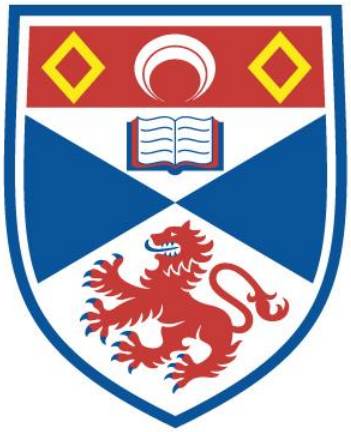
COMPUTING THE AQUEOUS SOLUBILITY OF ORGANIC DRUG-LIKE MOLECULES AND UNDERSTANDING HYDROPHOBICITY

James L. McDonagh

A Thesis Submitted for the Degree of PhD
at the
University of St Andrews



2015

Computing the Aqueous Solubility of Organic Drug-like Molecules and Understanding Hydrophobicity

James L. McDonagh

This thesis is submitted in partial fulfilment for the degree of Doctor of Philosophy

Sunday 4th January, 2015

School of Chemistry
University of St Andrews

Contents

Abstract

This thesis covers a range of methodologies to provide an account of the current state of the art and to develop new methods for solubility prediction. We focus on predictions of intrinsic aqueous solubility, as this is a measure commonly used in many important industries including the pharmaceutical and agrochemical industries. These industries require fast and accurate methods, two objectives which are rarely complementary. We apply machine learning in **Chapters 4** and **5** suggesting methodologies to meet these objectives. In **Chapter 4** we look to combine machine learning, cheminformatics and chemical theory. Whilst in **Chapter 5** we look to predict related properties to solubility and apply them to a previously derived empirical equation. We also look at *ab initio* (from first principles) methods of solubility prediction. This is shown in **Chapter 3**. In this chapter we present a proof of concept work that shows intrinsic aqueous solubility predictions, of sufficient accuracy to be used in industry, are now possible from theoretical chemistry using a small but diverse dataset. **Chapter 6** provides a summary of our most recent research. We have begun to investigate predictions of sublimation thermodynamics. We apply quantum chemical, lattice minimisation and machine learning techniques in this chapter.

Text which is **bold** represents internal hyperlinks; text which is ***bold and italic*** represents external hyperlinks. Italic text signifies important terminology.

For my Family and Friends

Declaration

1. Candidate's declarations:

I, James McDonagh, hereby certify that this thesis, which is approximately 45,000 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for a higher degree. I was admitted as a research student in September 2010; the higher study for which this is a record was carried out in the University of St Andrews between 2010 and 2014. I, James McDonagh received assistance in the writing of this thesis in respect of grammar and spelling, which was provided by Mr Leo Holroyd and Miss Rosanna Alderson.

Date signature of candidate

2. Supervisor's declaration:

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of Doctor of Philosophy in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date signature of supervisor(s)

3. Permission for publication: (to be signed by both candidate and supervisor)

In submitting this thesis to the University of St Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. I also understand that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that my thesis will be electronically accessible for personal or research use unless exempt by award of an embargo as requested below, and that the library has the right to migrate my thesis into new electronic forms as required to ensure continued access to the thesis. I have obtained any third-party copyright permissions that may be required in order to allow such access and migration, or have requested the appropriate embargo below. The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

No embargo on printed or electronic copy

Date signature of candidate
Date signature of supervisor(s)

Acknowledgements

I want to thank my supervisors Dr John Mitchell and Dr Tanja van Mourik who have provided support, guidance, encouragement and numerous useful discussions. Both have been excellent supervisors and I am grateful to have had such wonderful supervision. I would also like to thank SULSA for funding during my PhD; the EaStCHEM research computing facility for use of their computer clusters and the NSCCS for a kind computer time grant. Thanks also to Dr Herbert Früchtl and Dr Alexandra Simperler for computer support and many useful conversations. I am thankful to Professor Maxim Fedorov, Dr David Palmer and their group members for help with RISM and making me welcome on my visit to Leipzig. I am also grateful to Prof Graeme Day for supplying some additional pieces of software. I would also like to thank Dr Daniel Barker for allowing me become involved in his initiative education and out reach project 4273π. Thanks to Ms Kathleen Nicholson for maths support.

I am grateful to all of my friends and colleagues in rooms 208 and 150 in the Purdie building: Lukasz, Ragnar, Neetika, Alex, Nicolas, Luke, John, Leo, Rosie, Ava, Laz, Luna, Ludo, Rachael, Jose, Simon, Gregor, Jan and Hamse. The friendly atmosphere, great pub trips to the Whey Pat and Aikmans and many useful discussions have been fantastic. Special thanks to Leo and Rosie for proof reading sections of my thesis. I also thank the members of the University of St Andrews fencing club and real ale societies, in particular: Seb, Darren, Vicky, Kati, Anna, Hugh, Jia luen, Simone, Jess, Sophie and Ashley; John, Naomi, Marion, Jo, Martin, Luke, Ben, Sarah, Katie, Jonathan and Christian, you have all been a great support. I would like to thank those who helped me reach this point from Helsby high school and Bangor University. I would especially like to thank Mike, Sam, David (Fozz), Dave, Dan and Chris from Helsby high; and Andrew, Adam, Nick, Victoria, Ruth, Welsh Jon, John, Kyle, Dave, Steph, Dr Mike Beckett, Dr Lorrie Murphy, Dr Andrew Davies and Dr Greg Chass from Bangor University.

This would also not have been possible without the marvellous surgeons, doctors and nurses working under Mr Foubister from Ninewells hospital, who so fantastically reconstructed my right radius. Additionally, I thank Adam Legg for many hours of patient guitar tuition, the staff of the University of St Andrews sports centre and the staff of Avertical world climbing centre in Dundee for helping me get back to sport after breaking my arm. This provided some great physio-therapy.

Finally, I am truly indebted to my wonderful family and girlfriend for their continuous support. I wish to thank you all but particularly my parents Janet and Dave, brother Matt, girlfriend Rosie and grandparents Ron and Ivy, Michael and Dorothy, I could not have done this without your support.

Publications

- ***Melting Point Prediction by QSAR methods: Application to solubility prediction using the GSE*** McDonagh, J. L.; van Mourik, T; Mitchell, J. B. O. *Molecular Informatics*, Manuscript in preparation.
- ***University level practical exercises in bioinformatics benefit voluntary groups of pupils in the last two years of school*** Barker, D.; Alderson, R. G.; McDonagh, J. L.; Plaisier, H.; Comrie, M.; Duncan, L.; Muirhead, G.; Sweeney, S., *Open Review of Educational Research*, Manuscript in preparation.
- ***A Tutorial of Recent Approaches and Methods for Accurate Solubility Predictions.*** Skyner, R. E.; McDonagh, J. L.; Groom, C.R.; van Mourik, T.; Mitchell, J. B. O. *Physical Chemistry Chemical Physics*, Manuscript submitted.
- ***Elemental discoveries.*** McDonagh, J. L., *Science - books et al*; 345, (6194): 262; July 2014
- ***Uniting Cheminformatics and Chemical Theory to Predict the Intrinsic Aqueous Solubility of Crystalline Druglike Molecules.*** McDonagh, J. L.; Nath, N.; De Ferrari, L.; van Mourik, T.; Mitchell, J. B. O. *Journal of Chemical Information and Modelling*, 54, 844-856, 2014.
- ***First-Principles Calculation of the Intrinsic Aqueous Solubility of Crystalline Druglike Molecules.*** Palmer, D. S.; McDonagh, J. L.; van Mourik, T.; Mitchell, J. B. O.; Fedorov, M. V. *Journal of Chemical Theory and Computation*, 8, 3322-3337, 2012
- ***Enzyme Informatics.*** Alderson R. G.; De Ferrari L.; Mavridis L.; McDonagh J. L.; Mitchell J. B. O.; Nath N. *Current Topics in Medicinal Chemistry*, 12(17), 1911-1923 , 2012.

List of Figures

List of Tables

List of Equations

List of Abbreviations

σ	Standard deviation
R^2	Coefficient of determination
S_0	Intrinsic Solubility
A.U.	Atomic Units
ADMET	Absorption, Distribution, Metabolism, Excretion and Toxicity
BO	Born-Oppenheimer approximation
BZ	Brillouin Zone
CART	Classification and Regression Tree
CAS	Chemical Abstract Services
CCDC	Cambridge Crystallographic Data Centre
CDK	Chemistry Development Kit
CML	Chemical Mark-up Language
CSD	Cambridge Structural Database
CV	Cross Validation
DFT	Density Functional Theory
E-State	electrotopological state
FDA	USA Food and Drug Administration
GF	Gaussian fluctuation
GGA	generalised gradient approximation
GI	Gini index
GSE	General Solubility Equation
GTO	Gaussian-type orbital
HF	Hartree-Fock
HNC	HyperNetted Chain
HTVS	High Throughput Virtual Screening

IET	integral equation theory of liquids
InChI	IUPAC International Chemical Identifier
KS	Kohn and Sham
LDA	Local density approximation
logP	Logarithm to the base 10 octanol-water partition coefficient
LV	Latent Variables
MC	monte carlo
MD	molecular dynamics
MM	molecular mechanics
MOZ	Molecular Ornstein-Zernike equation
MP	Møller-Plesset
MP1100	Melting Point dataset containing 1100 molecules
MP2	Møller-Plesset second order perturbation theory
OZ	Ornstein-Zernike
PCF	Pair Correlation Functions
PCM	Polarisable continuum model
PLHNC	Partially Linearised HyperNetted Chain
PLS	Partial Least Squares or Projection to Latent Structures
QSAR	Quantitative Structure-Activity Relationships
QSPR	Quantitative Structure-Property Relationships
RDF	Radial Distribution Functions
RF	Random Forest
RISM	reference interaction site model
RMSD	Root Mean Square Deviation
S.I.	International system of units
SC	Solubility Challenge
SMILES	Simplified Molecular Input Line Entry System
SPC	Simple Point Charge model
STO	Slater-Type Orbital
SVM	Support Vector Machine
UC	Universal Correction

VDW	van der Waals
XML	eXtensible Mark-up Language

"There is a rhythm and a pattern between the phenomena of nature which is not apparent to the eye, but only to the eye of analysis; and it is these rhythms and patterns which we call Physical Laws."
Richard Feynman, 1964

Chapter 1

Introduction

"The important thing is not to stop questioning. Curiosity has its own reason for existing. One cannot help but be in awe when he contemplates the mysteries of eternity, of life, of the marvelous structure of reality. It is enough if one tries merely to comprehend a little of this mystery every day."

Albert Einstein, 1955

1.1 Motivation

If solubility prediction is searched on Google Scholar, over 200 search results for the last year (2013 - 2014) alone are returned.[1] Solubility prediction is important in numerous fields in and outside of chemistry such as environmental predictions,[2,3] biochemistry,[4] pharmacy,[5,6] drug-design,[7] agrochemical design,[8] and protein ligand binding.[9] Aqueous solubility is of fundamental interest owing to the vital biological and transportation functions played by water. In this chapter a discussion of solubility is presented, along with a general overview of the area of solvation prediction and modelling.

Besides the clear scientific interest in water solubility and solvent effects, accurate predictions of solubility can have implications of industrial importance. Such predictions save time and money in many chemical product development processes. In the pharmaceutical industry these predictions provide early stage viability screening of drug candidates.[10,11] *Quantitative Structure-Activity relationships* (*QSAR*), *Quantitative Structure-Property Relationships* (*QSPR*) and *data mining* have been successfully applied in this area for some time. These models provide efficient predictions of solubility and represent the current industry standard. However, such models can lack physical insight. As such, a purely theoretical method, capable of achieving similar levels of accuracy at an appropriate computational cost, would be a powerful research and development tool.

Our motivation here is to investigate QSAR/QSPR models and attempt to combine them with theoretical chemistry. We also wish to explore the possibility of a first principles prediction of solubility. As such in this thesis we investigate the

application of existing and novel prediction schemes to predict the *intrinsic aqueous solubility* of organic drug-like molecules. *Intrinsic aqueous solubility is defined as the solubility of an unionised species in a saturated solution* and is generally represented by the symbol S_0.[12] The intrinsic solubility is a particularly important quantity as it can be used to find the pH dependent profile and estimate the pKa. Throughout this thesis we therefore work with neutral molecules and assume such molecules are neutral at pH 7. From now on when solubility is quoted we will be referring to the intrinsic aqueous solubility unless otherwise stated. Solubility measurements are typically presented as the base 10 logarithm of solubility quoted in moles per litre ($Log_{10}S$ referred to units mol/L).

1.2 Solubility and its Implications

Solubility is a property with implications in many situations, from simply will compound x dissolve in solvent y, to our sensitivity to odour, which has been shown to be lower as hydrophobicity increases.[13]

There are two clear concepts which are subtly distinct from one another occurring during solvation: solubility and dissolution. A solution can be defined as a thermodynamically stable state in which an equilibrium exists between the solute and solvent. Solute molecules are transferred from the solute to the solvent and dispersed, eventually reaching a constant state, equilibrium. The concentration of a particular solute which can be dispersed within a solvent is known as the solubility, which is a property of the thermodynamic equilibrium. The rate of dissolving a solid is known as dissolution and is a kinetic property. These concepts are important for drug molecules as drug delivery is affected by dissolution whereas drug activity/availability is affected by solubility.[14] The two concepts are related to one another by the Noyes-Whitney equation.[15]

$$\frac{dW}{dt} = \frac{kA(C_s - C)}{L} \tag{1.1}$$

Equation 1.1: $\frac{dW}{dt}$ is the rate of dissolution, A is the solute surface area which is in contact with the solvent, C solute concentration in the bulk solvent at a given time, C_s is the solute concentration in the diffusion layer (given from the solubility of the molecule and under the assumption that the diffusion layer is saturated), k is the diffusion coefficient and L is the diffusion layer thickness.

Solubility is therefore influenced by interactions within the solute and interactions between the solution's constituents. As a result there are a large number of degrees of freedom within the system which impact solute solubility. A schematic representation of the solvation process is presented below in **Figure 1.1**.[14]

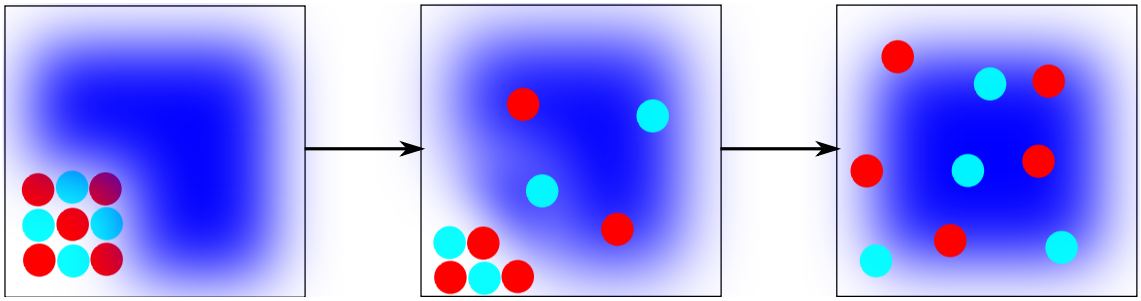


Figure 1.1: A pictorial representation of the solvation process.

In this section we present a brief overview of the factors affecting solubility and how solubility is experimentally determined. No experimental solubility determination has been attempted in this thesis. We will conclude this section with a brief description of where solubility determination affects the drug discovery pipeline. This is a pertinent industry for the application of the ideas presented through this thesis.

1.2.1 Factors Affecting Solubility

The vast majority of drug-like molecules are organic and hence here, we focus specifically on the factors affecting organic molecules.

1.2.1.1 Organic Crystals - Solid State Effects

As we stated above, solubility is affected by the strength of solute-solute interactions. Solutes that are only weakly bound together will tend to have a higher solubility, as the energy cost of breaking up the lattice is lower. Hence, amorphous structures generally have a higher solubility than crystalline materials. This is complicated by polymorphic effects. *Polymorphs* are alternative 3D crystalline arrangements of identical molecular units. Polymorphs can have substantially different physical properties including solubility. The classical example of this is that of Ritonavir,[16,17] in which a change in polymorph led to a drastic shift in solubility leaving the drug poorly available. This occurs in many drug-like molecules but rarely with the extreme results of Ritonavir.[18–20]

1.2.1.2 Temperature and Pressure Effects

Temperature affects solubility differently depending on the phase of the system being considered. Gases dissolve into solutions more readily at lower temperature; this is due to the second law of thermodynamics (*"In an isolated system, spontaneous processes occur in the direction of increasing entropy"*,[21]). Gases will disperse, hence become more disordered as temperature increases. For the same reason, the solubility of a solid will increase as the temperature is raised. It has been shown previously that changes in temperature of approximately $\pm 10°C$ from room temperature have little effect on solubility.[22] For gaseous solvation the partial

pressure of the gas is also a factor. The solution is in equilibrium with the surrounding gases as a result, should the composition of the surrounding gas change so does the equilibrium. This is perhaps best displayed as the effect that occurs when opening a can of carbonated soft drink.

