *Advances in Planar Lipid Bilayers and Liposomes*, Iglič, A., Kulkarni, C. V. Eds.; Vol. 18; Academic Press, 2013; pp 181-208.
26. *Dragon (Software for Molecular Descriptor calculation)*; Talete srl: 2015. <http://www.talete.mi.it/> (accessed 2016).

Chapter 4:
Novel Surfactant Descriptor – Potential
Amphiphilicity

4.1. Amphiphilicity and Surfactants

As mentioned in *Chapter 3: Understanding Polymer Detergent Properties via QSPR method*, surfactants are a key chemical in detergents. Surfactants are amphiphilic (or amphipathic) molecules containing lyophilic and lyophobic sections which are strongly attracted or repulsive to a specific solvent respectively [1, 2]. When the specific solvent is water, the sections are usually referred to as hydrophilic and hydrophobic sections respectively. When a surfactant dissolves in an aqueous phase, the weak attraction between the hydrophobic section and water molecules breaks the hydrogen bond interactions between the water molecules, distorting the structure of the aqueous phase. This distortion causes some of the surfactant molecules to be exposed to the interface between the aqueous phase and non-aqueous phase (e.g. air, lipophilic phase). The exposed surfactant molecules then arrange themselves such that the hydrophobic sections have the minimum contact with the water molecules (**Figure 4.1**). As a result, the hydrophobic sections predominantly point towards the non-aqueous phase, lowering the surface tension by reducing the dissimilarity between the two phases [1, 2]. Surfactants can be classified into different groups depending on the nature of the hydrophilic sections: anionic, cationic, zwitterionic and non-ionic (**Figure 4.2**).

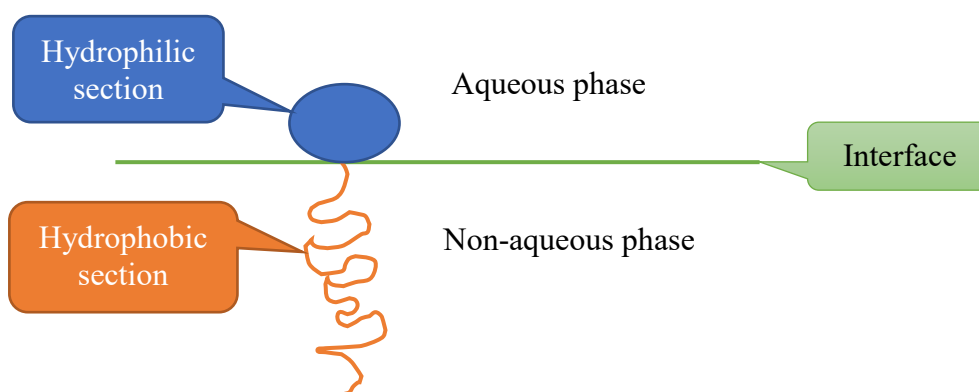


Figure 4.1. Illustration of the arrangement surfactant molecules take when exposed to the interface between aqueous phase and non-aqueous phase, where size of the sections is not to scale.

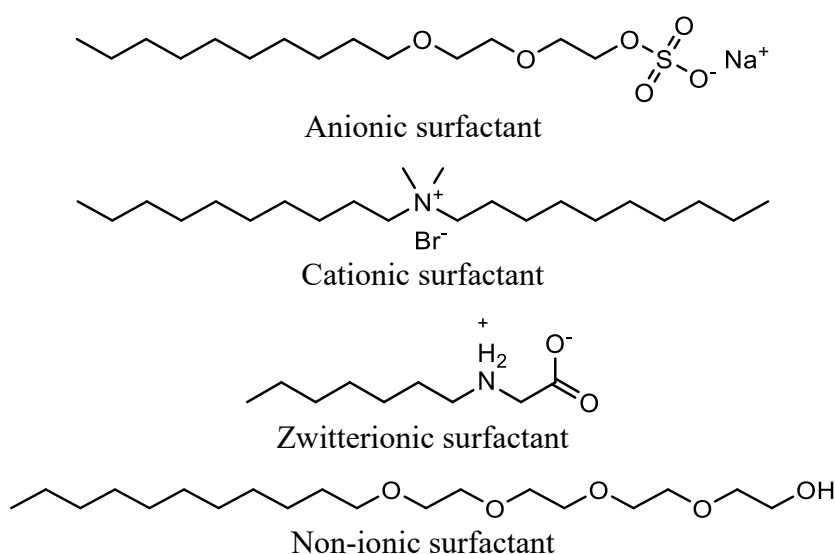


Figure 4.2. Examples of different types of surfactants.

It is known that amphiphilicity is dependent on the hydrophilicity and hydrophobicity of the molecule, and the relative position of the hydrophilic and hydrophobic sections [3]. There had been attempts to quantify amphiphilicity for small charged drug molecules and α -helices of proteins as amphiphilicity plays a critical role in biology [4, 5]. Similar to the mechanism of surfactants orienting themselves into position where the hydrophobic sections have minimum contact with water molecules, amphiphilic molecules in living matter have an inherent tendency to orient themselves in a suitable environment (e.g. a lipid bilayer) [4]. This ability to orient themselves into a suitable position within the environment holds the key for the structural organisation of living matter and how interactions can occur [4]. For α -helices of proteins, amphiphilicity is quantified by the hydrophobicity moment, defined as the magnitude of the vector sum of the hydrophobicity of the amino acid residues projected perpendicular to the axis of the helix [5].

Extending this idea further, conformation dependent amphiphilicity of a molecule can be defined as the amphiphilic moment, which is defined as the vector pointing from the centre of the hydrophobic domain to the centre of the hydrophilic domain, with the strength of the amphiphilic moment quantified by the vector length [6]. For small charged drug molecules, amphiphilicity is defined by amphiphilic moment, calculated by vector addition of individual atom/fragment contribution values and calibrated with known free energy of transfer of a compound from aqueous phase to air-water interface or into a micelle [4]. However, to the best of our knowledge, there is no literature that quantifies specifically the amphiphilicity of a surfactant molecule using calculated or measured molecular properties at present. Therefore, to attempt to quantify this property could provide a great step forward towards surfactant related modelling.

One of the most challenging areas in quantifying amphiphilicity is the relative position of the hydrophilic and hydrophobic sections. First, it requires the identification of the “boundary” between the hydrophilic and hydrophobic sections. Although this can be done manually using relevant chemistry knowledge [7], when this is a step within the analysis of library of hundreds of molecules or more, there is a necessity to automate this process. Once the “boundary” is identified, energetically reasonable conformations of the hydrophilic and hydrophobic sections need to be calculated. This can be a time-consuming step as an increasing number of rotatable bonds increases the number of local energetic minimum conformations the molecule can adopt [8], e.g. if a single rotatable bond can adapt 6 torsion angles, two rotatable bonds can give rise to $6^2 = 36$ conformations; three rotatable bonds can give rise to $6^3 = 216$ conformations. Also, it is difficult to calculate what fraction of the molecule would adopt which conformation as surfactants when used as cleaning product would be used in a solution of various concentrations and temperatures. Therefore, calculation of the relative position of the hydrophilic and hydrophobic sections is not the focus of this chapter due to the high time consumption and computational cost.

4.1.1. What is potential amphiphilicity

As the number of rotatable bonds is directly related to the number of possible conformations of a molecule, the number of rotatable bonds of the hydrophobic and hydrophilic sections can be considered as some sort of indicator for their flexibility of conformations [9]. The more flexible a molecule is, the higher the possibility that the conformations it can take arrange the hydrophobic and hydrophilic sections in such a way that the molecule can sit “well” at the interface with the hydrophobic section in the non-aqueous phase and hydrophilic section in the aqueous phase. However, as mentioned above, the calculation of the conformers can be very time consuming, especially for large data sets with molecules with many rotatable bonds. On the

other hand, if the hydrophobic and hydrophilic sections are considered separately and it is assumed that all the hydrophobic and hydrophilic sections can be positioned in the aqueous and non-aqueous phase respectively, which in practice may be hindered by the lack of flexibility (**Figure 4.3**), it is possible to calculate the potential amphiphilicity for the molecule.

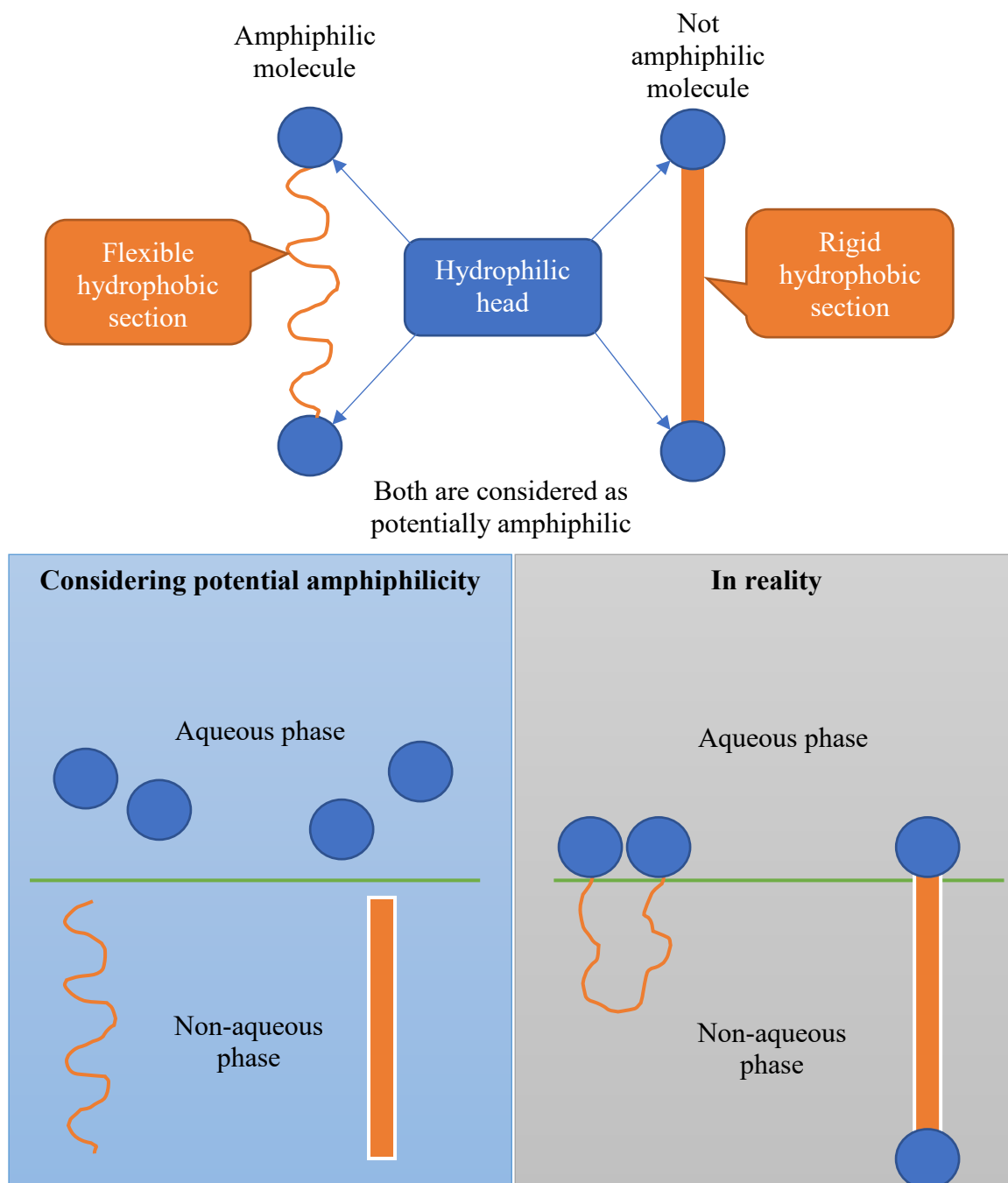


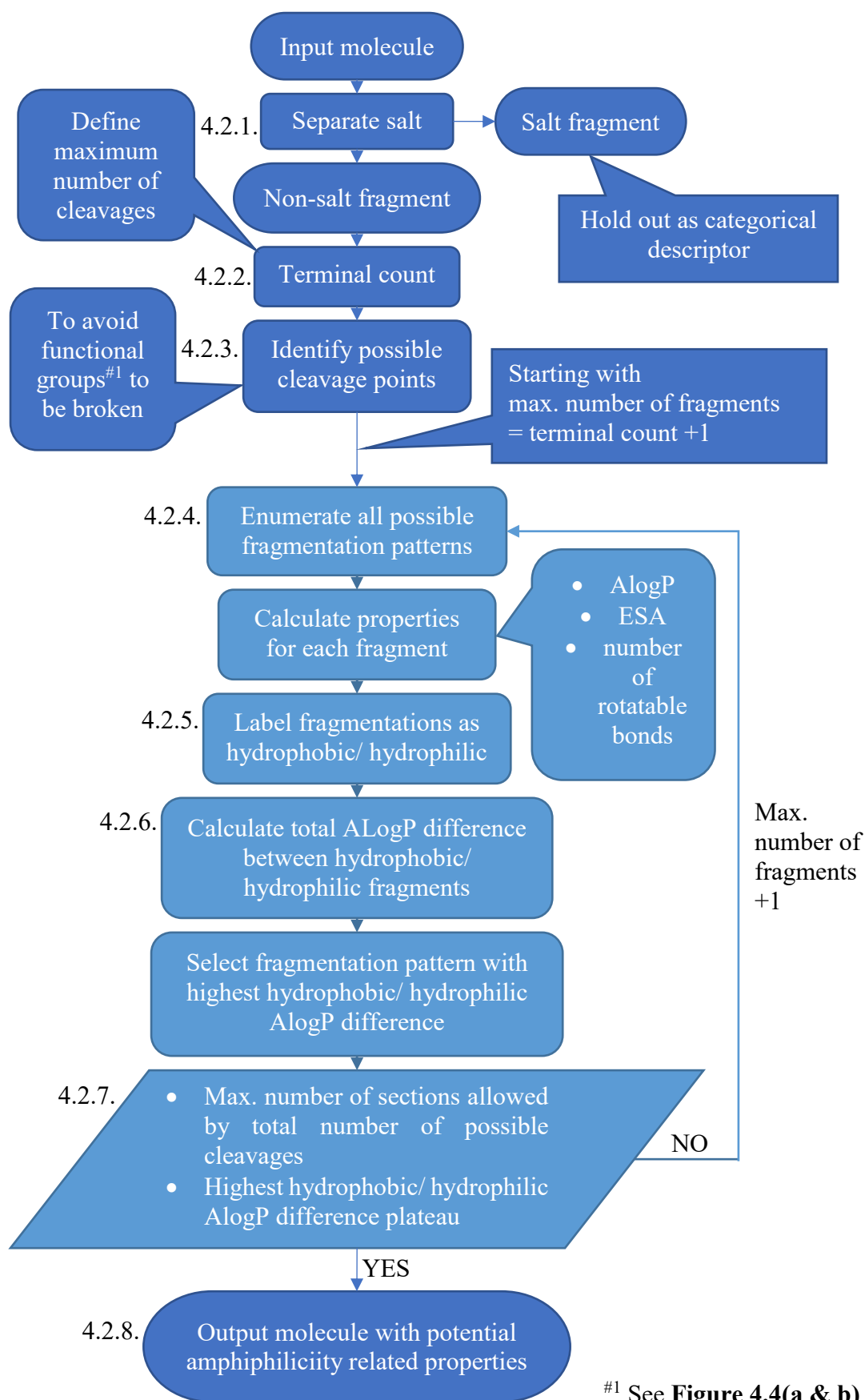
Figure 4.3. Illustration of the difference between actually amphiphilic molecule and potentially amphiphilic molecule, which in reality is not amphiphilic due to lack of flexibility.

In the calculation of the potential amphiphilicity, our hypothesis was that AlogP [10] and electronegativity surface area (ESA) of the hydrophilic and hydrophobic sections would be crucial to in the quantification of the hydrophilicity and hydrophobicity of the sections of the molecule. ESA here defined as the electronegativity of all bonds in the section \times surface area

of the section, where the electronegativity of a bond is calculated as the difference in electronegativity of the bonding atoms. The difference in AlogP between the hydrophilic and hydrophobic sections provides the basis of what amphiphilicity refers to, and we hypothesise is an important factor in the identification of the boundary between the hydrophilic and hydrophobic sections in this work (See 4.2. *Calculating Potential Amphiphilicity*). The number of rotatable bonds of the hydrophilic and hydrophobic sections are also included as an indication of the flexibility of the sections.

4.2. Calculating Potential Amphiphilicity

Potential amphiphilicity properties were calculated following **Scheme 4.1** within Pipeline Pilot [11]. Overall, the aim of the protocol is to identify the optimal position to cleave a molecule into hydrophilic and hydrophobic fragments which retain all important functional groups, identify the cleavage which give rise to the highest ALogP difference between the hydrophilic and hydrophobic fragments, and calculate properties for each fragment (**Table 4.1**). **Molecules 4.1 – 4.3** are used as examples for protocol exemplification.



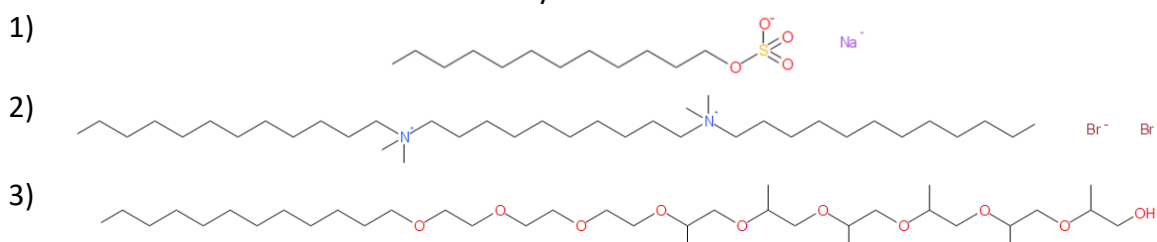
Scheme 4.1. Overall procedure for calculating potential amphiphilicity related properties.

Table 4.1. Description of potential amphiphilicity properties calculated

Potential amphiphilicity property	Description
Hydrophilic ALogP	Total ALogP of the hydrophilic fragments
Hydrophobic ALogP	Total ALogP of the hydrophobic fragments
Fragment ALogP difference	Difference between total ALogP of the hydrophilic fragments and the total ALogP of the hydrophobic fragments
Hydrophilic number of rotatable bonds	Total number of rotatable bonds of the hydrophilic fragments
Hydrophobic number of rotatable bonds	Total number of rotatable bonds of the hydrophobic fragments
Polar Electronegative Surface Area (ESA)	Total surface area ^a x electronegativity ^b of the hydrophilic fragments
Non-polar Electronegative Surface Area (ESA)	Total surface area ^a x electronegativity ^b of the hydrophobic fragments

^a Surface area is calculated using the Pipeline Pilot inbuilt 2D approximation

^b Electronegativity is calculated as the sum of the electronegativity differences of the two atoms of every bond in the molecule



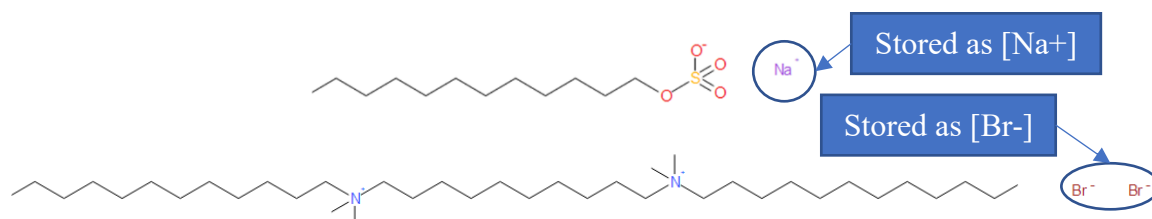
Molecule 4.1 – 4.3. Example molecules to illustrate main steps of the potential amphiphilicity descriptor calculator.

4.2.1. Input molecules

First, surfactant molecules are imported to the protocol as one of the following:

- SD file
- .txt file with SMILES
- .txt file with SMARTS
- .csv file with SMILES

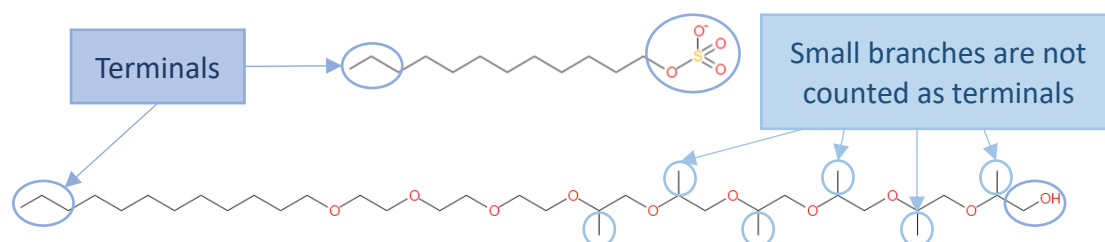
The imported molecules are filtered for salt fragments. The salt counter ion(s) of any molecules are extracted and stored as their SMILES. For any molecules with duplicated counter ions (**Example 4.1**), the duplication is removed.



Example 4.1. Molecule 4.1 and 4.2 with their salt extracted.

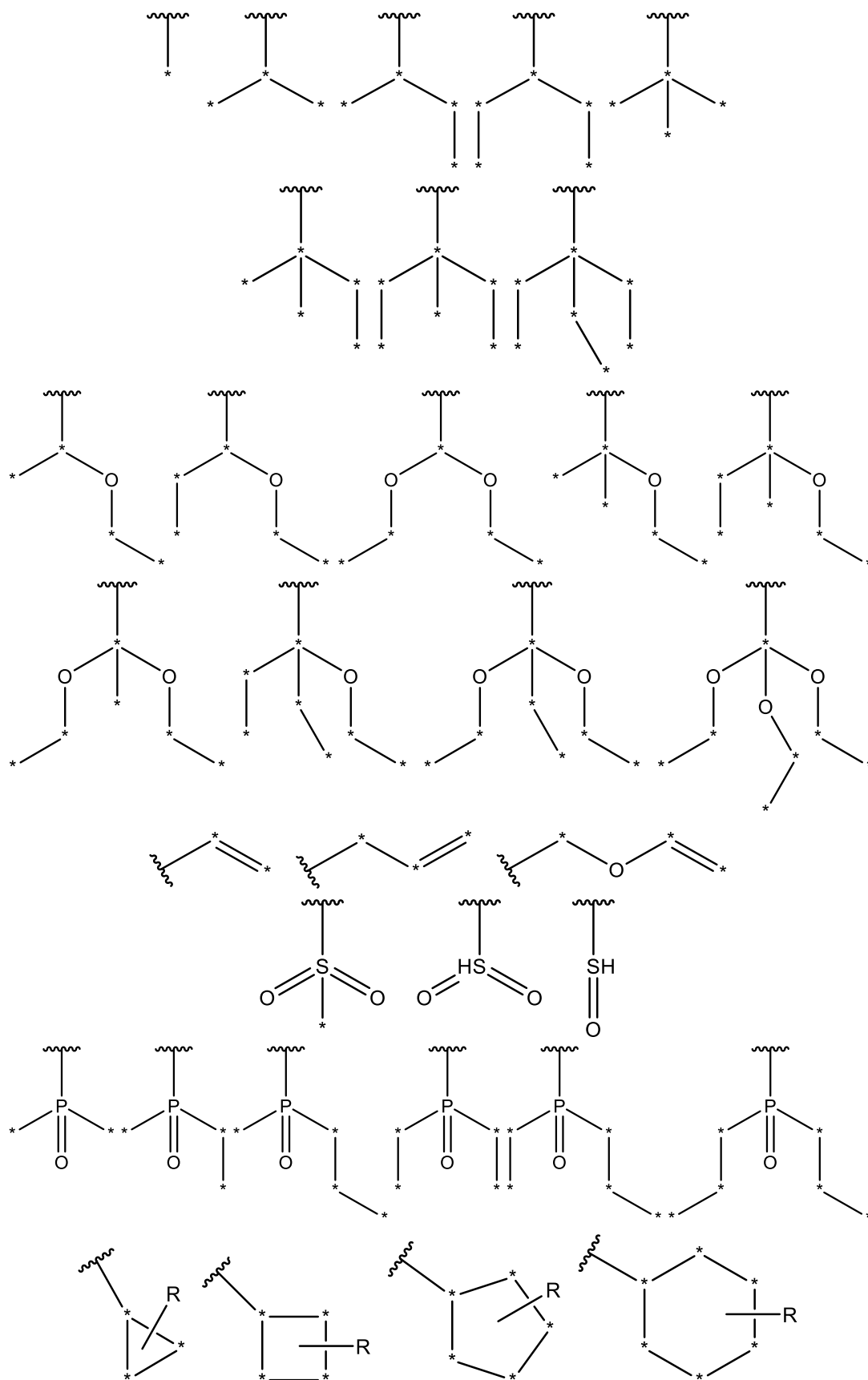
4.2.2. Terminal count

After removing the salt fragments, the number of “terminals” each molecule has is counted (**Example 4.2**). The number of terminals provides a starting point to how many fragments a molecule could be cleaved into in order to obtain the hydrophobic and hydrophilic sections. This is a procedure to reduce the computational cost when generating fragmentation patterns, especially with larger molecules. Within this protocol, terminals are defined as any groups of short branches, functional groups, rings with short branches (**Figure 4.4**). It is to note that short side branches up to 2 atoms in length or 2 atoms plus oxygen as points of attachment are disregarded as a terminal (**Figure 4.4**, R groups). This choice of criteria of terminal and short branches is due to the fact that it is unreasonable to expect short branches have any chance of partitioning themselves into a different phase (aqueous/non-aqueous) to the section they are attached to.



Example 4.2. Terminal count of **Molecule 4.1** and **4.3**.

a)



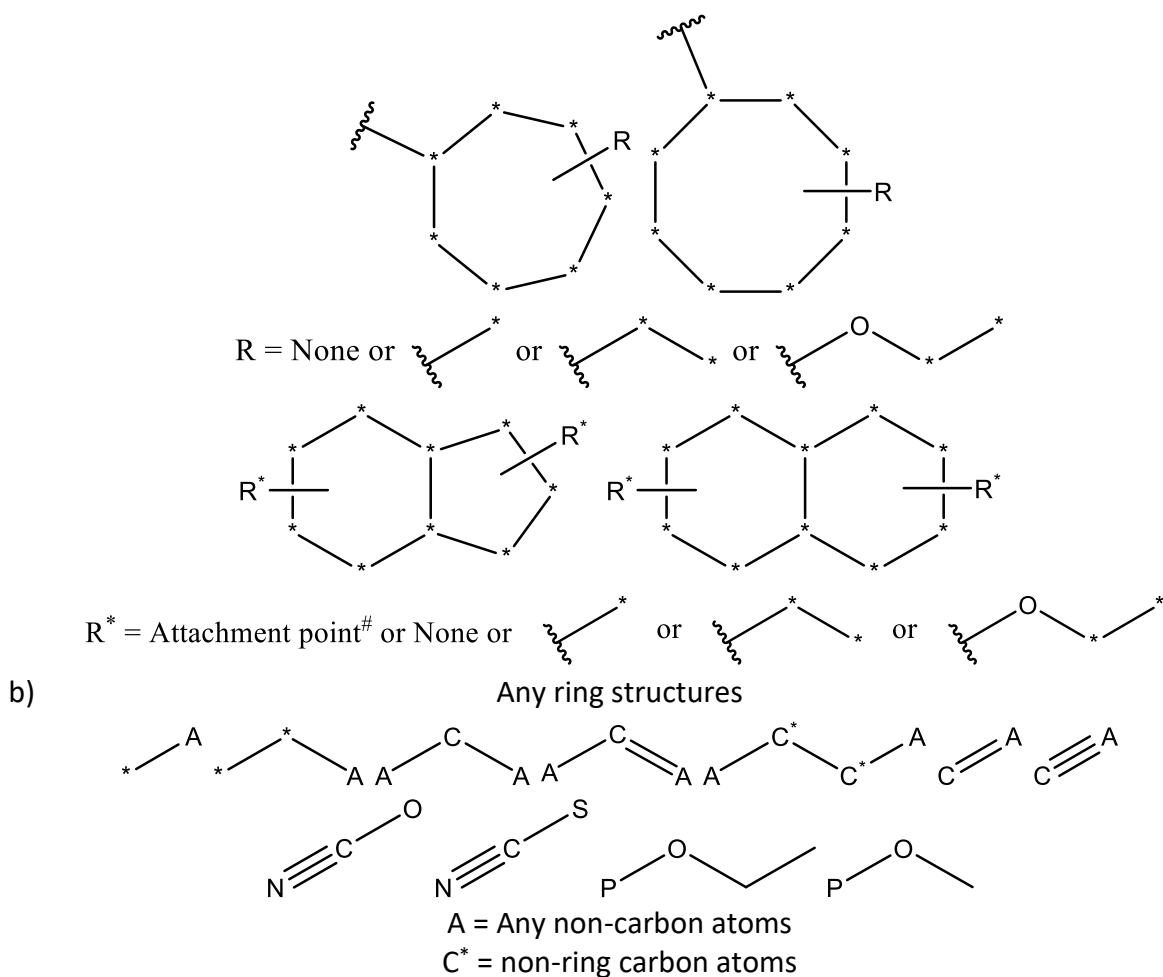
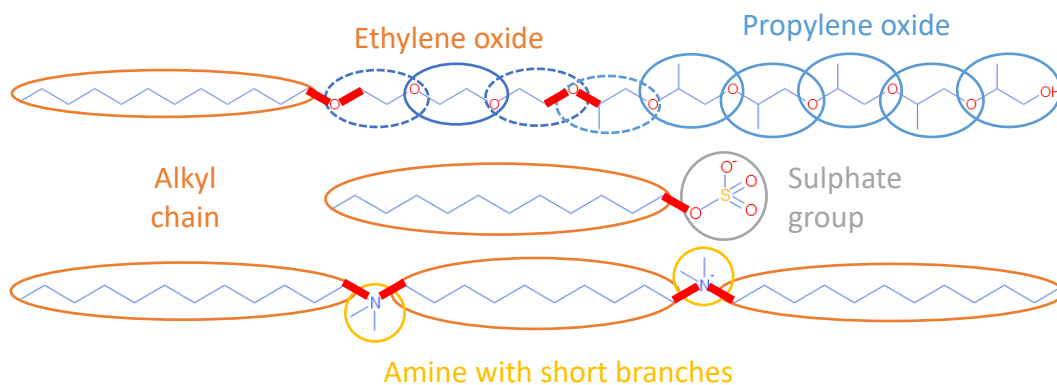


Figure 4.4. Structures of a) terminals and b) functional group substructures. Unless specified, all bonds can be single, double or triple bonds where atom valence allows. Multiple R groups can be present on the ring structures where atom valence allows. [#]Only one attachment point can be present in total for the infused ring structures to be a valid terminal.

4.2.3. Identifying possible cleavage points

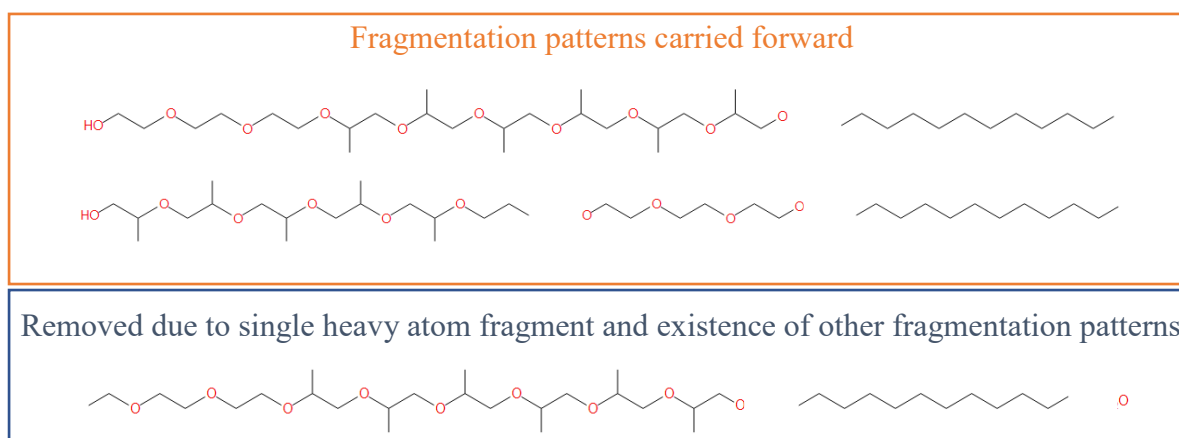
Next, possible cleavage points of the molecules are identified. Cleavage points are identified in such a way that functional groups and rings are retained where possible. In situation where without cleavage within terminal branches or a functional group leads to zero cleavage points, terminal branches and functional groups are allowed to be cleaved. Any repeat units, for example ethyl oxide or alkyl chains, are also retained except the bond connected to atoms other than the connecting repeat units (**Example 4.3**). AlogP values [10] are calculated at this point.



Example 4.3. Molecule 4.1 – 4.3 with their functional groups circled. Dotted circles indicate the repeat unit joined to atoms other than the connecting repeat units. Red highlighted bonds indicate the identified possible points of cleavage. Methyl groups of the propylene oxide repeat units are not cleaved as they are short side branches not disregarded as terminal in 4.2.2. *Terminal count*.