1.2.1.3 Ionisation and Ion effects

Although in this thesis we will focus on intrinsic solubility it is important when considering experimental data to be aware of the effects of ionisation. pH can have a significant effect on solubility. As many drug molecules are either weak acids or weak bases (have ionisable basic/acidic functional groups) the pH at which a solubility measurement is made is important. Weak acids and base have the following dissociation paths:[14]

$$HA + H_2O \leftrightarrow A^- + H_3O^+ \quad \therefore \quad pK_a = -log_{10}\left(\frac{[A^-]\cdot[H_3O^+]}{[HA]}\right)$$

$$B + H_2O \leftrightarrow BH^+ + HO^- \quad \therefore \quad pK_b = -log_{10}\left(\frac{[BH^+]\cdot[HO^-]}{[B]}\right)$$

HA represents an acidic drug molecule whilst B represents a basic drug molecule. The pK_a is the acid dissociation constant (quantifies the degree of dissociation at a given pH) and the pK_b, is the basic dissociation constant. The Henderson–Hasselbalch equation (**Equation 1.2**) allows one to calculate the solubility of weak acids and bases in solutions of differing pH values as a function of pH and pk_a.[14,23]

$$\begin{aligned} logS_{Total}^{Acidic} &= logS_0 + log(1 + 10^{pH-pK_a}) \\ logS_{Total}^{Basic} &= logS_0 + log(1 + 10^{pK_a-pH}) \end{aligned} \tag{1.2}$$

Equation 1.2: $logS_{Total}$ is the total solubility, $logS_0$ is the intrinsic solubility and the final term on the right hand side is the solubility of the ionised form.[14]

The *total solubility* above is therefore given as *the sum of the base 10 logarithms of the concentration of both ionised and non-ionised forms at thermodynamic equilibrium.* As a result this value is heavily influenced by the potential ionisation of a molecule. The intrinsic solubility may still be mildly affected. Experimentally, buffers are often used to provide some control over ionisation when solubility determinations are made. Often mixtures of buffers are used to control over a large pH range. The use of such buffers can however affect the solubility measurements. As the solubility of electrolytes is generally increased with increasing ionisation, with the opposite being true of non-electrolytes, buffer agents can interfere with this process potentially increasing the ionic strength of the solution and thus reducing the solubility of non-electrolytes.[24]

1.2.1.4 Co-Solvents

An additional feature to be aware of when considering solubility values in the literature is that of co-solvents. In the pharmaceutical industry these are often used to allow the evaluation of a molecule's solubility quickly. By mixing co-solvents with water the water hydrogen bond network is broken up, hence aiding the solvation of an organic non-polar molecule.

1.3 Experimental Solubility Determination

There are many sources of experimentally determined solubility in the literature.[25–29] However, large errors have been repeatedly found in the reporting of solubility values which limits the use of these data in terms of modelling and model validation.[24,30–33] It is necessary to have experimental conditions and purity information about the precipitate reported along with the solubility measurement in order for one to confidently minimise the error in standard datasets. Error estimates in literature suggest a minimum average error of approximately 0.6 logS units but commonly errors of up to 1.5 logS units have been reported.[32]

As a result of this, when solubility datasets are derived often the most reliable information comes from multiple solubility assays. Therefore, these datasets can often contain high levels of random error or noise from the different techniques which have been employed. There are a number of such techniques which are commonly applied; some common methods are briefly discussed below.

1.3.1 Thermodynamic Solubility Determination

We initially introduce thermodynamic methods of solubility determination. These methods focus of making solubility measurements at thermodynamic equilibrium. As equilibrium needs to be achieved in these methods they are often slow but generally provide very reliable and reproducible results.

1.3.1.1 Shake Flask Method

The shake flask method is a commonly applied technique and generally accepted as an accurate way to measure intrinsic solubility, especially of molecules which are poorly soluble. This is a thermodynamic technique to measure solubility. A sample of the solute is added to a buffer solution until saturation. The flask is then shaken until thermodynamic equilibrium is achieved. The dissolution profile should be investigated beforehand so that an optimal time for thermodynamic equilibrium to be established can be located. The precipitate is then extracted and the solution concentration measured commonly by *High Pressure Liquid Chromatography* (*HPLC*). The method suffers from a lack of confirmation that thermodynamic equilibrium has been reached even if the dissolution profile has been investigated and thus instead often relies upon long standing times (24 - 72

hours). This also makes it unsuitable for *High ThroughPut Screening* (*HTPS*). Additionally, we must be aware of the buffers used and the protocols used to extract the precipitate (is the solution temperature maintained and is there a chance for cross contamination?).[14]

1.3.1.2 Synthetic Method

The synthetic method uses a laser beam to monitor when the solid form has completely dissolved into the solvent. Known quantities of solute are added to a solvent which is continuously stirred at a constant temperature. When all of the solute has dissolved the laser's detector has its maximum value. The process is repeated until the maximum value significantly drops, hence the solution is saturated. The solubility is then calculated based upon the amount added to the solution.[34] This method is useful particularly when viscous solvents need to be used.[14]

1.3.1.3 CheqSol: Chasing Equilibrium

The CheqSol[35] method is a chasing equilibrium method. It applies acid-base titration to "chase" the equilibrium position. The method proceeds to use a small aliquot of buffered solution before dissolving the solute. The experiment is carried out under an inert atmosphere to exclude atmospheric gases. The solution is titrated with acid or base until a precipitate is detected by light scattering. Once a precipitate is detected, measured amounts of acid and base are added taking the solution from supersaturated to undersaturated. At the point at which this happens the intrinsic solubility can be determined. A single CheqSol run makes eight determinations, the average of which is taken as the intrinsic solubility. This method hence checks that the system has reached thermodynamic equilibrium and makes several solubility determinations in a run time of a few hours.[24,35] It has been reported that CheqSol can reduce errors to as low as 0.05 logS units over multiple runs.[36]

1.3.2 Kinetic Solubility Determination

Kinetic solubility determination is generally employed in industry where fast assessments are needed at the molecular development stage. Kinetic methods do not rely on the system being at thermodynamic equilibrium and hence are much faster but generally less reliable.

1.3.2.1 Turbidmetric Measurement

In the pharmaceutical industry measurements are required quickly and so thermodynamic rigour is bypassed in favour of speed. Turbidmetric solubility determination involves the addition of a DMSO solution containing the solute to buffered water at pH 7. Buffers are carefully selected to avoid interference in the measurements; where such interference is likely to occur counterions are added

to minimise them. Once precipitation occurs, which is detected generally by UV spectroscopy, the solubility is calculated using the known concentration of the solute in DMSO. This method fails to establish a thermodynamic equilibrium therefore making results difficult to reproduce. Additionally, as DMSO is used solid state characterisation is prevented as well as having the effect of generally increasing solubility. The method is fast and can have useful input when quick estimates are required, but it is not useful when accurate reproducible values are required.[14]

1.3.3 Solubility in the Drug-Development Process

We consider here the biological processes that occur to allow a drug molecule to be absorbed and transported. We then move on to consider the processes of drug discovery and why solubility is important.

1.3.3.1 Absorption and Distribution

Absorption of an orally administered drug occurs throughout the gastro-intestinal tract, predominately in the intestines. The absorption process involves an orally administered drug molecule having to pass through gastric fluids, cell membranes and often blood. These are very different environments including low pH (gastric acid has a pH range of 1.5 - 3.5), lipophilic and hydrophilic environments. As many drugs are weak acids and bases these environmental shifts can have a major impact on the molecules. For example, if a weak acid of $pK_a = 4.4$ is administered orally, its predominant form in the stomach will be unionised, hence enabling absorption in stomach mucus. However, for a weak base of $pK_b = 4.5$ the predominant form would be ionised therefore reducing the chance of mucus absorption. Due to the larger surface area the majority of absorption occurs in the intestines.[23]

Molecules can be absorbed in several ways in the intestine: *passive transport, facilitated passive transport, active transport* and *pinocytosis*. Passive transport involves the diffusion of a molecule across a membrane due to diffusion across a concentration gradient (high concentration GI tract to low concentration blood for example). The size and ionisation of a molecule also affects this rate. Facilitated passive transport involves the molecule reversibly binding to a transport molecule in order to cross a membrane it would otherwise be unable to cross. Active transport is a highly selective process which can allow movement against the concentration gradient. This method is generally only applicable to molecules bearing a similarity to native chemical structures and has an energy cost to the process. Pinocytosis is the process by which cells acquire and ingest fluid. Substances adsorbed to the cell membrane are engulfed by the cell in a vesicle formed from the cell membrane. Pinocytosis is therefore a process allowing adsorbed liquid droplets to pass to the cells interior in larger quantities than by passive or active transport. This process again costs energy.[23]

The rate at which these processes occur is important in order to keep a molecule's concentration high enough to be therapeutically active but not toxic. Administrative timings and dosages are specifically designed to do this. It is rare to find a drug

molecule with a logS below -6. The majority (estimated at 85%)[5,23] falling in to the range of -5 to -1 logS. This is generally due to issues related to barrier crossing.[5,23]

1.3.3.2 Excretion and Toxicity

The body's main excretion process is via the kidneys: water soluble materials are screened out of the body. Key materials such as salts, glucose and B vitamins are usually actively or passively reabsorbed in the renal tubes. Drug molecules, and metabolites of drug molecules, tend to be ionised and cannot diffuse back into circulation. As a result these products are excreted in the urine. The rate of urinal tract diffusion excretion is affected by the pH of the urine. Secretions from the kidneys aid in the diffusion of cationic and anionic metabolites to the urine.[37] If a drug or its metabolite is not excreted then this can cause a toxic build up in the body. Additionally, if a metabolite or drug completes with other substances for excretion, the use of one drug can lead to an increase in the retention time of another drug. This also has potential to cause toxic accumulations to occur in the body.[37]

1.3.3.3 The Drug-Development Process

The drug development and discovery process is a long multi-step process. **Figure 1.2** below highlights the key steps:

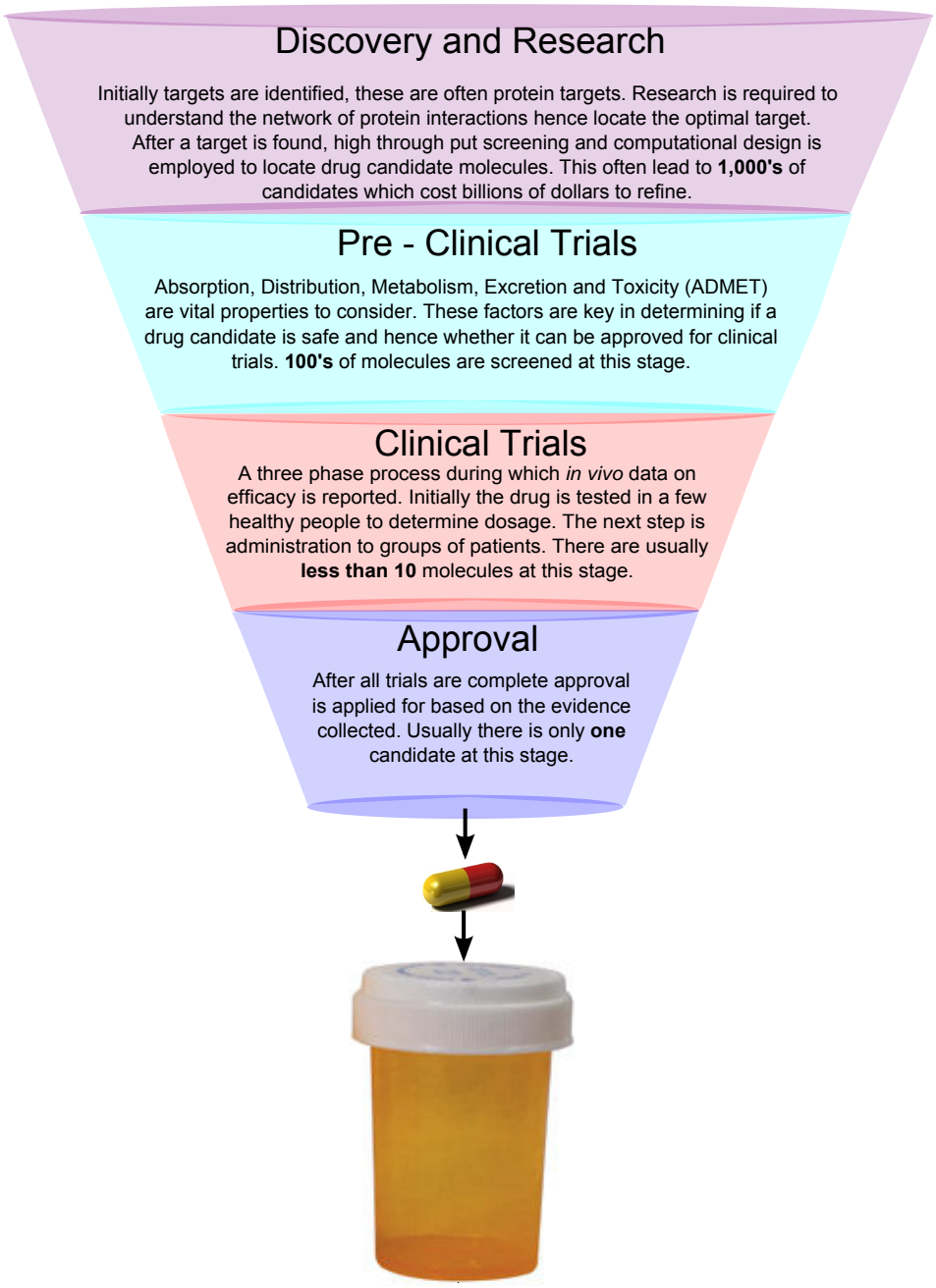


Figure 1.2: A summary diagram of the full drug discovery process.[38,39]

For orally administered drugs we are concerned with *pharmacokinetics - how a molecule is transported around the body* and *pharmacodynamics - a molecule's pharmaceutical action.* These two overarching terms are used to cover a multitude of factors. Solubility has an impact in both areas; in pharmacokinetics, where we are concerned with *Absorption, Distribution, Metabolism, Excretion and Toxicity* (*ADMET*) properties; solubility is a key parameter in determining the bio-availability of a drug molecule and its absorption, as discussed above. If the solubility is low then this means that the amount of the molecule that can be available in the gut and bloodstream will also be low, making formulation very difficult due to

potentially large variations in the results between patients.[40] In pharmacodynamics the solubility of a molecule can impact upon how easily a molecule binds to a target, hence how pharmaceutically active the molecule may be. For these reasons solubility is considered to be a fundamental physicochemical property. The USA's *Food and Drug Administration* (*FDA*) regulations require extra testing be carried out on any low solubility drug molecule.[41]

Whilst solubility is important all of the way through the drug discovery process, it is most extensively scrutinised at the early stages, the first two steps of **Figure 1.2**. These early stages encompass many steps requiring interdisciplinary groups and large capital expenditure. **Figure 1.3** below provides a general overview of the steps often found in the early stages of pharmaceutical development.

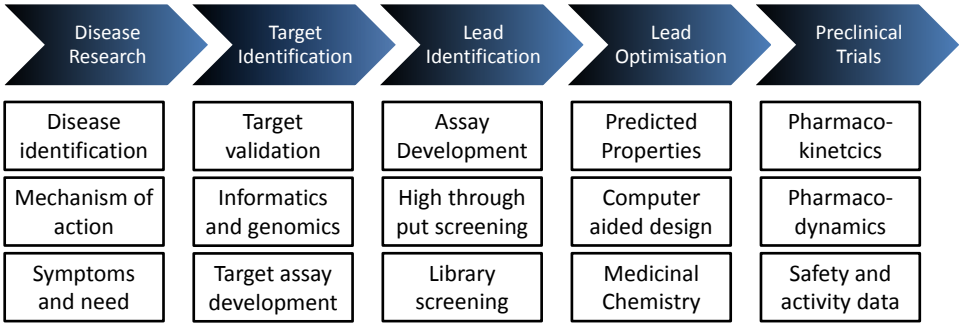


Figure 1.3: Common steps found in the early stages of drug discovery.