4.2.4. Enumerate fragmentation pattern

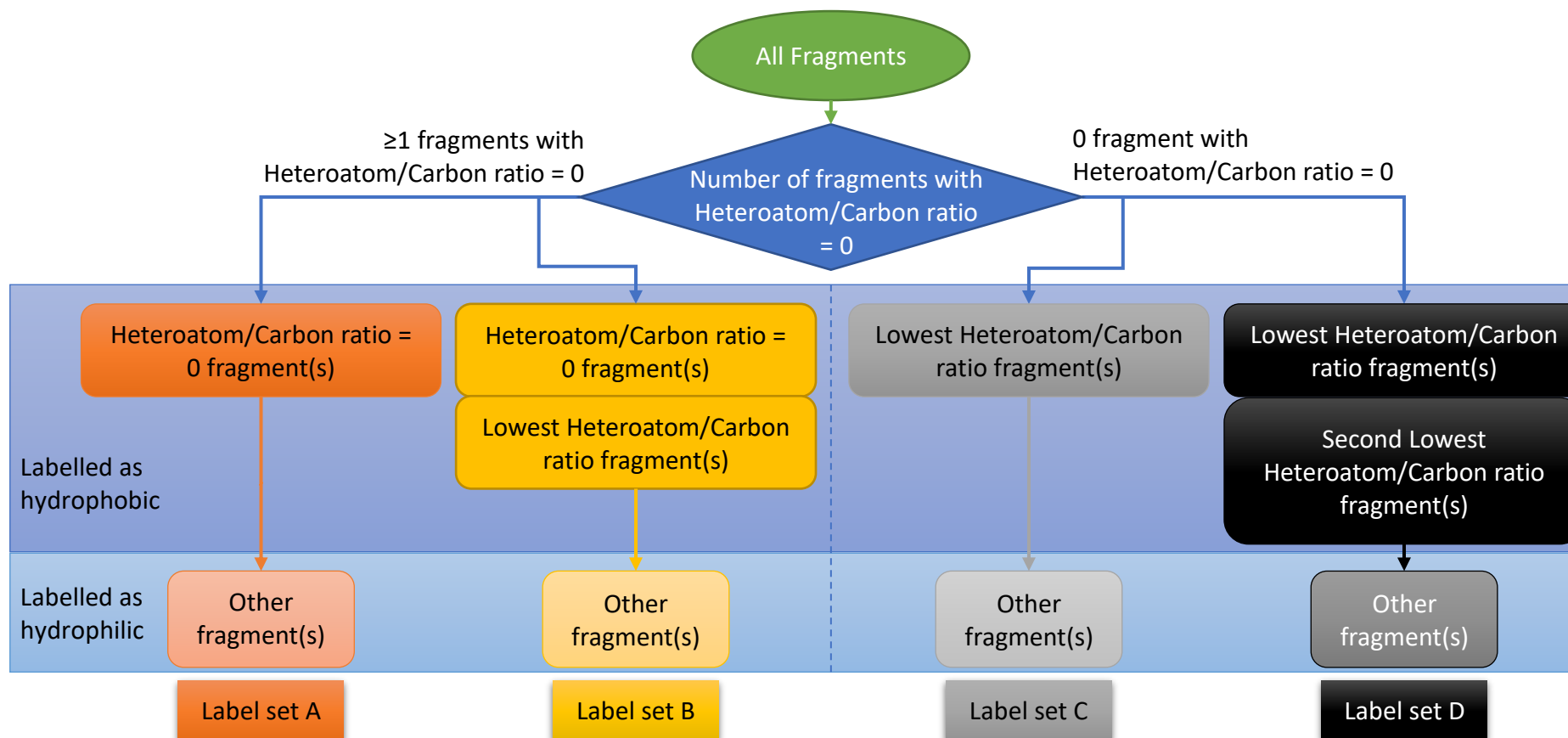
Once the possible cleavage points are identified, all possible fragmentation patterns with up to number of cleavages equal to the number of terminals counted in 4.2.2 *Terminal count* are enumerated. Any fragmentation patterns which give single heavy atom fragments are discarded, except when such fragmentation pattern is the only one possible (**Example 4.4**). The heteroatom to carbon ratio and sum of atomic AlogP score of each fragment are then calculated.



Example 4.4. Example cleavage patterns of **Molecule 4.3**.

4.2.5. Label fragments as hydrophobic/hydrophilic

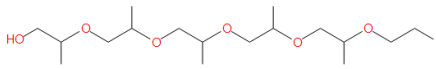
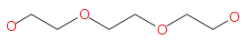
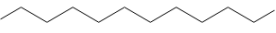
For each fragmentation pattern generated, the fragments are labelled as hydrophobic/hydrophilic following the decision tree in **Scheme 4.2 (Example 4.5)**. The decision tree is constructed in such a way that it is possible for fragments with low heteroatom to carbon atom ratio to be hydrophobic and any fragments which shares the same heteroatom to carbon ratio to have the same hydrophobic/hydrophilic label. It is noted that the labels are only valid if there are fragments labelled as hydrophobic and hydrophilic within the same fragmentation, i.e. if there are no fragments labelled as hydrophilic, the set of labels are invalid. For each fragmentation, there can only be up to two sets of labels.



Scheme 4.2. Decision tree for hydrophobic/hydrophilic label of fragments. For each fragmentation pattern, a decision is made upon whether the fragmentation pattern contains one or more fragments with a heteroatom/carbon ratio of 0 (i.e. hydrocarbons, halocarbon). If the fragmentation pattern contains one or more fragments with a heteroatom/carbon ratio of 0, two labelling options are available: Label set A – the fragments with heteroatom/carbon ratio of 0 labelled as hydrophobic, other fragments labelled as hydrophilic, Label set B – the fragments with heteroatom/carbon ratio of 0 and the fragments with lowest heteroatom/carbon ratio labelled as hydrophobic, other fragments labelled as hydrophilic. If the fragmentation pattern contains no fragments with a heteroatom/carbon ratio of 0, two labelling options are available: Label set C – the fragments with lowest heteroatom/carbon ratio labelled as hydrophobic, other fragments labelled as hydrophilic, Label set D – the fragments with lowest and second lowest heteroatom/carbon ratio labelled as hydrophobic, other fragments labelled as hydrophilic.

4.2.6. Hydrophobic/hydrophilic AlogP difference calculation

With the fragments from each fragmentation labelled, the AlogP difference between the hydrophobic and hydrophilic fragments can be calculated. In the cases where there are two labelling options available for the same fragmentation pattern, the AlogP difference for each option is calculated and the option with the largest AlogP difference is carried forward for analysis (**Example 4.5**). For each molecule, once the largest AlogP difference label set for each fragmentation pattern is determined, the fragmentation patterns are compared against each other. The labelled fragmentation pattern with the largest AlogP difference and the minimum number of fragments is selected. This helps to reserve the overall structure of the hydrophobic and hydrophilic sections for any fragment property calculation.

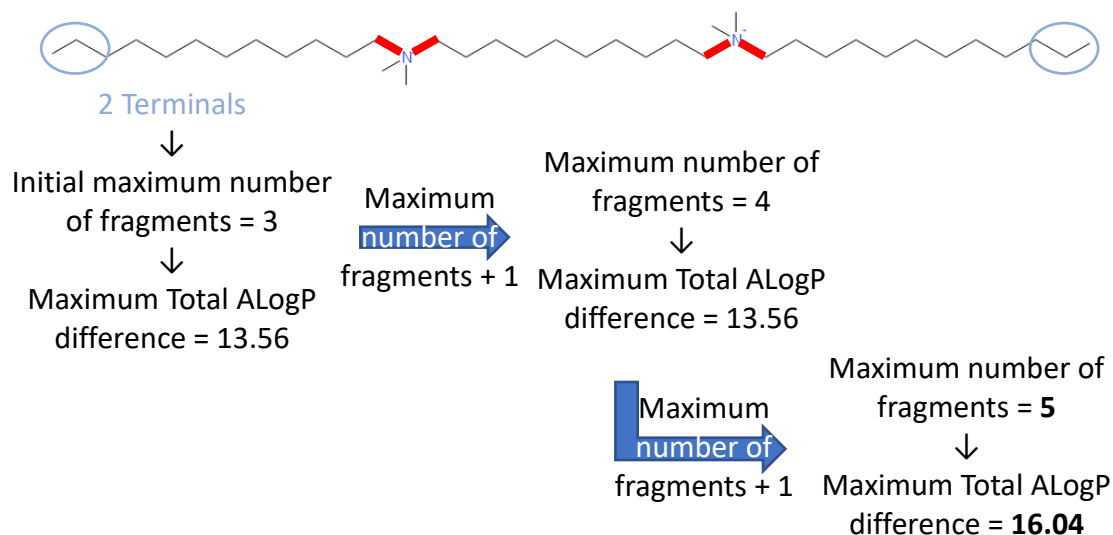
Fragment	Heteroatom/ Carbon ratio	AlogP	Label set A	Label set B
	0.33	1.17	Hydrophilic	Hydrophobic
	0.67	-0.55	Hydrophilic	Hydrophilic
	0	5.08	Hydrophobic	Hydrophobic
Total AlogP difference			4.46	5.70

Example 4.5. Example fragmentation pattern for **Molecule 4.3** and the hydrophobic/hydrophilic labelling of the fragments.

4.2.7. Loop for Increasing number of fragments

Some molecules require more fragments than the number of terminal count from 4.2.2. *Terminal count* suggests to capture their hydrophobic and hydrophilic sections. This occurs when one or more hydrophilic groups are sandwiched between hydrophobic groups, and vice versa. Therefore, 4.2.4. *Enumerate fragmentation pattern* to 4.2.6. *Hydrophobic/hydrophilic AlogP difference calculation* are looped with the maximum number of fragments increased by one each time until one of the following conditions is met:

- Maximum number of fragments allowed by the total number of possible cleavage points is achieved (**Example 4.6**)
- The same largest AlogP difference is observed three times in a row at 4.2.6. *Hydrophobic/hydrophilic AlogP difference calculation* of each loop to ensure a plateau of maximum total AlogP difference is reached

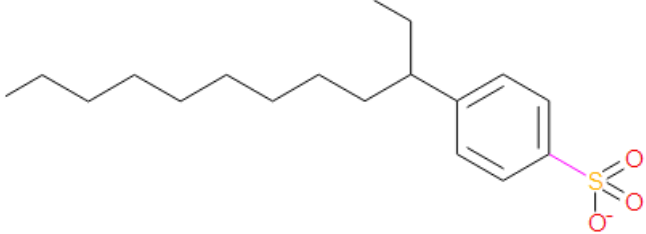
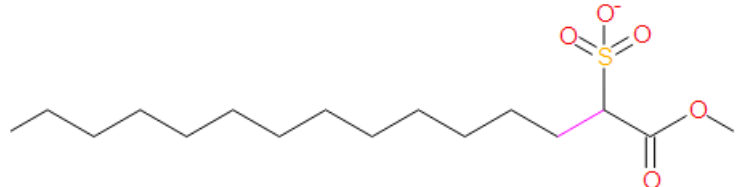


Example 4.6. increasing the number of fragments for **Molecule 4.2** in search for the fragments which capture best the hydrophobic and hydrophilic sections.

4.2.8. Export result

For the fragmentation pattern with the largest AlogP difference between hydrophobic and hydrophilic fragments and fewest cleavages at the end of 4.2.7. *Test to Increase number of fragments.* the potential amphiphilicity properties in **Table 4.1** are calculated. The fragments are exported as a SD file and the potential amphiphilicity properties exported in a .csv file. A html file is also generated capturing the potential amphiphilicity properties with the point of cleavage highlighted (**Example 4.7**).

CHAPTER 4: NOVEL SURFACTANT DESCRIPTOR – POTENTIAL AMPHIPHILICITY

Molecule	MoleculeNumber	TotalNum_Fragments	NumPhilicFragments	NumPhobicFragments
	1	2	1	1
	2	2	1	1

PhilicALogP	PhobicALogP	FragmentALogPDifference	TotalALogP	PhilicNumRotatableBonds	PhobicNumRotatableBonds	PhilicESA	PhobicESA	Salts
-2.7609	6.9789	9.7398	4.2180	0	10	140.61	306.44	[Na+]
-2.9068	5.9814	8.8882	3.0746	3	10	558.40	252.53	[Na+]

Example 4.7. Screenshot of HTML output of the potential amphiphilicity properties with point of cleavage highlighted in pink. The table had been split into two parts for clarity.

4.3. Critical Micelle Concentration (CMC), Amphiphilicity and Potential Amphiphilicity

Although the factors related to potential amphiphilicity, e.g. ALogP and number of rotatable bonds for the hydrophilic and hydrophobic sections, can be calculated using the above protocol currently there is no method to verify a calculated potential amphiphilicity value as there are no absolute reference points for amphiphilicity or potential amphiphilicity. However, as surfactant molecules are amphiphilic molecules, it is commonly accepted that typical surfactant phenomena, such as lowering interfacial tension and aggregation to form micelles, are characteristics of amphiphilic molecules [1, 2]. It is therefore reasonable to hypothesise a quantitative relationship between a quantitative amphiphilicity descriptor, such as potential amphiphilicity we propose here, and these surfactant phenomena. Critical micelle concentration (CMC) is a commonly used quantitative descriptor for surfactants. As described in *Chapter 3: Understanding Polymer Detergent Properties via QSPR method*, CMC describes the concentration where the surfactant molecules aggregate to form micelles [12], and this temperature dependent phase change due to this aggregation can be observed by the sharp change in physical properties, measurable via various different methods (e.g. surface tension, conductivity) [13]. This provides a route to evaluate the utility of the potential amphiphilicity descriptors by comparing the predictive performance of models which include these descriptors as an input to models which do not.

In addition, according to previous research within Unilever, there is reference to previous literature which defined a link between oil/water interfacial tension and CMC when looking at the free energy of micellisation [14]. This relation is important as these quantities are useful indicators of detergent performance and can be used to aid the design and selection of new, more effective and planet friendly surfactants, e.g. improved cleaning properties with smaller quantity of surfactant. To date, there are several different literature reports which indicate the use of molecular dynamic simulations to calculate interfacial energy reduction [15] and micellization [16, 17]. However, the computational cost of such simulations is still high and therefore is impractical to apply such calculations to large libraries for selection of surfactant candidates for further investigation and development. Molecular thermodynamic models are also calculatable for micellization [14, 18]. As this type of model can be evaluated quickly in comparison to molecular dynamics simulations, it is seemingly possible to use such models to perform selection of candidates from a large library, although the narrow chemical space the parameters of the models are adjusted for hinder such possibility. On the other hand, by applying QSPR models which use simple and quick to calculate molecular descriptors to identify potential candidates, the cost of the candidate selection process can potentially be greatly reduced.

4.3.1. CMC QSPR

There has been previous work in constructing CMC QSPR models using molecular properties [7, 19-22]. Often, these studies are restricted to a single type of surfactant, for example anionic surfactants or ethylene oxide surfactants only [7, 20, 23]. In this work, we aim to expand this and construct CMC QSPR models that can accommodate multiple types of surfactants, e.g. models which can predict CMC for sulfonic surfactants, ethylene oxide surfactants and amine surfactants. Nevertheless, the descriptors that were identified in the literature as important could provide some insight for choosing descriptors to calculate for our models.

Preselecting descriptors prior to QSPR model construction by hypothesising which descriptors are appropriate is an approach used in molecular thermodynamic modelling [14]. In this work, this approach is taken as the hypothesis that the potential amphiphilicity descriptors alone are

unlikely to work well unless other descriptors such as flexibility indicators (e.g. number of rotatable bonds, Kier flexibility index) are included. In addition, when evaluating the utility of the potential amphiphilicity descriptors via QSPR model comparison, it would be beneficial to select the range of descriptors which have already proven to be correlated to the observation by previous research. By using descriptors proven to be correlated, models with acceptable performance can be expected without the potential amphiphilicity descriptors when using dataset covering the same chemical space. This can then be used as a benchmark when comparing the performance of the models which includes potential amphiphilicity descriptors.

Within one of the previous studies [7], it was shown that including descriptors such as Kier and Hall connectivity index, Kier flexibility index and moment of inertia calculated for the hydrophilic and hydrophobic sections in addition to the molecular properties for the whole molecule contributes to a good CMC model by providing hydrophilic and hydrophobic specific information to the model. In the study, hydrophobic and hydrophilic sections are defined precisely for the molecule within the database [7]. In comparison, our potential amphiphilicity descriptors protocol enumerate the hydrophobic and hydrophilic fragments automatically. By including the fragment descriptors in the QSPR model construction and gauging the importance of such descriptors in well performing models, the utility of the protocol in enumerating the hydrophobic/hydrophilic sections can also be proven.

Within the previous QSPR studies, CODESSA was used in several of the studies to calculate the descriptors [7, 20, 22]. CODESSA calculated descriptors including average information content [24, 25], complementary information content [24, 25], fractional partial negative surface area [24], relative negative charge surface area [24], momentum of inertia [26], Kier & Hall connectivity index [27], Kier flexible index [24] were found to play a part in good CMC prediction [7, 20, 22, 28]. However, CODESSA is no longer available and therefore within this work, alternatives (CDK descriptor calculator [29] and alvaDesc [30]) were used to calculate the descriptors identified within those studies where possible.

4.4. Utility of Potential Amphiphilicity via QSPR

4.4.1. Surfactant library generation

In order to investigate the potential amphiphilicity descriptor and links to CMC, it was necessary to construct a surfactant library which contains various surfactant structures and their CMC. One challenging area in constructing such a library is that CMC values are method, equipment and temperature dependent. As changes in any of these experimental factors would change the accuracy of the observed value, it was necessary to record them in the library as well. In this project, a surfactant database was provided by Unilever containing CMC values from various sources, including previously published QSPRs [7, 31-35] (**Supporting Information 4.1**). Where possible, each entry was traced back to the original experimental source and experimental details including method, equipment and temperature were recorded. Any incomplete entries, i.e. where structure was missing and/or CMC value was missing were removed.

4.4.2. Construction of surfactant CMC QSPR models Part 1

4.4.2.1. Data preparation

The above database was divided into 3 subsets based on the following:

- Subset A: entries from a previous QSPR study [35], where traceable measured via surface tension at 25°C (45 entries)
- Subset B: entries with CMC measured via surface tension at 25°C (76 entries)

CHAPTER 4: NOVEL SURFACTANT DESCRIPTOR – POTENTIAL AMPHIPHILICITY

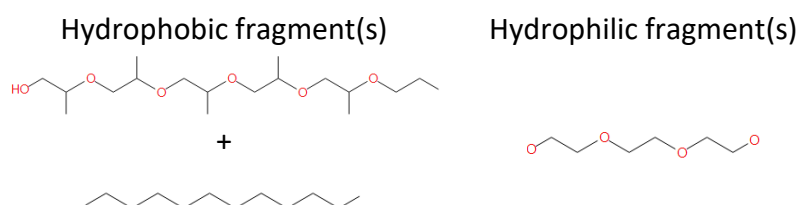
- Subset C: whole database (477 entries)

For all subsets, descriptors for each entry were calculated as follows:

- Functional class fingerprint [36] (FCFP) 2048 bits and number of rotatable bonds calculated using Pipeline Pilot 2017 [11]
- Potential amphiphilicity related properties calculated using Pipeline Pilot [11] protocol **Scheme 4.1**
- Kier & Hall connectivity index (SP), Fractional partial positive surface area (FPSA), Fractional partial negative surface area (FNSA), relative positive charge surface area (RPCS), relative negative charge surface area (RNCS), momentum of inertia (MOMI) for whole molecule and hydrophobic/hydrophilic fragments calculated using CDK descriptor calculator [29], based on the correlation proven in previous research
 - Calculated for the whole molecule and for the hydrophobic/hydrophilic fragments
 - Lowest energy 3D conformation calculated using DataWarrior [37] using default settings prior descriptor calculation
- Average information content (IC), complementary information content (CIC), Kier flexible index (PHI) calculated for whole molecule and hydrophobic/hydrophilic fragments calculated using alvaDesc [30], based on the correlation proven in previous research
 - Calculated for the whole molecule and for the hydrophobic/hydrophilic fragments

For Subset B and C, the type of salt counter ion is also included as a descriptor. In addition, for Subset C, the CMC measuring method and measurement temperature were also included as descriptors.

For the fragments, when there were 2 or more hydrophobic fragments or hydrophilic fragments, the sum of their descriptors was used (**Example 4.8**).



Number of rotatable bonds

$$23 + 11 = 34$$

9

Example 4.8. The number of rotatable bonds for the hydrophobic and hydrophilic fragments of **Molecule 4.3**

When a CMC entry was for a mixture of surfactants, the FCFP was aggregated before folding [38] into 2048 bits, the number of rotatable bonds were added together and the weighted sum of the other descriptors was taken (**Example 4.9**).

CHAPTER 4: NOVEL SURFACTANT DESCRIPTOR – POTENTIAL AMPHIPHILICITY

Example 4.9. The FCFP, hydrophilic fragment number of rotatable bonds and polar ALogP of a surfactant mixture entry (Observation ID 7 in Subset C)

Molecule	Fraction	FCFP	Hydrophilic fragment number of rotatable bonds	Hydrophilic ALogP
1	0.67	1 0 16 5 8 1872154524 -1096398038 203677720 1618154665 1186303932 -1272798659 136597326 260714409 -1276502889 -578281385 -453677277 -1133295320 -1959657804 -728595178 1175638033 -1577600103 -729123682	0	-3.3010
2	0.33	0 1 5 8 136597326 -1272798659 -1043339860 85389210 1872154524 260714409 565998553 -1143715940 136627117 -1577600103 1175638033 -55265897 -1307727540 -288711152 -16971222 339830961	3	-3.1100

CHAPTER 4: NOVEL SURFACTANT DESCRIPTOR – POTENTIAL AMPHIPHILICITY

Molecule	Fraction	FCFP	Hydrophilic fragment number of rotatable bonds	Hydrophilic ALogP
Combined	1.00	1	3	-3.2373
		0		
		16		
		5		
		8		
		1872154524		
		-1096398038		
		203677720		
		1618154665		
		1186303932		
		-1272798659		
		136597326		
		260714409		
		-1276502889		
		-578281385		
		-453677277		
		-1133295320		
		-1959657804		
		-728595178		
		1175638033		
		-1577600103		
		-729123682		
		0		
		1		
		5		
		8		
		136597326		
		-1272798659		
		-1043339860		
		85389210		
		1872154524		
		260714409		
565998553				
-1143715940				
136627117				
-1577600103				
1175638033				
-55265897				
-1307727540				
-288711152				
-16971222				
339830961				

↓
2048 bits
beginning
with 0000...

CHAPTER 4: NOVEL SURFACTANT DESCRIPTOR – POTENTIAL AMPHIPHILICITY

4.4.2.2. Data compilation

Once all descriptors were calculated, the data were imported into R [39] and each subset was separated into different datasets by the descriptor types as stated in **Table 4.2**. Note that descriptors calculated by CDK descriptor calculator and avlaDesc for hydrophobic and hydrophilic fragments are treated as an expansion to the descriptor calculated by CDK descriptor calculator and alvaDesc for the whole molecule and are never included in a dataset alone. This resulted in 20 datasets for each of the three subsets (A, B and C), giving a total of 60 datasets.

Table 4.2. The breakdown of the descriptors included for each dataset

Dataset	Potential amphiphilicity related descriptors	Functional Class fingerprint	Number of rotatable bonds for the whole molecule	Descriptors calculated by CDK descriptor calculator and alvaDesc for whole molecule	Descriptors calculated by CDK descriptor calculator and alvaDesc for hydrophobic and hydrophilic fragments
1	✓				
2		✓			
3				✓	
4				✓	✓
5	✓		✓		
6	✓	✓			
7	✓			✓	
8	✓			✓	✓
9		✓	✓		
10		✓		✓	
11		✓		✓	✓
12			✓	✓	
13			✓	✓	✓
14	✓	✓	✓		
15	✓		✓	✓	
16	✓		✓	✓	✓
17		✓	✓	✓	
18		✓	✓	✓	✓
19	✓	✓	✓	✓	
20	✓	✓	✓	✓	✓

4.4.2.3. Data Pre-processing

For each of the 60 data subsets (3 Subsets × 20 descriptor datasets), a training/test split was carried out with a 9:1 ratio using `caret::createDataPartition` to ensure the resulting training set and test set contain a similar ratio of observation values from each percentile based section of the overall observation values.

Using the training set, near zero variance predictor, predictors with low frequency ratio (*number of observation/5* or above) for the most common value over the second most common value, were removed due to the lack of information in these predictors.

Predictors with pair-wise absolute correlations over 0.9 were identified with `caret::findCorrelation`. For each predictor pair, the average correlation with the rest of the predictors was calculated and the predictor with the higher average correlation was removed to give the final version of the predictor set.

4.4.2.4. Algorithm selection

One linear (Partial least square, PLS), one non-linear (Support vector machine, SVM) and one tree-based (Random forest, RF) model algorithm were chosen to investigate their performance on the 60 data subsets, covering a simpler and more interpretable model (PLS) to more complex models (SVM, RF) for identifying the optimally performing methods [40]. In comparison to previous chapters where over five algorithms are selected, only three algorithms were selected here because the aim of the models is to prove the utility of the potential amphiphilicity descriptors protocol in providing useful descriptors and enumerating hydrophobic and hydrophilic sections. Also, apart from the potential amphiphilicity descriptors and molecular fingerprint, the descriptors are pre-selected by their correlation to CMC displayed in previous studies. Therefore, it is not necessary to search over a large number of algorithms for this comparative study. In an attempt to obtain more information on predictor-observation relationship and possible persistent outliers, defined as observation entries which are highlighted as outliers in multiple models and/or have extremely high RMSE in comparison to other entries, ten models using each modelling algorithm were built for each of the 60 data subsets, resulting in a total of 1800 models.

4.4.2.5. Model building and performance assessment

Model construction with the default tuning parameters and 10-fold cross validation on the training set for each predictor data sets (see 4.4.2.2. *Data compilation*) was performed in order to optimise the hyperparameters for each modelling algorithm against R^2 . For PLS, the predictors were centred and scaled within `caret::train()` using the `preProcess` option. The models were then tested using the test set.

Model performance was assessed using the following criteria suggested by Golbraikh and Tropsha and we defined that all needed to be fulfilled for a regression model to be deemed acceptable [41]:

- Cross-validated R^2 via internal resampling on training set > 0.5
- R^2 on test set > 0.6
- R^2 through origin (R_0^2) close to R^2
- $\frac{R^2 - R_0^2}{R^2} < 0.1$ or $\frac{R^2 - R_0'^2}{R^2} < 0.1$, where $R_0'^2$ is R_0^2 when observation and prediction are inversed
- And the corresponding $0.85 \leq k \leq 1.15$ or $0.85 \leq k' \leq 1.15$

where k is the gradient of the observation vs. prediction line of best fit and k' is the gradient of the prediction vs. observation line of best fit (See 1.2.7.1. *Regression performance metrics* for full definition). In addition to the above criteria, a more recent literature by Alexander, Tropsha and Winkler emphasised the importance of RMSE and suggested that for a predictive model, the following criteria needs to be fulfilled [42]:

- High R^2 on test set
- Low RMSE of test set predictions

Generally, R^2 on training set is higher than R^2 on test set as the data is seen by the model during construction. However, in the cases where R^2 on training set is calculated based on the model during the cross-validation stage, it is possible that the R^2 on training set is lower than the R^2 on test set as the hold-out data during the cross-validation stage is not seen by the model during construction.

These criteria were also adopted in order to identify the best models overall using Pareto sort for optimisation [43] (see 4.4.2.7. *Selection of best performing model*). With the above criteria in mind, Spearman's rank correlation coefficient (ρ) was also considered for model performance. Spearman's rank correlation coefficient is used to measure the strength and the direction of the association between the observation and prediction, which can have a value between -1 to 1 [44]. By looking at the absolute value, ρ is commonly interpreted as follows:

- < 0.20 = poor agreement
- $0.20 - 0.40$ = fair agreement
- $0.40 - 0.60$ = moderate agreement
- $0.60 - 0.80$ = good agreement
- $0.80 - 1.00$ = very good agreement

Models with moderate agreement or above are usually considered to be good. When a Spearman test is carried out to calculate ρ , an associated probability value (p-value) which measures the likelihood of the observed correlation is due to chance is also calculated. The closer the p-value is to zero, the more likely the observed correlation is not due to chance and therefore the more likely the correlation is valid.

4.4.2.6. Model robustness

Validation of the models for each of the data subset was carried out by random shuffling of the observations before training using `base::sample()`. This was repeated ten times for each of the built original models. Using the training set R^2 of the original models and these y-randomised models, the statistical significance of the original models for the training set was evaluated with the standard hypothesis testing method. Specifically, the robustness of the models was examined using the Z score statistics following **Equation 4.1** [45].

$$Z = \frac{\text{training } R_{\text{original}}^2 - \text{Avg}(\text{training } R_{\text{y-randomised}}^2)}{SD(\text{traininng } R_{\text{y-randomised}}^2)} \quad (\text{Equation 4.1})$$

If the original model was valid, the overall performance of y-randomised models should be greatly reduced in comparison, resulting in a high Z score. Models with Z scores of over 3 are considered as significant [45].

4.4.2.7. *Selection of best performing model*

Taking account of the literature and Z score, the follow modified criteria for validity was used to assess the successes of the tuned models in terms of their predictability and understanding the trend within the data:

- On training set
 - Cross-validated $R^2 > 0.5$
 - RMSE < 0.5
- On test set
 - (Adjusted) $R^2 > 0.6$
 - (Adjusted) R_0^2 close to R^2
 - $\left| \frac{R^2 - R_0^2}{R^2} \right| < 0.1$
 - RMSE < 0.35
 - Slope of R_0^2 regression line: $0.85 \geq k \geq 1.15$
 - $\rho > 0.80$
 - Z score > 3

Good performing models are identified by the high number of the above criteria fulfilled, and the best performing models are selected by putting the well performing models through a Pareto sort to find the models with the optimum performance across all criteria. Pareto optimisation is a multi-objective optimisation algorithm which identifies the models with the optimum performance where the criteria are maximised or minimised, i.e. maximum R^2 , minimum RMSE, etc [43]. For k, the Pareto optimisation is set to minimise $|1 - k|$. The optimal models are defined as those for which it is not possible to improve on one criterion without degrading at least one other.

Once the optimal models are identified, the variable importance is calculated using `caret::varImp`. The importance score calculation is dependent on the model algorithm; for PLS, it is based on the weighted sums of the absolute regression coefficients; for RF, it is based on the prediction difference for the out-of-bag data during cross validation with the variable permuted; for SVM, it is based on the relationship between each variable and the prediction.

4.4.2.8. *Constructed models*

After consolidating the performance of the models constructed using the same data subset and algorithm, it was found that only the models constructed using Subset A had excellent performance (**Table 4.3**). It was also seen that the performance drops as the data subset progress from Subset A, B to C (**Table 4.3 – 4.5**).

CHAPTER 4: NOVEL SURFACTANT DESCRIPTOR – POTENTIAL AMPHIPHILICITY

Table 4.3. Average performance of the models constructed using Subset A (Supporting Information 4.2)

Model No.	Dataset	Modelling Algorithm	Training R ² *	Training RMSE*	Test R ²	R ² – R ₀ ²	$\frac{ R^2 - R_0^2 }{R^2}$	Test RMSE	k	ρ	Z Score	Criteria fulfilled
A1	1	PLS	0.945	0.216	0.973	0.024	0.025	0.212	0.924	1.000	15.770	9
A2	1	RF	0.969	0.222	0.990	0.008	0.009	0.171	0.956	0.980	14.625	9
A11	4	RF	0.957	0.294	0.976	0.020	0.020	0.273	0.908	0.960	11.475	9
A13	5	PLS	0.934	0.231	0.970	0.028	0.028	0.216	0.922	1.000	11.025	9
A14	5	RF	0.965	0.227	0.988	0.010	0.010	0.183	0.953	0.960	12.384	9
A16	6	PLS	0.904	0.269	0.974	0.024	0.024	0.202	0.907	0.960	11.039	9
A17	6	RF	0.954	0.288	0.967	0.028	0.029	0.273	0.902	0.980	13.254	9
A20	7	RF	0.963	0.270	0.985	0.011	0.012	0.238	0.925	0.960	12.646	9
A23	8	RF	0.960	0.297	0.974	0.022	0.023	0.269	0.920	0.960	10.103	9
A32	11	RF	0.952	0.304	0.974	0.022	0.023	0.273	0.918	0.940	10.599	9
A38	13	RF	0.961	0.292	0.978	0.018	0.018	0.270	0.907	0.960	10.686	9
A40	14	PLS	0.913	0.278	0.964	0.032	0.034	0.215	0.921	0.960	15.046	9
A47	16	RF	0.961	0.294	0.975	0.020	0.021	0.269	0.930	0.960	5.159	9
A56	19	RF	0.962	0.266	0.988	0.010	0.010	0.224	0.924	0.960	10.609	9
A59	20	RF	0.960	0.293	0.975	0.020	0.021	0.279	0.901	0.960	10.958	9

Bold Model No.: Criteria optimised models

Bold: Criteria fulfilled

*Training R² and Training RMSE extracted from average of 10-fold cross validation

CHAPTER 4: NOVEL SURFACTANT DESCRIPTOR – POTENTIAL AMPHIPHILICITY

Table 4.4. Average performance of the models constructed using Subset B (Supporting Information 4.3)

Model No.	Dataset	Modelling Algorithm	Training R ² *	Training RMSE*	Test R ²	R ² – R ₀ ²	$\frac{ R^2 - R_0^2 }{R^2}$	Test RMSE	k	ρ	Z Score	Criteria fulfilled
B3	1	SVM	0.280	2.358	0.538	0.345	0.641	1.913	0.194	0.32	3.228	1
B16	6	PLS	0.507	1.923	0.488	0.438	0.898	1.630	0.398	0.48	9.334	2
B17	6	RF	0.449	2.021	0.417	0.474	1.138	1.919	0.229	0.66	7.040	1
B18	6	SVM	0.475	1.956	0.616	0.312	0.507	1.384	0.445	0.56	8.147	2
B29	10	RF	0.362	2.194	0.538	0.360	0.668	1.900	0.293	0.52	3.551	1
B31	11	PLS	0.518	1.909	0.567	0.365	0.643	1.558	0.456	0.4	3.567	2
B33	11	SVM	0.443	2.018	0.582	0.349	0.601	1.516	0.443	0.6	5.871	1
B41	14	RF	0.442	2.024	0.415	0.477	1.150	1.917	0.226	0.6	5.264	1
B42	14	SVM	0.483	1.947	0.588	0.340	0.578	1.395	0.449	0.56	3.980	1
B50	17	RF	0.371	2.189	0.563	0.337	0.599	1.854	0.303	0.52	4.604	1
B52	18	PLS	0.517	1.921	0.567	0.365	0.642	1.557	0.456	0.42	3.808	2
B55	19	PLS	0.512	1.913	0.574	0.365	0.636	1.535	0.475	0.5	6.679	2
B57	19	SVM	0.471	1.963	0.619	0.313	0.506	1.371	0.440	0.7	6.329	2
B58	20	PLS	0.517	1.895	0.567	0.366	0.645	1.552	0.460	0.5	6.535	2
B60	20	SVM	0.461	1.985	0.566	0.362	0.640	1.494	0.413	0.66	4.797	1

Bold: Criteria fulfilled

*Training R² and Training RMSE extracted from average of 10-fold cross validation

CHAPTER 4: NOVEL SURFACTANT DESCRIPTOR – POTENTIAL AMPHIPHILICITY

Table 4.5. Average performance of the models constructed using Subset C (Supporting Information 4.4)

Model No.	Dataset	Modelling Algorithm	Training R ² *	Training RMSE*	Test R ²	R ² - R ₀ ²	$\frac{ R^2 - R_0^2 }{R^2}$	Test RMSE	k	ρ	Z Score	Criteria fulfilled
C4	2	PLS	0.261	2.373	0.633	0.292	0.461	1.621	0.349	0.825	37.751	3
C5	2	RF	0.244	2.356	0.545	0.374	0.686	1.627	0.307	0.785	34.779	1
C6	2	SVM	0.285	2.361	0.624	0.296	0.475	1.578	0.316	0.825	40.737	3
C9	3	SVM	0.244	2.559	0.220	0.492	2.238	2.738	0.220	-0.018	36.295	1
C13	5	PLS	0.180	2.911	0.179	0.391	2.186	4.973	-0.127	0.587	23.181	1
C15	5	SVM	0.181	2.685	0.220	0.479	2.177	2.809	0.196	0.307	24.196	1
C30	10	SVM	0.461	2.166	0.375	0.389	1.037	2.470	0.383	0.533	17.832	1
C31	11	PLS	0.421	2.289	0.350	0.396	1.131	2.616	0.431	0.654	53.469	1
C33	11	SVM	0.486	2.115	0.406	0.369	0.908	2.408	0.425	0.579	27.356	1
C36	13	SVM	0.243	2.560	0.216	0.495	2.286	2.741	0.218	-0.022	29.916	1
C51	17	SVM	0.459	2.171	0.379	0.386	1.018	2.464	0.387	0.531	48.959	1
C52	18	PLS	0.421	2.294	0.350	0.396	1.131	2.617	0.431	0.647	32.513	1
C54	18	SVM	0.482	2.122	0.408	0.367	0.899	2.402	0.427	0.579	53.963	1
C57	19	SVM	0.475	2.136	0.395	0.376	0.953	2.430	0.411	0.413	61.919	1
C60	20	SVM	0.486	2.114	0.411	0.366	0.892	2.395	0.430	0.458	54.243	1

Bold: Criteria fulfilled

*Training R² and Training RMSE extracted from average of 10-fold cross validation

A Pareto sort of the 19 models constructed using Subset A which fulfil all nine criteria showed that **Model A2** was the optimum performing model (Test $R^2 = 0.99$, RMSE = 0.17, $\rho = 0.98$, $Z = 14.6$).

4.4.2.9. Predictor importance

On analysis of **Model A2**, it was found that out of the five descriptors (PhobicNumRotatableBonds, PhobicALogP, PhobicESA, PhilicALogP, PhilicNumRotatableBonds), only the hydrophobic predictors were deemed to be important in their relation to CMC (**Figure 4.5**). Out of the three hydrophobic descriptors, PhobicNumRotatableBonds (importance = 98.64 ± 2.62) was deemed to be the most important, followed by PhobicALogP (importance = 87.46 ± 10.56) and PhobicESA (importance = 62.59 ± 9.04). This is an unexpected phenomenon as the hydrophobic and hydrophilic balance is generally considered to be important to the amphiphilicity of a molecule. On analysis of the descriptors and structures of the surfactants in Subset A, the following hypothesis was drawn to in attempt to rationalise this observation. The surfactants in Subset A are all analogues of non-ionic ethylene oxide surfactant analogues. Although 11 out of 45 surfactants contains a phenol ring, they are all flexible straight chain surfactants. When comparing the size of hydrophobic and hydrophilic sections, the hydrophobic sections are generally the same size or smaller than the hydrophilic sections. However, during micelle formation, the hydrophilic sections are able to freely extend outwards towards the aqueous phase while the hydrophobic sections would cluster together towards the centre of the micelle (**Figure 4.6**). Therefore, the flexibility of the hydrophobic sections, indicated by the PhobicNumRotatableBonds, is key to how well the hydrophobic sections can cluster together to form the micelle, followed by the PhobicALogP which is an important factor in defining the concentration of the hydrophobic sections clustering together to form micelles. In addition, the absolute number and the range of the AlogP for the hydrophobic sections are larger than the hydrophilic sections.

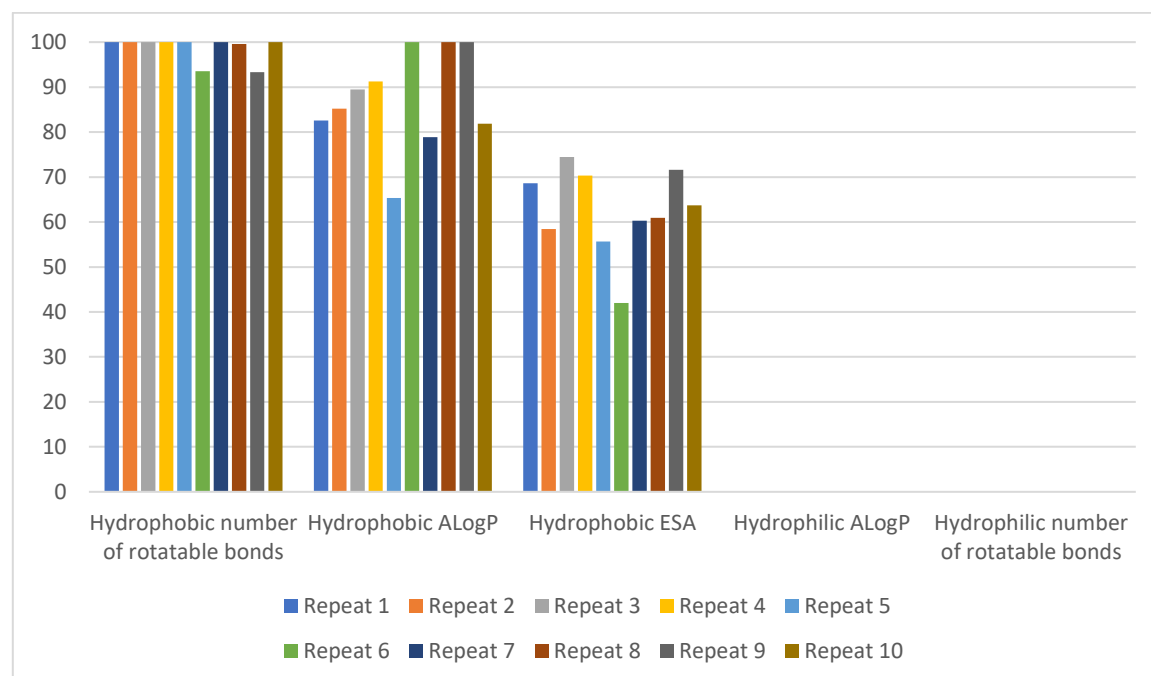


Figure 4.5. The predictor importance for all ten repeats of **Model A2**. For all ten repeats, the descriptors with importance are all descriptors for the hydrophobic sections.

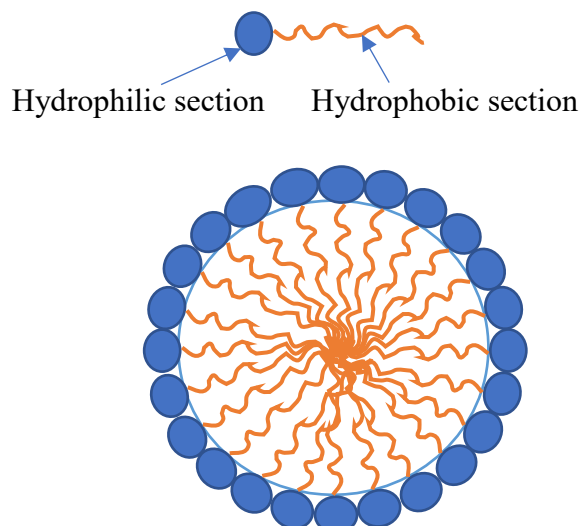


Figure 4.6. Possible position of hydrophobic and hydrophilic sections of a surfactant in a micelle.

4.4.3. Construction of surfactant CMC QSPR models Part 2

As there were no good models with $RMSE < 1$ on the training or test set for Subset B and C, several modifications to the QSPR model construction process were made. First, Subset C was split into the following subsets:

- Subset D: ionic surfactants only, defined by the presence of salt counter ions (399 entries)
- Subset E: non-ionic surfactants, defined by the absence of salt counter ions (78 entries)

Both subsets were then filtered and any entries with unknown source or measurement temperature were removed. Two different approaches were taken in an attempt to construct good models. One method was to start with the whole database and prune the entries which have $RMSE > 1$ using the model with the highest cross validated R^2 . The modelling and pruning were repeated until no entries in the model with the highest cross validated R^2 have $RMSE > 1$.

The other method was to cluster the whole database based on their similarity calculated using DataWarrior [37] and build the QSPR models from the largest clusters, remove persistent outliers defined as any entries with $RMSE > 1$ and add the next largest cluster for modelling. This was repeated until all entries had at least been added for modelling once. Different to previously, only one repeat of the model was constructed for each method before pruning/addition of entries. We took the first approach in priority and used the second method to verify the pruned entries as outliers. As a result, 210 entries were left in Subset D and 40 entries remaining in Subset E, and the final models constructed using the first method were taken for analysis.

Once outliers for Subset D and E were identified and pruned, pruned Subset D and E were combined to give Subset F (pruned Subset C, 250 entries). Models were constructed using Subset F to test the ability of the descriptor datasets in constructing QSPR models for ionic and non-ionic surfactants combined.

4.4.3.1. *Constructed models*

Using the pruned data sets 180 models were constructed in total (**Table 4.6 – 4.8**). Out of the 60 models constructed for the pruned Subset D, one model fulfilled all criteria (**Model D60**, Test $R^2 = 0.95$, RMSE = 0.35, $\rho = 0.90$, $Z = 79.3$) while three criteria optimised models (**Model D33, D51 and D57**, Test $R^2 > 0.94$, RMSE < 0.36, $\rho > 0.89$, $Z > 30.0$), fulfilling eight out of nine criteria, were found through Pareto sort of all metrics after inspection. It is to note that FCFP are included in all these models. In addition, SVM constructed most of the models with optimised criteria.

For pruned Subset E, four models were found to fulfil all criteria (**Model E2, E14, E20 and E44**, Test $R^2 > 0.99$, RMSE < 0.20, $\rho = 1$, $Z > 7.3$), within which two models (**Model E14 and E20**, Test $R^2 > 0.99$, RMSE < 0.20, $\rho = 1$, $Z > 11.5$) were found to be the most optimal through Pareto sorting. Caution needs to be taken when analysing these models as the high performance can be due to the small size of the Subset (40 entries) and hence the small coverage of chemical space. It is to note that these models all contain potential amphiphilicity descriptors. Different to pruned Subset D, RF constructed most models of the with the best performance.

Using the Subset F, one model was found to fulfil eight out of nine criteria (**Model F57**, Test $R^2 = 0.92$, RMSE = 0.34, $\rho = 0.86$, $Z = 59.2$), while two models fulfilling seven out of nine criteria were found to have their performance optimised against the criteria on inspection of the Pareto sort (**Model F17 and F33**, Test $R^2 > 0.93$, RMSE < 0.50, $\rho > 0.88$, $Z > 59.2$). However, it is noted that none of these models fulfil the training RMSE < 0.5 criteria, and only the model fulfilling eight out of nine criteria fulfils the test RMSE < 0.35 criteria. In comparison to the top models from Subset D and E, the performance of these models has decreased. It is to note that the better performing models for Subset F are constructed using RF or SVM.

CHAPTER 4: NOVEL SURFACTANT DESCRIPTOR – POTENTIAL AMPHIPHILICITY

Table 4.6. Performance of the top 15 models identified by Pareto sort, constructed using ionic surfactants (pruned Subset D) only (**Supporting Information 4.5**)

Model No.	Dataset	Modelling Algorithm	Training R ² *	Training RMSE*	Test R ²	R ² – R ₀ ²	$\frac{ R^2 - R_0^2 }{R^2}$	Test RMSE	k	ρ	Z Score	Criteria fulfilled
D12	4	SVM	0.883	0.574	0.893	0.098	0.109	0.498	0.886	0.762	67.480	5
D18	6	SVM	0.805	0.731	0.928	0.065	0.070	0.448	0.837	0.911	62.878	6
D26	9	RF	0.842	0.668	0.890	0.101	0.113	0.521	0.864	0.950	27.906	5
D27	9	SVM	0.783	0.794	0.924	0.067	0.073	0.474	0.784	0.857	28.336	6
D29	10	RF	0.849	0.653	0.862	0.126	0.146	0.565	0.882	0.968	71.089	5
D30	10	SVM	0.887	0.544	0.951	0.044	0.047	0.345	0.897	0.929	88.653	8
D33	11	SVM	0.929	0.455	0.944	0.051	0.055	0.360	0.917	0.887	29.927	8
D42	14	SVM	0.805	0.694	0.954	0.042	0.044	0.341	0.886	0.932	66.237	8
D50	17	RF	0.848	0.654	0.855	0.132	0.155	0.580	0.880	0.968	88.653	5
D51	17	SVM	0.888	0.538	0.951	0.045	0.047	0.346	0.897	0.929	98.795	8
D53	18	RF	0.868	0.605	0.784	0.197	0.252	0.720	0.852	0.971	29.927	5
D54	18	SVM	0.920	0.478	0.948	0.047	0.050	0.353	0.903	0.887	100.218	8
D55	19	PLS	0.786	0.760	0.822	0.162	0.197	0.651	0.870	0.922	46.804	5
D57	19	SVM	0.887	0.538	0.957	0.039	0.041	0.323	0.911	0.953	71.975	8
<i>D60</i>	20	SVM	0.928	0.461	0.949	0.046	0.049	0.346	0.910	0.899	79.332	9

Bold Model No.: Optimal models

Italic Model No.: Model fulfilling most criteria

Bold: Criteria fulfilled

*Training R² and Training RMSE extracted from average of 10-fold cross validation

CHAPTER 4: NOVEL SURFACTANT DESCRIPTOR – POTENTIAL AMPHIPHILICITY

Table 4.7. Performance of the top 15 models identified by Pareto sort, constructed using non-ionic surfactants (pruned Subset E) only (**Supporting Information 4.6**)

Model No.	Dataset	Modelling Algorithm	Training R ² *	Training RMSE*	Test R ²	R ² – R ₀ ²	$\frac{ R^2 - R_0^2 }{R^2}$	Test RMSE	k	ρ	Z Score	Criteria fulfilled
<i>E2</i>	1	RF	0.927	0.438	0.999	2.98E-04	2.99E-04	0.144	1.089	1	11.417	9
<i>E14</i>	5	RF	0.945	0.420	0.998	9.30E-04	9.32E-04	0.197	1.127	1	12.098	9
E17	6	RF	0.907	0.516	0.999	7.06E-04	7.07E-04	0.094	1.052	1	12.101	8
E18	6	SVM	0.925	0.758	0.999	6.78E-03	6.78E-03	0.308	0.790	0.8	12.227	7
<i>E20</i>	7	RF	0.928	0.448	0.995	3.76E-03	3.78E-03	0.152	1.043	1	11.542	9
E23	8	RF	0.879	0.597	0.998	2.04E-03	2.05E-03	0.071	0.976	0.8	6.025	8
E32	11	RF	0.854	0.713	0.999	1.00E-02	1.00E-02	0.379	0.744	0.8	4.728	6
E41	14	RF	0.894	0.532	0.998	1.37E-03	1.38E-03	0.080	1.036	1	8.470	8
E42	14	SVM	0.906	0.792	0.999	6.82E-03	6.83E-03	0.319	0.783	0.8	6.025	7
<i>E44</i>	15	RF	0.927	0.454	0.995	3.31E-03	3.33E-03	0.166	1.054	1	7.301	9
E47	16	RF	0.890	0.624	0.998	2.03E-03	2.03E-03	0.070	0.986	0.8	1.851	7
E53	18	RF	0.914	0.731	0.999	1.22E-02	1.22E-02	0.410	0.728	0.8	6.703	6
E55	19	PLS	0.884	0.708	0.917	7.13E-02	7.78E-02	0.414	0.961	1	12.101	7
E56	19	RF	0.908	0.569	0.999	2.48E-04	2.48E-04	0.151	1.070	1	12.227	8
E59	20	RF	0.861	0.730	0.996	1.59E-03	1.60E-03	0.165	1.077	0.8	8.649	8

Bold Model No.: Criteria optimised models

Italic Model No.: Model fulfilling most criteria

Bold: Criteria fulfilled

*Training R² and Training RMSE extracted from average of 10-fold cross validation

CHAPTER 4: NOVEL SURFACTANT DESCRIPTOR – POTENTIAL AMPHIPHILICITY

Table 4.8. Performance of the top 15 models identified by Pareto sort, constructed using Subset F (Supporting Information 4.7)

Model No.	Dataset	Modelling Algorithm	Training R ² *	Training RMSE*	Test R ²	R ² - R ₀ ²	$\frac{ R^2 - R_0^2 }{R^2}$	Test RMSE	k	ρ	Z Score	Criteria fulfilled
F4	2	PLS	0.546	3.163	0.988	0.001	0.001	1.796	1.408	0.800	36.513	6
F5	2	RF	0.616	2.682	0.930	0.002	0.002	1.966	1.342	1.000	54.534	6
F6	2	SVM	0.637	2.678	0.984	0.004	0.004	1.167	1.232	0.800	51.736	6
F14	5	RF	0.828	0.773	0.926	0.065	0.070	0.497	0.874	0.850	69.155	7
F17	6	RF	0.844	0.713	0.925	0.066	0.071	0.492	0.909	0.963	59.157	7
F18	6	SVM	0.850	0.819	0.955	0.039	0.040	0.402	0.879	0.898	46.577	7
F23	8	RF	0.861	0.706	0.891	0.096	0.108	0.595	0.855	0.890	58.288	6
F33	11	SVM	0.905	0.715	0.960	0.034	0.036	0.398	0.881	0.881	59.891	7
F41	14	RF	0.845	0.748	0.921	0.069	0.075	0.504	0.904	0.970	55.919	7
F47	16	RF	0.868	0.694	0.900	0.088	0.098	0.571	0.855	0.890	45.488	7
F54	18	SVM	0.892	0.737	0.962	0.032	0.034	0.400	0.868	0.881	69.155	7
F56	19	RF	0.804	0.826	0.893	0.094	0.105	0.590	0.859	0.970	31.624	6
<i>F57</i>	19	SVM	0.834	0.864	0.971	0.025	0.026	0.344	0.886	0.863	59.157	8
F59	20	RF	0.866	0.699	0.902	0.086	0.095	0.563	0.865	0.968	44.654	7
F60	20	SVM	0.899	0.718	0.957	0.037	0.039	0.395	0.890	0.892	45.714	7

Bold Model No.: Criteria optimised models

Italic Model No.: Model fulfilling most criteria

Bold: Criteria fulfilled

*Training R² and Training RMSE extracted from average of 10-fold cross validation

4.4.3.2. Predictor importance

4.4.3.2.1. Pruned Subset D

On inspection of the predictor importance of **Model D60** (**Supporting Information 4.8**), Phobic_MOMI.Y, Phobic_FPSA.2, FragmentALogPDifference, Phobic_SP.7, SP.1, MOMI.R, IC0, FPSA.2 and SP.7 were found to have an importance of over 70 (Importance = 100.00, 96.64, 96.24, 92.62, 84.79, 78.03, 77.36, 73.53 and 71.50 respectively, **Figure 4.7**). It is again interesting to see the several hydrophobic descriptors (Phobic_MOMI.Y, Phobic_FPSA.2 and Phobic_SP.7) have high importance while the most important hydrophilic descriptor is PhilicALogP (Importance = 15.27, **Figure 4.7**). However, different to **Model A2**, the difference between the hydrophobic and hydrophilic sections of the surfactants are accounted through the FragmentALogPDifference descriptor (Importance = 96.24).

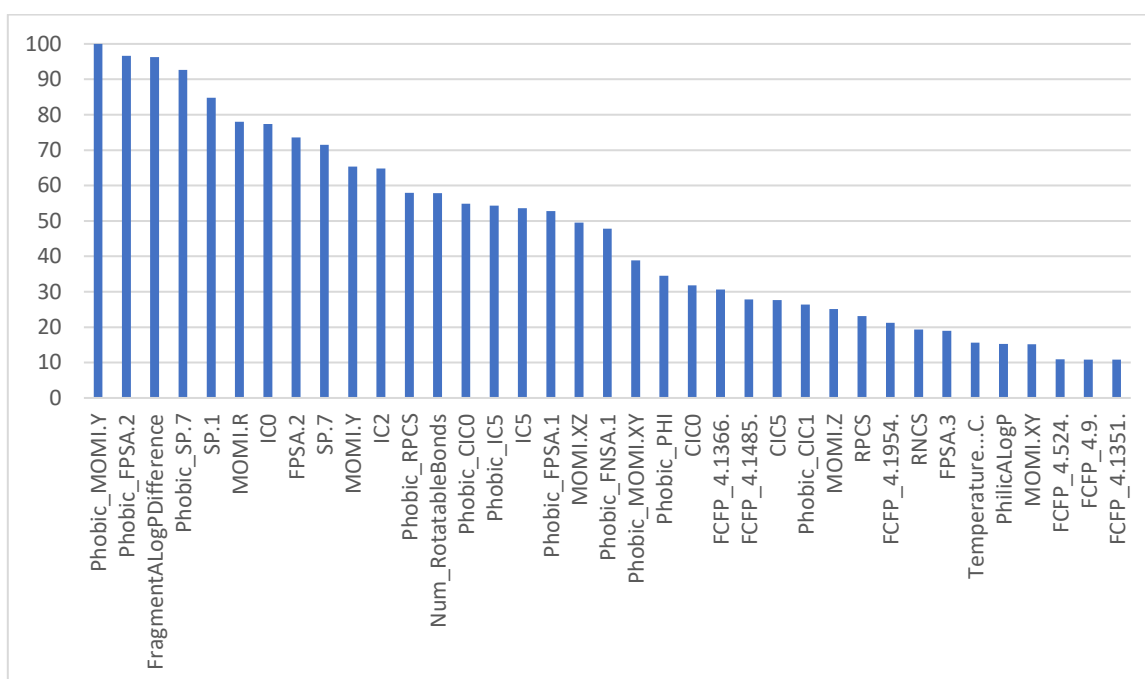


Figure 4.7. The predictor importance of **Model D60** with an importance of over 10.00. Out of the 37 predictors with an importance of over 10.00, nine of the predictors are identified with high importance (importance > 70.00).

When considering the criteria optimised models (**Model D33**, **D51** and **D57**) in addition to **Model D60**, it was found that SP.1, MOMI.R, IC0 and FPSA.2 have importance of over 70 across the four models, while SP.7 and MOMI.X have importance of over 70 in three models (**Supporting Information 4.9**).

When inspecting the predictors with importance over 70 in the four top models of pruned Subset D, the following can be rationalised for their high importance. First, SP.1 is the Kier and Hall's connectivity index of order 1 which indicates the bimolecular interaction possibilities of a molecule when considering individual bonds [27] (**Figure 4.8**). When a micelle is forming, surfactant molecules would come into close proximity and interact with each other to form a micelle. Therefore, it is natural for a descriptor which indicates this interaction possibility to be important in predicting CMC. Next, MOMI.R is the radius of gyration. In chemistry, this is a parameter characterising the size of a particle of any shape [26] (**Figure 4.9**). This is an important factor to micelle formation and the shape of the molecules indicates how well they can pack together to form a micelle. Following that, IC0 information

content of order 0 measures molecular symmetry, where the diversity of the elements are accounted for [24, 25, 46] (Figure 4.10). If a molecule is asymmetrical when considering its elements, there is a higher probability of a large difference in hydrophobicity/phility across the molecule, leading to a large amphiphilicity, which is desired for surfactants. Finally, FPSA.2 is the fractional charged partial positive surface area, calculated as the total charge weighted positive surface area divided by total molecular solvent-accessible surface area [24]. In the pruned Subset D, about 22% of the surfactants are cationic. It is possible that due to this it was necessary to include a positive charge related descriptor to distinguish between the cationic and anionic surfactants, and FPSA.2 was the one which was the most suitable.

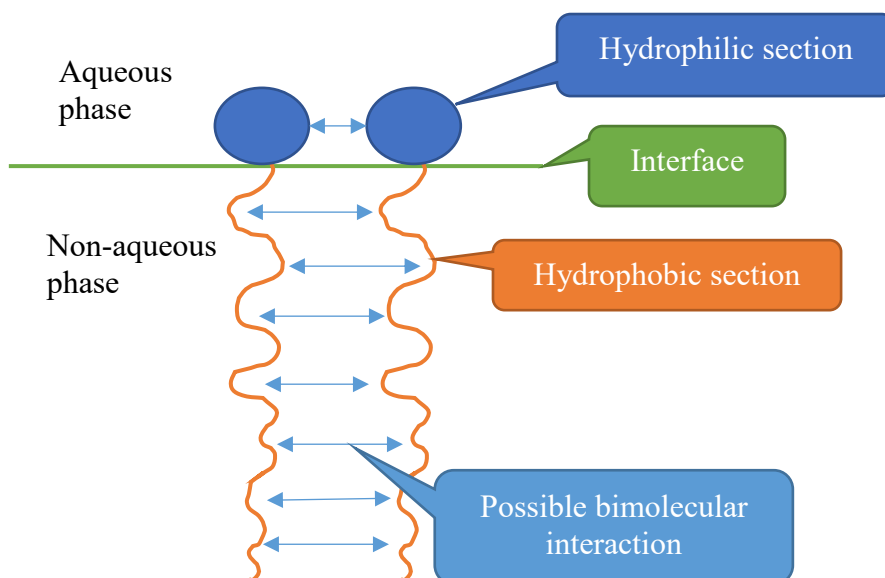


Figure 4.8. Illustration indicating the possible bimolecular interaction between two of the same surfactant molecules on the interface between aqueous phase and non-aqueous phase.

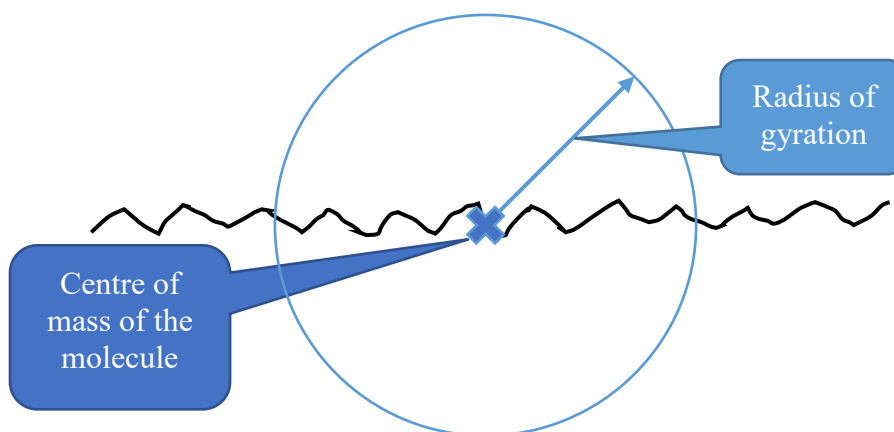


Figure 4.9. Illustration indicating radius of gyration of a molecule, calculated as the root mean square average of the distance of all scattering elements from the centre of mass of the molecule [47].

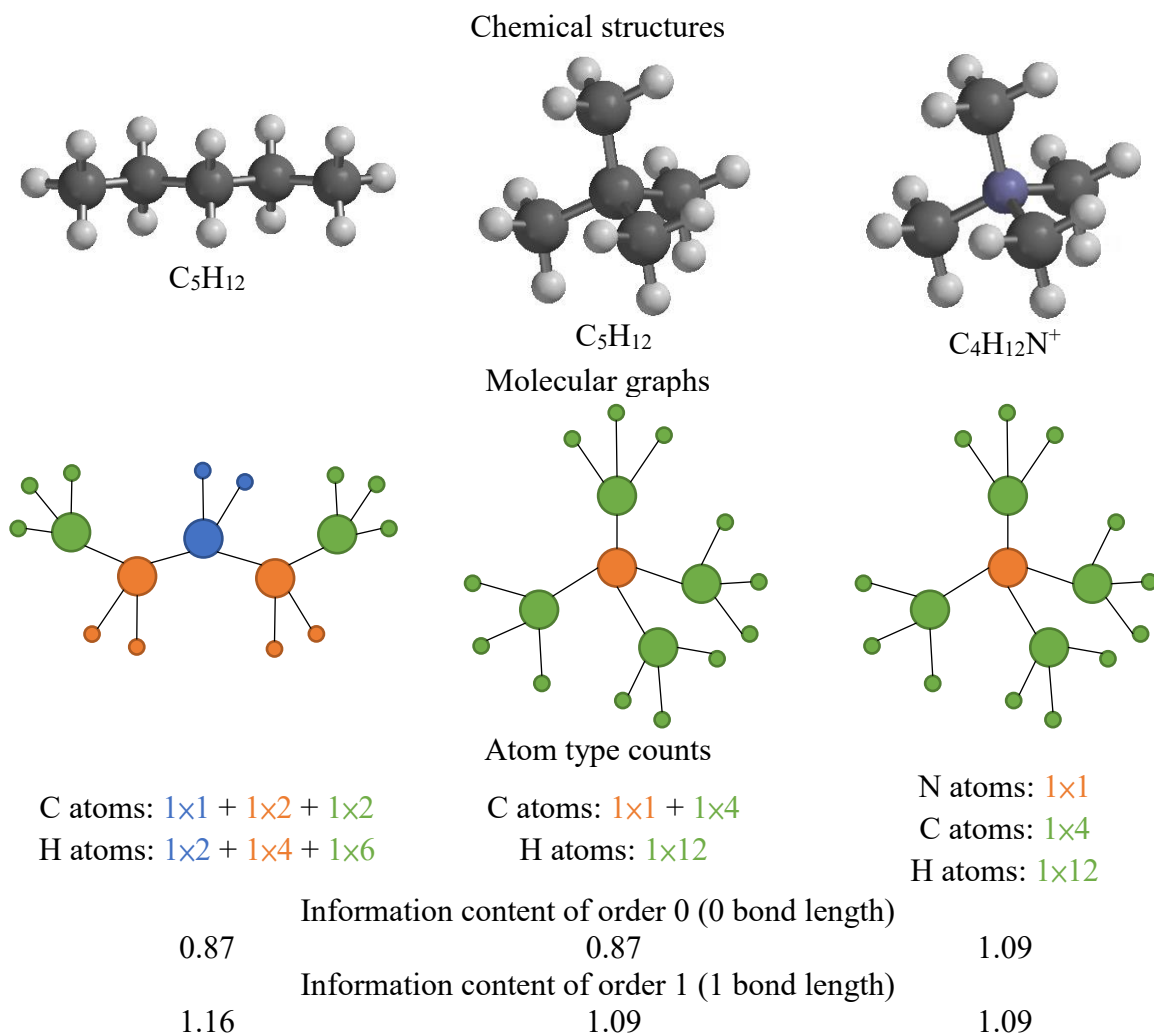


Figure 4.10. Illustration displaying the difference in information content of order 0 and 1 for two C_5H_{12} isomers, pentane (left) and neopentane (middle), and tetramethylammonium (right).

Although SP.7 and MOMI.X were not have important in all four top models of pruned Subset D, they still post high importance in a majority of them. Their importance can be rationalised as follows. SP.7, similar to SP.1 is the Kier and Hall's connectivity index of order 7 which indicates the bimolecular interaction possibilities of a molecule [27]. In comparison to SP.1 which accounts for the interaction possibility over a short length (1 bond), this accounts for the interaction over a longer length (7 bonds). MOMI.X is the moment of inertia along the X axis (**Figure 4.11**). Although dependent on the alignment prior to descriptor calculation, this describes how much torque it takes to spin the molecule along the X axis, and therefore indicates the shape of the molecule [24, 26]. The shape of the molecule is important to how well it can pack to form the micelle. Variation due to alignment is minimised by generating the 3D conformation for all molecules using the same process (see 4.4.2.1. *Data preparation*).

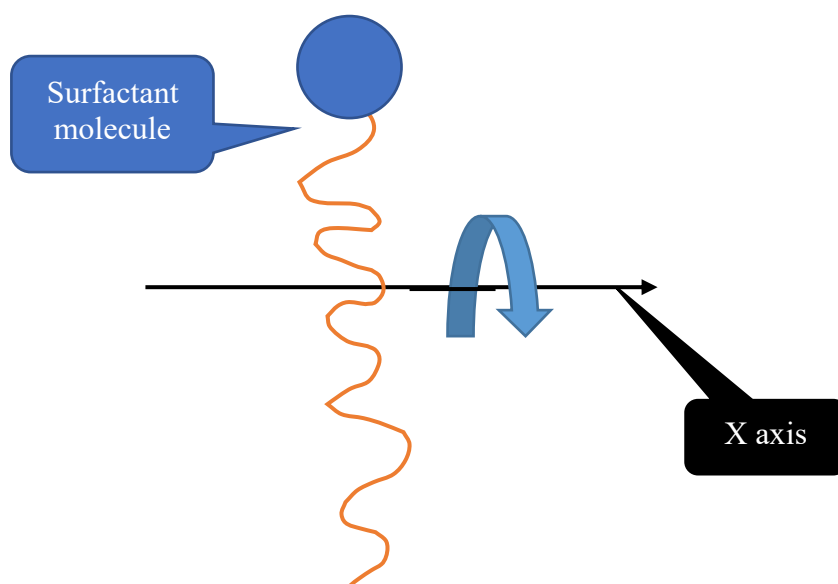


Figure 4.11. Illustration of moment of inertia along the X axis for a surfactant molecule.