We can see in this diagram that computational methods are applied throughout: Bioinformatics is applied in the early stages of target identification, for proteomics and structural and functional genomics. Cheminformatics and bioinformatics are applied to library screening; this could be for processes such as matched molecular pairs screening for how single structural substitutions can affect a compound's activity. Computational chemistry and cheminformatics are then applied to molecular design and property prediction. These methods form part of a multicomponent optimisation to produce a single molecule which is capable of reaching and acting pharmaceutically on the target, therefore benefiting the patient. Solubility is one component of this multi parameter optimisation. Given the pharmaceutical industry's increasing expenditure and given the reducing number of candidates making it past clinical trials, it is industrially important to make optimal use of these techniques. Around 40%[42] of new drug molecule candidates are estimated to be effectively insoluble making them poor candidates. It is this high attrition rate which makes solubility prediction such an important area of research.[5,23,42]

1.4 Solubility Modelling

We have discussed above why solubility prediction is important and how it is measured experimentally. Methods to predict aqueous solubility were fairly recently tested in a blind challenge called the *Solubility Challenge*[28] which was run in 2008. This challenge provided accurate solubilities for 132 molecules and challenged

participants to predict a specific 32 of them blind, given the other 100. The dataset is made freely available on-line from the following reference.[43] The challenge had over 100 entries and concluded that no single method consistently produced the best solubility prediction. Correct predictions are considered to be within ±0.5 logS unit or 10% of the raw S value.[44] The percentage of correct predictions from participants ranged between molecules from 2% for Probenecid to 80.8% for Imipramine. In this section we introduce some of the ideas behind solubility prediction and modelling. Principally we introduce thermodynamic cycles for solubility modelling as well as the conventions which accompany them.

1.4.1 Thermodynamic Cycles for Solubility Prediction

The ideal situation would be to be able to model the direct transfer of a crystal structure to an aqueous solution. This would be replicating the physical, energetic and entropic changes which occur upon solvation. Practically this is not possible. This is due to the immense number of degrees of freedom open to the system. "Brute force" calculations may eventually be able to probe solubility directly but for the time being indirect calculations of solvation properties remain our best course. In this thesis there are two alternative thermodynamic cycles used. The first models solubility via the gaseous phase and the second via a hypothetical super cooled liquid state. The relations are best demonstrated diagrammatically below in **Figure 1.4**. In both methods, the cycles begin by breaking up the crystal before hydrating the individual molecules.

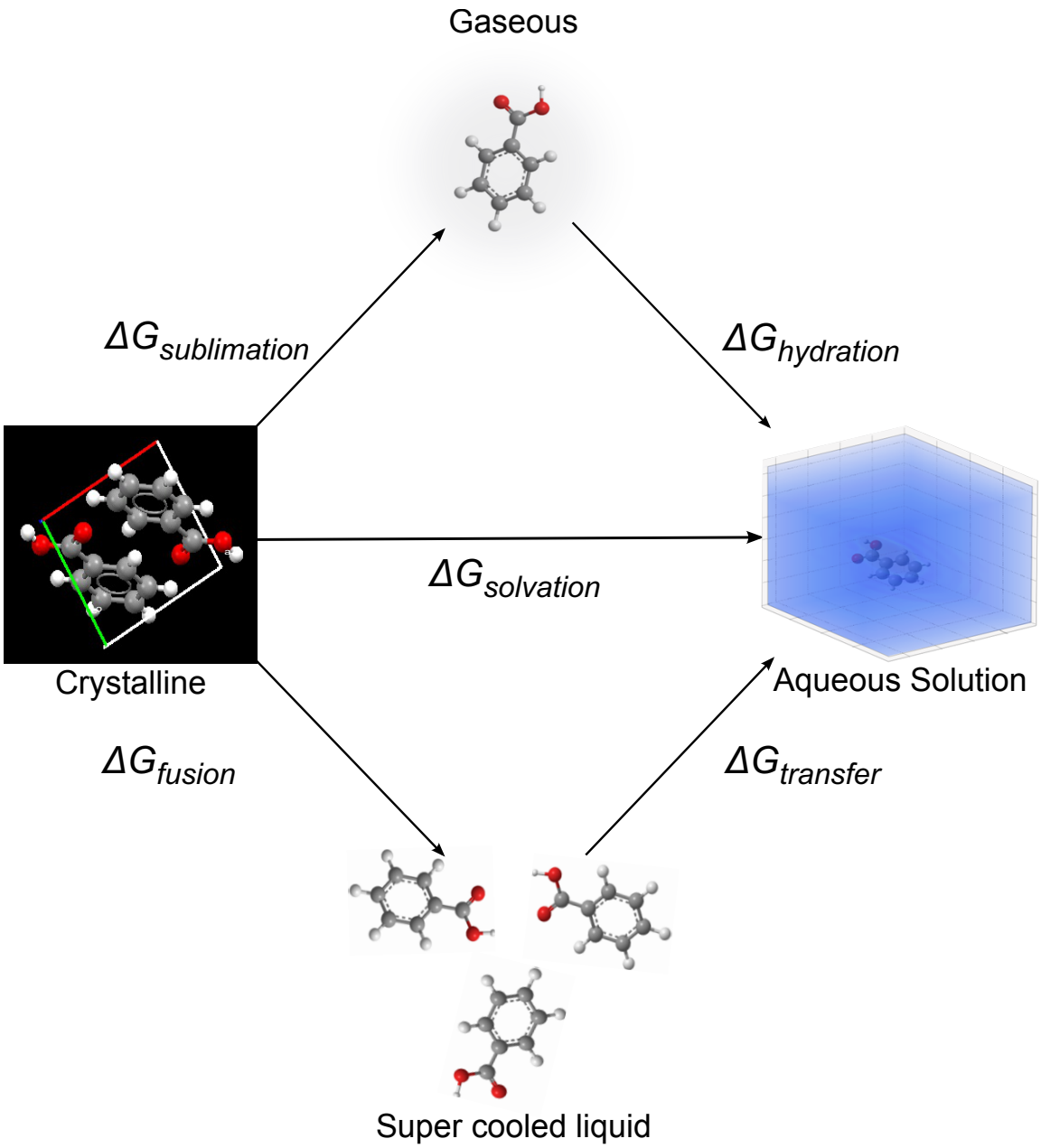


Figure 1.4: Thermodynamic cycles of solubility prediction

Both cycles have been applied successfully previously in different research groups.[6,45] In this work we focus on the former cycle which goes via the gaseous phase, see **Figure 1.4**. We apply the other cycle when using the *General Solubility Equation* (*GSE*)[46,47] in **Chapter 5**. The discussion here focuses on the use of the thermodynamic cycle via the gaseous state.

1.4.2 Standard States and Conventions

To avoid confusion, presented here is an overview of the standard state conventions used in this thesis. Sublimation energies are generally calculated in the 1 atmosphere (atm) standard state, as this is the state used by experimentalists. Solvation free energies on the other hand are commonly given in the standard state defined by Ben-

Naim of 1 mol/L with a fixed centre of mass.[48–50] In this thesis ΔG° corresponds to the 1 atm standard state, whereas ΔG^{*} corresponds to the Ben-Naim 1 mol/L standard state. The difference between the two standard states is a constant 1.89 kcal/mol (7.91 kJ/mol). All tabulated final solubility predictions presented in this thesis will be in the 1 mol/L standard state.

1.4.3 Thermodynamics and Solubility

A general ideal relationship exists in order to calculate intrinsic aqueous solubility. The thermodynamic relationship is below in **Equation 1.3** and provides us with a direct calculation of the Gibbs free energy of solution (ΔG_{sol}):

$$\Delta G^{*}_{solution} = \Delta G^{*}_{sublimation} + \Delta G^{*}_{hydration} = \ -RTln(S_0 V_m) \tag{1.3}$$

*Equation 1.3: Equation to calculate the Gibbs energy of solution ($\Delta G_{solution}$) in the 1 mol/L standard state (Ben-Naim terminology 1 mol/L with a fixed centre of mass)[50] represented by *. S_0 is the intrinsic solubility, V_m is the crystalline molar volume, R is the gas constant (8.314 $JK^{-1}mol^{-1}$) and T is the temperature in Kelvin (K).*

The Gibbs energy of sublimation (ΔG_{sub}) accounts for the breakup of the lattice; it corresponds to the energy required to take a single molecule from the crystal to the gas phase. The final chapter of this thesis discusses several methods to calculate ΔG_{sub} as well as the enthalpy and entropy of sublimation (ΔH_{sub} and ΔS_{sub} respectively). An introduction to these calculations is provided in **Section 3.1.2**.

The Gibbs energy of hydration (ΔG_{hyd}) provides the energy of transferring a single molecule from the gaseous state to the solution. This term has been extensively studied. Recent work involving the *Reference Interaction Site Model* (*RISM*) (**Section 2.2.12.4**) has focused on the accurate predictions of ΔG_{hyd}.[51,52] Historically ΔG_{hyd} has also been a property of interest dating back to 1803 with the generation of *Henry's law*. The law states a proportional relationship between gaseous partial pressure and the solution concentration of a specific component, at a constant temperature.

$$P = k_H C \tag{1.4}$$

Equation 1.4: Henry's Law. P is the partial pressure of the gas, k_H is the Henry's law constant in units of pressure over concentration, and C is the concentration in solution.[21]

Henry's law constants are experimentally determinable quantities which have been used to predict solubility as the inverse of the Henry's law constant multiplied by the partial pressure of the gas. Henry's law is somewhat idealistic breaking down away from equilibrium and is not applicable if the solute is not volatile at a temperature at which the solvent is a liquid, otherwise the solute will not enter a vapour phase hence no equilibrium can be established.[21]

In the second thermodynamic cycle the Gibbs energy of fusion (ΔG_{fus}) is a metric of the energy required to melt a substance at its melting point. This term is also accounting for the loss of the crystal. In the hypothetical supercooled liquid state the Gibbs energy of transfer (ΔG_{trans}) is the energy required to transfer one molecule from a solution of itself to an aqueous solution. These quantities are approximated later in the GSE using the melting point as a descriptor for ΔG_{fus} and the *logarithm to the base 10 of the partition coefficient between n-octanol and water (log P)* as a descriptor for the ΔG_{trans}.[53] This equation's principles are derived upon several assumptions:

- The solute's crystalline form is not affected by the presence of the solvent.
- Walden's rule ($\Delta S_{melting} = 56.6\ JK^{-1}mol^{-1}$ Empirical rule derivied from coal tar derivatives which can be considered as rigid organics.[54]) gives the entropy of melting.
- The change in the heat capacity of the solid and the liquid are negligible.
- Log P is equal to the ratio of the solubilities of the solute in octanol and water.
- Solute molecules are completely miscible in n-octanol.

These assumptions are generally valid for organic drug-like molecules.[47]

1.5 Computational Property Predictions

Molecular property prediction is an expansive field incorporating complex methodologies working from the top down, using bulk descriptors for predicting material properties, or bottom up, using a quantum mechanical calculations in order to predict bulk material properties from first principles.

Within this thesis we have applied these concepts to predict condensed phase properties, either those of crystalline drug-like organic solids or aqueous solutions of organic drug-like molecules. We have offered a first principles prediction of solubility as a proof of concept with a small dataset and combined theoretical chemistry and cheminformatics providing physically sound descriptors for empirical models. Here we describe the current state of the art and where our work fits into an overview of the area.

Solid state modelling and crystal structure prediction are important topics within computational chemistry, incorporating a wide variety of methodologies from chemistry and physics. These methodologies have been well described in the literature, and seen widespread development more recently.[55–59] The progress in this area is perhaps best displayed in the computationally predicted crystal structures from the CCDC's blind tests.[60–63] We can broadly separate solid state modelling methods into one of two theoretical approaches: Firstly, the introduction of *ab initio* quantum chemical modelling, directly applying quantum mechanics to solid state modelling. Secondly, we can apply classical force fields or fitted models parametrised to experimental or quantum chemical data. Both methods have seen advancement and widespread use.[58,64–67] More recently, we have seen new advancements in dispersion corrections for periodic DFT. These advances have

provided evidence that a first principles method is able to correctly predict the energetic ordering of some polymorphs.[59] In our work we have begun to expand into these fields, hence, much of the future work emanating out of this project will be in this direction. We have explored empirical melting point predictions and a variety of methods to predict ΔH_{sub}, ΔS_{sub} and ΔG_{sub}. Our work provides the groundwork for future exploration of the sublimation process by a variety of methods.

Solvent modelling and solubility prediction have also seen methodological improvements in recent years. Large systems can now be modelled with some level of quantum mechanical detail, using QM/MM or linear scaling DFT. These are useful abilities for pharmaceutical development in protein modelling. Recent advances have seen thermodynamic mapping of entropic and enthalpic terms within binding pockets,[68] which is a useful tool when analysing the binding action and method for drug molecules. We have additionally seen improved models for the solvent which offer more physical descriptions of the solvent at reasonable costs.[51,52,69] These models have been tested and evaluated in the work presented in this thesis on a small dataset. We present a proof of concept work of a computationally effective first principles prediction of solubility.[31] We go further combining some of these methods with informatics methods.[33]

Chapter 2

Theory and Methods

"Most people wouldn't know a wave function if they tripped over one, but almost everyone has heard of the uncertainty principle."

Chad Orzel, 2010

2.1 Cheminformatics and Machine Learning

QSAR and QSPR are models developed from *a priori* information. The models are essentially correlations between structural aspects of the molecule and physical properties. The primary assumption of such models is that structurally similar molecules have similar properties. As a result, we can train a model to define a relationship on a given training dataset and use it to predict properties of unseen molecules which contain similar structural features. These methods are not generally applicable, i.e. a reasonable prediction cannot be expected if the test molecules differ notably from the training molecules. Machine learning methods can be used to identify and build models which correlate these structural properties with the physical data. The machine learning methods available vary greatly in methodology and relative complexity. Some are discussed below. Typically a molecule is represented in a machine readable format, such as the *Simplified Molecular Input Line Entry System (SMILES)* [70–72] or the *IUPAC International Chemical Identifier (InChI)*.[73] These are input into a machine learning method which learns a model that correlates the structural features with the physical properties. The model is then used to make predictions of molecular properties based on a set of chemical structure descriptors.[74]

2.1.1 Molecular Input

The input of molecular structures as machine readable formats presents its own challenges. Numerous formats and representations have been developed for different applications from web based searching to chemical identification. The initial challenge is that ideally the format's input syntax should be canonical, i.e. have

a single unique string to determine each molecule. Depending on the purpose and complexity of the format, additional challenges arise such as, how to represent 3D structure in strings, present stereochemistry and determine equivalent chemical structures.[74–76]

SMILES are a commonly used format because they are human readable, (**Appendix A**) whilst also being machine readable. The problem is that the original SMILES are not a canonical method. SMILES is a hydrogen suppressed notation that has no strict definition of the point at which one should begin to read or write the structure from. This can lead to problems when calculating molecular descriptors, as often one SMILES string can be read with a different protonation state in different programs. Extensions to the SMILES scheme exist and are now commonly used. These extensions make SMILES a canonical format, and are generally based on the *Morgan algorithm*.[77] The Morgan algorithm defines the starting point as the point possessing the highest connectivity value. The connectivity value is initially the number of connected atoms but is iteratively increased by summing the connectivity values of bonded partners until the maximum number of different connectivity values is reached.

InChI are promoted as a worldwide non-proprietary standard chemical identifier.[76] InChI are designed to be canonical and have been integrated into a number of larger databases for use as search tools with the InChIkey system.[76] Additionally, they are a compact method of storing structural information and transferring it electronically. InChI strings are technically human readable and writeable, although more difficult to understand than SMILES (**Appendix A**). Each is a string beginning with "InChI=" and a version number, following from this is a string separated into multiple layers by "/". The string InChI always holds the molecular formula and a connectivity layer, then additional flags adding for example stereochemistry information. InChI are a canonical representation however, there is a known problem in version 1 of the InChI method that if a tautomer is represented in different ways, i.e. as a resonance structure or as a fixed charged structure, different InChI are sometimes produced. The developers hope to rectify this in version 2.[76]

A range of other formats exist, although many are now becoming obsolete. Other formats still in common use are *Chemical Abstract Services* (*CAS*),[78,79] the *Chemical Mark-up Language* (*CML*)[80] and *MOLfiles*. The CAS provides a unique database of millions of publicly available chemical structure entries. CAS classifies each with a unique number which itself contains no chemical information.[79] CML is designed for the transfer of information over the internet. It is built on the *eXtensible Mark-up Language* (*XML*) as an application and can deal with reaction mechanisms and structures.[80,81] MOLfiles are simple files containing just a header and a connection table. MOLfiles are now most widely used as a section of input for a larger file such as *SDfiles* which are files made up of a MOLfile and contain physical data related to the molecule.[75,76]

2.1.2 Descriptors

Descriptors come in many different forms and are used to represent physical features of chemical structures. They are generally single numerical values which hold some

information about a physical property of a specific molecule.[74] Descriptors can be simple properties such as the molecular weight or the number of a specific atom type or they can be a prediction with corresponding experimental values, such as the *octanol-water partition coefficient* (P). They can also be derived from classical or quantum chemistry. Clearly the cost to calculate different descriptors can vary dramatically. For example, the time needed to calculate the relative energy of a set of molecules using quantum chemistry will be substantially greater than calculating the number of H atoms in each molecule. It is generally true that descriptors offering higher levels of refinement incur a higher computational cost.[74] There are thousands of different types of molecular descriptors and numerous pieces of software to calculate them. Here, we discuss a few common descriptors from 2D and 3D molecular inputs. Several were used in this work, calculated using the *Chemistry Development Kit* (*CDK*)[82] from SMILES representations of molecules. The CDK is an open source cheminformatics Java library. Tools[83,84] have been developed to interface the CDK with common programs such as Microsoft Excel[83,85] and R[84,86].

The most basic descriptors are those concerned with straightforward counting. These can be the count of a particular atom type, features like H bond acceptors and donors or particular structural features (rotatable bond, aromatic rings, aliphatic C's etc). Other common descriptors are molecular weight and sub-structure weight, generally calculated using sub-structure searches. Rarely will simple counts be sufficient to discriminate molecular properties; so models will usually combine these descriptors with more complex ones.

Popular more complex descriptors include topological indices, fingerprints and predicted physicochemical properties. Topological indices describe branching, shape or size of a molecule via a single value. Early versions from Wiener[87] (**Equation 2.1**) involved calculating the distances between atom pairs and summing the number of bonds between them. This gives a descriptor for molecular branching:

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j>i}^{n} D_{ij} \tag{2.1}$$

Equation 2.1: D_{ij} is the separating distance between i and j.

Also in common use today, are the *chi molecular connectivity indices*. These are descriptors for molecular branching and size. The descriptors come from initial work by Randic[88] which was later generalised by Kier and Hall.[89] The descriptors account for valence electronic state and number of hydrogens bonded to an atom (Randic's original formulation was H suppressed). Additionally, this measure is no longer pairwise, as in Wiener's approach, instead a continuous sum runs over a selected path length, measured in number of bonds.

Following from this work Kier and Hall[90] also generated the *kappa shape indices*. Shape indices compare the number of paths to alternative so called 'extreme structure', where all atoms are bonded to one another or form a linear chain. Shape indices do not contain any information about the atom type. A ratio is then calculated giving a measure of how close to one of these extremes the actual molecules shape is. One other important topographical index is the *electrotopological index*, from the work of Hall, Mohney and Kier.[91] Such descriptors provide

information on the electronic and topological character of atoms in a molecule, defined as the *electrotopological state* (*E-State*) of an atom.

Combined together the above descriptors provide a molecular descriptor encoding limited information on electronic structure and shape of a molecule. Other descriptors have also been created to provide atomic level information, such as *BCUT*, which provides information on polarizability and atomic charges.[92]

Molecular fingerprints were originally designed to speed up sub-structure searching but have since found use as molecular descriptors. A binary fingerprint encodes specific structural information. As the sub-structures often relate to the biological activity of a molecule, it is likely that these descriptors provide information confirming the presence of sub-structures of interest in the main structure.[74] Extensions to 3D fingerprints have occurred in recent years.[93] Alternatives or additions to molecular fingerprints are atom pair[94] or topological torsions. Atom pair descriptors are defined for all pairs of atoms in the molecule and encode the shortest path between them. They hold information on the element type, number of non-hydrogen atoms bonded directly to both atoms and the number of bonding π electrons. As a result they provide finer granularity than molecular fingerprints. Topological torsions follow similar conventions to atom pairs but over four centres.[74]

A final, commonly used, set of descriptors are those of predicted physicochemical properties. Predicted properties provide inherent information to any model about the likely activities of the overall molecule. They take the form of additive models assigning a group or atom a particular value and summing them together. One, common property that is used as a descriptor is logP. Experimentally, logP the base 10 logarithm of the ratio of a molecule's solubility in water and octanol. It is an important descriptor in defining drug lead compounds, offering information on the bioavailability of a potential drug molecule in aqueous and organic solvents.[95] There are a number of definitions of predicted logP. Rekker[96,97] defined the following equation (**Equation 2.2**) which has formed the basis of many group additivity models:

$$logP = log\left(\sum_{i=1}^{n} a_i f_i + \sum_{j=1}^{m} b_j F_j\right) \tag{2.2}$$

Equation 2.2: Rekker's logP equation, a_i labels a fragment and f_i is the contribution to the sum of the fragment, b_j labels the number of incidences of the correction factor F_j s application.

ClogP[95,98,99] is an example of a group additivity model. The model fragments a molecule on the basis of establishing isolated carbons i.e. those possessing only single bonds to heteroatoms. The model considers these atoms hydrophobic and those groups containing heteroatoms to be polar fragments. The values assigned are based on a relatively small library of fragments for which logP has been experimentally measured. A number of atomic additivity methods have also been described in the literature,[74,100] two examples being AlogP[101–103] and XlogP.[104] These methods have

a similar form to the group method described above, with the notable difference of sums running over atoms rather than fragments.

2.1.3 Machine Learning Models

Having produced a range of descriptors we now require that these can be correlated with known results. Machine learning/data-mining can be very good tools for correlating such multi-dimensional data to an activity. A few of the algorithms that machine learning models can use are discussed here, although it should be noted that many others are available.

2.1.3.1 Random Forest

Random Forest (*RF*), is an ensemble learning method that generates a forest of decision trees. The method follows a general workflow of:

- Selecting at random a sample of the training molecules with replacement.
- Growing a tree to its maximum extent on the basis of the best split achievable from a random subset M_{try} of the given descriptors (generally taken to be $[No.of descriptors]^{\frac{1}{2}}$ in classification and 1/3 no. of descriptors for regression).[105]
- Repeat the above two steps until a sufficiently large number of trees are generated.

This leads to a forest of a number of trees (M_{tree}).[105] Each tree has the structure of a main initial input/*root node (parent node)* and two sub nodes (*child nodes*). The splitting then continues giving the first two child nodes two child nodes each until no further splitting can be made, at which point the child node is known as a terminating node (*leaf node*) (**Figure 2.1**). This method of continually separating the data is known as recursive partitioning.[74,106] There are a number of ways to define a criterion at which to split the data. One of the most common is that of the *classification and regression tree (CART) algorithm* known as the *Gini index*(*GI*). The GI estimates the impurity in a child node if a split were to occur from the parent node on the basis of a specific variable. The GI holds its maximum value when the split would place equal amounts of data in both child nodes, hence, offering no real differentiability in the data. The GI holds its minimum value when splitting places all data in one of the child nodes. This means that the descriptor is highly discriminatory and offers a favourable splitting. This process eventually leads to those input data which share similar predictor values meeting in the same leaf node. This method is applied to classification. In regression the root mean square error is minimised. The predicted value is assigned as an average prediction of all training data occupying the same leaf node.

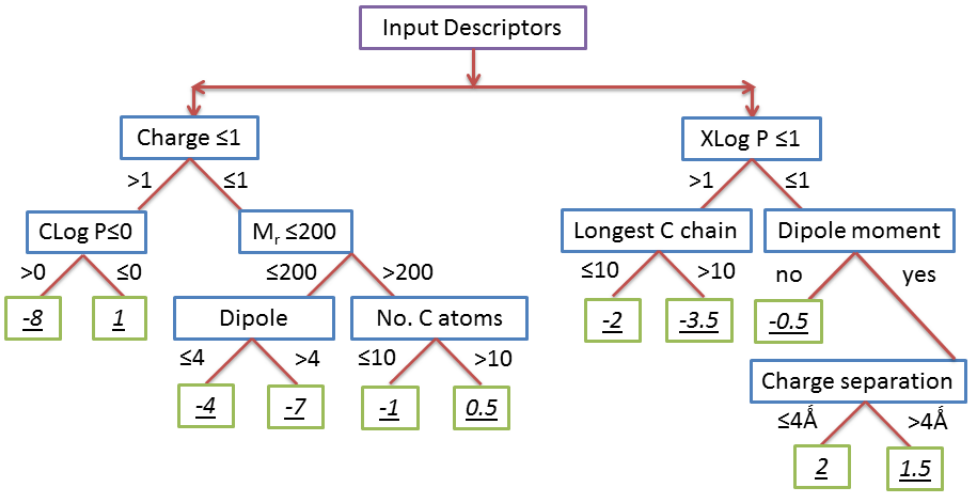


Figure 2.1: Random Forest: Illustrative example. The purple box is the root node. Blue boxes are parent and child nodes. Green boxes are leaf nodes.

2.1.3.2 Support Vector Machines

Support Vector Machines (*SVM*) is another algorithm and it allows classification and regression by separation and projection to a higher-space. SVM locates a surface which separates the data in the optimum manner, by mapping the data to a feature space in which it is separable.[74] The mapping to a higher order space allows the data to be used to produce a predictive function utilising the support vectors, which are the points closest to the separating surface. An illustrative example for classification is shown in **Figure 2.2**.

Figure 2.2: Support Vector Machine: an illustrative example of an SVM. Non-linear data plotted on the left; while a SVM hyperplane defining a separating surface is shown on the right in green.

SVM can also be used for regression. A parameter ϵ is defined as a margin of acceptable error. The predictive function aims to predict the y responses to the x input variables within this ϵ margin, whilst making the function as flat as possible, therefore avoiding over fitting. Ideally all of the points will lie within the ϵ boundary.[33,107] This implies that a function exists which approximates the

regression with the accuracy of ϵ. This is not always the case and hence slack variables are introduced (ζ) to provide some flexibility to the model, see **Figure 2.3** below.[108] These slack variables control the "hardness" of the boundaries: If $\zeta \leq 1$ it means the data point is inside the margin and on the correct side of the hyperplane, however, when $\zeta > 1$ then the point was erroneously predicted and a penalty is incurred for a bad prediction. The optimisation becomes a case of minimising the error penalty incurred by having $\zeta > 1$ but also having large enough margins to maximally accommodate the data.

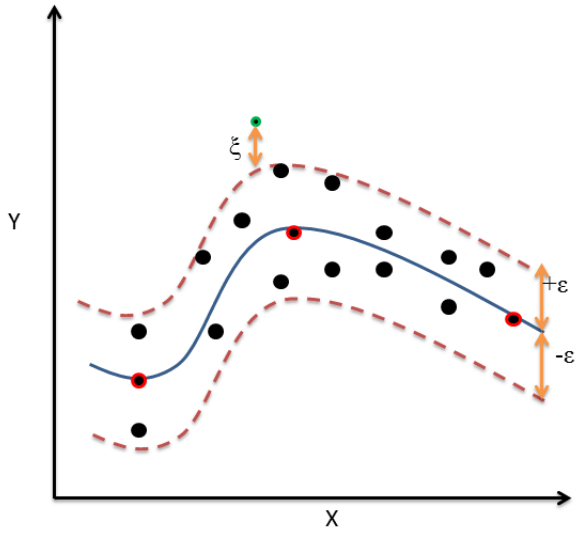


Figure 2.3: Support vector machine regression: An illustrative example showing ϵ margins and ζ error term. Dots with red outlines are the support vectors.

2.1.3.3 Partial Least Squares (Projection to Latent Structures)

Partial least squares (*PLS*),[109,110] can be considered as a classification algorithm which works by deflation to *latent variables* (*LV*). This method is also sometimes known as *projections to latent structure* and was originally applied in the social sciences, but has found widespread use in cheminformatics. The method attempts to explain the co-variation in the independent variable (x) and dependent variable (y). To achieve this, the dependent variable equation is cast in latent variables (**Equation 2.3** and **Equation 2.4**).

$$y' = a_1 t_1 + a_2 t_2 \ldots a_n t_n \tag{2.3}$$

Equation 2.3: Partial least squares: dependent variable as a sum of the products of LV (t_1) and their coefficients (a_1).[74]

Latent variables are linear combinations of the independent variables x_i and a suitable weight b_{ij} (**Equation 2.4**).

$$t_i = \sum_i b_{i1}x_1 + b_{i2}x_2 \ldots b_{iN}x_N \tag{2.4}$$

Equation 2.4: Latent Variables: linear combination of independent variables.[74]

The first LV (t_1) is a linear combination of independent variables which jointly provides a good explanation of the variance in the independent variable and, when scaled by its coefficient (a_1), gives a suitable approximation to the dependent variable. The independent variables within the first LV are then removed and additional iterations of the above process generate more LV. As a result there are always less LVs than there are independent variables helping to avoid over fitting.[74] The overall process is cast into a matrix formulation.

Independent variables are stored in the matrix X and dependent variables are stored in the matrix Y. The next step is to generate two sets of weights (w and c) which maximise the covariance of X and Y when a linear combination of the columns of X and Y are taken.

$$\overrightarrow{t} = Xw \qquad\qquad \overrightarrow{u} = Yc \tag{2.5}$$

Equation 2.5: Vectors maximising the covariance of X and Y. Constraints are placed on the procedure; orthogonality, $t^Tt = 1; u^Tu = 1$ and $; t^Tu$ is maximal.[111]

When the constraints in the caption of **Equation 2.5** are satisfied, the first LV is located. The descriptors comprising the LV's are subtracted from X and Y; the model of $\overrightarrow{t}$ vs $\overrightarrow{u}$ should now give a good approximation. A diagrammatic interpretation of this process is shown in **Figure 2.4**. The process is repeated until X becomes a null matrix.[111]

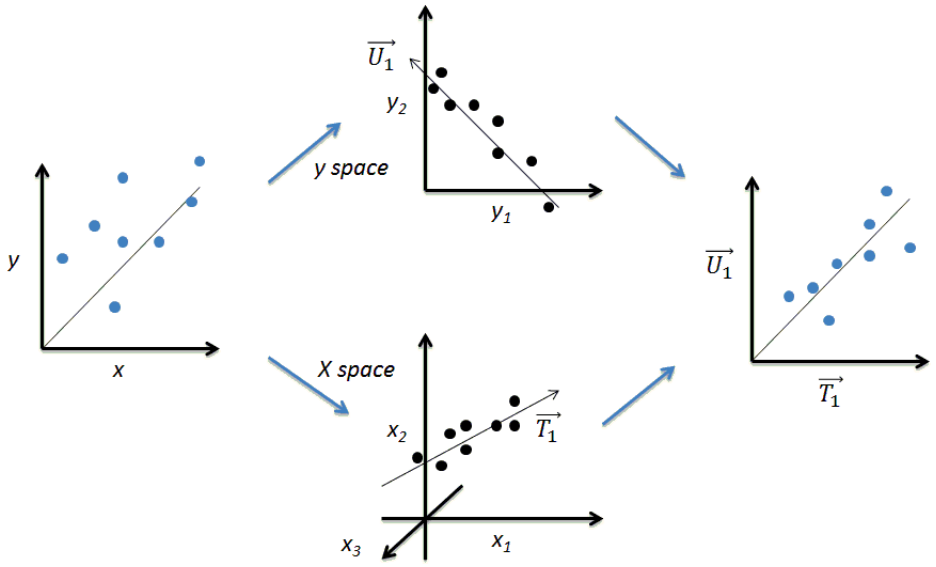


Figure 2.4: Partial Least Squares: An illustrative example, shows the original x and y model converted to a model made of LVs.[74]

2.2 Computational Theoretical Chemistry

Computational theoretical chemistry is the science concerned with the application of *in silico* models, whether statistical, classical or quantum, to chemical systems of interest. These methods contrast with those of informatics as they are physics based. The results of these calculations are highly interpretable, within the constraints of the physics they are based upon, and can aid in understanding physical processes. The calculations discussed here, however, are notably more computationally expensive than their informatics counterparts, as the real physical equations (or approximations thereof) are solved by numerical computational methods. In this section we outline the methods used in this thesis which range from quantum mechanical, which are discussed first, to those which are classical or statistical in nature.

2.2.1 Quantum Chemistry

Due to the size and mass of an electron we are unable to make electronic structure calculations utilising classical mechanics. On such a scale quantum mechanics must be employed. In this section we introduce theory and a discussion of methods which aim deal with problems of this nature.