4.4.3.2.2. Pruned Subset E

On inspection of the predictor importance of **Model E2, E14, E20** and **E44** (**Figure 4.12, Supporting Information 4.10**), a similar phenomenon to the predictor importance of **Model A2** was observed. The hydrophilic descriptors were either removed due to high correlation during 4.4.2.3. *Data Pre-processing* or have 0 importance. In all four models, PhobicESA has an importance of 100 and PhobicALogP has an importance of over 70 in three (**Model E2, E14** and **E44**, importance = 71.2, 76.0 and 85.0 respectively). The next most important predictor which can account for the hydrophilic section of the surfactants is FragmentALogPDifference, although the importance is below 51 (50.6 – 29.9). On inspection of the pruned Subset E, we see a very similar range of structures to Subset A. Except one entry, all entries in the pruned Subset E are ethylene oxide surfactant analogues. Due to the nature of the data, there is also a large overlap of entries between Subset A and pruned Subset E (29 entries). Here again, the hydrophobic sections are generally the same size or smaller than the hydrophilic sections, and they are all highly flexible. Different to **Model A2**, PhobicESA which describes the surface area in relation to the total electronegativity of the hydrophobic section have the highest importance. This is important to how the hydrophobic sections interact with each other when forming micelles. Extending from this, the other highly important predictor, PhobicALogP, contributes in defining the concentration at which the intermolecular hydrophobic sections cluster together to form micelles.

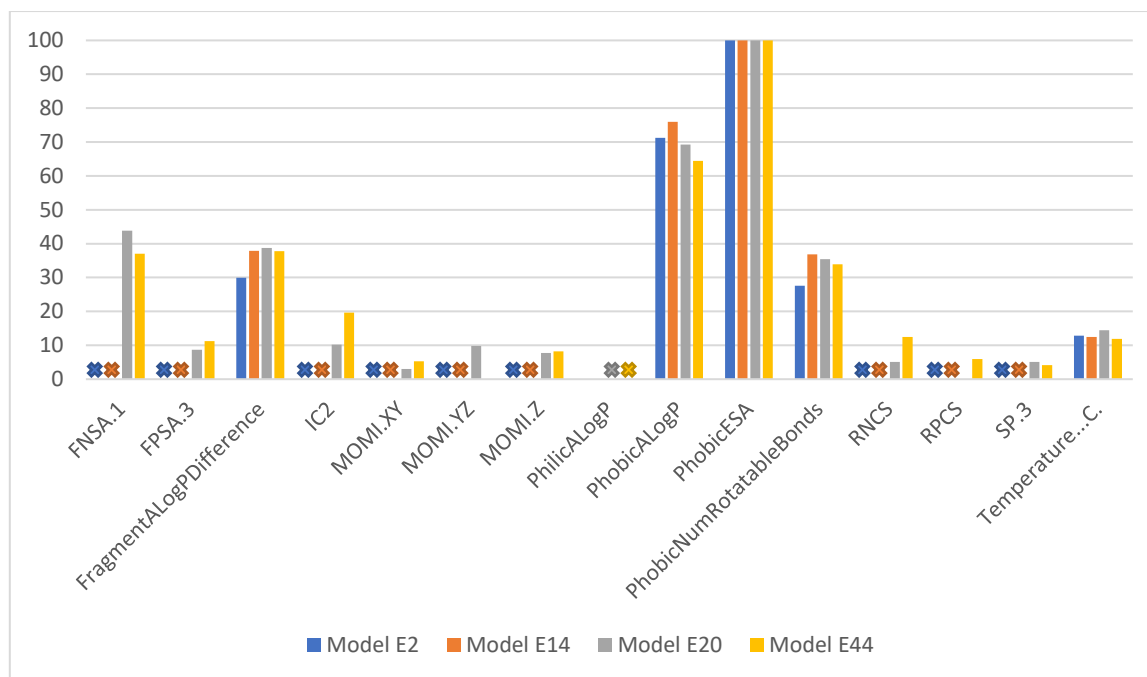


Figure 4.12. The predictor importance of **Model E2, E14, E20 and E44**, with the absence of the descriptor in the inputting predictor list of the model indicated by the corresponding colour × sign.

4.4.3.2.3. Subset F

When the pruned Subset D and E were combined to give Subset F, the best model (**Model F57**) had PhobicNumRotatableBonds, FragmentALogPDifference and PhobicALogP as its highly important predictors (Importance = 100.0, 92.6 and 88.6 respectively, **Figure 4.13, Supporting Information 11**). PhobicNumRotatableBonds indicates how flexible the hydrophobic section of the molecule is, which is important in determining how well the molecules can cluster together to form micelles. FragmentALogPDifference indicates the difference in ALogP between the hydrophobic and hydrophilic sections, which is crucial to determining how well the molecule can sit at the oil-water interface. Similar to **Model A2** and the top performing models for pruned Subset E, PhobicALogP holds the key in defining the concentration of the hydrophobic sections clustering together to form micelles.

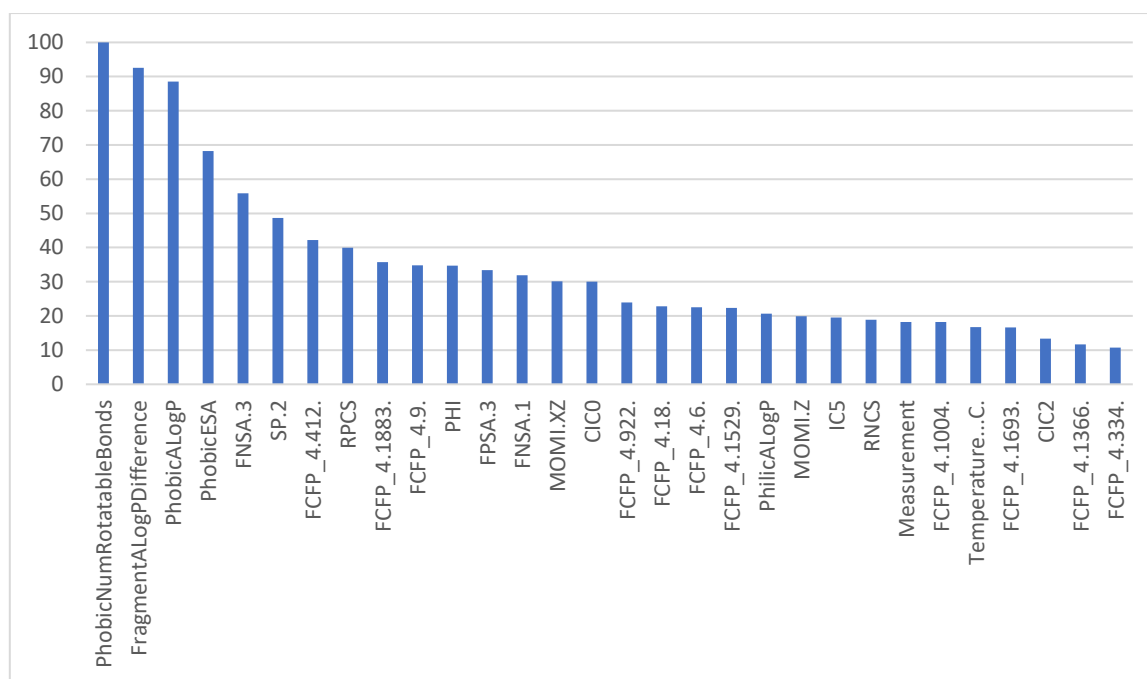


Figure 4.13. the predictor importance of **Model F57** with an importance of over 10.00. Out of the 30 predictors with an importance of over 10.00, three of the predictors are identified with high importance (importance > 70.00).

In addition to **Model F57**, **Model F17** and **F33** also performed well for Subset F. When inspecting their predictor importance (**Supporting Information 4.12**), **FragmentALogPDifference**, **PhobicALogP** and **Phobic_MOMI.Y** are again seen to have high importance (Importance = 100.0 and 93.6 for **Model F17** and Importance = 100.0 for **Model F33** respectively). The rationale behind their high importance mirrors the good performing models for pruned Subset D and E, in that **FragmentALogPDifference** accounts for the relativity between the hydrophobic and hydrophilic sections, **PhobicALogP** contributes in defining the concentration at which the hydrophobic sections cluster together to form micelles, and **Phobic_MOMI.Y** describes how much torque it takes to spin the hydrophobic section along the Y axis [24, 26]. **Phobic_MOMI.Y** therefore indicates the size and shape of the hydrophobic section of the surfactant, which is important to how well the molecule can pack to form the micelle.

Different to the other good performing models, **Model F17** and **F33** have some highly important predictors that have not been seen before, although the rationale can easily be mirrored from previously identified highly important predictors. **PhilicESA** is a highly important predictor for **Model F17** (Importance = 94.2). it describes the surface area in relation to the total electronegativity of the hydrophilic section of the molecule. This is important to how the hydrophilic sections interact with each other when forming micelles, especially when the hydrophilic sections are not linear and the formal charge is strong (e.g. sulphate groups). **Phobic_MOMI.R** and **Phobic_IC5** are highly important predictors for **Model F33** (Importance = 80.8 and 70.4). They respectively describe the size [26] and the element-related symmetry of the hydrophobic section [24, 25]. The size of the hydrophobic section affects how well molecules can pack together to form micelles, while the element-related symmetry can affect the intermolecular interaction when the hydrophobic sections are in close proximity due to electronegativity.

4.4.3.3. *Outlier analysis*

After identifying and rationalising the important predictors for the highly performing models, any outliers were identified as being 0.5 log unit between the observation and prediction. This value is the estimated measurement error anticipated for the variation of experimental CMC measurement method and temperature [48].

4.4.3.3.1. *Pruned Subset D*

On analysis of the top models for pruned Subset D (**Model D33, D51, D57 and D60, Figure 4.14**), eight entries were found to be outliers for all four models, while one is an outlier in three, four were in two and three were in one (**Table 4.9**). On inspection of the eight outliers for all four models, there were 4 cationic and 4 anionic (**Table 4.10**). Six entries were measured at 25°C, with the remaining two at 28°C and 40°C. Five out of eight of the measurement methods were via conductivity, although their equipment varies where it was identifiable (**Table 4.10**). Three of the eight outliers were sulphur containing (one sulphonate and two sulphuric), and one entry contains a 4-cyanopyridine (**Table 4.10**). The outlier that has 0.5 log unit between the observation and prediction in three of the top models for pruned Subset D is a sulphonate containing surfactant measured by conductivity at 25°C. The four outliers for two of the models were mainly measured by conductivity, with one of them calculated via micelle aggregation number. one of the conductivity measured entries was measured at 40°C, with the others measure at 25°C. Here again, one of the outliers is a sulphonate containing surfactant. Two of the three outliers in only one of the models were sulphonate containing surfactants, both measured at 28°C via surface tension method. The other outlier was an anionic surfactant measured by conductivity at 25°C.

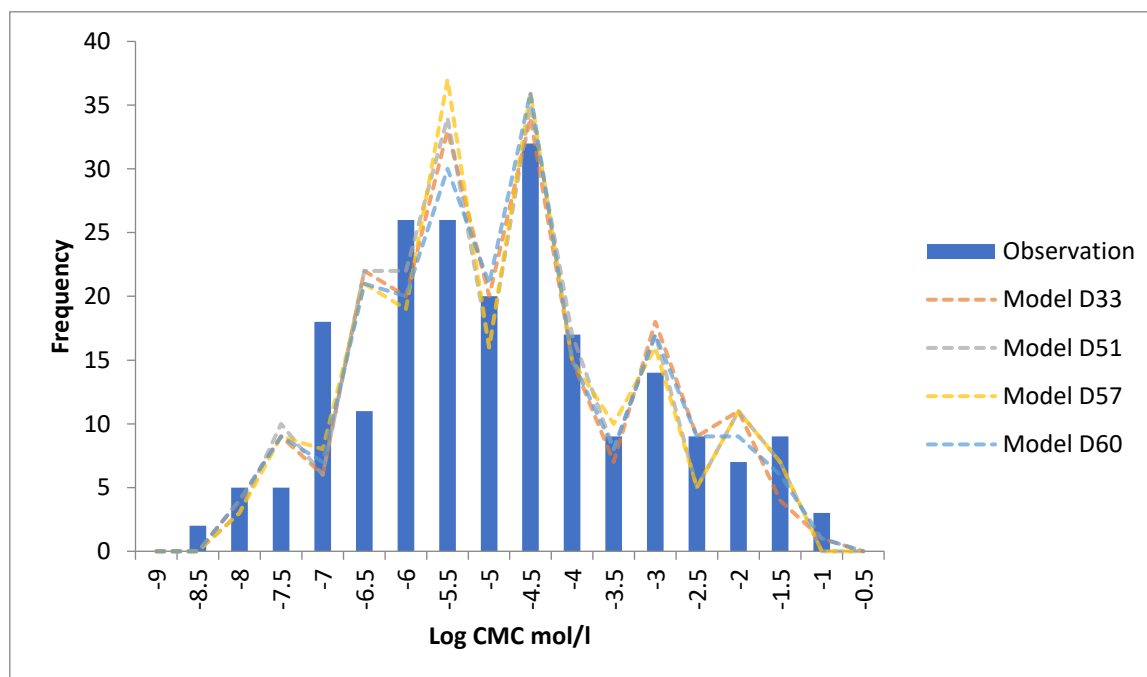


Figure 4.14. Histogram of the distribution of the observations and prediction by **Model D33, D51, D57 and D60**.


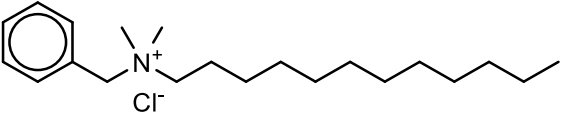
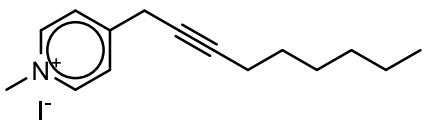
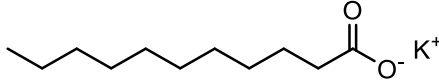
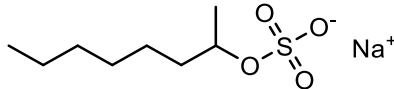
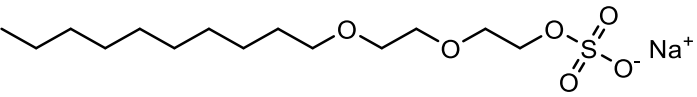
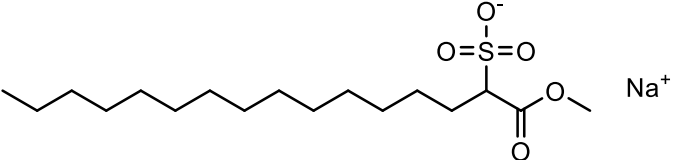
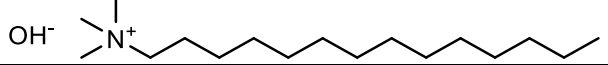
Table 4.9. The observation and prediction of the outliers for the top models of pruned subset D (Model D33, D51, D57 and D60)

Entry No.	Observation (Log CMC mol/l)	Prediction (Log CMC mol/l)			
		Model D33	Model D51	Model D57	Model D60
81	-3.75	-3.92	-4.81	-4.86	-3.92
82	-2.70	-2.87	-4.08	-4.39	-2.87
89	-2.73	-3.46	-5.23	-5.38	-3.34
93	-3.62	-4.19	-3.79	-3.79	-4.13
102	-2.06	-2.59	-3.21	-3.51	-2.60
119	-4.34	-3.39	-3.61	-3.68	-3.62
199	-2.32	-2.97	-3.09	-3.01	-3.03
203	-1.71	-2.25	-2.45	-2.42	-2.33
227	-4.41	-5.04	-5.33	-5.12	-5.04
292	-6.35	-5.83	-5.84	-5.87	-5.83
355	-3.40	-3.98	-4.01	-3.97	-3.98
357	-3.96	-4.37	-4.13	-4.18	-4.49
360	-4.90	-4.61	-4.38	-4.52	-4.73
492	-4.93	-5.56	-5.55	-5.53	-5.44
494	-6.01	-5.32	-5.57	-5.57	-5.53
515	-7.01	-6.22	-6.68	-6.71	-6.31

Bold: prediction over 0.5 log unit difference to observation

CHAPTER 4: NOVEL SURFACTANT DESCRIPTOR – POTENTIAL AMPHIPHILICITY

Table 4.10. Details of outliers common across the four top models of pruned subset D (Model D33, D51, D57 and D60)

Molecule	Entry No.	Structure	Measurement method	Equipment	Temperature (°C)
4.4	89		Micelle aggregation number (Fluorescence Probing method)	Unknown	25
4.5	102		Conductivity	Kyoto conductometer	24.85 ± 0.01
4.6	119		Conductivity	Philips PW 9501/01 conductivity meter	25.0 ± 0.1
4.7	199		Refractive index	Rayleigh-Haber-Löwe type of interferometer	25.000 ± 0.003
4.8	203		Conductivity	Unknown	40.00 ± 0.05
4.9	227		Conductivity	Pye type 1170 conductance bridge	25
4.10	355		Surface tension (Du Noüy ring technique)	Unknown	28
4.11	492		Conductivity	Unknown	25

From the above information, the models seem to be poor at predicting the CMC for sulphonate containing surfactants and surfactants measured via conductivity. However, this is due to the high proportion of outliers of the ionic surfactants containing sulphonate (54 out of 210 entries, with five entries being outliers) and measured via conductivity (135 out of 210 entries, with 10 entries being outliers). It is noteworthy that the salt counter ion, method of measurement and the temperature of measurements were not identified as important predictors (< 20). Nonetheless, in order to overcome this weakness, more CMC data on sulphonate containing surfactants and surfactants measured via conductivity could help.

In addition to the above analysis, the outliers in **Table 4.10** were inspected against the database to find similarities and difference between the outliers and the ones well predicted. It was found that **Molecule 4.4** and **4.5** both have a similar structure within the database which were identified as outliers in two of the top models. **Molecule 4.6** shares no alike structure with other entries as the structures from the same source contain no triple bonds. For these entries, addition of similar structure may improve the predictability of these structure. On the other hand, although **Molecule 4.7 – 4.10** share similarities with other well predicted structures within the subset, **Molecule 4.7 – 4.10** have the shortest alkyl chains of like molecules where any shorter alkyl chain analogues were pruned. This suggests a current limitation to the predictability of the structures with short alkyl chains. In order to overcome this limitation, addition of entries with the same surfactant structure to **Molecule 4.7 – 4.10** varying in salt counter ion, method of measurement and temperature may improve their predictability by providing similar data. Different to the above, when **Molecule 4.11** was inspected, it was found that the surfactant structure disregarding the salt counter was exactly the same as seven other entries. The range of observed Log CMC mol/l ranged between -5.4 to -5.7, greatly differing from the -4.9 observed for **Molecule 4.11**. In addition, an entry with observed -5.4 Log CMC mol/l was found to have the same salt counter ion (OH⁻) to **Molecule 4.11** and was also measure via conductivity at 25°C, suggesting possible error within the source data.

4.4.3.3.2. Pruned Subset E

On the other hand, possibly due to the general high similarity in structure of the pruned Subset E, only one entry was found to be an outlier across all top models (**Model E2, E14, E20 and E44, Figure 4.15**), with another entry as outlier in one model (**Model E14**). **Figure 4.16** presents the similarity of all structures in pruned subset E with each other (ECFP₄ fingerprint, Tanimoto similarity). The entry found to be an outlier across all top models was octylphenol ethoxylate with 12 ethylene oxide repeat units (DP12, **Molecule 4.12**) where the method of measurement was unspecified. Although the distance between the observation and prediction is not much further than 0.5 log unit (0.77 – 0.86), this was not unexpected as other surfactants of the same series with longer hydrophilic sections (16 – 55 ethylene oxide repeat units) were pruned and this surfactant has the longest hydrophilic section of the unpruned entries. This suggests that the length 12 ethylene oxide units is the border line between length of the hydrophilic section beyond which an accurate prediction cannot be made.

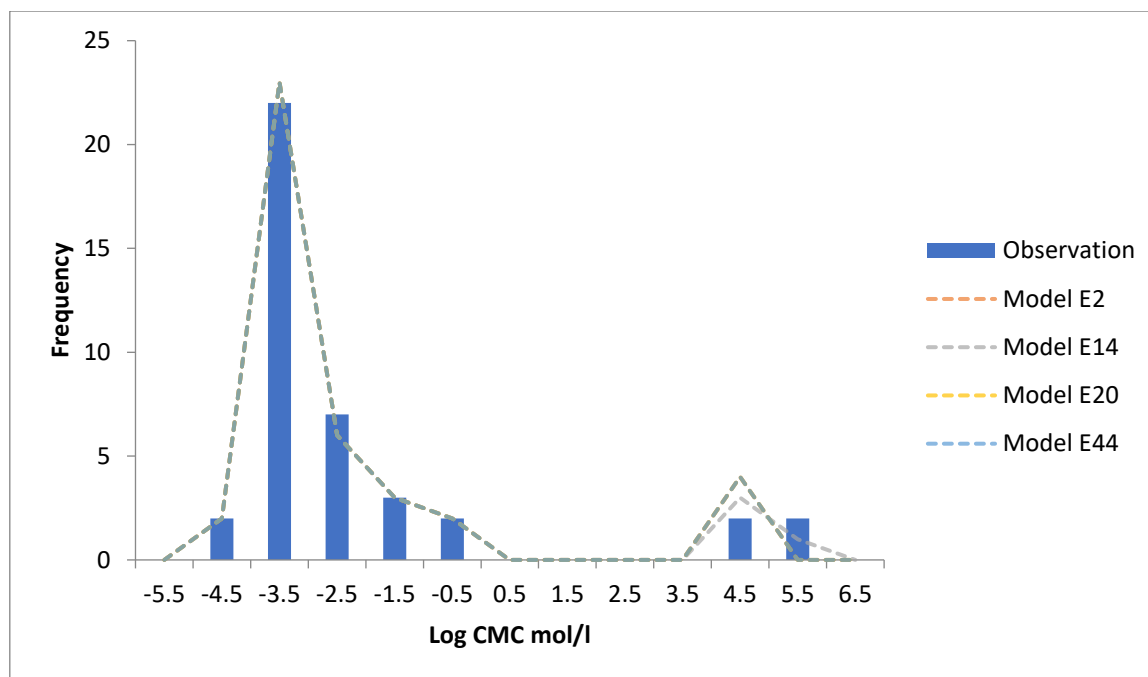


Figure 4.15. Histogram of the distribution of the observations and prediction by **Model E2, E14, E20 and E44.**

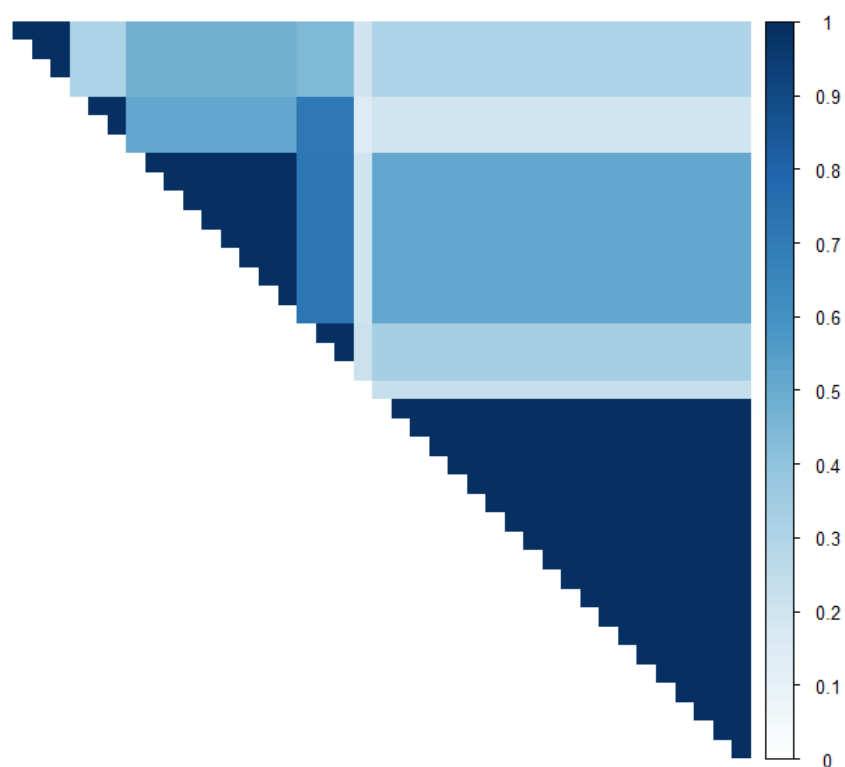
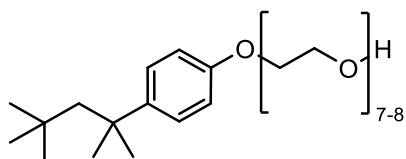


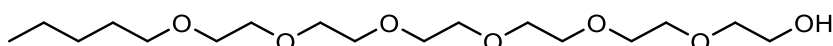
Figure 4.16. Correlation plot of the Tanimoto coefficients between the 40 structures within pruned Subset E using ECFP_4 fingerprint. Tanimoto coefficients on the structures themselves are omitted.

The entry that was found to be an outlier in **Model E14** in addition to octylphenol ethoxylate was $C_6(EO)_6$ (**Molecule 4.13**). It is to note that the distance between the observation and

prediction in the other models was also close to 0.5 log unit (0.42 – 0.45). This entry was measured via surface tension through the drop volume method at 25°C. However, most surfactants of the same series had not been pruned.



Molecule 4.12



Molecule 4.13

Molecule 4.12 – 4.13. Outliers for top models of pruned Subset E (**Model E2, E14, E20 and E44**).

4.4.3.3.3. Subset F

With the combined pruned Subset D and E, more outliers were expected as now the structures contain ionic and non-ionic species. As expected, a larger number of outliers were observed. Four entries were found to be outlier in all three of the top performing models (**Models F17, F33 and F57, Figure 4.17**), there were 13 outliers between two of the models and 20 outliers for one model only. When comparing these outliers to the outliers for the top models of pruned Subset D and E, only 22% of the outliers overlap. Here again octylphenol ethoxylate DP12 is an outlier across all three of the top performing models, however, most of the related analogues of the same series where the measurement method was unspecified are also outliers in two out of three models. Where some of the outliers do not overlap, their analogues are noted to be outliers in some occasions. This suggests that as there is a change of the overall structure of the QSPR model constructed, the outliers change but the series of surfactants, related by structure, measurement method, temperature or other descriptors, prone to producing outliers remain the same. Therefore, again it would be important to obtain more CMC data related to the outlier prone surfactant series in order to construct models which can predict accurately for these surfactant series.

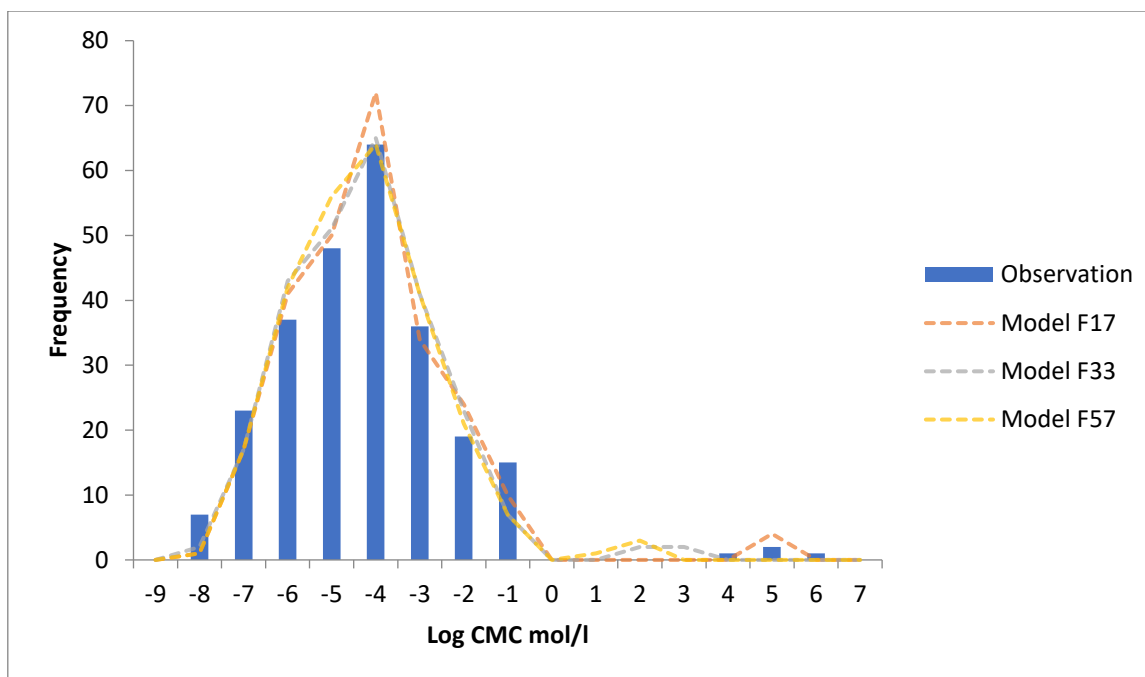
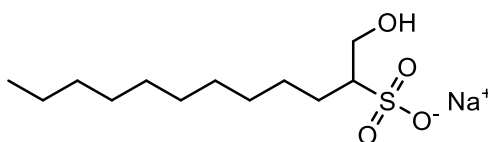


Figure 4.17. Histogram of the distribution of the observations and prediction by **Models F17, F33 and F57.**

4.4.3.4. Analysis of pruned entries

As a result of pruning, a total of 227 entries were pruned from the 477 entries of Subset C. Out of these pruned entries, 149 entries were without measurement temperature. These entries are mainly where the original data source cannot be accessed (purchase required or material not available online, 102 entries), failure to identify the reported CMC record in the original data source (36 entries) or the original source was unknown (six entries).

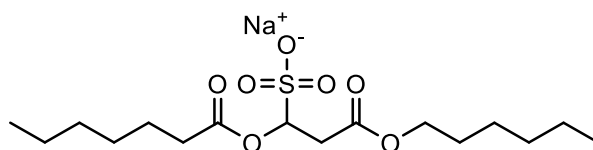
Within the remaining 78 pruned entries, 54 entries were ionic surfactants (21 cationic and 33 anionic). Looking closely at the method and the temperature of the pruned ionic surfactants (**Table 4.11**), the majority of the pruned entries were measured using conductivity, at 25°C (15 entries) and 40°C (eight entries). However, when looking at the entries pruned, the calculation method and conductivity at 30°C and 50°C stands out to have over 50% of the recorded entries pruned. Looking closer at the entries obtained via calculation method, only one of the entries was not pruned. This entry has the shortest hydrophobic section out of this series of surfactant CMC (**Molecule 4.14**).



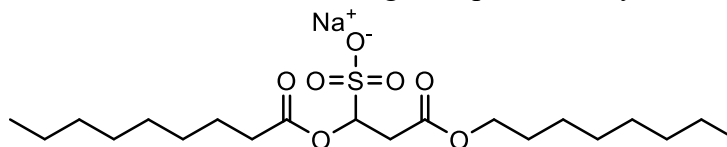
Molecule 4.14. the unpruned entry with CMC obtained through calculation.

For the two entries obtained via conductivity at 30°C, they are from different sources. Looking at the source where the pruned entries were referenced from, it contained four more entries where the CMC was measured at 25°C, with one being a structural analogue (**Molecule 4.15**) of the pruned entry (**Molecule 4.16**). Based on this, it is possible to conclude there were insufficient entries which share the similarity in structure (mean Tanimoto coefficient against

Subset F = 0.30 ± 0.10 , **Figure 4.18**), measurement method and temperature for the models to capture and predict its CMC.



Molecule 4.15. analogue of pruned entry



Molecule 4.16. pruned entry

Molecule 4.15 – 4.16. the pruned entry with CMC obtained through conductivity at 30 and its not pruned structural analogue.

Table 4.11. The breakdown of the pruned entries of Subset D (ionic surfactant entries of Subset C) Numbers quoted as (number of pruned entries)/(number of total entries at specified temperature)

Method	Temperature (°C)					
	25	27	28	30	40	50
Calculation	3/4	-	-	-	-	-
Calorimetry	3/10	-	-	-	-	-
Conductivity	15/93	-	-	1/2	8/59	1/2
Micelle aggregation number	6/17	-	-	-	-	-
Refractive index	1/7	-	-	-	-	-
Sound velocity	4/13	-	-	-	-	-
Surface tension	8/31	2/6	2/7	-	-	-

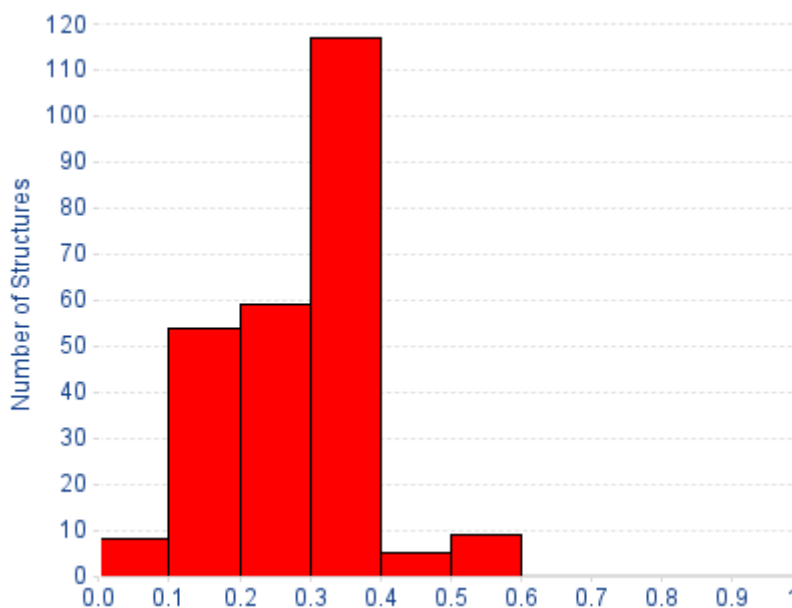
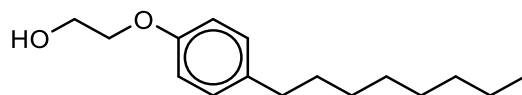


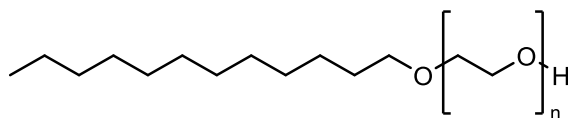
Figure 4.18. Histogram of the Tanimoto coefficients between **Molecule 4.16** and Subset F.