The *time dependent, non-relativistic Schrödinger equation* (**Equation 2.6** and **Equation 2.7**) is one of the foundations of quantum mechanics. It describes the time evolution of a quantum system. A wavefunction (Ψ) is defined which contains all information it is possible to know about a system; this is a fundamental postulate of quantum mechanics. An operator is constructed, in this case the *Hamiltonian operator* ($\hat{H}$) **(Equation 2.7)** which acts upon the wave function to predict the observable property of interest, here the energy of the system.

$$\hat{H}(r,t)\Psi(r,t) = i\hbar\frac{\partial}{\partial t}\Psi(r,t) \tag{2.6}$$

Equation 2.6: The time dependent Schrödinger equation. $i = \sqrt{-1}$, H=Hamiltonian operator, Ψ=wavefunction, $\hbar = \frac{h}{2\pi}$, $\partial/\partial t$ is the partial derivative with respect to time.

$$\hat{H}(r,t) = \hat{T}(r) + \hat{V}(r,t) \tag{2.7}$$

Equation 2.7: Hamiltonian operator H is the sum of T(r), kinetic energy dependent on position vector r and V(r,t), potential energy dependent on the position vector r and time t.

For calculations where a time independent potential energy is defined (*V(r,t) = V(r)*), it is possible to use the *time independent, non-relativistic Schrödinger equation*, given in operator notation in **Equation 2.8**. This notation cleverly disguises the complexity of the problem at hand. The equation takes the form of an *eigenvalue* problem (**Appendix F**), where Ψ is the eigenfunction and E is the eigenvalue.

$$\hat{H}\Psi = E\Psi \tag{2.8}$$

Equation 2.8: The time independent Schrödinger equation. H is the Hamiltonian operator, Ψ is the Eigenfunction (wave function) and E is the Eigenvalue

2.2.2 The Hamiltonian

The *Hamiltonian operator* ($\hat{H}$) is the differential quantum mechanical operator to extract energies from a system appropriately described by a given wave function. The operator takes the following form defined in international standard units (S.I.) (**Appendix B**) :

$$\hat{H} = -\sum_i \frac{\hbar^2}{2M_e}\nabla_i^2 - \sum_k \frac{\hbar^2}{2M_k}\nabla_k^2 - \sum_i \sum_k \frac{e^2 Z_k}{r_{ik}} + \sum_{i<k} \frac{e^2}{r_{ij}} + \sum_{k<l} \frac{e^2 Z_k Z_l}{r_{kl}} \tag{2.9}$$

Equation 2.9: i and j subscripts are for the electrons; k and l are for nuclei. M_e and M_k are electron and nuclear mass respectively. e and Z are electron and atomic number respectively. $\hbar$ is the reduced Planck's constant $\hbar = \left(\frac{h}{2\pi}\right)$. ∇^2 is the Laplacian operator (second derivative operator). Finally r_{zx} is the separation of two particles

The first two terms in **Equation 2.9** characterise the kinetic energy of the electrons and nuclei respectively, the final three terms represent the potential energy of the electron nuclear interaction, electron electron interactions and nuclear nuclear interactions respectively. Currently any solutions to the equation would be functions of both the nuclear and electronic coordinates. Given that protons and neutrons are approximately 1800 times heavier compared to the mass of an electron,[112] it is routine to invoke the *Born-Oppenheimer* (*BO*) *approximation*. The approximation suggests that in the time frame of electronic relaxation nuclear motion can be considered as static.[112,113] This decoupling results in the nuclear kinetic energy term of the Hamiltonian being neglected and the nuclear repulsion becomes a constant. This defines a new Hamiltonian of the form of **Equation 2.10**.

$$\hat{H}_{electronic} = -\sum_i \frac{\hbar^2}{2M_e}\nabla_i^2 - \sum_i \sum_k \frac{e^2 Z_k}{r_{ik}} + \sum_{i<k} \frac{e^2}{r_{ij}} \tag{2.10}$$

Equation 2.10: The electronic Hamiltonian

In turn, this defines the electronic Schrödinger equation (**Equation 2.11**), for which solutions now use the nuclear coordinates as parameters and the electronic coordinates are variables.[113,114]

$$(\hat{H}_{electronic} + V_{NN})\psi_{electronic} = E\psi_{electronic} \tag{2.11}$$

Equation 2.11: Electronic Schrödinger equation under the Born-Oppenheimer approximation. V_{NN} is the constant nuclear nuclear repulsion

From now on we will refer to only the electronic terms unless otherwise stated, hence drop the sub-script electronic.

A wavefunction is required upon which the Hamiltonian will act, this function must suitably describe the motion and whereabouts of the particles, in this case electrons. The *wavefunction* (Ψ) is a function with no formal physical interpretation. It is down to the operator to query the wave function and determine an observable property. A recognised interpretation of quantum mechanics (*Born Interpretation*) is that the modulus of the square of the wavefunction ($|\Psi^2|$) is the probability density. In the next section we discuss how suitable approximations of the wave function are generated.

2.2.3 Basis Set Approximation

Before moving into the technical details of any given method it is important to consider the basis set approximation. This approximation allows for the use of a restricted number of functions to be considered when an approximate wave function is being constructed. Following from this section, several methods which utilise this approximation are discussed. These methods aim to approximately solve the Schrödinger equation or a variant of it. The basis set approximation is born out of practicality as calculation time scales polynomially with the number of basis functions. The result is that sets of selected and optimised basis functions have been produced to provide an incomplete, but still fairly accurate, approximate wavefunction. A potentially infinite number would be required to produce an exact wave function. Broadly these basis sets are created and optimised on three counts:[112,113]

1. Bearing computational efficiency in mind the number of basis functions should be small enough to be efficient.
2. Larger basis sets should where possible be generated from functions which can most easily be calculated by the computer.
3. Function must be of chemical relevance. The functions should map to the probability density of the electrons, so have a greater amplitude where the electron's probability density is highest.

2.2.3.1 Atom Centred Basis Sets

There have been several suggested functional forms for basis sets to take. The first is that of the *Slater-Type Orbital* (*STO*). These functions were initially chosen for the fact that they bear great similarity to that of the hydrogenic atomic orbitals. However, these functions lack an analytical expression for many of the integrals that are required, **(Section 2.2.4.1)** such as the two electron integrals **(Equation 2.18)**. As a result Boys proposed the use of Gaussian functions as atomic orbitals which allowed for an analytical solution to these integrals. These orbitals are known as *Gaussian-type orbitals* (*GTO*). The difference here is in the radial decay, in STO's this is e^{-r} where as in GTO's this is e^{-r^2} The usual form of these GTO's is as follows:[112]

$$\psi(x,y,z;\alpha,i,j,k) = \left(\frac{2\alpha}{\pi}\right)^{\frac{3}{4}} \left[\frac{(8\alpha)^{i+j+k} i!j!k!}{(2i)!(2j)!(2k)!}\right]^{\frac{1}{2}} x^i y^j z^k e^{-\alpha(x^2+y^2+z^2)} \tag{2.12}$$

Equation 2.12: A general form of some GTO in Cartesian coordinates. α is a parameter that moderates the width of the function. i, j and k are positive integers which operate to alter the form of the function in Cartesian coordinates, i.e. s-type p-type orbital etc.[112]

When the indices i, j and k all equal 0 the form of the function is spherically symmetric, this is called an s-type function. If i, j or k equal one a single node along one axis appears, this is known as a p-type orbital. The chemical relevance of this is clear in that these functions naturally display a likeness to hydrogenic orbitals by varying these coefficients. Unfortunately GTO's also come with the disadvantage that they display the incorrect radial form (they are exponential in r^2). Additionally, GTO's have a steeper gradient hence dropping off quicker than the more natural choice of STO's. To prevent and minimise this undesirable effect but keep the very desirable computational efficiency linear combinations of GTO's are taken to best approximate an STO. This can be seen in **Figure 2.5**.[112]

$$\psi(x,y,z;\{\alpha\},i,j,k) = \sum_{a=1}^{N^{gaussians}} c_a \phi(x,y,z;\alpha_a,i,j,k) \tag{2.13}$$

Equation 2.13: Contracted Gaussian formulation. $N^{gaussians}$ is the number of Gaussian functions and c is a parameter optimising the shape and maintaining the normalisation of the basis function.[112]

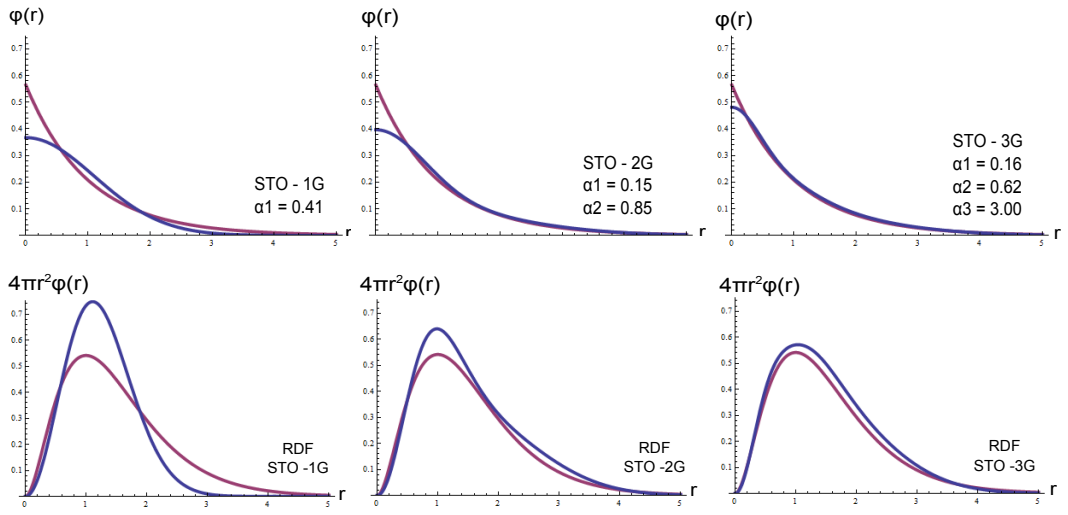


Figure 2.5: Contracted GTO's approximating an STO for the 1s H orbital. The red line is the STO and the blue is the GTO.[115] *α Values taken from reference.*[116]

The STO-3G basis set is a so called single - ζ basis set. This implies that each orbital is described by a single function which is a linear combination of GTO's,

three in this case. This means there is a single function defined to represent 1s, 2s, 2p ...and so on orbitals. We phrase this as minimal basis set; it contains the minimum number of functions required, clearly not even close to the infinite basis. In order to recover some flexibility in the basis we look to multiple - ζ basis sets. A double - ζ basis set would have two functions per orbital, i.e. two blue lines, one for each of the functions and both fully optimisable in **Figure 2.5**. Thus this allows extra flexibility to better optimise each orbital in different environments. Example of these kinds of basis sets are cc-pVnZ, where n = D double, T triple etc, which are from the work of Dunning and co-workers.[117] The acronym means correlation-consistent polarised core and valence n-ζ.[112]

Given that the core orbitals play little role chemically and largely remain similar in different chemical environments relative to the valence orbitals, it was realised that more flexible valence orbitals presented a greater gain than more flexible core orbitals. This lead to the development of split-valence orbital basis sets. Some of the most widely used basis sets of this form are those of Pople *et al.*[118] These basis sets are represented by an acronym of the form 6-31G. The values here refer to the number of primitive Gaussians which go into producing that function. The first number is the function representing the core orbitals, made up of in this case 6 primitive Gaussians. The numbers after the hyphen gives the number of functions to describe the valence, in this case it is a double - ζ valence; the first function is made up of 3 contracted Gaussians and the second is a single Gaussian function.[112]

Additional functions can be added to these basis sets in order to account for the molecular environment more effectively. More flexibility is added to the basis set by adding basis functions from higher orbitals. These are known as polarisation functions. A schematic representation is provided below **(Figure 2.6)**. These functions allow for much greater flexibility in the wave function definition. Whilst atoms can be well represented with simply the basis functions presented above, the molecular system has a dependency on multiple atomic positions and hence requires greater degrees of freedom.

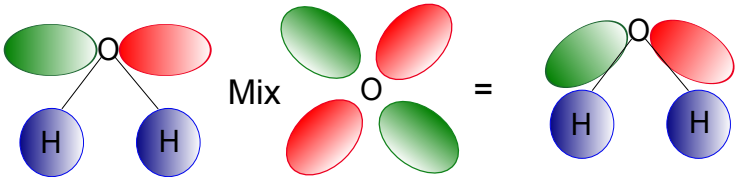


Figure 2.6: Polarisation basis functions from mixing with an orbital of a higher angular momentum

2.2.3.2 Plane Wave Basis sets

An alternative choice of basis functions is that of *plane waves*, regularly used in periodic calculations. These basis sets are a sum of sine and cosine functions. As periodic systems such as crystals or metals are vastly extended systems compared to single molecules, the use of such sine and cosine functions becomes a clear choice, owing to their infinite range. The molecular orbitals in these systems tend to group together becoming bands. These bands can be represented by sets of orbitals expanded in the basis of plane wave functions:[113]

$$\psi(x) = Acos(Kx) + Bsin(Kx) \equiv \psi(x) = Ae^{iKx} + Be^{-iKx}$$
$$\chi_K(r) = e^{iK\cdot r} \tag{2.14}$$

Equation 2.14: Top: Plane wave function represented as the sum of sine and cosine functions. Bottom: Molecular orbital expanded in the basis of plane wave functions[113]

K in the above equations is the wave vector which relates directly to the energy ($E = \frac{1}{2}K^2$). In this sense K determines the energy of the wave in terms of the frequency of oscillation. K also relates to the unit cell via the translation vector (t) defined as $K \cdot t = 2\pi m$ where m is a positive integer value. The size of the basis set therefore, is defined by K. Generally speaking, plane wave basis sets contain many more functions that STO and GTO basis sets. Plane waves can be applied to single molecule systems in a suitably large cell to avoid interaction with its periodic neighbour. However, due to the K's relationship with the unit cell translation vector this means a large number of plane waves must be used, hence, GTO's in this case are a more efficient choice. Plane waves excellently model delocalised and slowly varying electron densities. However, in the core of an atom electron density is firmly localised and occasional deep oscillations of valence orbitals means that K would have to be extremely high to represent the core. It is therefore common to use *pseudo-potentials* to describe the core region and hence reduce the maximum value of K and the basis set size.

2.2.3.3 Pseudo-Potentials

Pseudo-potentials (PP), allow for the core orbitals to be neglected from explicit calculation and instead to be represented by a single function which leaves the valence orbitals unaffected. The generation of a PP initially requires calculating the all-electron wavefunction of the atom. Replace the valence orbitals with a node-less set of pseudo orbitals. Then replace the core orbitals with a fitted set of analytical functions of the nuclear electron separation. The function is fitted such that it does not change the valence orbitals and matches to the wavefunction at that point (r_c). Finally additional parameters are fit such that when a calculation is run the pseudo-orbitals are equivalent to those in the all-electron calculation.[113]

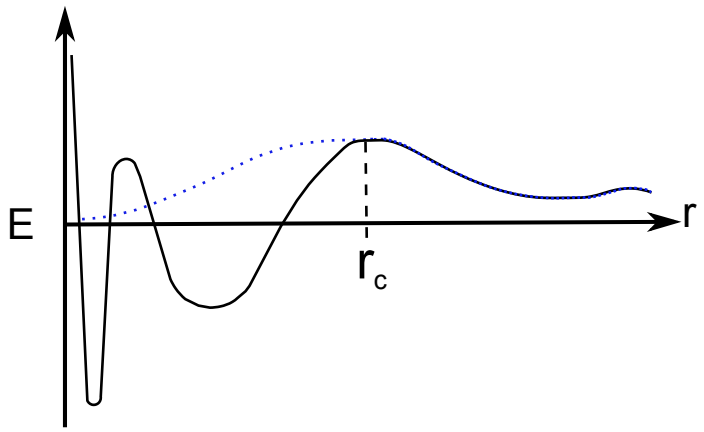


Figure 2.7: Pseudo-molecular orbital schematic, the all electron function is shown in black and the pseudo-molecular orbital is shown in blue.[113]

There are several different forms of PP. The norm-conserving PP requires that in addition to the matching form at r_c, the integral of the square of the PP and the original wave are the same. These PPs tend to be rather restrictive covering only a small core region hence still requiring relatively large values of K. Ultrasoft PP have also been proposed which relax the additional constraint. These allow for a much smaller value of K to be used.[113]

2.2.4 Wavefunction Methods

There are a couple of approaches to reaching approximate solutions to the Schrödinger equation. *Ab initio* methods aim to solve the equation without external parametrisation. Semi-empirical methods aim to solve the equations after simplifying them by the introduction of some empirical data (calculated or experimental). All of these methods use computationally intensive algorithms and methods in order to solve the equations.

2.2.4.1 Hartree-Fock

The founding member of all of these methods is the Hartree-Fock method (HF), which is itself an *ab initio* method. If we review the electronic Hamiltonian **(Equation 2.10)**, we note in the third term $\frac{e^2}{r_{ij}}$. This term implies a correlation between the electrons, in that the interaction between electron i and electron j depends on their joint positions. The correlated nature of electrons makes such a many body problem intractable; as a result the HF method opts for a *mean field approximation* (**Figure 2.8**). This approximation treats each electron individually; each electron experiences an interaction with an averaged field representing the electrons, including itself.[112,113,119] Utilising this method each electron can be described by a one electron wavefunction otherwise known as an orbital. Within the HF theory, we can then construct a molecular wave function as a product of the one electron functions. This is known as the *Hartree product*:[112,113,120]

$$\Psi(r_1, r_2, r_3 \dots r_n) = \phi_1(r_1)\phi_2(r_2)\phi_3(r_3) \dots \phi_n(r_n) \tag{2.15}$$

Equation 2.15: The Hartree product.[112]

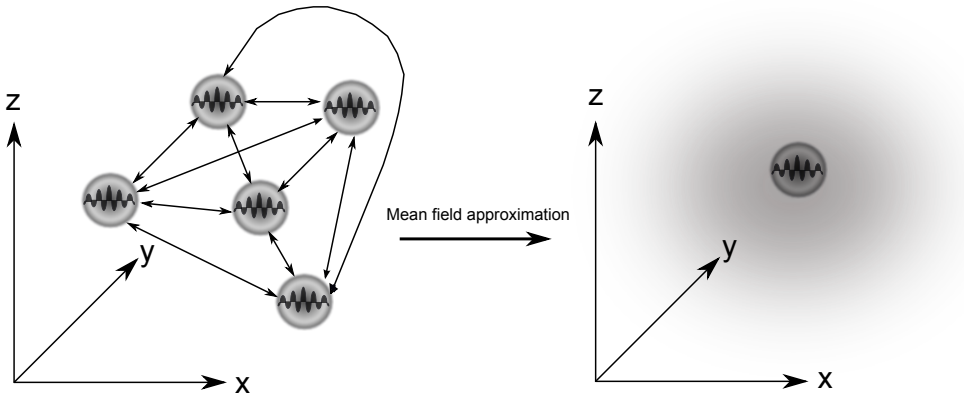


Figure 2.8: A diagrammatic representation of the mean field approximation and electron correlation.

However, **Equation 2.15** has a major problem. One of the fundamental principles of quantum mechanics, *the Pauli exclusion principle*, set a requirement for all fermion wave functions, of which electrons are one species of fermion, to be antisymmetric with respect to interchange. Unfortunately the Hartree product does not satisfy the antisymmetry requirement for the space and spin coordinates of the electrons. Electrons have spin of $\pm$ half defined as α spin and β spin. If spin is included in the orbital definition then the orbitals are defined as spin orbitals (χ_i). The molecular wave function must be anti-symmetric with respect to interchange of an electron's space and spin coordinates. Additionally, the Hartree product also requires that the electrons are distinguishable; this is not allowed in quantum mechanics as electrons are by definition indistinguishable particles.

These issues can be solved mathematically by representing the wave function in a *Slater determinant* **(Equation 2.16)** (**Appendix C**). The HF method uses a single Slater determinant to represent the molecular wave function. The columns of a single Slater determinant represent the atomic orbitals and the rows represent the electron coordinates. As a result each electron is at some point placed in each orbital, hence making them indistinguishable. The pre-factor is a normalisation.[112,113,119,121]

$$\Psi(1, 2, 3 \dots n) = \frac{1}{\sqrt{n!}} \begin{vmatrix} \chi_1(1) & \chi_2(1) & \chi_3(1) & \dots & \chi_n(1) \\ \chi_1(2) & \chi_2(2) & \chi_3(2) & \dots & \chi_n(2) \\ \chi_1(3) & \chi_2(3) & \chi_3(3) & \dots & \chi_n(3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \chi_1(n) & \chi_2(n) & \chi_3(n) & \dots & \chi_n(n) \end{vmatrix} \tag{2.16}$$

Equation 2.16: n electron Slater determinate.[112]

Having represented the wavefunction compactly, in terms of a Slater determinant, we must now move to solve for the orbitals that minimise the energy. The solution method utilises the *variational principle* (a prescription to reach the lowest energy available based on a wavefunction constructed of the given functions **Appendix E**) . This principle allows us to vary the parameters used to generate the orbitals in order to minimise the electronic energy ($E_{electronic}$) **(Equation 2.15)**.

$$E_{electronic} = \int \Psi^* \hat{H}_{electronic} \Psi dr$$
$$E_{electronic} = \langle \Psi | \hat{H}_{electronic} | \Psi \rangle \tag{2.17}$$

Equation 2.17: Top: Electronic energy evaluation. Bottom: Electronic energy evaluation in Dirac bra-ket notation

The bottom **Equation 2.17** uses the convenient Dirac bra-ket notation which implies the integral explicitly stated in the top of **Equation 2.17**. The Hartree-Fock energy equation can then be restated in terms of integrals over 1 or 2 electrons:

$$E_{HF} = \sum_i \langle \psi_i | h | \psi_i \rangle + \frac{1}{2} \sum_{ij} |ii|jj| - |ij|ji| \tag{2.18}$$

Equation 2.18: HF energy evaluation.[112]

The left hand term of **Equation 2.18** gives the one electron integrals. This term accounts for each electrons kinetic energy (motion of the electron) and potential energy (attraction from nuclear charges). The right hand term of **Equation 2.18** contains the two electron integrals and takes into account two distinct physical contributions to the energy. The first, represented by $|ii|jj|$, is the Coulombic interaction of the electron in the orbital χ_i with the mean field of all other electrons. It can be defined as the *Coulombic operator* J_{ij}. The second, $|ij|ji|$, represents the purely quantum mechanical exchange energy shown by the exchange of spin orbital subscripts i and j. In operator notation the *exchange operator* is K_{ij}. This now allows us to minimise the energy with respect to the orbitals. The *Hartree-Fock equations* **(Equation 2.19)** are now definable and are once again recognisable as eigenvalue problems:[119,121]

$$f(x_1)\chi_i(x_1) = \epsilon_i \chi_i(x_1) \tag{2.19}$$

Equation 2.19: The Hartree-Fock equations: f here stands for the Fock operator ***(Equation 2.20)*** *and ϵ_i is the expectation value for the ith spin orbital.*[112]

$$f(x_1) = h(x_1) + \sum_j [J_j(x_1) - K_j(x_1)]$$
$$f(x_1) = h(x_1) + V_{HF} \tag{2.20}$$

Equation 2.20: The Fock operator; V_{HF} is the Hartree-Fock potential defined by the Coulombic and exchange operators.[112]

As the HF equations contain a dependency on the orbitals in the Fock operator, they hence require their own result to solve the equations, so must be solved iteratively. This iterative solution process is known as the *self consistent field*. Following from the previous section **(Section 2.2.3)** we need to cast the molecular orbitals making up the wavefunction in terms of basis functions (ϕ).[113]

$$\chi_i = \sum_{\alpha}^{f_{basis}} c_{\alpha i}\phi_\alpha \tag{2.21}$$

Equation 2.21: Molecular orbital represented in a basis set.[113]

Now applying **Equation 2.21** to **Equation 2.20** we can restate the HF equations as follows:

$$F_i \sum_{\alpha}^{f_{basis}} c_{\alpha i}\chi_\alpha = \epsilon_i \sum_{\alpha}^{f_{basis}} c_{\alpha i}\chi_\alpha \tag{2.22}$$

Equation 2.22: HF equation using a atomic orbital basis.[113]

A matrix equation can be defined from this for closed shell systems. This matrix equation is known as the Roothan-Hall equations, where the HF equations have been set in an atomic orbital basis and grouped together in matrix notation:[113]

$$FC = SC\epsilon$$
$$F_{\alpha\beta} = \langle\chi_\alpha|F|\chi_\beta\rangle \tag{2.23}$$
$$S_{\alpha\beta} = \langle\chi_\alpha|\chi_\beta\rangle$$

Equation 2.23: Roothan-Hall equations.[113] *S is the overlap matrix between basis functions. F is the Fock matrix containing the Fock operator results. C is a coefficient matrix.*

This defines the HF equations in an atomic orbital basis for a closed shell system. The Fock matrix elements contain the integrals of the one electron operator and the sum over the two electron integrals.

2.2.4.2 Post-Hartree-Fock

A number of different methodologies have been proposed as advances on HF theory. As discussed previously, HF is limited in accuracy as it cannot account for the electron correlation.

$$E_{exact} = E_{HF} + E_{correlation} \tag{2.24}$$

Equation 2.24: The Hartree-Fock limit

From here, advances on HF are generally attempts to capture at least some aspect of this correlation energy. One such group of methods is based on perturbation theory. These methods generically offer a perturbation to the Hamiltonian, expressing the eigenfunctions and eigenvalues as a Taylor expansion in terms of the perturbation.[112]

$$O = O^{(0)} + \lambda V$$

Equation 2.24: Operator simplification in perturbation theory where $O^{(0)}$ is the simplified operator, V is is the perturbative operator, λ takes values between 0 and 1 mapping $O^{(0)}$ to the original operator O and $o^{(0)}$ is the zeroth order eigenvalue.[112]

Here we will focus on a particular form of perturbation theory, that of Møller and Plesset (MP),[122] in which the methods are labelled as MPx, where x is equal to the expansion order (MP2 = Møller-Plesset second order perturbation theory). Møller - Plesset perturbation theory takes as its starting point the Schrödinger equation and then applies the prescription of perturbation theory, as described above. In this case the generic operator shown in **Equation 2.24** is replaced with the Hamiltonian $O^{(0)} \rightarrow H^{(0)}$. They proposed suitable forms for the $H^{(0)}$ and V operators, in which $H^{(0)}$ is a sum of the single electron Fock operators. V is then the difference between the full Hamiltonian operator $\hat{H}$ and $H^{(0)}$. By this prescription it is possible to define perturbative methods truncated at any arbitrary order. In practice, it appears that the second order truncation is the first to offer improvement over HF. This correction to the energy is the leading term in the electron correlation. It is given as a sum over doubly excited determinants which are generated by promoting two electrons, which reside in occupied molecular orbitals in HF, to virtual orbitals. The sums are limited to avoid double counting **Equation 2.25**.[112,113,123]

$$E_{MP2} = \sum_{i}^{occupied} \sum_{j>i}^{occupied} \sum_{a}^{virtual} \sum_{b>a}^{virtual} \frac{[\langle \chi_i \chi_j | \chi_a \chi_b \rangle - \langle \chi_i \chi_a | \chi_j \chi_b \rangle]^2}{\epsilon_i + \epsilon_j - \epsilon_a - \epsilon_b} \tag{2.25}$$

Equation 2.25: MP2 energy correction. χ represents the molecular orbitals. Occupied orbitals are labelled by subscripts i and j virtual orbitals are labelled with subscripts a and b. ϵ represents the eigenvalues of the orbital.[112,113,123]

Other post-HF methods have been developed generally taking account of additional Slater determinants. All of these methods in essence provide a systematic way to approach the solution of the Schrödinger equation. These generally involve the use

of a number of Slater determinants. This is helpful as we know how to approach a solution to the Schrödinger equation, however, for all practical purposes these methods can quickly become too expensive to use on real systems of interest.

2.2.5 Density Functional Theory

Density functional theory (DFT) is now widely regarded as the work-horse of quantum chemistry. Unlike wavefunction theory, DFT aims to use the physically observable quantity of the electron density to define a system. To achieve this, functionals are employed; these are functions whose arguments are also functions, also known as functions of functions.

The history of DFT can be traced back as far as 1927[112] and the work of Thomas[124] and Fermi,[125] who derived a kinetic energy functional from a *uniform electron gas* (Thomas-Fermi model). Following this, Slater[126], Dirac[127] and Bloch[128] all generated similar expressions for the exchange energy (Tomas-Fermi-Dirac model). These expressions were combined with classically derived potential energy expressions and QM corrections, taking the electron density as their argument. This was the beginning of DFT. These initial models, however, suffered from serious flaws from a chemical stand point; chemical bonds were not predicted. This error is associated with the kinetic and exchange energy functionals due to the uniform electron gas assumption. Largely these models were used as empirical methods, finding application in periodic systems, before its modern formulation by Hohenberg, Kohn and Sham.[129,130]

Hohenberg and Kohn[129] provided two theorems. The first is the existence theorem, which states: The external potential is uniquely defined by the total electron density **(Appendix G)**.[129,131] This powerful statement provides a direct mapping of the electron density to the energy of a system. Moreover, if the electron density uniquely defines the external potential, it thus also determines the Hamiltonian and hence wavefunction. A perceptive view of why the electron density uniquely defines the system is attributed to E. B. Wilson:[113,132]

1. Integrating the total electron density defines the total number of electrons
2. The peaks in the density, known as cusps, are centred on the positions of the nuclei
3. Cusp height provides information on the nuclear charge of the nucleus responsible for it.

The second theorem, provides a variational theory, similar to that of the wave function methods **(Equation 2.26)**. This provides a means to optimise the system for the best trial/candidate density. **(Appendix G)**.

$$E_0(\rho) \leq E_0(\rho') \tag{2.26}$$

Equation 2.26: DFT variational theorem. ρ is the true ground state density and ρ' is the approximate density

These two theorems provide the rigorous mathematical base upon which DFT is built. DFT's main advantage over wavefunction methods is its implied simplicity. Wave function methods utilise the complicated wavefunction which in principle contains 4N variables (spin and position) per electron. DFT, being based on the complete electron density, in principle depends on only 3 variables (position) independent of system size. This is due to DFT being free of molecular orbitals, as opposed to wave function methods. The benefits are immediately obvious in terms of calculation time. However, although these theorems prove the existence of a functional capable of assessing a systems energy exclusively from the electron density, the functional remains unknown. Without a prescription to reach such a functional, models have been developed to approximate it, hence linking the electron density and energy.

DFT functionals, within computational chemistry, generally arise from the work of Kohn and Sham (KS). [130] In this ground breaking piece of work, Kohn and Sham elaborate on a *self-consistent field* method for DFT, analogous to that in HF theory. The method recalls the use of orbitals and posits the system as a fictitious, non-interacting set of orbitals (i.e. a system in which the electrons are charge neutral fermions lacking a Coulombic interaction) within an effective potential. This allows the kinetic energy to be separated into two terms: Firstly the non-interacting kinetic energy, which contains the vast majority of the kinetic energy, and secondly a small correction factor for the quantum nature of electrons. The introduction of orbitals is at the price of independence of system size. By including orbitals DFT increases in complexity to 3N variables for each electron. This was the critical step in Kohn and Sham's method. Assuming that electrons do not interact with one another it is possible to recover the vast majority of the energy of the system exactly.[112,113,133]

We can imagine, from the existence theorem, that provided the external potential remains consistent with the real system then density will also remain consistent with the real system. Therefore, we can calculate the exact solution for a non-interacting system by enforcing the potential has some dependence on the degree of interaction. If there is no interaction, and the system is non-degenerate, the solution is a single Slater determinant of the molecular orbitals (ϕ). Within this fictitious non-interacting system the kinetic energy can be exactly calculated as the first term in **(Equation 2.27)**.[132]

$$E_{non\ int\ DFT}[\rho] = T_{non-int}[\rho] + V_{Ne}[\rho] + J_{ee}[\rho] \tag{2.27}$$

Equation 2.27: Kohn Sham DFT energy functional for a non-interacting system of electrons. $T_{non-int}$ is the kinetic energy of the non-interacting system. V_{Ne} is the nuclear electron potential energy (external potential). J_{ee} is the electron electron classical Coulomb interaction.[112,113,132]

Equation 2.27 contains terms, respectively, referring to the kinetic energy of the non-interacting system ($T_{non\ int}$), the nuclear electron interaction or external potential (V_{Ne}) and the classical electron electron repulsion (J_{ee}). From the classical electron electron repulsion arises an un-physical self-interaction similar to HF in the mean field approximation. In HF this exactly cancels with the exchange energy.