Looking at the 24 temperature recorded pruned non-ionic surfactants, they were all measured at 25°C (**Table 4.12**). The majority of the pruned entries are measured by surface tension, in particular 15 of them share the same source. Only three entries came from the sources that contain entries which are not pruned. For one of them, the pruned entry was the shortest ethylene oxide analogue of the $C_8Ph(EO)_n$ series ($n = 1$, **Molecule 4.17**). The inability of the models to accurately predict this analogue could be due to the lack of non-ionic surfactants in the database which shares the same ethylene oxide chain and CMC measurement method and temperature in the pruned database to provide the necessary information.



Molecule 4.17. The shortest ethylene oxide analogue of the $C_8Ph(EO)_n$ series ($n = 1$).

For the other two entries, they were the longer alkyl chain analogues of the $C_n(EO)_8$ series ($n = 12$ and 14 , **Molecule 4.18**). However, there are analogues with alkyl chain length shorter, longer, or in between them and none of these analogues were identified as outliers in the top models for Subset F and E ($n = 10, 11, 13$ and 15). For the $n = 12$ analogue, there is another entry in the pruned database which shares the same structure. Both were originally included in the database as although their measurement method and temperature are the same, the equipment used was different and the resulting CMC are not the same ($\log CMC = -4$ and -9.55 mol/l). As there are no indications to which is the correct value, they were both included to see whether models would help identify through consideration of other analogue's predictions. On the other hand, the $n = 14$ analogue did not have any duplicate entries. Therefore, similar to $C_8Ph(EO)_1$, the reasonable rationale for the inability of the models to accurately predict this is the lack of non-ionic surfactants in the database which shares the same hydrophobic alkyl chain and CMC measurement method and temperature in the pruned database to provide the necessary information.



Molecule 4.18. pruned analogues of the $C_n(EO)_8$ series ($n = 12$ and 14).

Table 4.12. The breakdown of the pruned entries of Subset E (non-ionic surfactant entries of Subset C), all recorded at 25°C

Method	Number of pruned entries	Number of temperature recorded entries
Calorimetry	1	1
Surface tension	19	55
Unspecified	4	8

4.4.4. Comparison with present CMC QSPR models

When comparing the top performing models constructed using pruned Subsets D and E with previous CMC QSPR studies [7, 19, 21-23, 35], our models have comparable performance (**Table 4.13**). It is important to remember when comparing the models, that as the databases used are different, the chemical spaces covered by each of the studies are different. Especially when comparing the Subset D top models with the ionic models from previous studies, we have to take care that our database contains both anionic and cationic surfactants (162 and 48 entries respectively) while some of the previous studies only contain anionic surfactants.

Table 4.13. Performance comparison of good performing models constructed using pruned Subset D and E and previous QSPR studies

Surfactant Type	Training R ²	Test R ²
Ionic (pruned Subset D)	0.89-0.93	0.94-0.96
Non-ionic (pruned Subset E)	0.93-0.95	0.99
Mixed (Subset F)	0.83-0.91	0.93-0.97
Anionic [7, 21, 23]	0.88-0.98	0.90-0.99
Ionic [19, 32]	0.95-0.99	0.94-0.99
Non-ionic [22, 35]	0.95-0.99	0.94

The important point here is that in our models, they contain predictors which are calculated either directly by the potential amphiphilicity protocol (**Scheme 4.1**) or using the hydrophobic and hydrophilic sections identified by the protocol. This demonstrates that the protocol was successful in identifying the hydrophobic/philic boundary of surfactants and output properties that through QSPR can be related to CMC, which is related to amphiphilicity. It is also to note that in comparison to some of the previous studies where high computational cost descriptors such as solvation energy [7] and hardness [20] are required, our models use properties which can be calculated quickly. For example, for a library of around 500 molecules, the descriptors can be calculated within minutes (PC used: Intel(R) Core(TM) i5-6500 CPU @ 3.20GHz 3.19 GHz, 32GB RAM).

4.5. Conclusion

In the search to quantify amphiphilicity for surfactants, a protocol which is capable to automatically identify the hydrophobic and hydrophilic sections of a surfactant has been constructed, outputting properties which could be related to amphiphilicity. As there are no quantitative amphiphilicity values these properties can be compared against, we constructed CMC QSPR models using calculated properties. As a result, we were able to construct good performing models for ionic surfactants (Test R² > 0.94, RMSE < 0.36, ρ > 0.89, Z > 30.0) and non-ionic surfactants (Test R² > 0.99, RMSE < 0.20, ρ = 1, Z > 7.3). Within these models, hydrophobic and hydrophilic descriptors were identified as highly important to predicting CMC (See 4.4.3.2. *Predictor importance*). This demonstrates that the protocol is successful in properties related to amphiphilicity implying that the calculation of the hydrophobic/philic sections was useful. Especially, it was noted that descriptors related to the hydrophobic sections of the surfactants is seen to be much more important than the hydrophilic sections.

4.6. Future Work

In order to further test and improve the potential amphiphilicity protocol in its ability to automatically identify the hydrophobic and hydrophilic sections, surfactant molecules of various structures, especially molecules which are not analogues of the ones used in this study, can be used. The protocol can also be further developed to identify if a molecule is a possible surfactant molecule or not by collecting data on the difference in outputs for surfactant molecules and non-surfactant molecules. In addition, expansion of the database used for QSPR to enrich the data for CMC measured using different methods and temperatures, preferably via surface tension or conductivity due to their consistency with the top performing models, can possibly increase the accuracy of the QSPR models, and further verification of the performance of the current QSPR models can be done by using an external validation set.

4.7. References

1. Holmberg, K.; Jönsson, B.; Kronberg, B.; Lindman, B. *Surfactants and Polymers in Aqueous Solution*, 2nd ed. 2003.
2. Rosen, M. J.; Kunjappu, J. T. *Surfactants and interfacial phenomena*; John Wiley & Sons, 2012.
3. Georgiou, G.; Lin, S.-C.; Sharma, M. M. Surface-Active Compounds from Microorganisms. *Bio/Technology* **1992**, *10* (1), 60-65. DOI: 10.1038/nbt0192-60.
4. Fischer, H.; Kansy, M.; Bur, D. CAFCA: a Novel Tool for the Calculation of Amphiphilic Properties of Charged Drug Molecules. *CHIMIA International Journal for Chemistry* **2000**, *54* (11), 640-645.
5. Eisenberg, D.; Weiss, R. M.; Terwilliger, T. C. The helical hydrophobic moment: a measure of the amphiphilicity of a helix. *Nature* **1982**, *299* (5881), 371-374. DOI: 10.1038/299371a0.
6. Cruciani, G.; Crivori, P.; Carrupt, P. A.; Testa, B. Molecular fields in quantitative structure-permeation relationships: the VolSurf approach. *Journal of Molecular Structure: THEOCHEM* **2000**, *503* (1), 17-30. DOI: [https://doi.org/10.1016/S0166-1280\(99\)00360-7](https://doi.org/10.1016/S0166-1280(99)00360-7).
7. Katritzky, A. R.; Pacureanu, L.; Dobchev, D.; Karelson, M. QSPR Study of Critical Micelle Concentration of Anionic Surfactants Using Computational Molecular Descriptors. *Journal of Chemical Information and Modeling* **2007**, *47* (3), 782-793. DOI: 10.1021/ci600462d.
8. Goodman, J. M.; Chemistry, R. S. o. *Chemical Applications of Molecular Modelling*; Royal Society of Chemistry, 1998.
9. Caron, G.; Digiesi, V.; Solaro, S.; Ermondi, G. Flexibility in early drug discovery: focus on the beyond-Rule-of-5 chemical space. *Drug Discovery Today* **2020**, *25* (4), 621-627. DOI: <https://doi.org/10.1016/j.drudis.2020.01.012>.
10. Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. *The Journal of Physical Chemistry A* **1998**, *102* (21), 3762-3772. DOI: 10.1021/jp980230o.
11. *Chemistry Collection: Basic Chemistry User Guide, Pipeline Pilot Release 16.5.0.143*; Accelrys Software Inc.: San Diego, 2016. (accessed 2018).
12. Helenius, A.; McCaslin, D. R.; Fries, E.; Tanford, C. [63] Properties of detergents. In *Methods in Enzymology*, Vol. 56; Academic Press, 1979; pp 734-749.
13. Chakraborty, T.; Chakraborty, I.; Ghosh, S. The methods of determination of critical micellar concentrations of the amphiphilic systems in aqueous medium. *Arabian Journal of Chemistry* **2011**, *4* (3), 265-270. DOI: <https://doi.org/10.1016/j.arabjc.2010.06.045>.
14. Nagarajan, R.; Ruckenstein, E. Theory of surfactant self-assembly: a predictive molecular thermodynamic approach. *Langmuir* **1991**, *7* (12), 2934-2969. DOI: 10.1021/la00060a012.
15. Li, C.; Pu, H.; Zhang, S.; Zhao, J. Effect of Nanoparticles and Surfactants on Oil/Water Interfacial Tension: A Coarse-Grained Molecular Dynamics Simulation Study. In *SPE/AAPG/SEG Unconventional Resources Technology Conference*, 2019.
16. Fujiwara, S.; Itoh, T.; Hashimoto, M.; Horiuchi, R. Molecular dynamics simulation of amphiphilic molecules in solution: Micelle formation and dynamic coexistence. *The Journal of Chemical Physics* **2009**, *130* (14), 144901. DOI: 10.1063/1.3105341 (accessed 2021/11/16).
17. Faramarzi, S.; Bonnett, B.; Scaggs, C. A.; Hoffmaster, A.; Grodi, D.; Harvey, E.; Mertz, B. Molecular Dynamics Simulations as a Tool for Accurate Determination of Surfactant Micelle Properties. *Langmuir* **2017**, *33* (38), 9934-9943. DOI: 10.1021/acs.langmuir.7b02666.

18. Khoshnood, A.; Lukanov, B.; Firoozabadi, A. Temperature Effect on Micelle Formation: Molecular Thermodynamic Model Revisited. *Langmuir* **2016**, *32* (9), 2175-2183. DOI: 10.1021/acs.langmuir.6b00039.
19. Barycki, M.; Sosnowska, A.; Puzyn, T. Which structural features stand behind micelization of ionic liquids? Quantitative Structure-Property Relationship studies. *Journal of Colloid and Interface Science* **2017**, *487*, 475-483. DOI: <https://doi.org/10.1016/j.jcis.2016.10.066>.
20. Gaudin, T.; Rotureau, P.; Pezron, I.; Fayet, G. New QSPR Models to Predict the Critical Micelle Concentration of Sugar-Based Surfactants. *Industrial & Engineering Chemistry Research* **2016**, *55* (45), 11716-11726. DOI: 10.1021/acs.iecr.6b02890.
21. Hu, J.; Zhang, X.; Wang, Z. A review on progress in QSPR studies for surfactants. *Int J Mol Sci* **2010**, *11* (3), 1020-1047. DOI: 10.3390/ijms11031020 PubMed.
22. Katritzky, A. R.; Pacureanu, L. M.; Slavov, S. H.; Dobchev, D. A.; Karelson, M. QSPR Study of Critical Micelle Concentrations of Nonionic Surfactants. *Industrial & Engineering Chemistry Research* **2008**, *47* (23), 9687-9695. DOI: 10.1021/ie800954k.
23. Roberts, D. W. Application of Octanol/Water Partition Coefficients in Surfactant Science: A Quantitative Structure-Property Relationship for Micellization of Anionic Surfactants. *Langmuir* **2002**, *18* (2), 345-352. DOI: 10.1021/la0108050.
24. Todeschini, R.; Consonni, V.; Mannhold, R.; Kubinyi, H.; Folkers, G. *Molecular Descriptors for Chemoinformatics: Volume I: Alphabetical Listing / Volume II: Appendices, References*; Wiley, 2009.
25. *Molecular Descriptors Guide*; U.S. Environmental Protection Agency, 2020.
26. McNaught, A. D.; Wilkinson, A. *IUPAC. Compendium of Chemical Terminology, 2nd ed. (the "Gold Book")*. Blackwell Scientific Publications, Oxford (1997), (accessed 08/06/2021).
27. Kier, L.; Hall, L.; Act, C. The Meaning of Molecular Connectivity: A Bimolecular Accessibility Model. *Croatica Chemica Acta* **2002**, *75*.
28. Ivanciuc, O. CODESSA Version 2.13 for Windows. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 405-406.
29. *CDK Descriptor Calculator GUI*; 2018. <http://www.rguha.net/code/java/cdkdesc.html> (accessed 2018).
30. Mauri, A. alvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints. In *Ecotoxicological QSARs*, Roy, K. Ed.; Springer US, 2020; pp 801-820.
31. Cui, Z. G.; Canselier, J. P.; Zhou, X. Q. Mixed adsorption and surface tension prediction of nonideal ternary surfactant systems. *Colloid and Polymer Science* **2005**, *283* (5), 539-550. DOI: 10.1007/s00396-004-1183-3.
32. Fatemi, M.; Konuze, E.; Jalali-Heravi, M. Prediction of critical micelle concentration of some anionic and cationic surfactants using an artificial neural network. *Asian Journal of Chemistry* **2007**, *19*, 2479-2489.
33. Ghasemi, J.; Saaidpour, S. Quantitative Structure-Micellization Relationship Study of Cationic Surfactants Using Ordinary Least Squares Regression. *Journal of Sciences (islamic Azad University)* **2009**.
34. Hines, J. D.; Thomas, R. K.; Garrett, P. R.; Rennie, G. K.; Penfold, J. A Study of the Interactions in a Ternary Surfactant System in Micelles and Adsorbed Layers. *The Journal of Physical Chemistry B* **1998**, *102* (48), 9708-9713. DOI: 10.1021/jp983154y.
35. Yuan, S.; Cai, Z.; Xu, G.; Jiang, Y. Quantitative structure-property relationships of surfactants: prediction of the critical micelle concentration of nonionic surfactants. *Colloid and Polymer Science* **2002**, *280* (7), 630-636. DOI: 10.1007/s00396-002-0659-2.

36. Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Journal of chemical information and modeling* **2010**, *50* (5), 742-754, Article. DOI: 10.1021/ci100050t.
37. Sander, T.; Freyss, J.; von Korff, M.; Rufener, C. DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis. *Journal of Chemical Information and Modeling* **2015**, *55* (2), 460-473. DOI: 10.1021/ci500588j.
38. Sawada, R.; Kotera, M.; Yamanishi, Y. Benchmarking a Wide Range of Chemical Descriptors for Drug-Target Interaction Prediction Using a Chemogenomic Approach. *Molecular Informatics* **2014**, *33* (11-12), 719-731. DOI: 10.1002/minf.201400066.
39. R: What is R? <https://www.r-project.org/about.html> (accessed 26/10/2021).
40. Kuhn, M.; Johnson, K. *Applied predictive modeling*; Springer New York, 2013. DOI: <https://doi.org/10.1007/978-1-4614-6849-3>.
41. Golbraikh, A.; Tropsha, A. Beware of q²! *Journal of Molecular Graphics and Modelling* **2002**, *20* (4), 269-276. DOI: [http://dx.doi.org/10.1016/S1093-3263\(01\)00123-1](http://dx.doi.org/10.1016/S1093-3263(01)00123-1).
42. Alexander, D. L. J.; Tropsha, A.; Winkler, D. A. Beware of R²: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *Journal of Chemical Information and Modeling* **2015**, *55* (7), 1316-1322. DOI: 10.1021/acs.jcim.5b00206.
43. Deb, K.; Agrawal, S.; Pratap, A.; Meyarivan, T. A Fast Elitist Non-dominated Sorting Genetic Algorithm for Multi-objective Optimization: NSGA-II. In *Parallel Problem Solving from Nature PPSN VI*, Berlin, Heidelberg, 2000//, 2000; Schoenauer, M., Deb, K., Rudolph, G., Yao, X., Lutton, E., Merelo, J. J., Schwefel, H.-P., Eds.; Springer Berlin Heidelberg: pp 849-858.
44. Akoglu, H. User's guide to correlation coefficients. *Turk J Emerg Med* **2018**, *18* (3), 91-93. DOI: 10.1016/j.tjem.2018.08.001 PubMed.
45. Shen, M.; LeTiran, A.; Xiao, Y.; Golbraikh, A.; Kohn, H.; Tropsha, A. Quantitative Structure–Activity Relationship Analysis of Functionalized Amino Acid Anticonvulsant Agents Using k Nearest Neighbor and Simulated Annealing PLS Methods. *Journal of Medicinal Chemistry* **2002**, *45* (13), 2811-2823. DOI: 10.1021/jm010488u.
46. Sabirov, D. S.; Shepelevich, I. S. Information Entropy in Chemistry: An Overview. *Entropy* **2021**, *23* (10), 1240.
47. Rhodes, G. Chapter 9 - Other Diffraction Methods. In *Crystallography Made Crystal Clear (Third Edition)*, Rhodes, G. Ed.; Academic Press, 2006; pp 211-235.
48. Winter, J. Personal Communication.

Chapter 5:
Visualisation of Chemical Functionality for a
Chemical Library

CHAPTER 5: VISUALISATION OF CHEMICAL FUNCTIONALITY FOR A CHEMICAL LIBRARY

5.1. Chemical functionalities and pharmacophore

A functional group is defined by IUPAC as an atom, or a group of atoms that has similar chemical properties wherever it occurs in different compounds [1]. The functional groups of a potential drug candidate determine its pharmacodynamic and pharmacokinetic effects, which affects its possible route of administration, mechanism of action, route of metabolism and elimination, toxicity and tendency to cause adverse effects [2]. There are many functional groups, and broadly speaking, they can be classed into hydrophobic and hydrophilic functionalities. Hydrophilic functionalities can include heterocycles, hydrogen-bond donors, acceptors, and positively and negatively charged atoms, while hydrophobic functionalities include aliphatic chains, carbon rings, and π -systems and aromatic rings [2]. When the molecular features, e.g. functionalities of a small molecule, are in a particular spatial arrangement required for specific interactions with its biological target and its activities, they can be defined as pharmacophores of the molecule [1, 3, 4]. The concept of pharmacophores is used widely in computer-aided drug design. It is used extensively in virtual screening, *de novo* design, lead drug candidate optimisation and multitarget drug design, whether in a ligand-based or structure-based fashion [4, 5].

When designing a chemical library, especially for fragment-based lead discovery for the pharmaceutical industry, it is important for a chemical library to cover a wide range of pharmacophores, and hence chemical functionalities, over a diverse 3D “chemical” space allowing a wide range of interactions and properties to be explored [6]. By knowing which chemical functionality is present on which part of a molecule, it is possible to predict the kind of interactions that the molecule could partake in.

5.2. Visualisation of Chemical Functionality for a Molecule

There had been various approaches to identify and visualise the pharmacophores and chemical functionalities of molecules within chemical libraries, and measure their diversity in 3D space. Pharma is a pharmacophore search tool which identifies and calculates pharmacophores' features of molecules as coordinate-frame independent relation triangles (**Figure 5.1**) [7]. Using the pharmacophore feature triangles and other calculated spatial indices, matches between query pharmacophore features and chemical libraries can be found. As this approach uses relation triangles, it is necessary for the molecule or the query pharmacophore features to contain at least three pharmacophores to form the querying relation triangle. This tool had been adapted into an online interface – ZINCPharma [8] for searching purchasable compounds with desired pharmacophore spatial relation of the ZINC database, a comprehensive collection of commercially available, biologically relevant compounds suitable for screening [9].

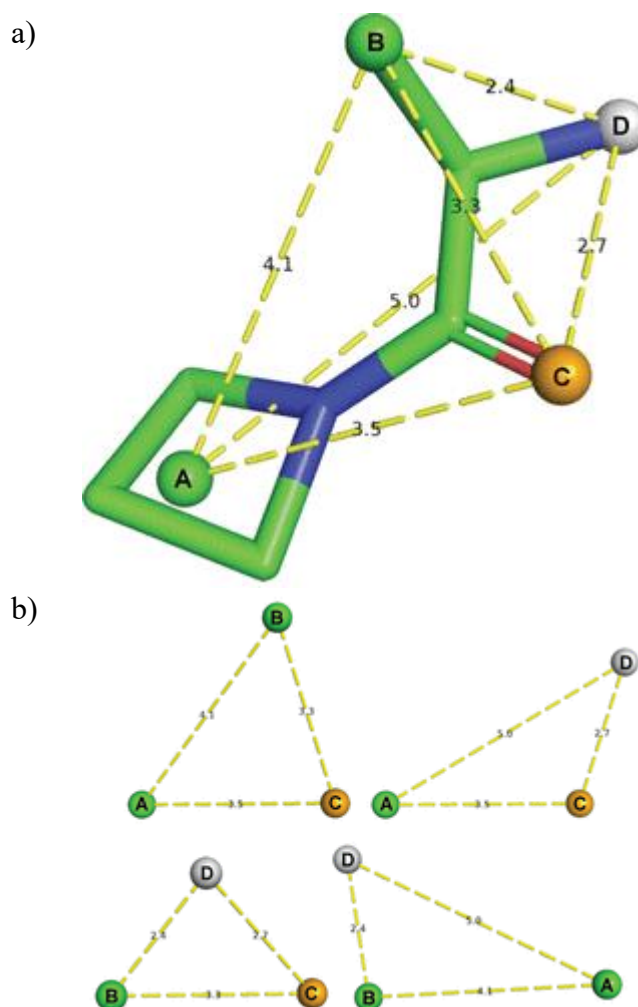


Figure 5.1. a) The pharmacophore features (lettered spheres) identified using user-configurable definitions for (2R)-2-amino-1-(azetidin-1-yl)propan-1-one. b) The collection of identified features is decomposed into coordinate-frame independent triangles. Taken from [7].

HookSpace is a program which assesses the diversity of a chemical library using the spatial relationships between pairs of functional groups within each molecule [10]. These functional groups require a carbon head and can be user defined. In the literature, acetate, N-methyl acetamide, methanol, methyl ammonium, phenyl, fluoride, chloride, iodide, bromide, thio ether and phosphono ether were defined for investigation (**Figure 5.2**). HookSpace calculates the spatial relationship between each pair of functional groups within each molecule by the following steps (**Figure 5.3**) [10]:

- Step 1. Orientate the molecule such that one of the functional groups (head 1 – tail 1) lies along the positive x-axis, with the head atom (head 1) at the origin
- Step 2. Rotate the molecule around the x-axis so that the head atom of the second functional group (head 2) was on the xy-plane with the head-to-tail vector (head 2 → tail 2) pointing in the positive z-direction
- Step 3. Coordinate of the second head atom (head 2) was stored
- Step 4. Repeat step 1 – 3 with the second functional group along the x-axis

HookSpace can present its result for analysis in various way: a histogram for distribution of the distance between the carbon heads of each pair of functional groups (**Figure 5.4.a**), a tile plot which display the number of functional groups present at each xy-position (**Figure 5.4.b**),

a 3D bar chart representing the frequency of each combination of functional groups (**Figure 5.4.c**), and a defined space dependent quantitative HookSpace Index representing the percentage of non-zero positions in the xy-plane. Currently, there are several limitations for HookSpace. For HookSpace to compute any result for a molecule in the library, the molecule is required to contain at least two of the defined functional groups. It also does not present the functional groups in 3D space and most importantly is restricted to the single rigid conformation entered into the program.

Name of group	Description	Structure
ACET	Acetate	
ACAM	<i>N</i> -methyl acetamide	
ACA2	<i>N</i> -methyl acetamide	
MEOH	Methanol	
MAMM	Methyl ammonium	
PHER	Phenyl	
FLUO	Fluoride	
CHLO	Chloride	
IODI	Iodide	
BROM	Bromide	
THIO	Thio ether	
PHOS	Phosphono ether	

Figure 5.2. Functional groups defined in [10], taken from [10]. For *N*-methyl acetamide, two definitions were available and were treated separately as there are two carbon heads existing within the functional group.

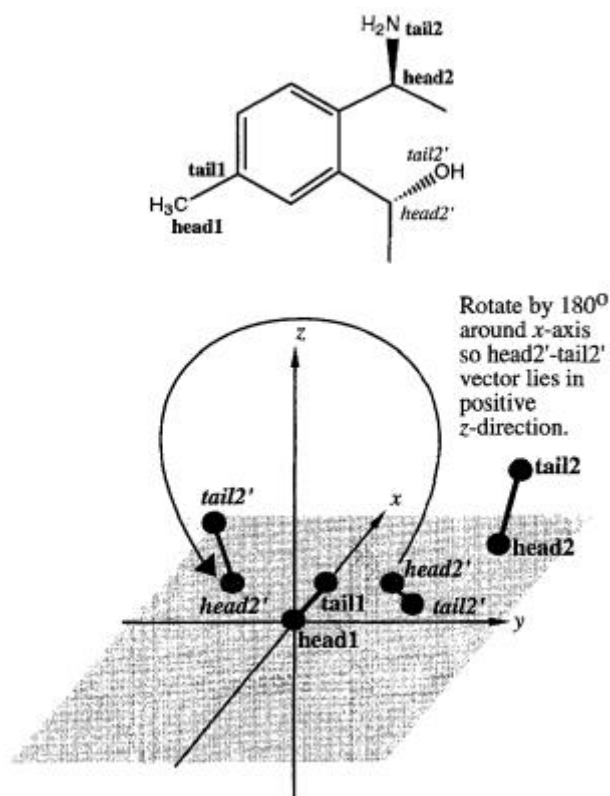


Figure 5.3. A schematic illustration of the calculation of functional group geometries, illustrated using (R)-1-(2-((S)-1-aminoethyl)-5-methylphenyl)ethan-1-ol as an example. Taken from [10].

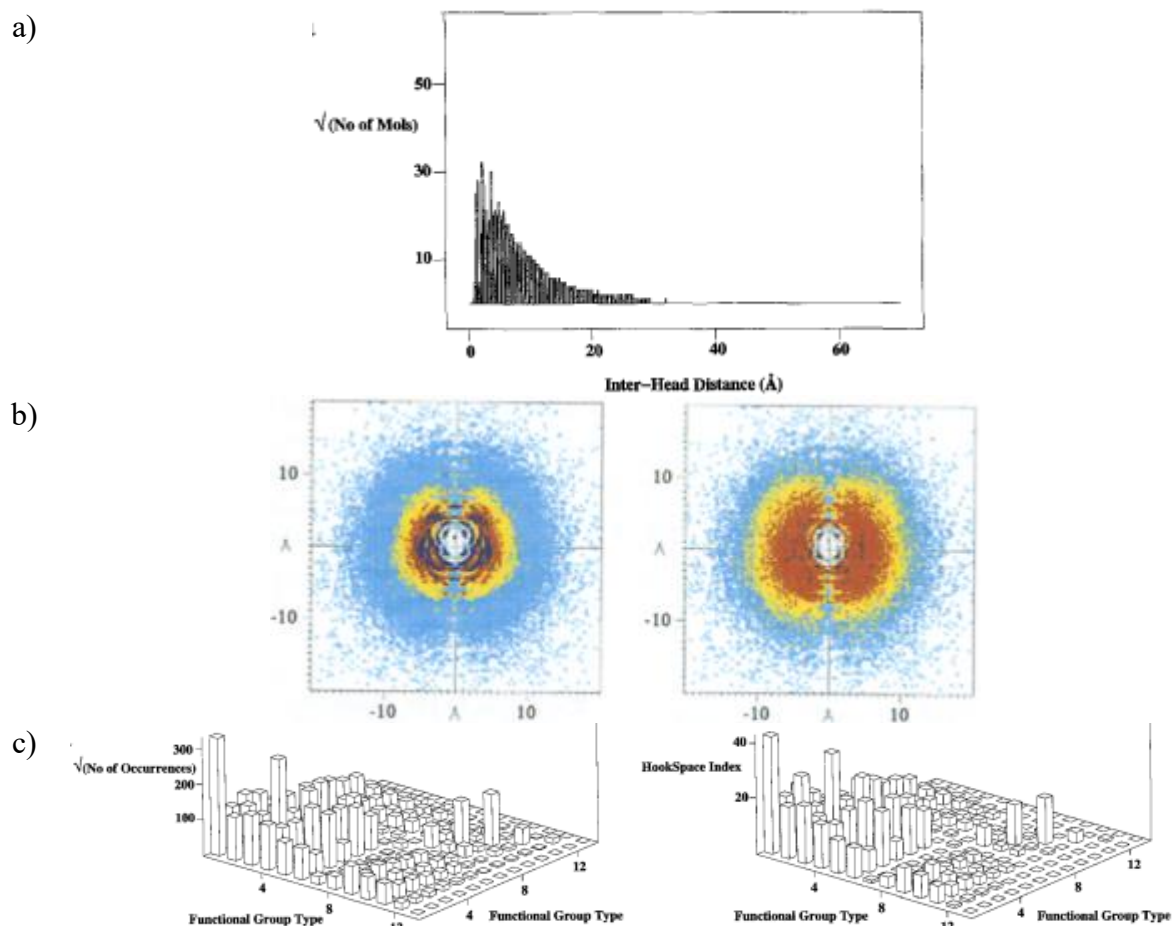


Figure 5.4. Example HookSpace results for the filtered CSD database analysed within [10], which contains 45887 molecules. a) histogram of the head – head distance between functional group pairs; b) tile plot displaying how many functional groups have that particular geometry (left) and how many different types of functional group have that particular geometry (right), where each tile is 0.2 Å in the x- and y-direction; c) 3D bar chart showing the number of occurrences (left) and HookSpace Index (right) for each type of functional group pair.

Gridding and partitioning (GaP) is a computational method designed for classification and selection of monomers for a combinatorial library [11]. Combinatorial libraries are collections of molecules synthesised taking into account all possible combinations of the contributing building blocks – monomers, small molecules with generally two reactive sites (**Figure 5.5**) [12, 13]. GaP takes into consideration that monomers are small and therefore might not contain multiple functional groups. It is based on the idea that 3D space can be gridded into cells when each axis is split into a given number of ‘bins’. It then takes a common feature for a combinatorial library, whether a functional group or atom, as the origin and tracks the position of the pharmacophores with free rotation around the x-axis for each conformation a monomer can take in 3D space (**Figure 5.6**). This information is recorded using a binary code for cells that a pharmacophore ‘hits’. This binary code can easily be used to analyse where the pharmacophores are present or absent, however, the length of this code is dependent on the total size of the analysing space and the size of the ‘bins’ each axis is split into. This makes it difficult to analyse the code without appropriate aid. In addition, as this method is designed for the monomers for combinatorial libraries, it does not consider the conformation hindered by substituents when developing into potential drug candidate molecules [14]. Nonetheless, this

method had been incorporated into screening procedures in pharmaceutical companies such as GlaxoSmithKline for matching 3D pharmacophoric patterns [15].

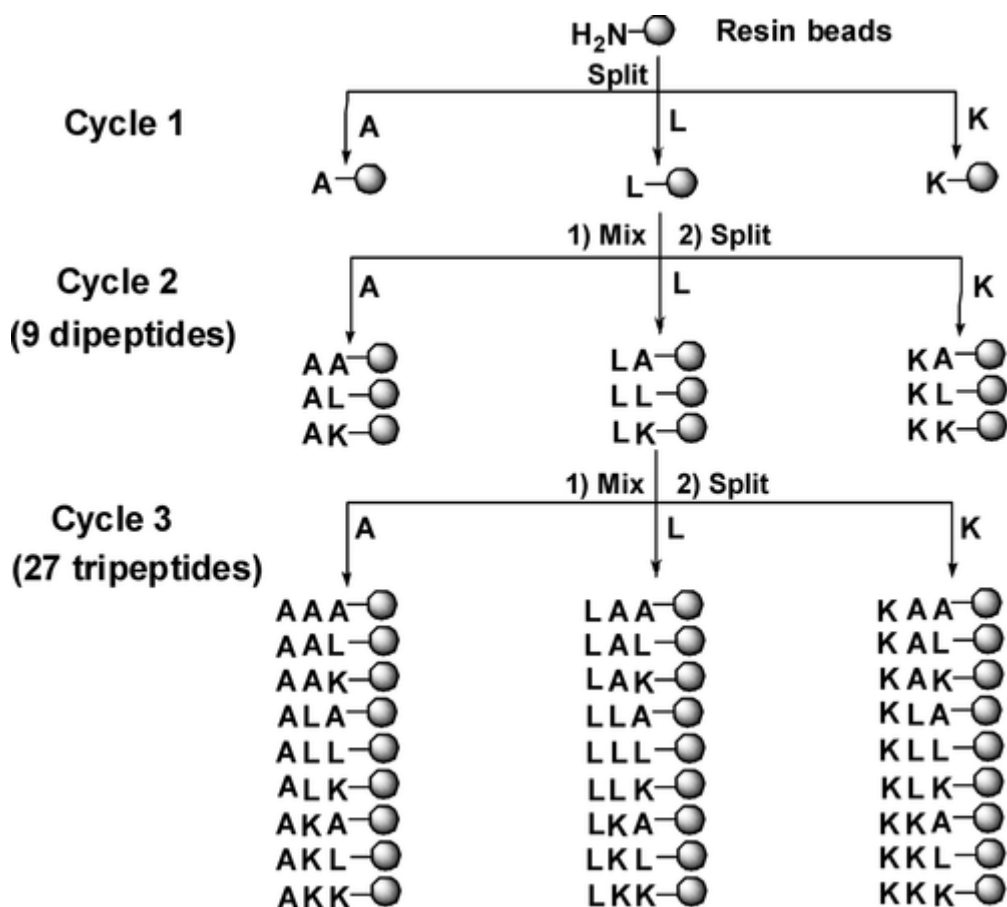


Figure 5.5. A schematic illustration of an example combinatorial library generation via “split-mix synthesis” method, one of the various methods of combinatorial library generation. Taken from [13].

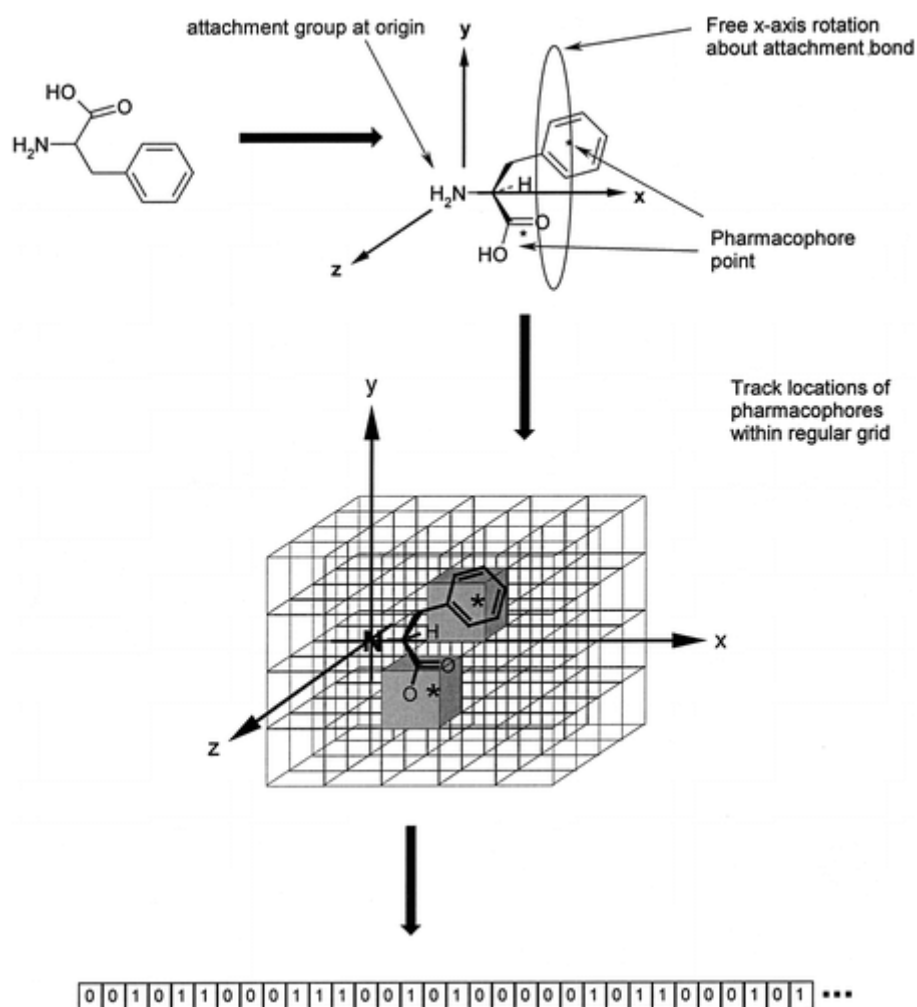


Figure 5.6. A schematic illustration of the GaP process, illustrated using phenylalanine as an example. Taken from [11]. This molecule contains two pharmacophoric groups (aromatic ring and acid). The molecule is oriented with the primary alkylamine attachment group at the origin and the adjacent bond along the x-axis. For each conformation the molecule is permitted free rotation about the x-axis, tracking those pharmacophore cells that are “hit”.

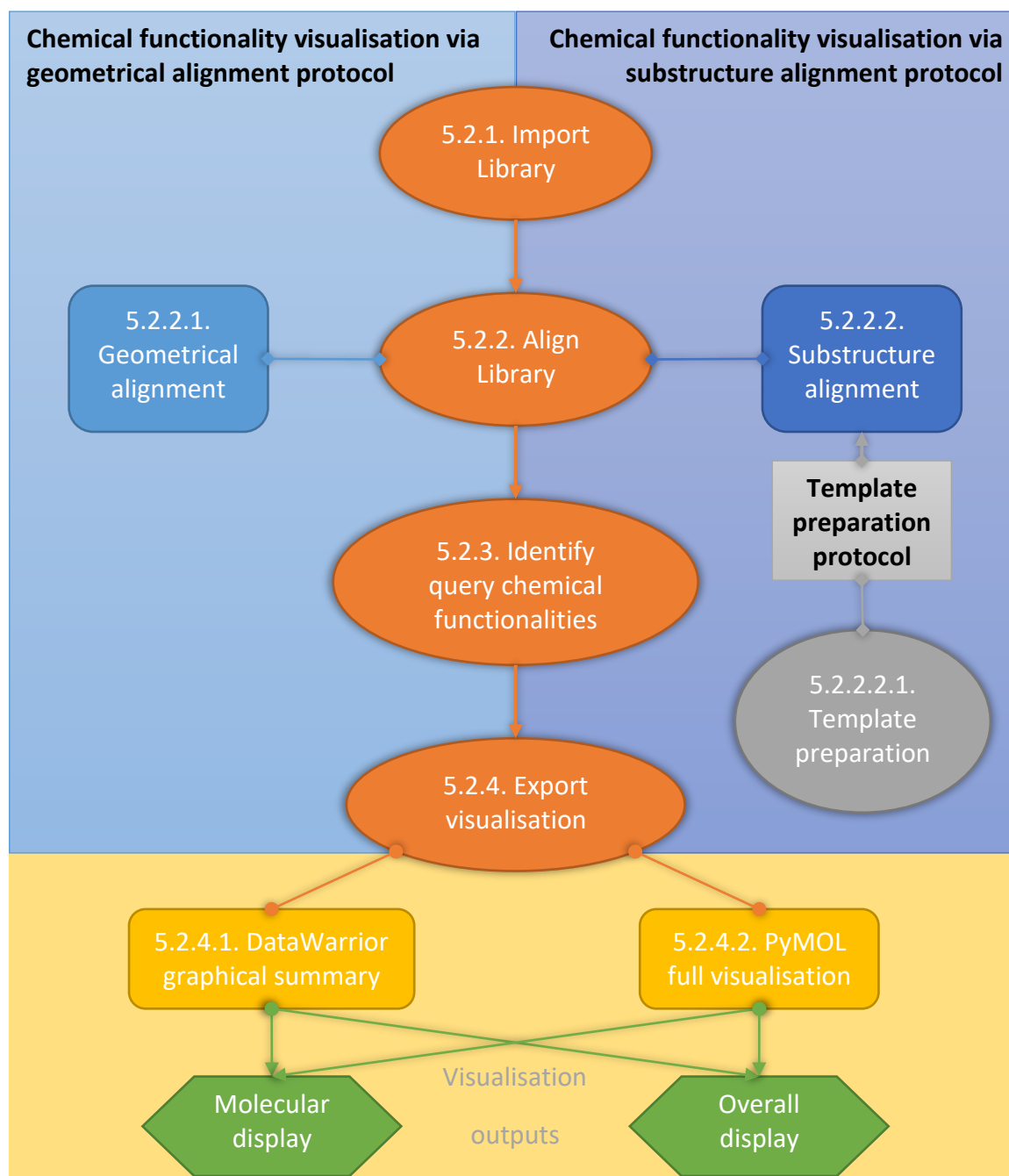
Each of the above methods comes with their own advantages and disadvantages. Pharma allows a user to define the pharmacophore of interest and calculates the pharmacophore feature relations without the complexity of library alignment as the relation triangles are coordinate-frame independent. On the other hand, as the calculation of relation triangles of the identified pharmacophore is necessary, at least three of the defined pharmacophores are required to be present in the analysed molecule. HookSpace is able to present graphically the predefined functional group pair information for an analysed library graphically. However, its applicability is restricted to molecules with at least two of the defined functional groups being analysed. The defined functional groups are also required to have a carbon head and functional groups with multiple carbon heads are treated as separate functional groups. GaP is able to consider the conformations a monomer can take and analyse the location of the pharmacophore groups in relation to an attachment point. At the end of the calculation, it outputs an analysing space size dependent binary code for each monomer analysed, which is not necessarily straight forward to understand without prior knowledge.

In addition to the above advantages and disadvantages, the above tools/methods all miss the ability to visualise the query pharmacophores of the whole, or a selected part of the library in 3D space. In order to visually aid prediction of likely interactions of molecules and fragment libraries and provide an “easy to grasp” 3D visualisation for reporting and presentation purposes, a series of protocols have been developed on Pipeline Pilot to visualise the chemical functionalities of molecules within a chemical library.

In this chapter, the process to visualise the chemical functionalities within Pipeline Pilot will be explained and presented along with the pros and cons of the different visualisation methods used to create the 3D view of the chemical functionalities. Examples of the visualisation outputs are displayed using a series of example molecules and an in-house library for Liverpool ChiroChem Ltd. (LCC) [16], a company founded in 2014 with their expertise in synthesising chirally-pure compounds as 3D-rich building blocks for small molecule drug discovery.

5.3. Process to Visualise Chemical Functionality for a Chemical Library

The process of visualising the 3D chemical functionality for a chemical library consists of: import library, align library, identify query chemical functionalities and export visualisation. In this project, Pipeline Pilot protocols have been generated for these processes as shown below (**Scheme 5.1**).



Scheme 5.1. Main processes to visualise chemical functionalities for a chemical library. Black text: name of protocols. Orange ovals: main processes. Rounded rectangles: options of the linked process. Green hexagons: visualisation display options.

5.3.1. Import Library

In order to create visualisation of chemical functionalities for chemical libraries that can be used for comparison, any imported libraries need to be processed. Within the main protocols, the chemical library is imported as a SD file, preferably with 3D coordinates. The molecules are standardised regarding their stereo, charges and π -systems as follows:

1. If the imported molecules contain 3D coordinates, the stereochemistry of the atoms and bonds are assigned using the 3D coordinates; if the imported molecules only contain 2D coordinates, the up/down bond markings are used to assign their stereochemistry

2. Any stereochemical marking on non-stereo atoms or bonds are removed
3. The formal charges of common functional groups, e.g. nitro groups, are standardised
4. π -systems are converted to kekulised representations
5. Any duplicate molecules are identified using canonical Simplified Molecular Input Line Entry System (SMILES) and removed
6. Calculate 3D coordinates if 3D coordinates are not available
7. All molecules are then energy minimised using the Clean forcefield [17]

Sometimes visualisation of chemical functionalities of a library is desired based upon reference to a certain molecular structure (**Figure 5.7**). In such cases, reference molecular structure can be chosen and imported at this stage as SD files.

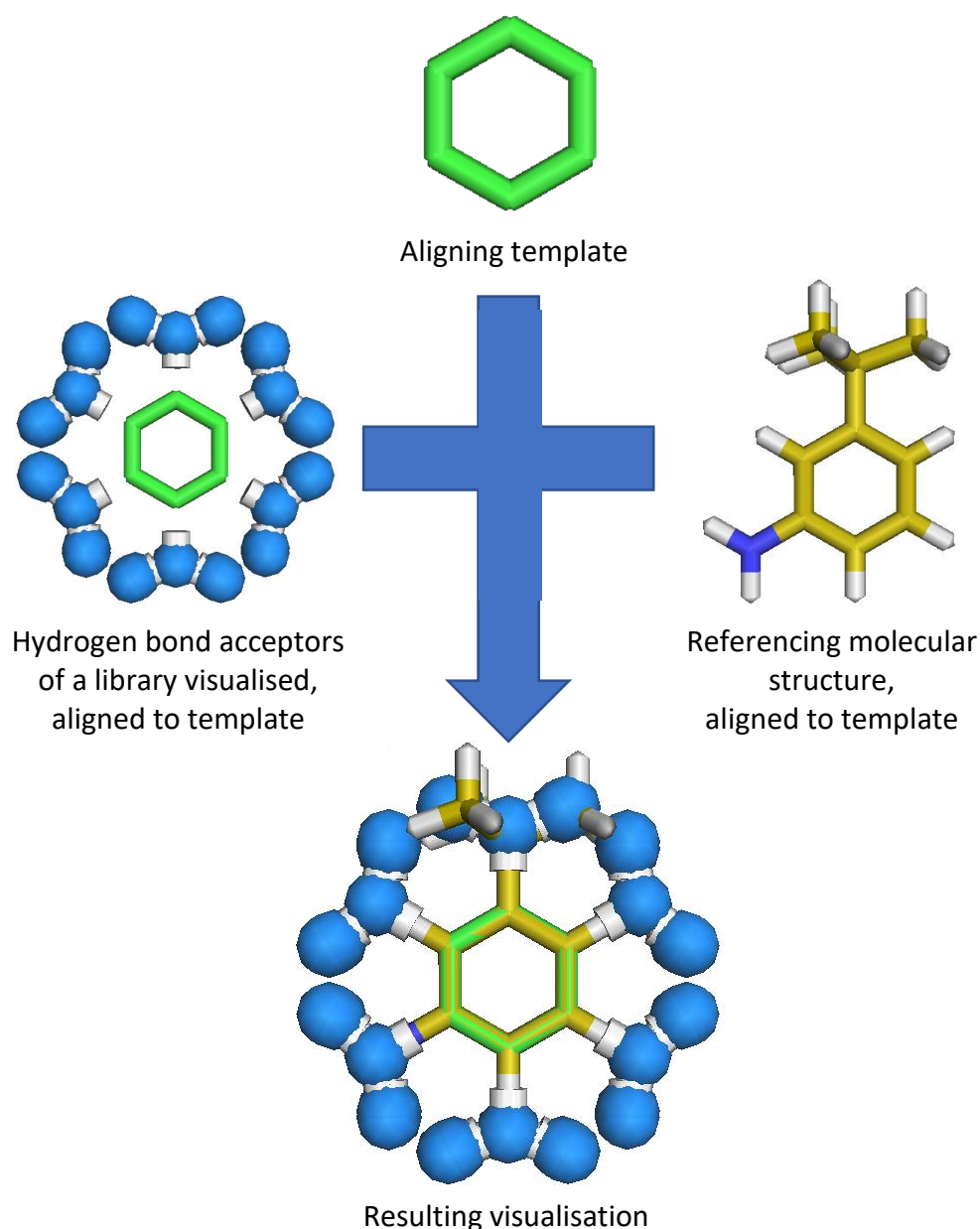


Figure 5.7. Visualisation of chemical functionalities of a library with reference to a specific molecular structure

5.3.2. Align library

The imported library is then aligned. Within this project, we provide 2 methods of alignments: geometric alignment by the plane and line of best fit of the heavy atoms within the molecules, or substructure alignment against template scaffold(s).

5.3.2.1. Geometric alignment

When geometric alignment is the selected method, the plane of best fit, followed by the line of best fit on the plane of best fit of all the heavy atoms within each molecule is identified. Geometrically aligning molecules using the coordinates of the heavy atoms is an approach taken similar to Plane of Best Fit in characterising the 3-dimensionality of molecules [18], where the average distance of the heavy atoms from the plane of best fit is used to quantify the 3-dimensionality of the molecules. The transformation required for the plane of best fit to align to the xy-plane and the line of best fit to align with the x-axis are calculated and is applied to all the atoms within the molecule (**Figure 5.8**). The benefit of this method that this is applicable to any molecules. However, it is rare for any substructures to be similarly aligned within the library and therefore may not be optimal for comparing different libraries.

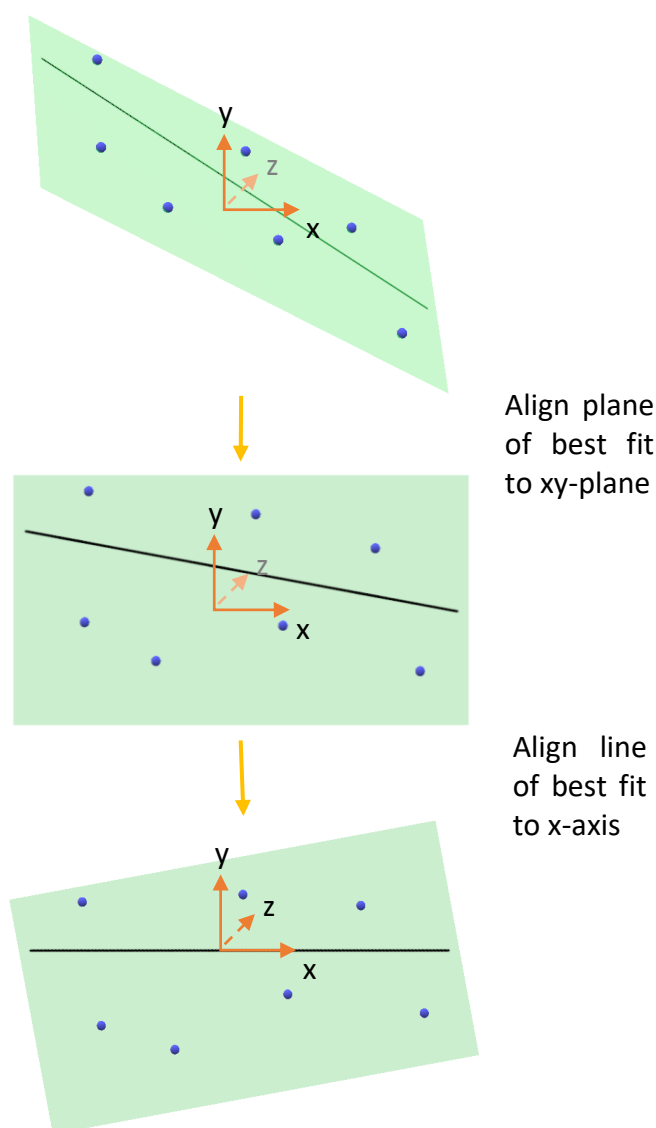


Figure 5.8. Schematic illustration of aligning a set of coordinates to the xy-plane and x-axis.

5.3.2.2. *Substructure alignment*

When substructure alignment is the selected method, a list of templates containing scaffolds which the chemical library is to be aligned against is required. These templates need to be aligned against each other following unified criteria and capture the rotational axis of the scaffolds to be usable for comparisons. The details for the criteria and preparation of the templates are described in 5.2.2.2.1 *Template preparation*.

With the prepared template list imported, molecules in the chemical library to be visualised (and referencing molecular structure if imported) are aligned against the template list. Where there are multiple template substructures within the molecule, the substructure with the most rings and atoms nearest to the centre of the molecule is selected as the aligning substructure.

Once aligned, for any template substructures containing rotational axes identified using chemical structure viewer Jmol [19] (C₂, C₃, C₄, C₅, C₆ and C₈), the aligned molecules are rotated accordingly (**Figure 5.9**). This is to capture all the possibilities of the molecule aligning to the specific template substructure.

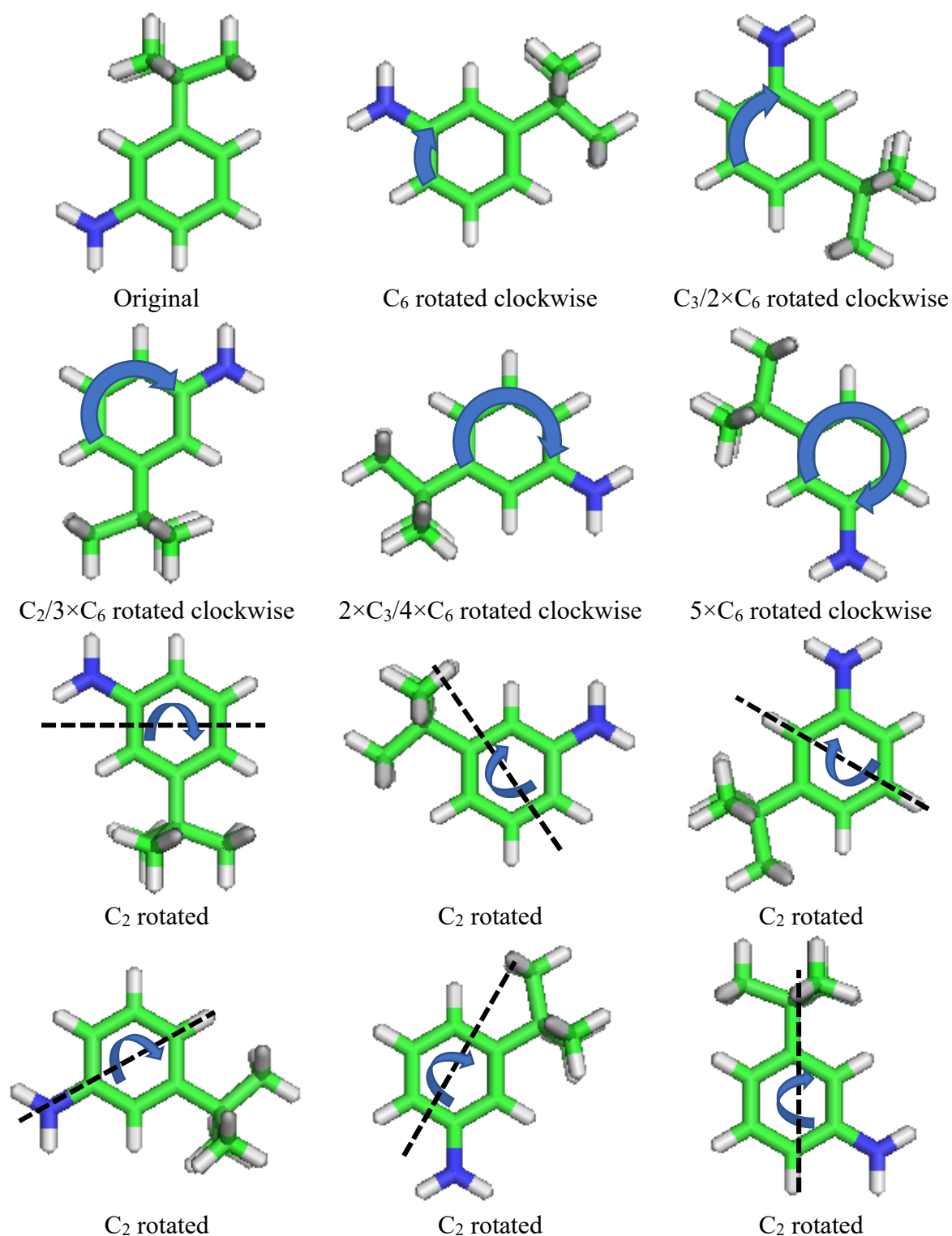


Figure 5.9. Example of a molecule with benzene scaffold rotated according to the C_2 , C_3 and C_6 rotational axis of benzene.

The benefit of this method is that chemical functionalities are visualised in relation to a selected substructure and therefore different chemical libraries can be compared against each other provided the alignment templates used are identical or structurally similar. However, this method is not applicable to structures without ring scaffolds and the ring scaffold within the

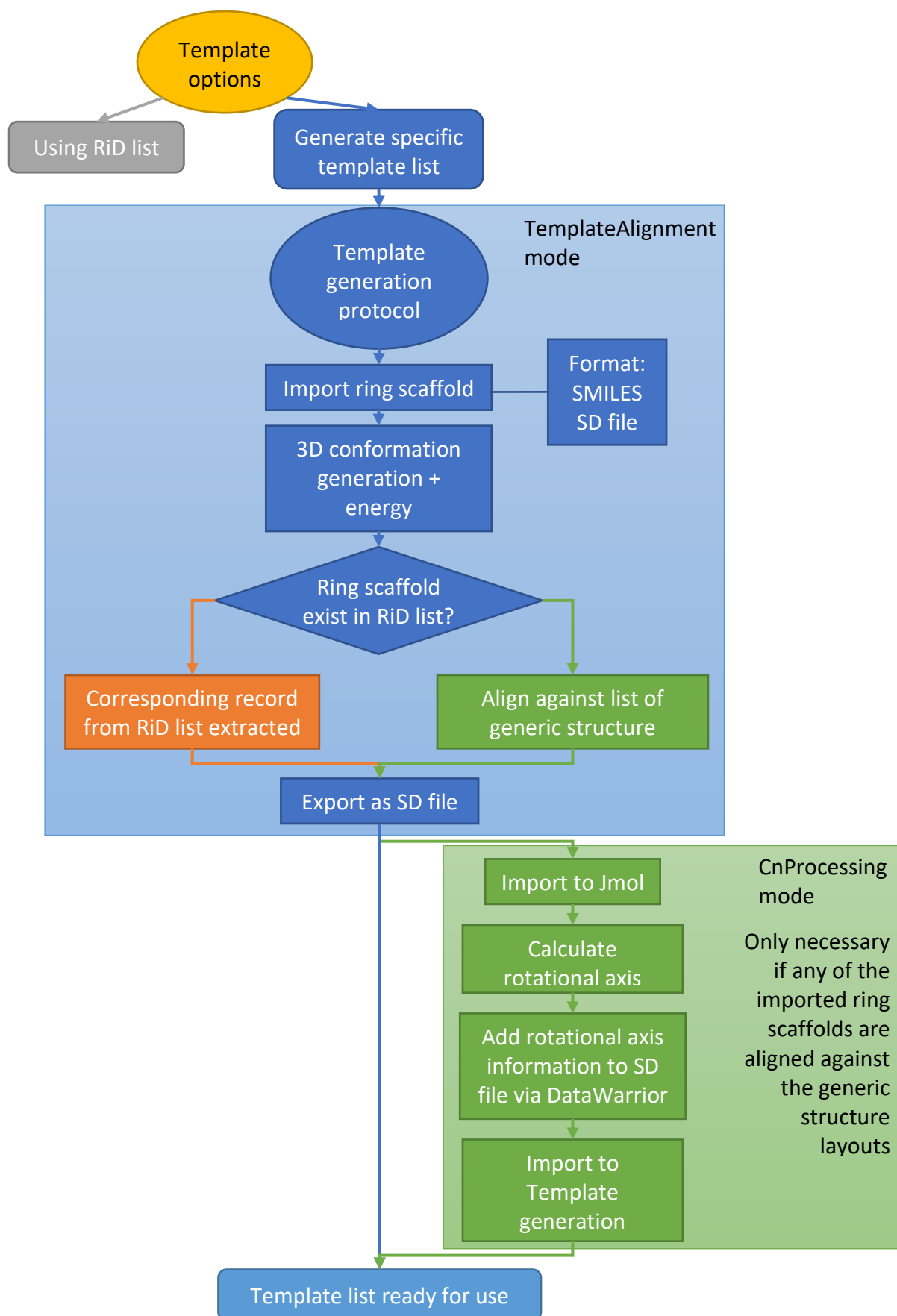
molecules must exist in the alignment template list for alignment to occur. Any molecules not aligned are removed from further calculations.

As the template scaffolds are the core of the substructure alignment, preparation of them to make sure they are aligned against each other and contain the necessary rotational axis information is required.

5.3.2.2.1. Template preparation

There are two options available for the template scaffolds (**Scheme 5.2**):

- i) using the prepared template list generated from Rings in Drugs [20] (RiD) list, or
- ii) a specific template list using the Template generation protocol.



Scheme 5.2. A flowchart showing the process of template preparation on default settings.

As Rings in Drugs captured all the U.S. Food and Drug Administration approved drug ring scaffolds in literature publication to date [20], the RiD list provides a ready prepared scaffold list for alignment, containing any rotational axis information generated from JMol where available, which captures many ring scaffolds likely to appear in a chemical library.

On the other hand, when using the Template generation protocol to prepare a template list, it allows users to select specifically individual ring scaffolds which suits the visualising chemical library the best for a given use. However, extra steps might be required to manually add rotation axis information generated from JMol where necessary as rotational axis information cannot be obtained within Pipeline Pilot. The procedure is explained below.

The template generation protocol first imports the selected ring scaffolds as SMILES or SD file, and the 3D conformations are calculated and energy minimised using the Clean forcefield [17]. Next, it compares the selected ring scaffolds against the RiD list. If a selected ring scaffold is identical to a scaffold within the RiD list, the corresponding record containing the aligned scaffold and any rotational axis information is selected for exportation.

On the other hand, if the selected ring scaffold does not exist in the RiD list, it is aligned against a list of generic structure layouts (**Figure 5.10**) following the criteria below where possible:

- First, when a structure contains multiple rings, the largest ring is placed towards the negative xy-direction with other rings extending towards to positive x- or positive xy-direction so the structure is as linear as possible; similar ring scaffolds are to be aligned in a similar fashion;
- Secondly, when a structure contains heteroatoms, at least one of the heteroatoms are to be pointing towards the positive y- or positive x-direction where it does not violate the previous criterion.

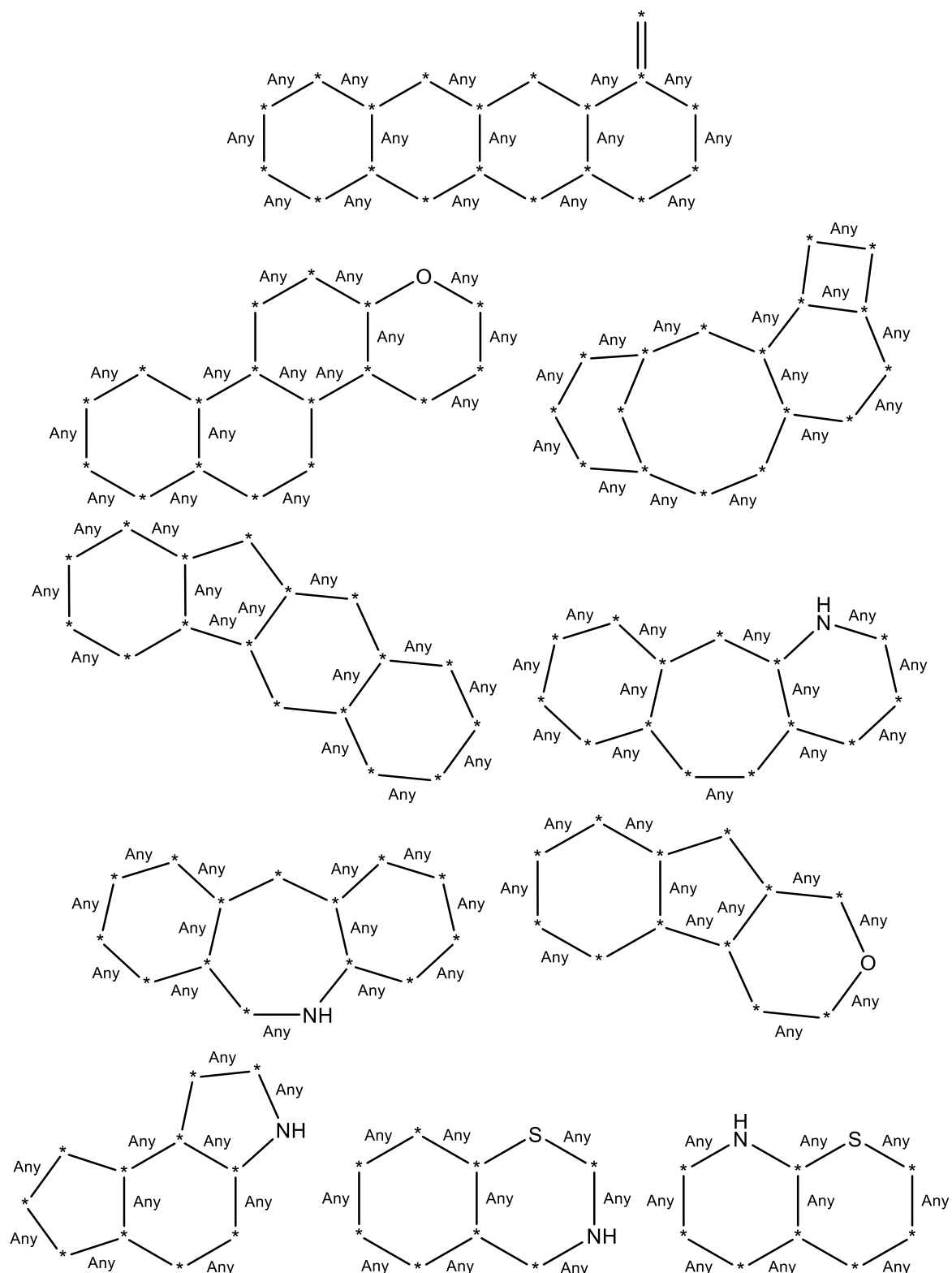


Figure 5.10. Examples of generic structure layouts templates are to be aligned against.

Once aligned, the alignments are exported as a SD file and imported into Jmol for rotational axis calculation (**Figure 5.11**). The rotational axis information is then added to the SD file manually using DataWarrior and processed using the Template generation protocol for correct formatting. This allows the substructure alignment protocol to correctly interpret the rotational

axis information and rotate molecules aligned to the template accordingly. The RiD list was originally prepared using the Template generation protocol.