To account for the interacting system, we need to add corrections to include the quantum nature of the system. This is made up of a correction to the kinetic energy and the potential energy of electron electron interactions. This is all included into a term which can be added to the end of **Equation 2.27**; the exchange and correlation functional (E_{xc}). Altogether this gives a KS DFT energy functional:

$$E_{DFT}[\rho] = T_{non\ int}[\rho] + V_{ne}[\rho] + J_{ee}[\rho] + E_{xc}[\rho] \tag{2.28}$$

Equation 2.28: Kohn Sham DFT energy functional

The difference between E_{DFT} and the exact energy (E_{exact}) defines the exchange and correlation energy functional E_{xc}.

$$\begin{aligned} E_{xc}[\rho] &= (T[\rho] - T_{non\ int}[\rho]) + (E_{ee}[\rho] - J_{ee}[\rho]) \\ E_{xc}[\rho] &= \Delta T[\rho] + \Delta V_{ee}[\rho] \end{aligned} \tag{2.29}$$

Equation 2.29: Kohn Sham DFT E_{xc} functional definition.[113]

The exchange and correlation functional therefore contains the corrections to the kinetic energy, i.e. the kinetic energy contribution due to electron interactions, and secondly, a correction to the potential energy for electron exchange and correlation. In principle this contains all warranted corrections, hence removing undesirable features such as self-interaction. It is important to note, that there is no reduction in the generality or exactness of this formulation over the orbital free formulation of Hohenberg and Kohn.[129] Having created a functional we require a method to locate the orbitals that minimise the energy. To do this a KS operator (h^{KS}) is applied in a pseudo-eigenvalue equation:

$$h_i^{KS} = -\frac{1}{2}\nabla_i^2 - \sum_k^{nuclei} \frac{Z_k}{|r_i - r_k|} + \int \frac{\rho(r')}{|r_i - r'|} dr' + \frac{\delta E_{xc}}{\delta \rho} \tag{2.30}$$

Equation 2.30: Kohn Sham Hamiltonian[112,132]*. The terms on the right hand side are the kinetic energy of the non-interacting system, the nuclear electron interaction, the electron electron interaction and finally the exchange correlation potential (the functional derivative of E_{xc}) respectively.*

$$h_i^{KS}\psi_i = \epsilon_i \psi_i \tag{2.31}$$

Equation 2.31: Pseudo-eigenvalue equation from Kohn Sham DFT[112]

The molecular orbitals (ψ) are expressed within a basis set $\{\phi\}$ and the molecular orbital coefficients are determined by solution of a secular equation as in HF. The density is needed for solution of the secular equation, but the density is based upon the orbitals resulting from the secular equation. The solution is therefore found

iteratively, as in HF. Some similarities exist between DFT and HF. A key difference is that HF is approximate by construction, a first step to a full solution, but DFT is a formally exact theory as described above. However, due to the form of the E_{xc} functional being unknown, the functionals we use are approximations of the exact theory.[112,132] As a result many DFT functionals suffer from incomplete treatment, leading to errors such as self-interactions. Perdew provided a "Jacob's ladder" of DFT, **Figure 2.9**.[134] It represents the various approximations of DFT functionals and a guide to their relative success.

Figure 2.9: Jacob's ladder of DFT approximations.[134]

There is no systematic approach to DFT functional improvements, due to a lack of prescription to find the exact exchange and correlation functional. Approximate KS DFT employs a range of functionals with differing levels of theoretical justification and complexity, often the most unlikely functionals have proven most successful.[132] The first of these approximations is the *local density approximation* (*LDA*) . The basis of the LDA is the uniform electron gas as mentioned earlier in relation to the Thomas-Fermi DFT kinetic energy functional. LDA is the basis of most modern approximate DFT exchange and correlation functionals. A uniform electron gas is one in which there is no net electrical charge, i.e. the background is a positively charged distribution of the same magnitude but opposite sign to that of the electrons. The number of electrons (N) and the volume (V) are assumed to approach infinity whilst the density is still a finite quantity holding the same value everywhere within the system($\rho = \frac{N}{V}$).[132] This corresponds nicely to an ideal simple metal, indeed this approximation has been applied successfully in solid state physics for many years. **Figure 2.10** below, shows a pictorial representation of the LDA approximation:

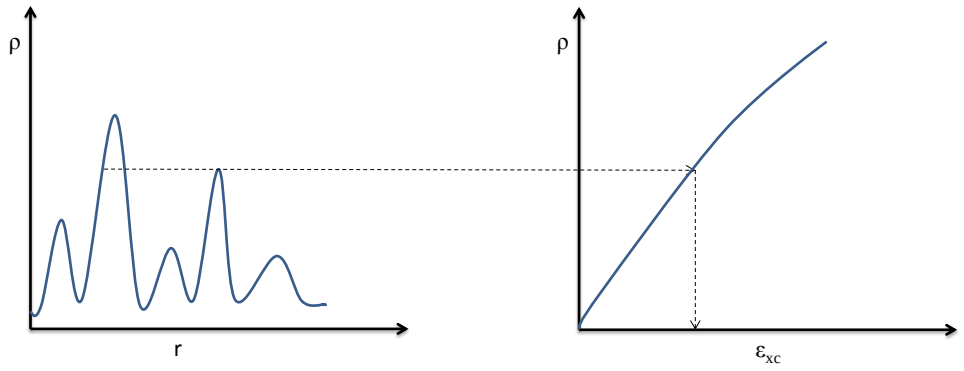


Figure 2.10: A pictorial representation of the local density approximation. This diagram represents the idea of the varying electron density on the left, being approximated by the uniform electron gas by mapping the density to same level and taking the corresponding ϵ_{xc} value. Inspired by the lectures and notes of N.M. Harrison.[135]

Unfortunately this functional provided few applications in computational chemistry. This is due to molecules having rapidly changing density, as is shown in the diagram, hence not close to the approximation of a uniform electron gas.[132]

The second level of approximation is the *generalised gradient approximation* (*GGA*). This model builds on the LDA adding a variable, which is the first derivative of the density, i.e. the gradient of the density. This was the method that introduced DFT as a useful method in computational chemistry. Other attempts to add a gradient correction had previously occurred but had poor results.[113] Many popular GGA functionals have been produced. An early, and still popular, exchange energy functional is that from Becke. It is known as B or B88.[136] This is a fitted functional which has seen extensive use. Other fitted functionals were produced such as the HCTH functionals which were fitted to large experimental datasets.[137] Correlation energy potentials, such as that from Lee, Yang and Parr (LYP), were developed at the same time.[138] This functional contained four fitted parameters. Other popular developments from this time were those of Perdew and Wang (PW86, PW91),[139–141] and Perdew-Burke-Ernzerhof (PBE).[134] These are linked exchange and correlation functionals based off the same model. Each iteration from PW86 to PBE can be treated as a refinement to the previous functionals. PBE is still widely employed, especially in solid state calculations. Further extensions to GGA models have been suggested, including *meta-GGA*'s which either contain a dependency on the orbital kinetic energy or a Laplacian (second derivative $\nabla^2(\rho)$) of the electron density. It is common that meta-GGA's use the orbital kinetic energy term as it is generally more numerically stable than the $\nabla^2(\rho)$.[113]

The final level is that of hybrid functionals. It is debatable whether these should be placed at this point of the ladder, owing to their often heavily empirical nature. Nevertheless, hybrids have been used successfully in many applications. These are methods which are empirically fitted and contain some portion of exact exchange i.e. HF exchange. The well known B3 functional from Becke is a three parameter functional fitted to experimental values and containing 20% exact exchange.[142]. This functional has been one the most extensively used combining several existing DFT exchange and correlation functionals such as B3LYP, B3PW91 and O3LYP. B3LYP

is the most extensively used DFT functional in computational chemistry and has the functional form of **Equation 2.32**.

$$E_{xc}^{B3LYP} = (1-a)E_x^{LSDA} + a(E_x^{exact(HF)}) + b\Delta E_x^{B88} + (1-c)E_c^{LSDA} + c(E_c^{LYP}) \quad (2.32)$$

Equation 2.32: B3LYP functional form. The usual values are $a \simeq 0.2$, $b \simeq 0.7$ *and* $c \simeq 0.8$.[113,142]

Other hybrids have also been produced such as PBE0. This adds some portion of HF exact exchange to the existing PBE GGA functional. Hybrids have recently been extended to double hybrids including a portion of the MP2 correction to the correlation energy.[143]

From the generation of GGA's and hybrids in the 1990's, DFT has become a vital resource to computational chemistry. DFT usually provides a level of accuracy above HF and approaching the MP2 level in terms of wavefunction methods, and does so at a fraction of the CPU time cost. Due to the lack of an exact exchange and correlation functional though the full use of DFT is limited and *ad hoc* corrections are now routinely applied to correct it and provide better chemical accuracy. *Dispersion* corrections are currently a major development field. These provide a correction to the DFT correlation energy which generally does not predict the weak attraction due to the van der Waals (VDW) forces. Several methods to incorporate some correction to dispersion have been developed including semi-empirical corrections, such as those from Grimme, Tkatchenko and Scheffler. These are fitted models which generally provide a post-calculation correction to the energy.[144–146] Hybrid-meta-GGA functionals, which include parameters to natively account in an approximate way for dispersion, the Minnesota M05 and M06 family of functionals fall into this category.[147] New functionals are still being developed, but at a slower pace than previously.[148]

2.2.6 Crystallography

Periodic systems are systems that can be described continuously using a *unit cell*. The systems can be 1D, a polymer, 2D, a surface, and 3D, such as a crystal. The cell is characterised by three primitive cell vectors ($\vec{a_1}$, $\vec{a_2}$, $\vec{a_3}$) and the three angles separating the vectors (α, β, γ). Varying these gives seven unique combinations shown in **Table 2.1**. These are combined with nets **(Figure 2.11)** defining unique positions within the cell to place an atom or molecule. Together, the nets and cells form the *Bravais lattices*. The Bravais lattices represent 14 unique 3D combinations of cells and nets, which when translated by the lattice vector $\vec{t}$ (**Equation 2.33**) cover the entire space hence reproducing the structure.

$$\vec{t} = n_1\vec{a}_1 + n_2\vec{a}_2 + n_3\vec{a}_3 \quad (2.33)$$

Equation 2.33: Translational lattice vector

Lattice System	$\lvert a_1\rvert$, $\lvert a_2\rvert$, $\lvert a_3\rvert$ relation	α, β, γ relationship	Lattice Centring
Cubic	$\lvert a_1\rvert = \lvert a_2\rvert = \lvert a_3\rvert$	$\alpha = \beta = \gamma$	P I F
Tetragonal	$\lvert a_1\rvert = \lvert a_2\rvert \neq \lvert a_3\rvert$	$\alpha = \beta = \gamma$	P I
Orthorhombic	$\lvert a_1\rvert \neq \lvert a_2\rvert \neq \lvert a_3\rvert$	$\alpha = \beta = \gamma$	P I F C
Hexagonal	$\lvert a_1\rvert = \lvert a_2\rvert \neq \lvert a_3\rvert$	$\alpha = \beta = 90°\gamma = 120°$	P
Trigonal	$\lvert a_1\rvert = \lvert a_2\rvert = \lvert a_3\rvert$	$\alpha = \beta = \gamma \neq 90°$	P
Monoclinic	$\lvert a_1\rvert \neq \lvert a_2\rvert \neq \lvert a_3\rvert$	$\alpha = \beta = 90°\gamma \neq 90°$	P C
Triclinic	$\lvert a_1\rvert \neq \lvert a_2\rvert \neq \lvert a_3\rvert$	$\alpha \neq \beta \neq \gamma \neq 90°$	P

Table 2.1: Bravais lattices

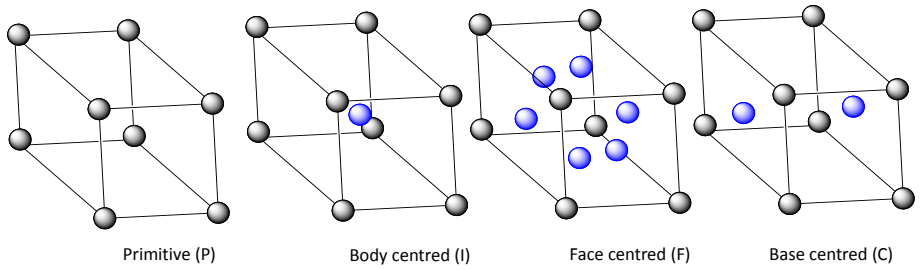


Figure 2.11: The possible lattice centrings in the Bravais lattice systems

The concept of reciprocal space is an important one for periodic systems as it can provide a simplification to the problem. The reciprocal of a cell is also defined by three vectors ($\vec{b_1}$, $\vec{b_2}$, $\vec{b_3}$) which are derived from the primitive lattice vectors of the real space cell and following orthonormality (**Appendix D**).

$$\vec{b_1} = 2\pi\frac{\vec{a_2}\times\vec{a_3}}{\vec{a_1}(\vec{a_2}\times\vec{a_3})} \quad \vec{b_2} = 2\pi\frac{\vec{a_3}\times\vec{a_1}}{\vec{a_2}(\vec{a_1}\times\vec{a_3})} \quad \vec{b_3} = 2\pi\frac{\vec{a_1}\times\vec{a_2}}{\vec{a_3}(\vec{a_1}\times\vec{a_2})} \quad \vec{a_i}\vec{b_j} = 2\pi\delta_{ij} \quad (2.34)$$

Equation 2.34: Relation of reciprocal space lattice vectors and real space lattice vectors.

A cubic cell in real space remains a cubic cell in reciprocal space, but with its cell lengths scaled as: L in real space $\rightleftharpoons \frac{2\pi}{L}$ in reciprocal space. We can analogously define a vector in the reciprocal space ($\vec{K}$) to that in real space($\vec{r}$). This is known as a wave vector as it has units of $\frac{1}{Length}$.

2.2.7 The Electronic Structure of Crystals

Bloch's theorem states that the electronic wave function in a periodic structure can be written as a product of a plane wave and periodic function, with the same periodicity as the unit cell. This is known as a Bloch state or *Bloch orbital*.

$$\psi_{n,k}(r) = e^{iK.r}\upsilon_n(r) \tag{2.35}$$

Equation 2.35: Bloch orbital. $\psi(r)$ is a one-electron wavefunction. The first term on the right hand side represents a plane wave with K being the wave vector, $i = \sqrt{-1}$ and r is the position. The second term on the right hand side is the cell periodic function.

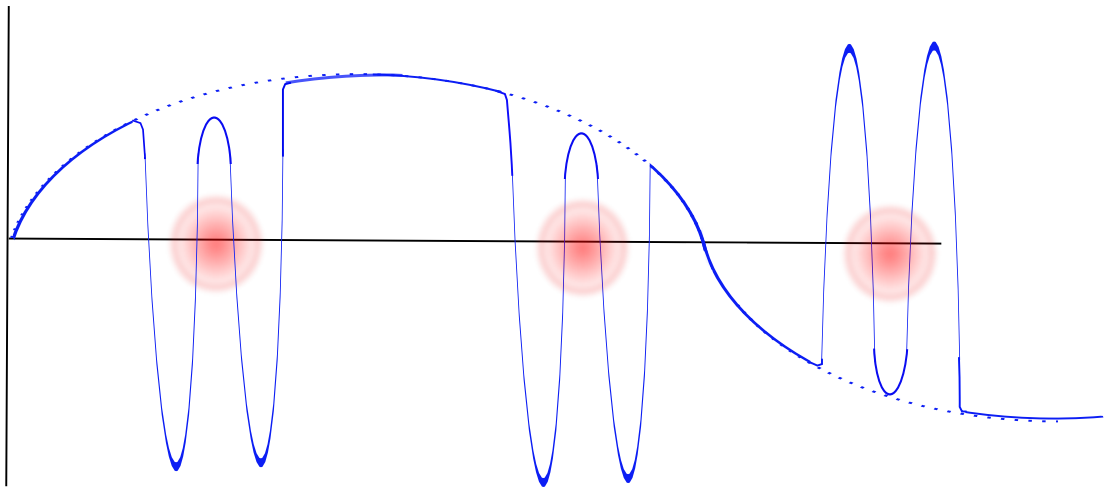


Figure 2.12: Schematic of the real portion of a Bloch wave: The dotted line represents the plane wave. The red spheres are atoms and the oscillating waves are the cell periodic functions.[149]

We can expand the cell periodic function ($\upsilon(r)$) in a basis set. As this is a periodic function the natural choice of basis set is that of plane waves, though it can also be done using GTO basis functions. **(Section2.2.3)**[113]

$$\upsilon_n(r) = \sum_{\alpha}^{N_{basis}} c_{n\alpha}\chi_{\alpha}^{Planewave}(r)$$

$$\psi_{n,k}(r) = e^{iK.r}\sum_{\alpha}^{N_{basis}} c_{n\alpha}\chi_{\alpha}^{Planewave}(r) \tag{2.36}$$

Equation 2.36: Bloch orbitals. Top: Cell periodic function expanded in a basis of plane waves. Bottom: Bloch function having being expanded in a basis of plane waves.