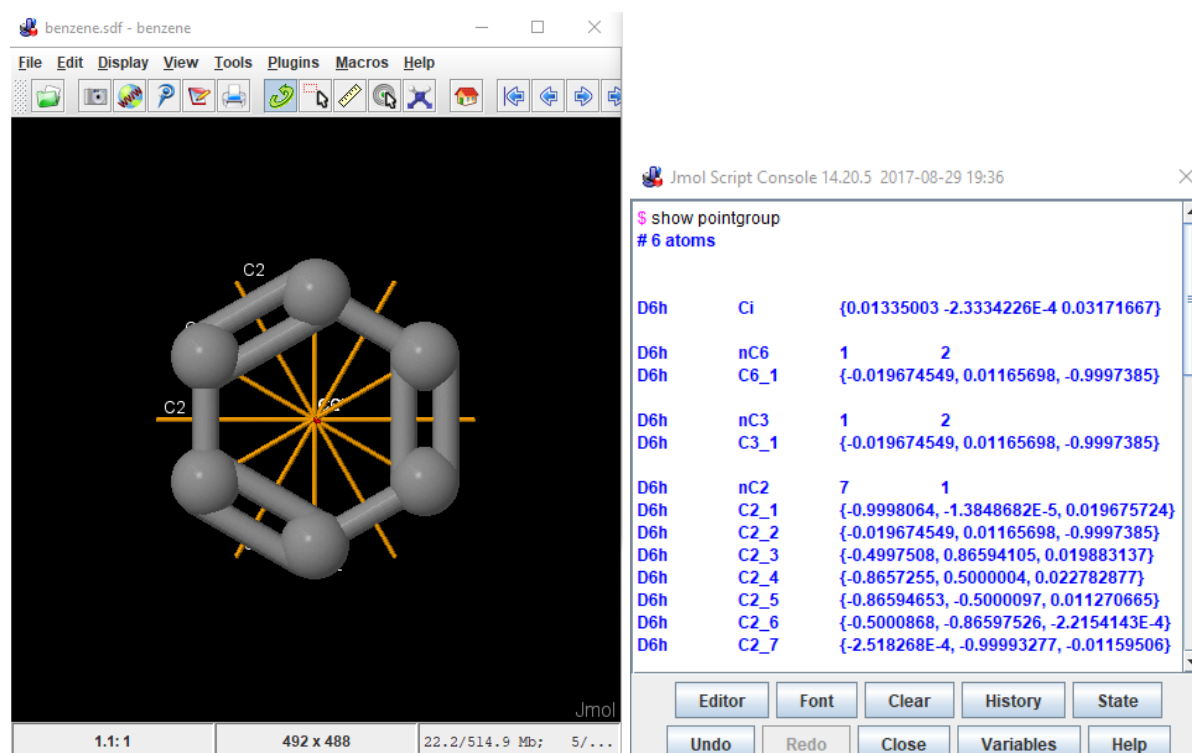


Figure 5.11. Jmol displaying rotational axis information of benzene.

5.3.3. Identify query chemical functionalities

Once the molecules in the library are aligned, they are analysed for chemical functionalities using SMiles ARbitrary Target Specification (SMARTS), a language used for describing molecular patterns and properties [21, 22]. Depending on the end user's needs, different combinations of query chemical functionalities can be selected. Within the protocol, chemical functionalities are identified as **Table 5.1** and **Figure 5.12**, where some of the SMARTS definitions used originate from Daylight Chemical Information System, Inc. [23]. Once identified, the atoms and coordinates of the chemical functionalities are stored, ready to be exported.

CHAPTER 5: VISUALISATION OF CHEMICAL FUNCTIONALITY FOR A CHEMICAL LIBRARY

Table 5.1. The definitions of the chemical functionalities identifiable within the protocol (for SMARTS definition see **Supporting information 5.1**)

Chemical functionality		Definition
Hydrogen-bond donor	Heteroatoms	Coordinate of the hydrogen bond donor heteroatom
	Hydrogens	Coordinate of the hydrogen bond donor hydrogen
Hydrogen-bond acceptors		Coordinate of the hydrogen bond acceptor atom
Charged atoms	Positive	Coordinate of the positively charged atom
	Negative	Coordinate of the negatively charged atom
Hydrophobics	Terminals	Maximum length from point of attachment = 3 Range of size: 1-7 atoms Branching possible (Figure 1a)
	Aliphatic chains	Range of size: 1-3 atoms Branching possible (Figure 1b)
	Rings	Range of size: 3-8 atoms
	Aromatic rings	Range of size: 3-8 atoms
	sulphur	
Pi systems		The structure of the pi systems

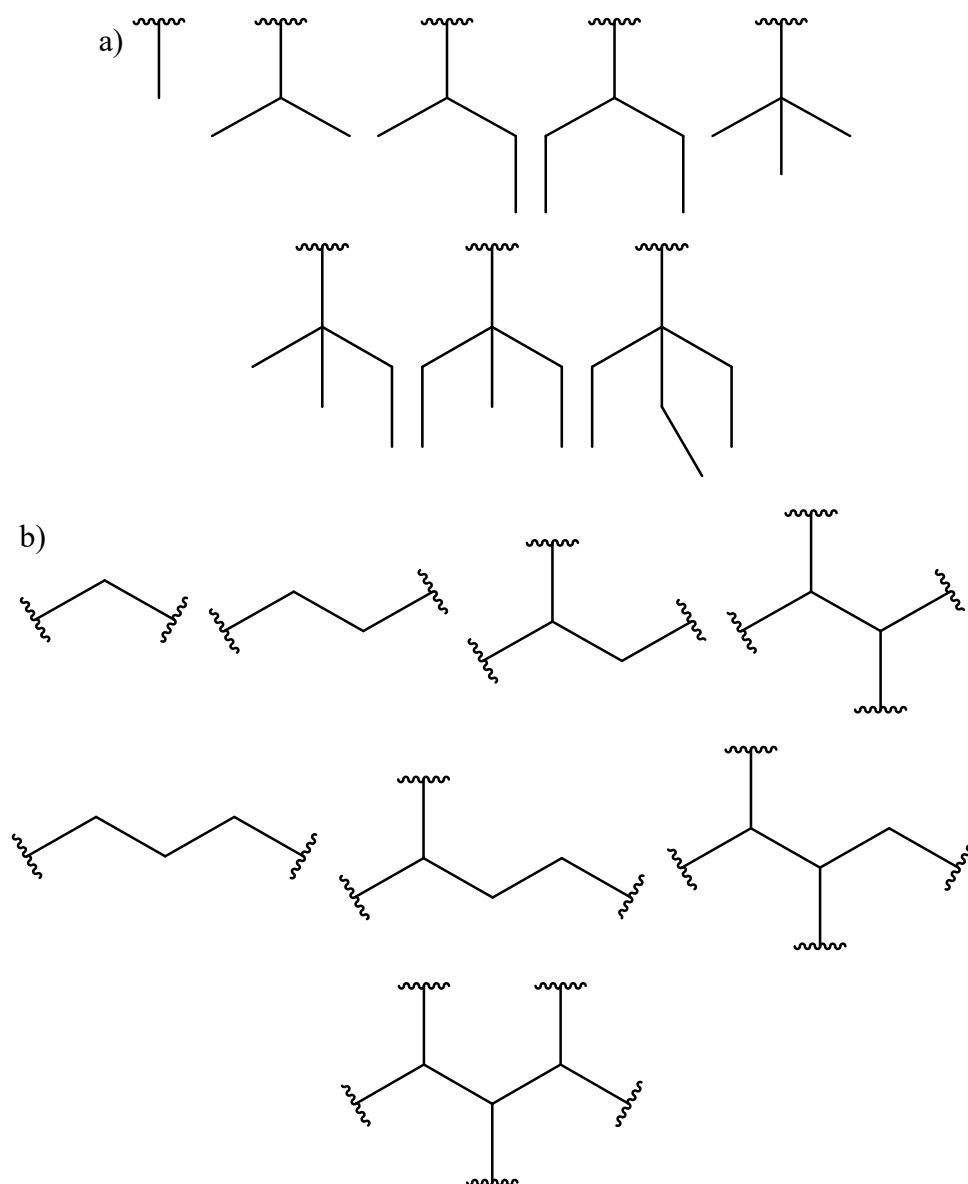


Figure 5.12. Structures of a) hydrophobic terminals and b) aliphatic chains.

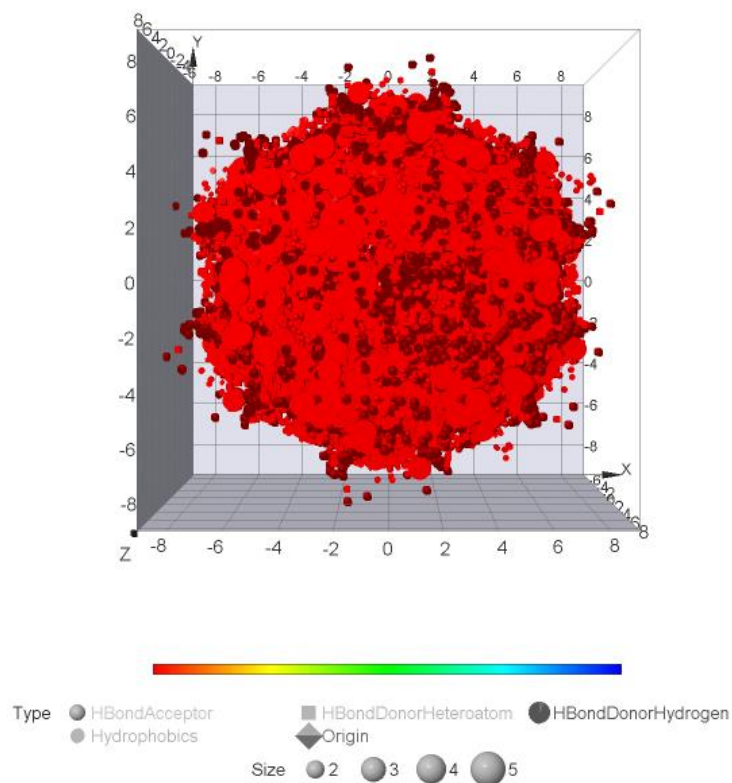
5.3.4. Export visualisation

The stored chemical functionality information is exported, ready for visualisation. Within this work, we provide two visualisation methods: DataWarrior graphical summary and PyMOL full visualisation.

5.3.4.1. *DataWarrior graphical summary*

The DataWarrior graphical summary is obtained by importing the stored chemical functionality information into DataWarrior and processing it using the macro provided with the protocol. It is to note that the coordinates of the chemical functionality information for this visualisation are binned into 0.5\AA bins with conformer shape calculation error consideration for clearer imaging and easier identification of chemical functionality coordinate duplication by the same molecule (**Figure 5.13**). When only one template is used in the substructure alignment method, or when a reference molecule is imported, the individual atoms of the structure are represented in grey and labelled with their atomic symbol. When multiple templates are used in the substructure alignment method, a grey shape locates the origin (0, 0, 0).

a)



b)

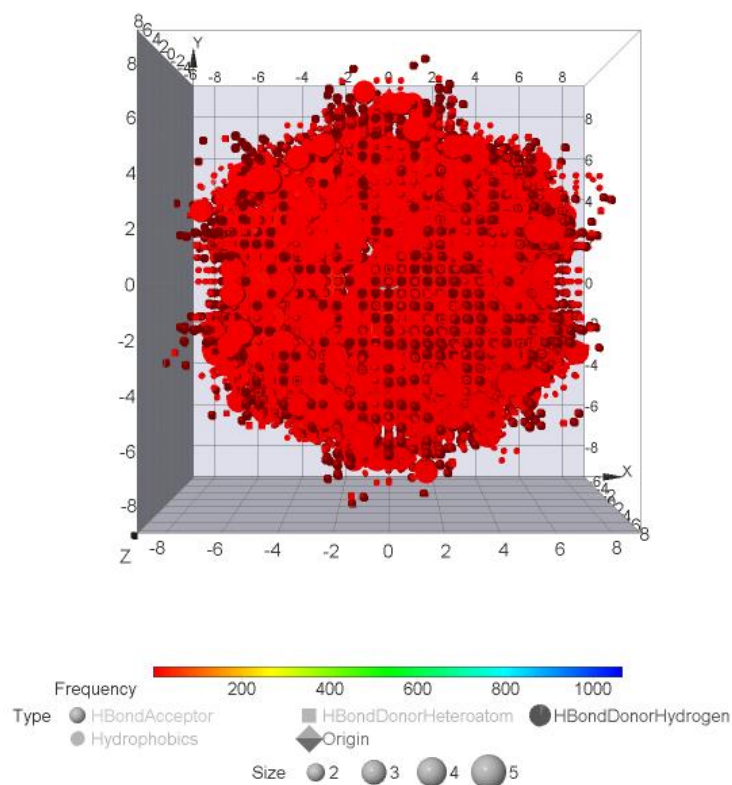


Figure 5.13. Location of the four indicated chemical functionality with respect to centre of alignment of molecule (in Å). Difference in image clarity a) without and b) with binning of DataWarrior graphical overall summary using the LCC in-house library. Without binning, overlaps of functionality cannot be identified and therefore frequency scale is not available.

There are two options of graphical summaries: an overall summary (**Figure 5.14a, Supporting Information 5.2**) which also provides the frequency of a chemical functionality seen for each bin and does not disclose any of the molecular structure within the library, or a traceable summary (**Figure 5.14b, Supporting Information 5.3**) which by selecting a chemical functionality coordinate will highlight the molecular structures within the library which provides the chemical functionality at the specified coordinate.

The DataWarrior graphical summary provides a quick way to render images of the chemical functionalities which can be saved for viewing later. However, as it is a graphical summary, the generated image does not necessarily resemble the shape of structures of the chemical functionalities identified. The shapes used in the graphical summary are also dependent on the number of different types of chemical functionalities identified and therefore can be difficult to interpret initially.

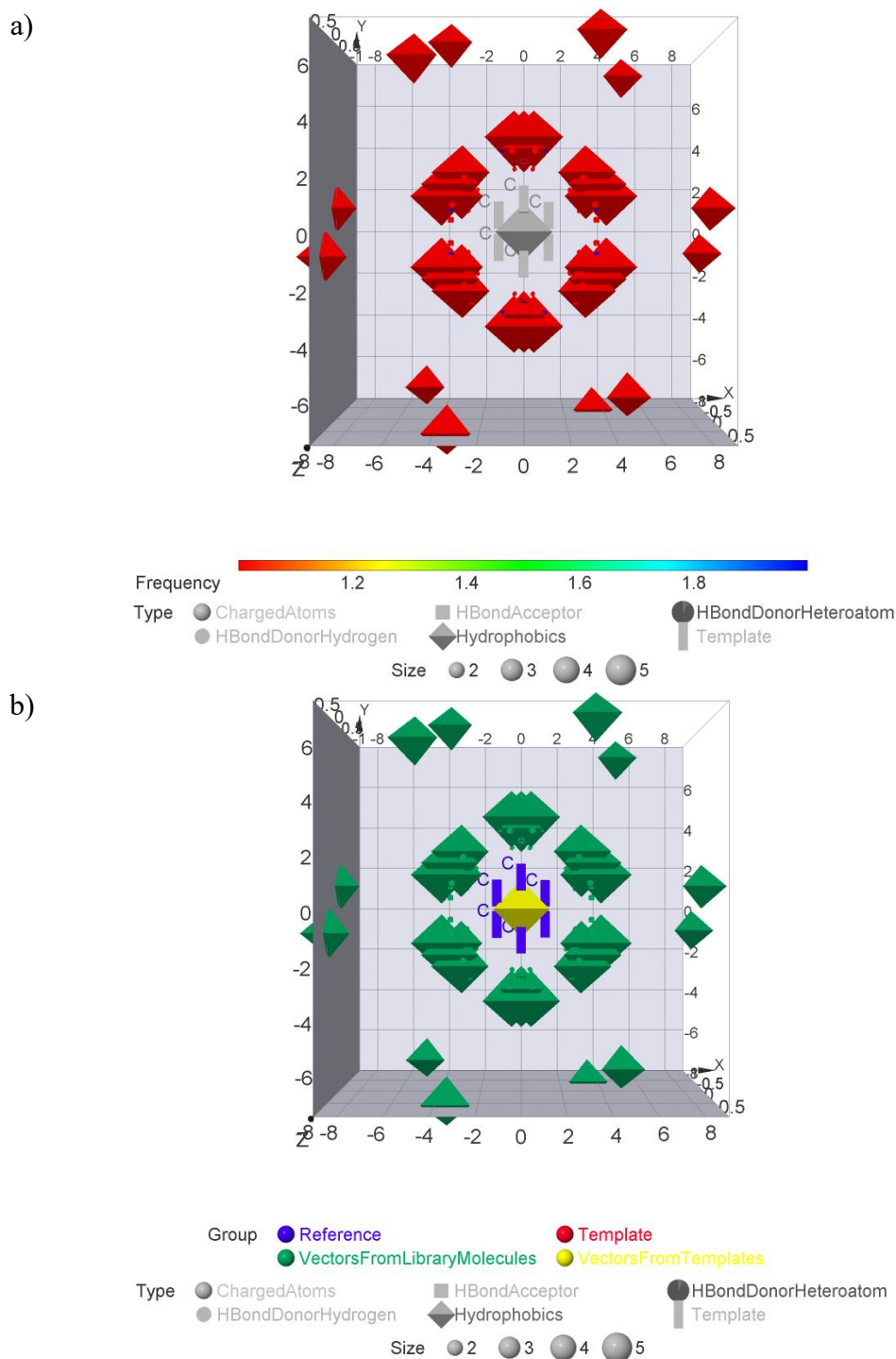


Figure 5.14. Example of DataWarrior graphical a) overall summary (**Supporting Information 5.2**) and b) traceable summary (**Supporting Information 5.3**). In both summaries, chemical functionality groups are represented with different 3D shapes, noted by Type in the legend, and the size (number of atoms) of the group is represented with the according size. In the overall summary (a), the number of times a particular chemical functionality is seen at a particular coordinate is represented by the Frequency scale. In the traceable summary (b), the colour of the data points represents whether the data points represent the referencing structure, or if the particular chemical functionality stems from the templates used or the analysing library. The traceable summary (b) allows selection of chemical functionalities in the 3D graph and tracing back to the relevant structures.

5.3.4.2. *PyMOL full visualisation*

There are two options for the PyMOL full visualisation.: Overall visualisation or Molecular visualisation.

Overall visualisation disregards the individual molecular structures of the chemical library and outputs the overall image of the chemical functionalities available within the library. Molecular visualisation keeps the individual molecular structures of the chemical library and overlays them with the chemical functionalities.

These visualisations are generated by running the PyScript generated at the end of the protocol in PyMOL initially, which is then saved as PySession for viewing later. The resulting visualisation (**Figure 5.15**) represents the chemical functionalities as shown for a simple example in **Table 5.2**. For the Overall visualisation, when only one template is used in the substructure alignment method, or when a reference molecule is imported, these structures are represented in green. On the other hand, when multiple templates are used in the substructure alignment method, a green sphere locates the origin (0, 0, 0).

Both PyMOL visualisation methods provide more of a chemical structure/functionality view of the chemical functionalities than the ones DataWarrior can provide. However, these methods are high in computational cost, and the time required for generating and rendering this visualisation increases as the chemical library size increases. For example, a list of four molecules only takes minutes to render but a library of approximately 12000 molecules can take up to 15 hours to render but once saved as a PyMol session can be loaded very quickly (PC used: Intel(R) Core(TM) i5-6500 CPU @ 3.20GHz 3.19 GHz, 32GB RAM).

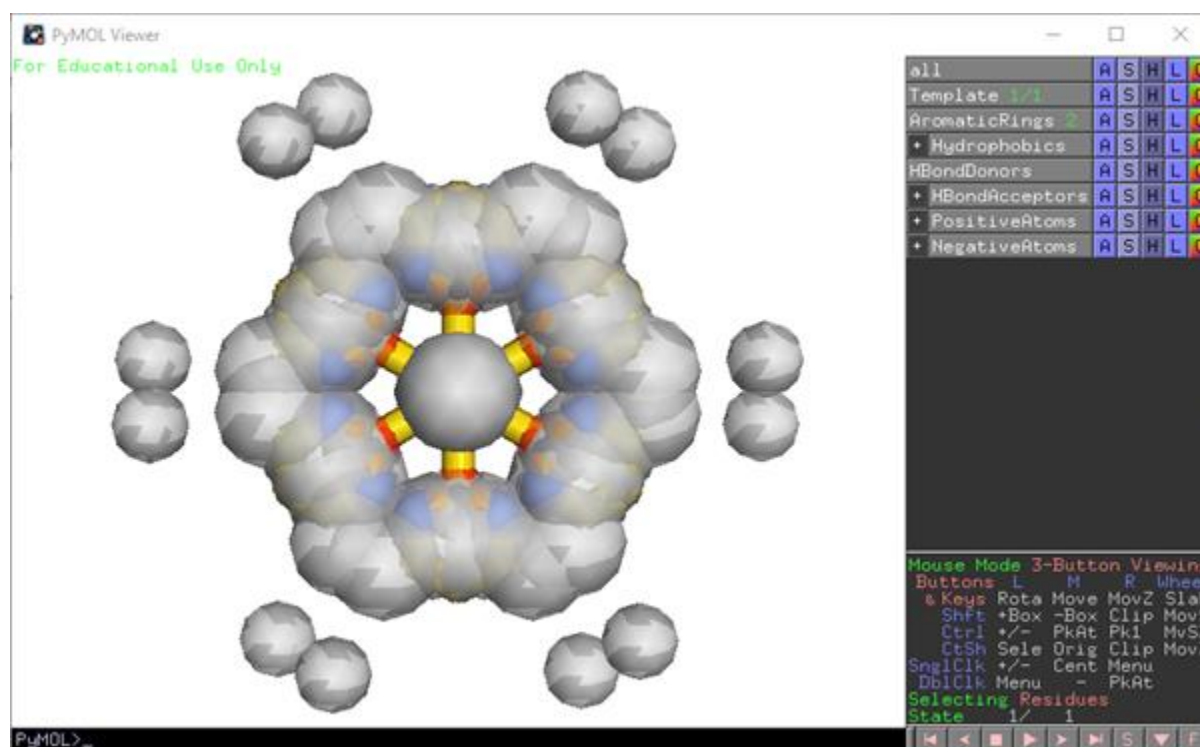
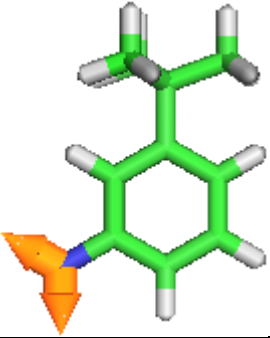
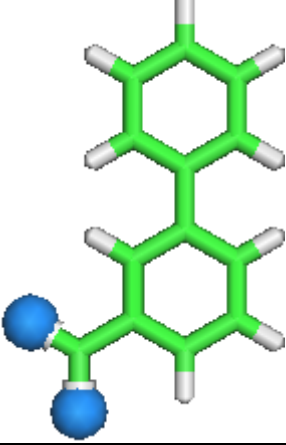
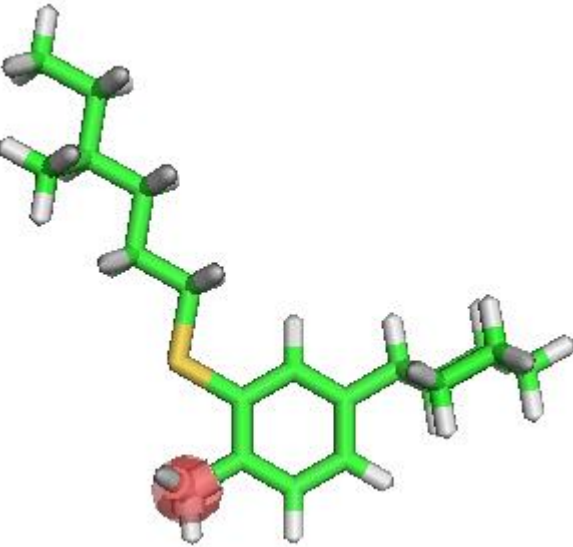
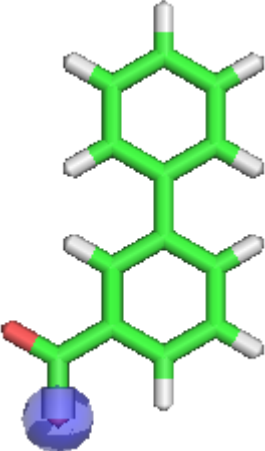
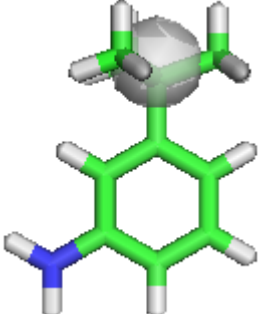
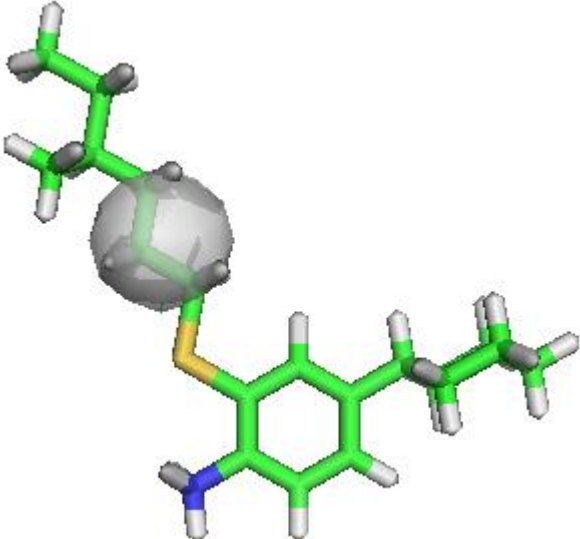
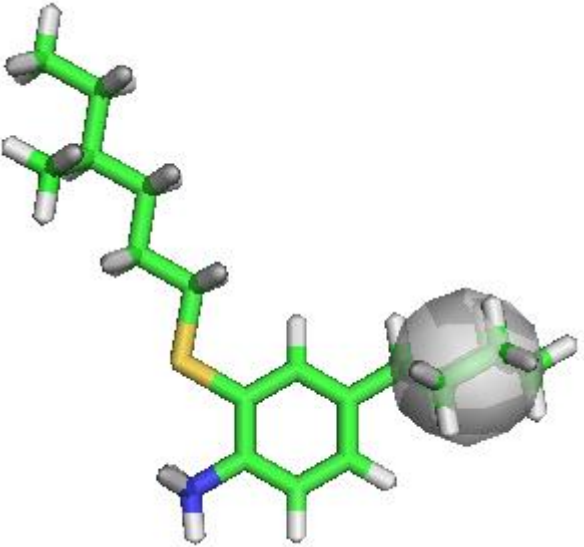
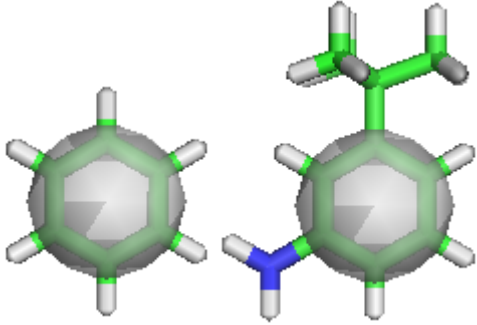
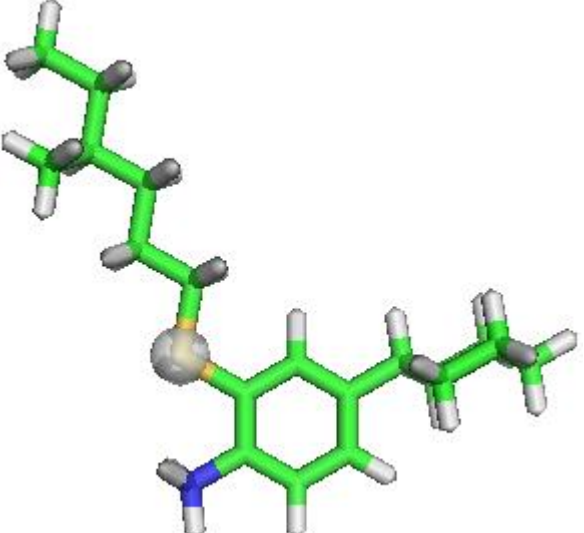
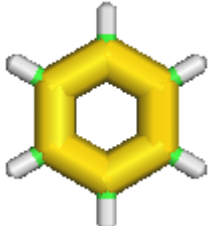


Figure 5.15. An example of PyMOL visualisation output.

Table 5.2. The definitions of the chemical functionalities identifiable within the protocol

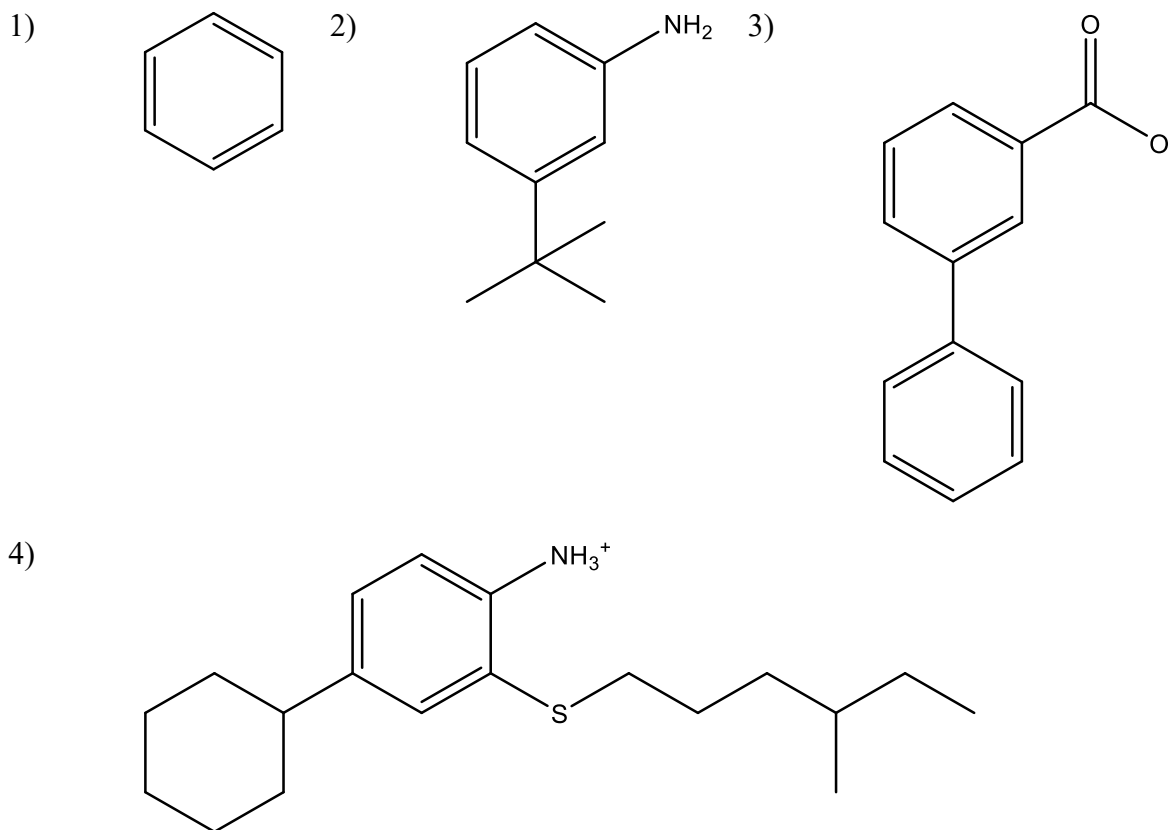
Chemical functionality	Visualization	Example
Hydrogen-bond donor	Orange arrow from hydrogen-bond donor heteroatom to hydrogen	
Hydrogen-bond acceptors	Light blue sphere with white cylinders showing non-accessible areas due to bonds	
Positively charged atom	Red sphere with white cylinders showing non-accessible areas due to bonds	

Chemical functionality	Visualization	Example
Negatively charged atom	Blue sphere with white cylinders showing non-accessible areas due to bonds	
Hydrophobic terminals	Light gray sphere of various size dependent on the number of atoms within the group	
Aliphatic chains	Light gray sphere of various size dependent on the number of atoms within the group	

Chemical functionality	Visualization	Example
Carbon rings	Light grey sphere of various size dependent on the number of atoms within the group	
Aromatic rings	Light grey sphere of various size dependent on the number of atoms within the group	
Hydrophobic Sulphur	Light grey sphere	
Pi systems	Yellow structures	

5.4. Example of Visualised Chemical Library

Using example **Molecules 5.1 – 5.4** and the LCC in-house fragment library, the different visualisation options available and their effect on the visualisation output are illustrated below.



Molecule 5.1 – 5.4. Example molecules used to illustrate the visualisation options.

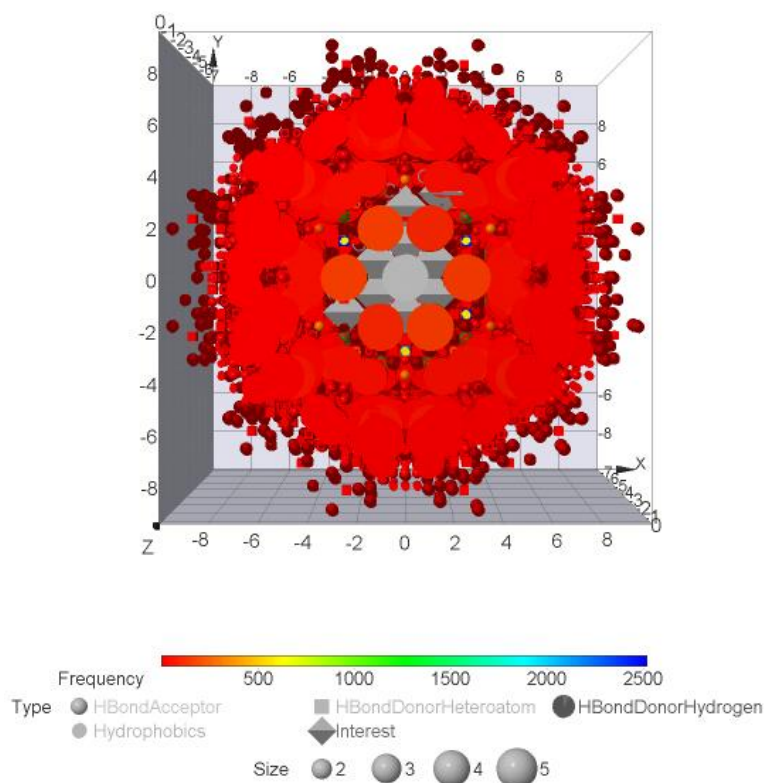
5.3.1. Settings within the protocol

There are four main settings within the Chemical functionality visualisation protocols which affects the display of the visualisation outputs:

- i) Molecule of interest
- ii) Visualisation Type
- iii) Output Type
- iv) Features.

Molecule of interest is an option to add a molecule of interest as reference to the visualisation (**Figure 5.16, Supporting Information 5.4 and 5.5**). This allows comparison of the chemical functionalities of the analysing library and the structure of the molecule of interest.

a)



b)

For Educational Use Only

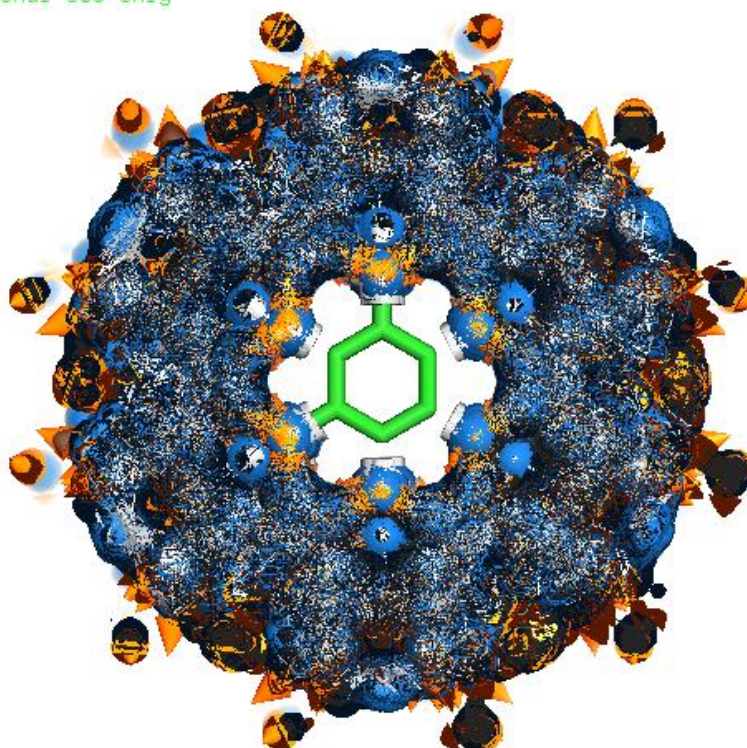


Figure 5.16. Molecule 5.2 used as molecule of interest for visualisation of LCC in-house fragment library in the a) DataWarrior overall summary (Supporting Information 5.4) and b) PyMOL overall visualisation (Supporting Information 5.5). Both images are the cross-sections of the visualisation where Molecule 5.2 is visible, with the atoms represented by grey octahedrons in the former and the molecule is in green for the latter, surrounded by the identified chemical functionalities.

As the DataWarrior graphical summary is quick to render visualisation, it is a standard in the visualisation type selection. On the other hand, as the PyMOL full visualisation is much heavier in computational cost to render, the visualisation type setting determines whether the PyMOL full visualisation will be calculated.

Output type determines which of the visualisation options within the visualisation outputs are being calculated. As explained within 5.2.4.1. *DataWarrior graphical summary* and 5.2.4.2. *PyMOL full visualisation*), the overall summary/visualisation does not disclose any molecular structures within the library analysed (**Figure 5.17, Supporting Information 5.6 and 5.7**), and therefore is suitable for companies, in the example case LCC, to display the chemical functionality of their library without exposing their structures.

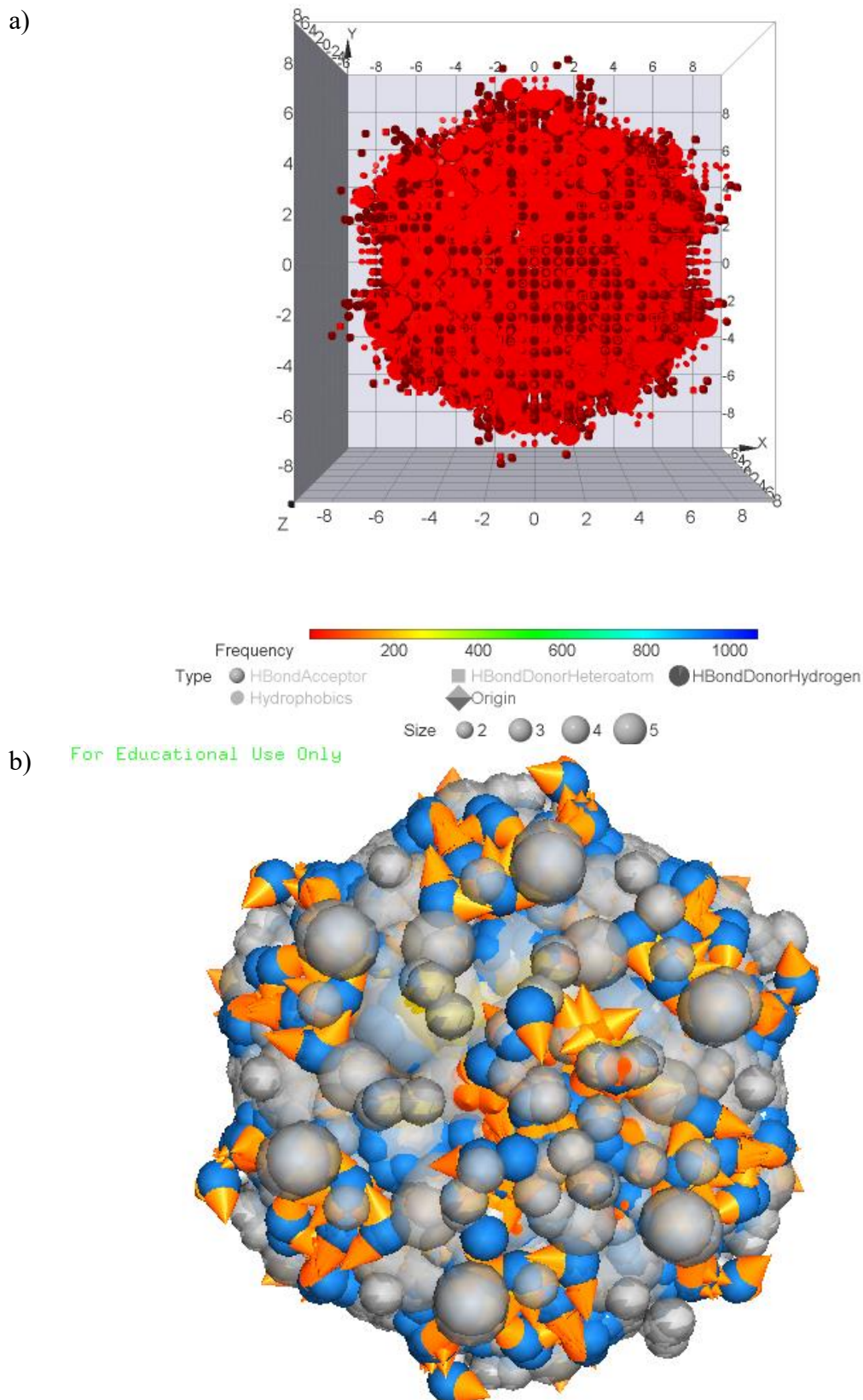


Figure 5.17. a) DataWarrior overall summary (**Supporting Information 5.6**) and b) PyMOL overall visualisation using the shape and colour as described in **Table 5.2 (Supporting Information 5.7)** of the LCC in-house fragment library, where individual structures are not available.

On the other hand, the for DataWarrior graphical traceable summary and PyMOL molecular visualisation (**Figure 5.18, Supporting Information 5.3 and 5.8**) display the chemical functionalities along with the individual molecule, allowing easy understanding of which molecules display which functionalities.

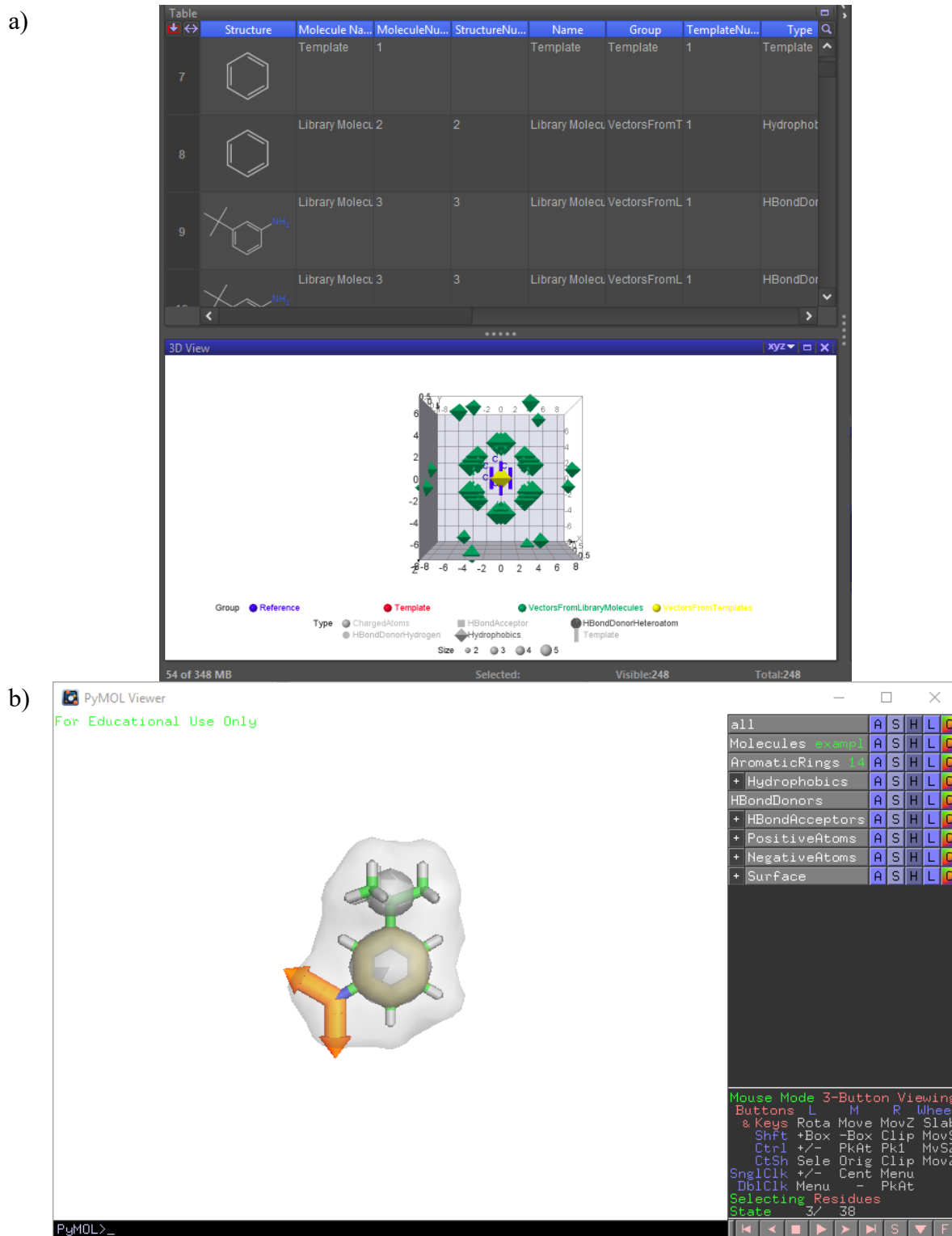


Figure 5.18. a) DataWarrior traceable summary (**Supporting Information 5.3**) and b) PyMOL molecular visualisation (**Supporting Information 5.8**) of Molecule 5.1 – 4, where individual structures are available.

The feature option determines which categories of chemical functionalities are to be calculated. There are five categories of chemical functionalities selectable: hydrogen bond donor, hydrogen bond acceptor, charged atoms, π -system and hydrophobics (Detail see **Table 5.1**). Only the selected chemical functionalities would be calculated and it is suggested to calculate all categories as within the individual visualisation, there are options to turn the display of selected chemical functionalities off (Detail see 5.3.2. *Options within each visualisation output*). However, it is to note that for large libraries, this might lead to heavy computational cost, especially for the PyMOL full visualisation generation.

5.3.2. Options within each visualisation outputs

5.3.2.1. DataWarrior overall summary

Figure 5.19 shows the DataWarrior overall summary of **Molecule 5.1 – 4** (**Supporting Information 5.2**). Within area A, the following options are available to determine visibility of the functionalities displayed in area B where available:

- Reference point/atom position outline
- Chemical functionalities from the template structure
- Chemical functionalities from the library visualising
- Chemical functionality categories (**Table 5.1**)

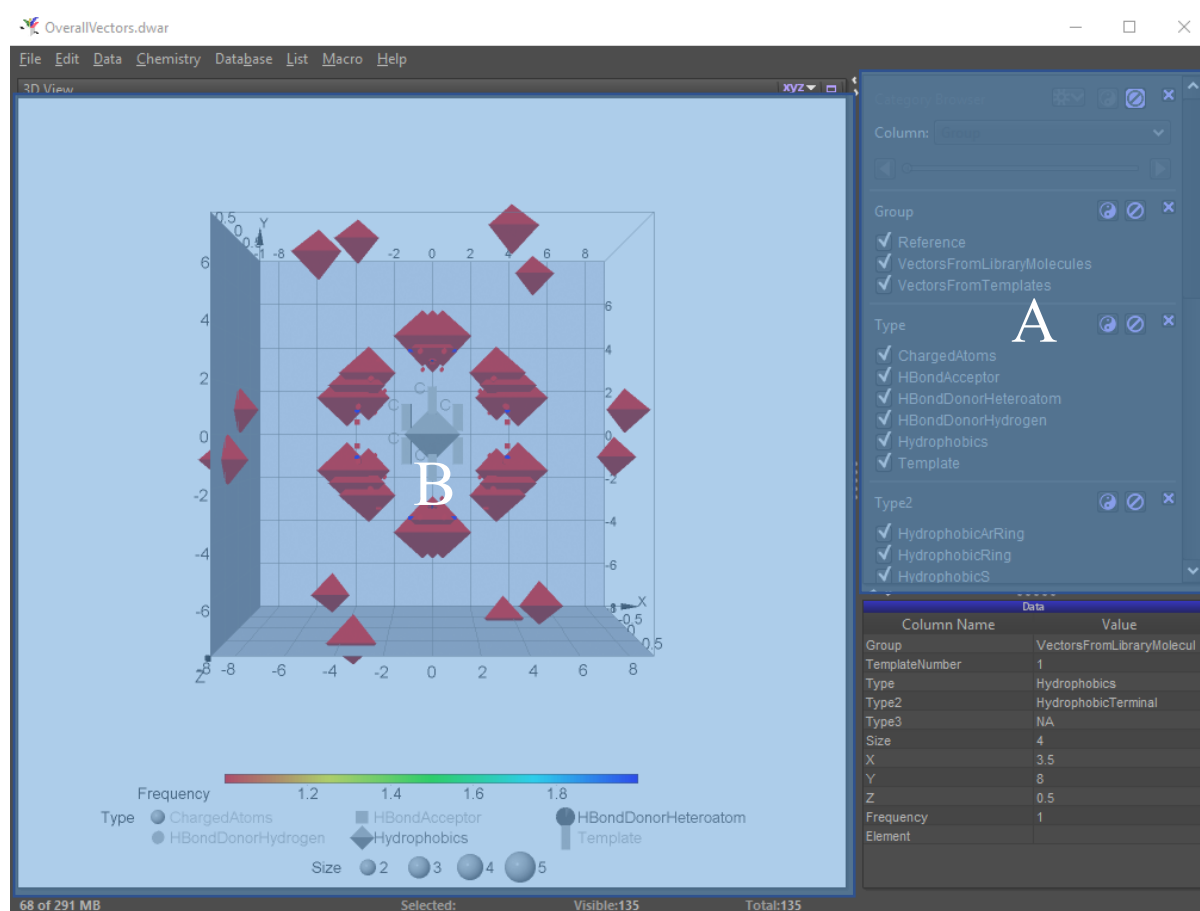


Figure 5.19. DataWarrior overall summary of **Molecules 5.1 – 4** (**Supporting Information 5.2**), where area A contains the options to adjust the visibility of the functionalities displayed in area B.

5.3.2.2. *DataWarrior traceable summary*

Figure 5.20 shows the DataWarrior traceable summary of **Molecule 5.1 – 4 (Supporting Information 5.3)**. Within section A, the following options are available to determine visibility of the functionalities displayed in section B where available:

- Reference point/atom position outline
- Chemical functionalities from the template structure
- Chemical functionalities from the library visualising
- Chemical functionality categories (**Table 5.1**)

By selecting chemical functionalities within section B, the corresponding molecules entries are highlighted in section C. These molecules can then be extracted by creating a new file from the highlighted entries.

The screenshot shows the DataWarrior software interface. At the top is a menu bar with 'File', 'Edit', 'Data', 'Chemistry', 'Database', 'List', 'Macro', and 'Help'. Below the menu is a table with columns: Structure, Molecule Na..., MoleculeNu..., StructureNu..., Name, Group, TemplateNu..., and Type. The table contains four rows of molecule data. A large white letter 'C' is overlaid on the table. To the right of the table is a settings panel with sections for 'Molecule Name', 'Name', 'Group', and 'Type'. Each section has a list of checkboxes. A large white letter 'A' is overlaid on this panel. Below the settings panel is a 'Data' table with columns 'Column Name' and 'Value'. The 'Data' table shows 'Structure' and '3D-Structure' with corresponding molecular images. A large white letter 'B' is overlaid on the '3D-Structure' image. At the bottom of the interface is a 3D view window showing a ball-and-stick model of a molecule on a grid. A large white letter 'B' is overlaid on this view. At the very bottom, there are status indicators: '54 of 348 MB', 'Selected:', 'Visible:248', and 'Total:248'.

Figure 5.20. DataWarrior traceable summary of **Molecule 5.1 – 4 (Supporting Information 5.3)**, where area A contains the options to adjust the visibility of the functionalities displayed in area B, and area C contains the table of molecule entries.

5.3.2.3. *PyMOL overall visualisation*

Figure 5.21 shows the PyMOL overall visualisation of **Molecule 5.1 – 4 (Supporting Information 5.9)**. Section A lists the PyMOL objects, named following the chemical functionality categories (**Table 5.1**) available within the visualisation. By clicking on the objects, visualisation of the objects in section B can be triggered on and off. Some of the lists, for example Hydrophobics, can be expanded to show further breakdowns of the chemical functionalities by subtypes and size.

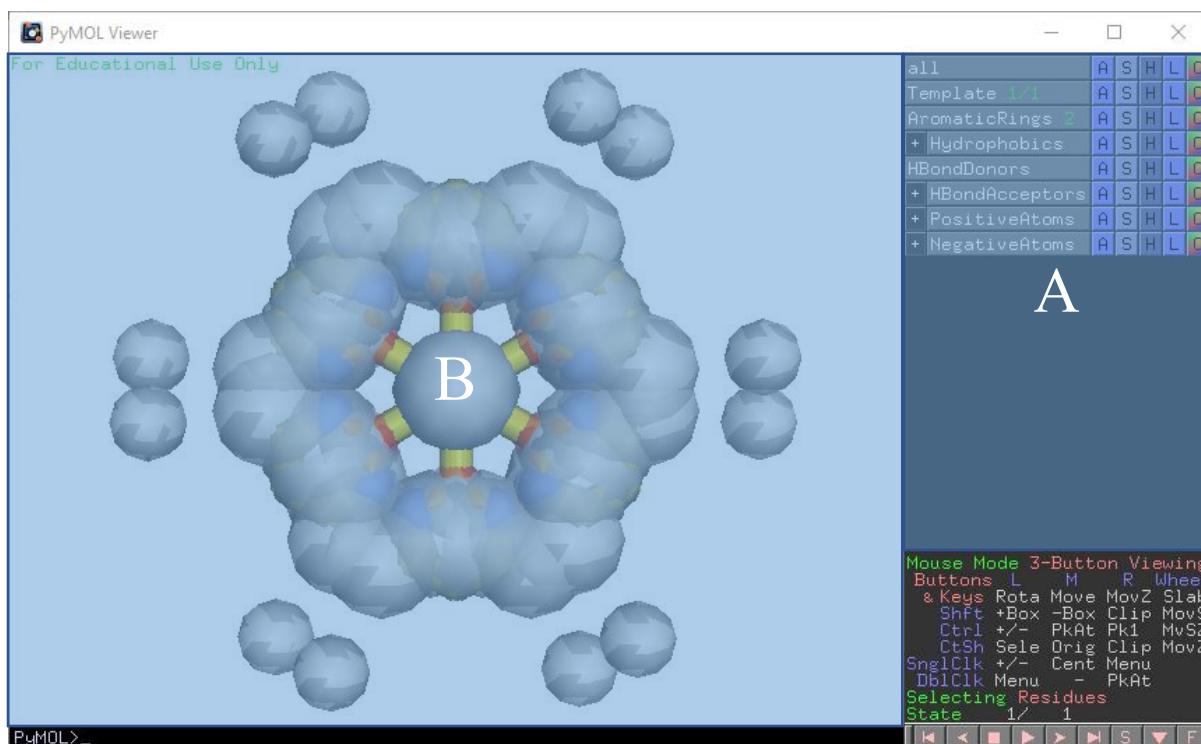


Figure 5.21. PyMOL overall visualisation of **Molecule 5.1 – 4** (**Supporting Information 5.9**), where area A contains the options to adjust the visibility of the functionalities displayed in area B.

5.3.2.4. *PyMOL molecular visualisation*

Figure 5.22 shows the PyMOL molecular visualisation of **Molecule 5.1 – 4** (**Supporting Information 5.8**). Section A lists the PyMOL objects, named following the chemical functionality categories (**Table 5.1**) available within the visualisation. By clicking on the objects, visualisation of the objects in section B can be toggled on and off. Some of the lists, for example hydrophobics, can be expanded to show further breakdowns of the chemical functionalities by subtypes and size. Within section C is the control for moving between molecules of display.

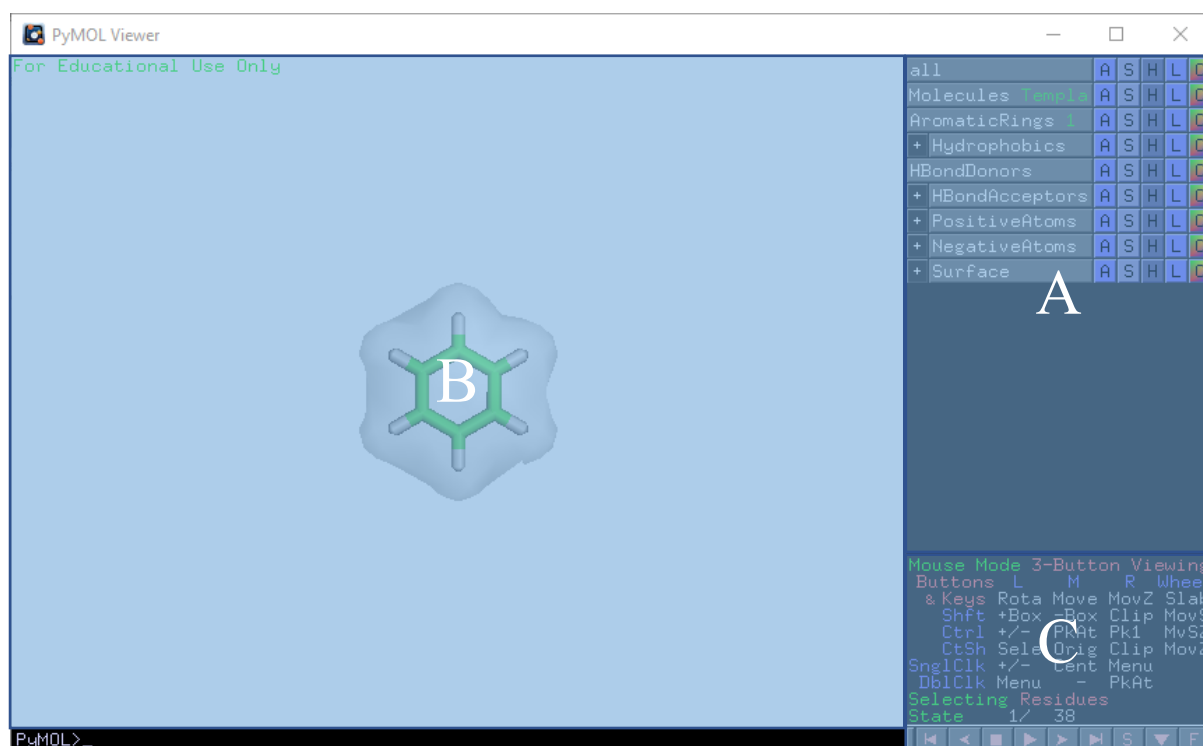


Figure 5.22. PyMOL molecular visualisation of **Molecule 5.1 – 4 (Supporting Information 5.8)**, where area A contains the options to adjust the visibility of the functionalities displayed in area B, and area C contains the controls for moving between molecule of display.

5.4. Further Work

This work was the start of research which has developed further to an open innovation project with LCC and Abbvie called Project Vector where the aim was to quantify the efficiency of a library at covering 3D space. Project Vector uses the same fundamental steps to identify the vectors of various chemical functionalities of a fragment library, then uses the vector information to calculate vector efficiency, a novel quantity for quantifying the efficiency of a library at covering a defined 3D space with its chemical functionalities. Using the concept of vector efficiency, allows the analysis and comparison of the diversity and the 3-dimensionality of chemical functionalities of fragment libraries. Much work had been spent in this project at developing the Vector efficiency calculation algorithm and the most recent work had been on translating the protocol to an open platform KNIME for accessibility for a wider range of end users. Efforts have also been put toward testing the protocol with a wider range of libraries and preparation for publication.

In addition, the applicable chemical space of the protocols can be expanded in theory by extending the substructure alignment algorithm from ring scaffold to any substructures. However, complications related to rotation axis and plane of alignment arise when the substructure become too simple, e.g. C-C, and therefore more work is required in order to apply this.

5.5. References

1. McNaught, A. D.; Wilkinson, A. *IUPAC. Compendium of Chemical Terminology, 2nd ed. (the "Gold Book")*. Blackwell Scientific Publications, Oxford (1997), (accessed 08/06/2021).
2. Harrold, M. W.; Zavod, R. M. *Basic Concepts in Medicinal Chemistry*; American Society of Health-System Pharmacists, 2013.
3. Bajorath, J. Pharmacophore. In *Encyclopedia of Cancer*, Schwab, M. Ed.; Springer Berlin Heidelberg, 2011; pp 2849-2852.
4. Roy, K.; Kar, S.; Das, R. N. Chapter 10 - Other Related Techniques. In *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*, Roy, K., Kar, S., Das, R. N. Eds.; Academic Press, 2015; pp 357-425.
5. Yang, S.-Y. Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug Discovery Today* **2010**, *15* (11), 444-450. DOI: <https://doi.org/10.1016/j.drudis.2010.03.013>.
6. Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discovery Today* **2002**, *7* (17), 903-911. DOI: [https://doi.org/10.1016/S1359-6446\(02\)02411-X](https://doi.org/10.1016/S1359-6446(02)02411-X).
7. Koes, D. R.; Camacho, C. J. Pharmer: Efficient and Exact Pharmacophore Search. *Journal of Chemical Information and Modeling* **2011**, *51* (6), 1307-1314. DOI: 10.1021/ci200097m.
8. Koes, D. R.; Camacho, C. J. ZINCPharmer: pharmacophore search of the ZINC database. *Nucleic acids research* **2012**, *40* (Web Server issue), W409-W414. DOI: 10.1093/nar/gks378 PubMed.
9. Irwin, J. J.; Shoichet, B. K. ZINC--a free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling* **2005**, *45* (1), 177-182. DOI: 10.1021/ci049714+ From NLM.
10. Boyd, S. M.; Beverley, M.; Norskov, L.; Hubbard, R. E. Characterising the geometric diversity of functional groups in chemical databases. *Journal of Computer-Aided Molecular Design* **1995**, *9* (5), 417-424. DOI: 10.1007/BF00123999.
11. Leach, A. R.; Green, D. V.; Hann, M. M.; Judd, D. B.; Good, A. C. Where are the GaPs? A rational approach to monomer acquisition and selection. *Journal of Chemical Information and Modeling* **2000**, *40* (5), 1262-1269. DOI: 10.1021/ci0003855 From NLM.
12. Saini, A.; Verma, G. Chapter 10 - Peptoids: tomorrow's therapeutics. In *Nanostructures for Novel Therapy*, Ficai, D., Grumezescu, A. M. Eds.; Elsevier, 2017; pp 251-280.
13. Aina, O. H.; Liu, R.; Sutcliffe, J. L.; Marik, J.; Pan, C.-X.; Lam, K. S. From Combinatorial Chemistry to Cancer-Targeting Peptides. *Molecular Pharmaceutics* **2007**, *4* (5), 631-651. DOI: 10.1021/mp700073y.
14. Agrafiotis, D. K.; Lobanov, V. S.; Salemme, F. R. Combinatorial informatics in the post-genomics era. *Nature Reviews Drug Discovery* **2002**, *1* (5), 337-346. DOI: 10.1038/nrd791.
15. Hann, M. M.; Oprea, T. I. Pursuing the leadlikeness concept in pharmaceutical research. *Current Opinion in Chemical Biology* **2004**, *8* (3), 255-263. DOI: <https://doi.org/10.1016/j.cbpa.2004.04.003>.
16. LiverpoolChiroChem. <https://www.liverpoolchirochem.com/> (accessed 26/01/2022).
17. Hahn, M. Receptor Surface Models. 1. Definition and Construction. *Journal of Medicinal Chemistry* **1995**, *38* (12), 2080-2090. DOI: 10.1021/jm00012a007.
18. Firth, N. C.; Brown, N.; Blagg, J. Plane of Best Fit: A Novel Method to Characterize the Three-Dimensionality of Molecules. *Journal of Chemical Information and Modeling* **2012**, *52* (10), 2516-2525. DOI: 10.1021/ci300293f.

19. *Jmol: an open-source Java viewer for chemical structures in 3D*. <http://www.jmol.org/> (accessed 07/01/2022).
20. Taylor, R. D.; MacCoss, M.; Lawson, A. D. G. Rings in Drugs. *Journal of Medicinal Chemistry* **2014**, 57 (14), 5845-5859. DOI: 10.1021/jm4017625.
21. Sayle, R. *1st-class SMARTS patterns*. Daylight, 1998. <https://www.daylight.com/meetings/summerschool98/course/basics/ref/sayle/> (accessed 07/01/2022).
22. Daylight Chemical Information Systems, I. *SMARTS Tutorial*. https://www.daylight.com/dayhtml_tutorials/languages/smarts/ (accessed 07/01/2022).
23. Daylight Chemical Information Systems, I. *SMARTS Examples*. https://www.daylight.com/dayhtml_tutorials/languages/smarts/smarts_examples.html (accessed 07/01/2022).

LIST OF SUPPORTING INFORMATION

List of Supporting Information

Items available to download from <http://datacat.liverpool.ac.uk/1674/>

Chapter 2: Predicting Ames Mutagenicity

Supporting Information 2.xlsx

- Contains **Supporting Information 2.1a – c, 2.4, 2.5, 2.7 and 2.8**

Supporting Information 2.zip

- └ **Supporting Information 2.2.sdf**
- └ **Supporting Information 2.3.sdf**
- └ **Supporting Information 2.6.docx**

Chapter 3: Understanding Polymer Detergent Properties via QSPR method

Supporting Information 3.xlsx

- Contains **Supporting Information 3.1 – 3.5**

Chapter 4: Novel Surfactant Descriptor – Potential Amphiphilicity

Supporting Information 4.xlsx

- Contains **Supporting information 4.1 – 4.12**

Chapter 5: Visualisation of Chemical Functionality for a Chemical Library

Supporting Information 5.xlsx

- Contains **Supporting Information 5.1**

Supporting Information 5.zip

- └ **Supporting Information 5.2.dwar**
- └ **Supporting Information 5.3.dwar**
- └ **Supporting Information 5.4.dwar**
- └ **Supporting Information 5.5.pse**
- └ **Supporting Information 5.6.dwar**
- └ **Supporting Information 5.7.pse**
- └ **Supporting Information 5.8.pse**
- └ **Supporting Information 5.9.pse**
- Note: **Supporting Information 5.2 – 5.4 and 5.6** requires DataWarrior, **Supporting Information 5.5, 5.7 – 5.9** requires PyMOL

LIST OF SUPPORTING INFORMATION

Pipeline Pilot Protocols

CFMProtocol

- └ Alignment_Structures.sdf
- └ Chemical functionality mapping using geometrical alignment.ppxml
- └ Chemical functionality mapping using substructure alignment.ppxml
- └ MoleculeVectorMacro.dwam
- └ OverallVectorMacro.dwam
- └ ReadMe.txt
- └ Rings_in_Drugs.sdf
- └ Template generation.ppxml
- └ UserManual(Thesis).docx
- Contains files related to the protocols described in *Chapter 5: Visualisation of Chemical Functionality for a Chemical Library*
- Note: Pipeline Pilot protocols (.ppxml files) may require the import of components within Self Written.zip to run successfully

Potential amphiphilicity calculator.ppxml

- Protocol described in *Chapter 4: Novel Surfactant Descriptor – Potential Amphiphilicity*
- Note: May require the import of components within Self Written.zip to run successfully

Self Written.zip

- Contains user written components required by the protocols described in *Chapter 4: Novel Surfactant Descriptor – Potential Amphiphilicity* and *Chapter 5: Visualisation of Chemical Functionality for a Chemical Library* to run successfully