Using Bloch's theorem we can employ *periodic boundary conditions*. These enforce that at the point r if we travel through space by a translation of $\vec{t}$, we will be at an equivalent point within the overall structure, but simply one unit cell away from where we started. In this case the periodic portion of the wavefunction at the two points will be equivalent only changing in the plane wave component. This immediately reduces a seemingly intractable problem of an infinite system, hence,

infinite number of electrons, to a single unit cell as this cell can be used to recreate the entire structure by translation.[113]

We can however simplify the problem further. If we take the reciprocal of the unit cell and join the lattice points, then bisect the connecting lines, we can define a new cell. This is known as the first *Brillouin zone* (*BZ*) of the real space lattice or *Wigner-Seitz* cell of the reciprocal lattice. These are primitive cells as they contain only a single lattice point. They also cover the entire structure making them ideal choices for reciprocal space unit cells. Hence, by working in reciprocal space it is not necessary to even fully calculate the unit cell; it is in fact enough to only calculate the first BZ. This has greatly simplified the calculation we require, by taking us from a set of integrals over an infinite system to a set of integrals over the first BZ. We can finally apply a grid of points to the first BZ called K points. K points are a set of points within the first Brillouin zone positioned to represent the symmetry of the system, an important K point is the gamma point, which is the centre point of the 1st Brillouin zone. At each of these points the appropriate calculation is made hence approximating the integral over the entire BZ.[113]

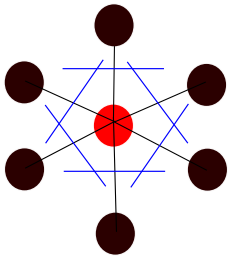


Figure 2.13: Schematic of a Wigner-Seitz cell. The red sphere represents the single lattice point enclosed in the WS cell with the blue bisecting line representing the edges of the WS cell. The black spheres are the next set of lattice points away from the red lattice point. Inspired from the following reference.[150]

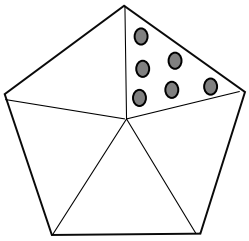


Figure 2.14: A 2D schematic of a K-point grid. Using the symmetry of the cell it is sufficient to sample over only one section of the Brillouin zone. The K points, grey circles, represent the points where the system is evaluated to approximate the integral over the entire zone.

2.2.8 Crystal Lattice Simulation

We can model crystal lattice structures using model potentials and electrostatic potentials. These models are cheaper than periodic DFT and can lead to good results more efficiently. The program DMACRYS[65] is an example of such a scheme. Here a fitted repulsion and dispersion potential is applied along with an electrostatic

potential calculated by *distributed multipole analysis* (*DMA*), originally developed by A. Stone.[151] DMA calculates an multipole moments which allow the electrostatic potential to be accurately calculated. The multipole moments are determined directly from the density matrix and basis functions of a prior quantum chemical calculation. The charge overlap between the basis functions can be expanded as a multipole around the overlap. For example two s orbitals overlapping produces a point charge, an s and p function overlap has components of a point charge and a dipole.[113]

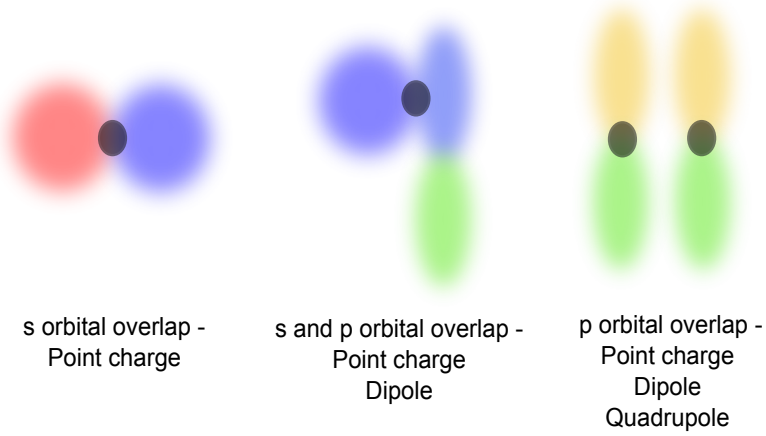


Figure 2.15: Depiction of charge overlap with different basis functions.[123]

This expansion in principle is limited if all distributed multipoles are produced for every pair of basis functions. However, pragmatically this involves too many multipole centres and for general purposes multipole centres are restricted to the atomic nuclei and some points along bonds. This relocating of multipole centres prevents the convergence to a limited number of terms however, generally the terms of higher order contribute little to the now infinite sum so can be ignored. If Gaussian basis functions are used the centring is given by **Equation 2.37**.

$$R_{cent} = \frac{\alpha R_1 + \beta R_2}{\alpha + \beta} \tag{2.37}$$

Equation 2.37: Distrusted multipole analysis centring from Gaussian basis functions. R_i are the coordinates of two nuclei upon which the basis functions are centred. α and β are the exponents of the Gaussian functions.[113]

In addition to the multipoles representing the electrostatic interactions DMACRYS employs an empirical repulsion and dispersion potential. This is fitted to experimental data from crystal structures and given in the form of a *Buckingham potential.*

$$U^{mn} = \sum_{m \in i \; n \in j} A_{ij} e^{-B_{ij} R_{ij}} - \frac{C_{ij}}{R_{ij}^6} \tag{2.38}$$

Equation 2.38: Where i and j label the atoms in the molecule and m and n label a molecule respectively. This Buckingham potential defines an intermolecular potential.

Equation 2.38 shows the functional form of the Buckingham potential. A, B and C are empirical constants.[152–158]Here, C quantifies the attractive portion of the potential relating to dispersion forces. A and B model represent the repulsive barrier due to steric interactions. The functional form of R^{-6} is theoretically more justifiable in this function, as it is the leading term in the dispersion energy. Exchange-repulsion is modelled by the exponential form. The exponential function remains finite even when r tends to 0, therefore this potential risks running into a 'Buckingham catastrophe'. This is a spurious artefact of the potential form in which nuclear fusion erroneously occurs as a result of strong dispersive and electrostatic forces capable of overcoming the repulsive barrier.[65,123]

$$A_{ij} = (A_{ii} A_{jj})^{\frac{1}{2}} \qquad B_{ij} = \frac{1}{2}(B_{ii} + B_{jj}) \qquad C_{ij} = (C_{ii} C_{jj})^{\frac{1}{2}} \tag{2.39}$$

Equation 2.39: The combining rules of the Buckingham potential empirical parameters. A and C geometric mean, B arithmetic mean.

The mixing rules shown in **Equation 2.39** combine the empirical parameters related to an atom interacting with some other atom of another molecule. The parameters are derived from hetero-atomic interactions to reproduce the lattice constants and experimental sublimation data. The parameters are defined for homo-atomic interactions, for example if $Z = A, B$ or C, then Z_{ii} is the parameter of element i interacting with element i in another molecule and the same for Z_{jj}. The mixing rules are then used to locate Z_{ij} which is then the average of the interaction of the elements i and j. An example Buckingham potential is plotted below.[65]

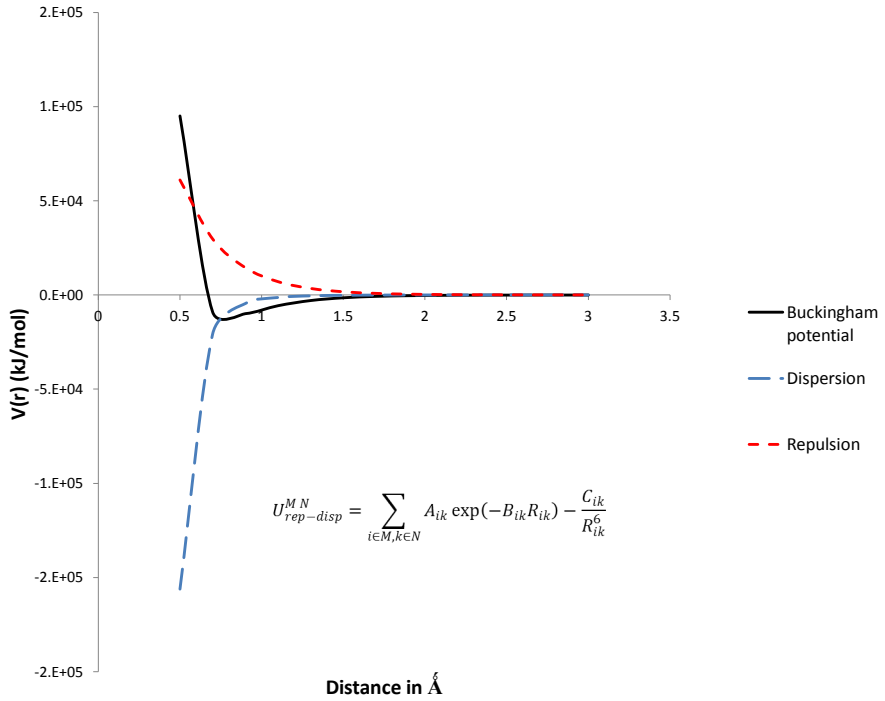


Figure 2.16: Intermolecular Buckingham potential function for c-c, A,B and C values taken from the FIT potential in DMACRYS.

2.2.9 Lattice Minimisation

Within DMACRYS, the electrostatic portion of the lattice energy is evaluated using an Ewald summation up to rank two multipoles (charge-charge, charge-dipole and dipole-dipole) and direct summation for higher order terms. The repulsion and dispersion interactions are evaluated as described above in the Buckingham potential. Additional terms such as induction can also be account for using a similar scheme. The minimisation routine involves centralising the forces, torques and second derivatives of each molecule to the centre of mass of each molecule. Changes to the crystal structure can be expressed as a multi-component vector δ. This vector contains components related to molecular translation and rotation (three components each) and six components related to strain modelled as a bulk deformation of the crystal structure. DMACRYS then minimises the intermolecular energy as a function of a small change (r) in the crystal structure. This is conveniently given as a power series:[65]

$$U_{intermolecular}(r') = U_{intermolecular}(r) + \delta^T \cdot g + \frac{1}{2}\delta^T \cdot W \cdot \delta \qquad (2.40)$$

Equation 2.40: A power series for lattice energy minimisation in DMACRYS.[65]

2.2.10 Phonon Modes

Phonons are described as quanta of sound and regarded much the same as a photon being a quanta of light. Phonons are vibrations of interacting elastic materials, in

this case crystals. These vibration represent excited vibrational states within the system.

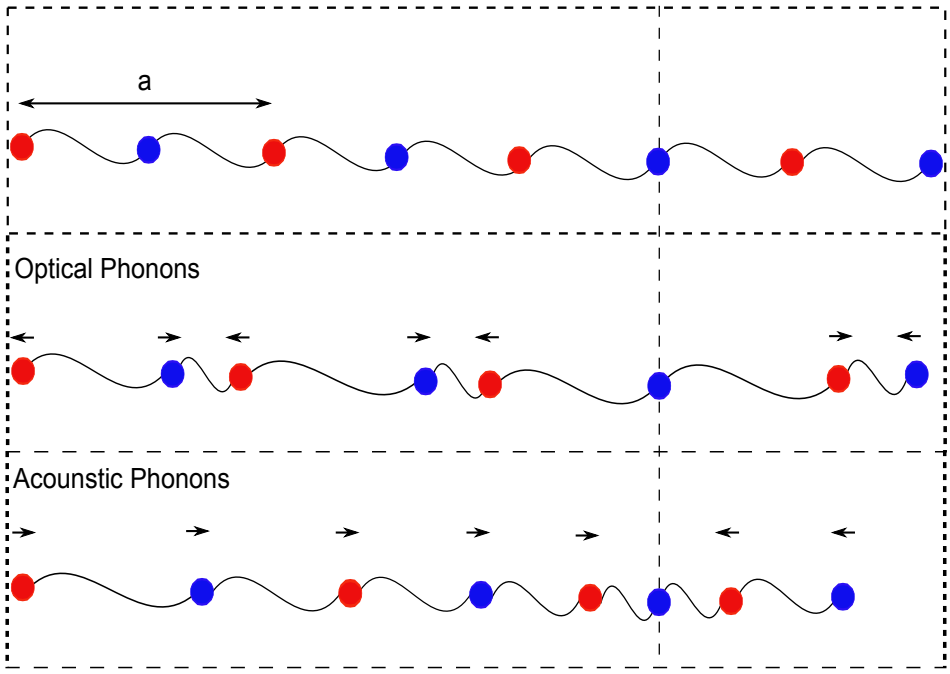


Figure 2.17: Phonon modes: The top image shows evenly spaced spheres representing atoms with a being repeat unit distance. The lower schemes show the two different types of phonon modes optical and acoustic in a simple two element model.[159]

Figure 2.17 shows the two different phonon modes possible in a system composed of two or more different atoms. The middle scheme shows the out of phase vibrations known as optical phonons, these vibrations show a non-zero frequency at the centre of the Brillouin zone (gamma point)(**Figure 2.18**). The lower scheme represents the vibrations of the acoustic phonons. These modes are more similar to waves in air or water moving atoms in one region closer together and atoms other regions further apart. These modes have a zero-frequency at the gamma point (**Figure 2.18**).

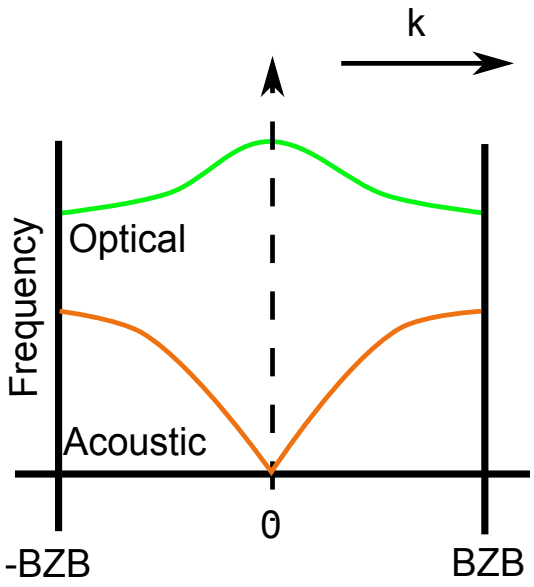


Figure 2.18: Phonon modes:BZB is an abbreviation for Brillouin zone boundary. The image s hows the frequency of both phonon modes related to the Brillouin zone centre.[159]

The phonon modes in this work are used to calculate the entropy of the crystal and are related to the internal energy via the Helmholtz free energy (F) (**Equation 2.41**).

$$F = U - TS \tag{2.41}$$

$$F = U + \frac{1}{2}\sum_i h\nu_i + kT\sum_i ln\left(1 - e^{-\frac{h\nu_i}{kT}}\right)$$

Equation 2.41: Helmholtz free energy. h is Planck's constant, k is the Boltzmann constant, ν is the frequency of the vibration, T is the temperature in kelvin, U is the internal energy, V is the volume and S is the entropy.[160]

DMACRYS calculates the optical phonons at the gamma point only. Contributions from acoustic and optical phonons outside of the gamma point are approximated using a hybrid Debye-Einstein approximation,[65] where the Einstein approximation assumes a single frequency of vibration for all atoms, hence each atom can be represented by decoupled 3D quantum harmonic oscillators. Debye's approximation calculates a theoretical maximum frequency of vibration based on the number density of an atom and the speed of sound in the crystal.[160] This enables the calculation of the density of states of the phonons away form the gamma point. The gamma point approximation becomes better for large unit cells as the calculation takes place in reciprocal space, hence large objects become smaller. The entropy of the crystal is then calculated as the partial derivative of the free energy with respect to the temperature at a constant volume **Equation 2.42**.

$$S = -\left(\frac{\partial F}{\partial T}\right)_v \tag{2.42}$$

Equation 2.42: The entropy calculated as the partial derivative of the free energy with respect to temperature at constant volume.[160]

2.2.11 Gas Phase Entropy Contributions

All gas phase entropy contributions were calculated by Gaussian 09 using statistical thermodynamics. The partition function of a given component can be used to determine the entropy contribution of the given component using the **Equation 2.43**, which is used to calculate molar quantities assuming ideal gas behaviour.[161]

$$\begin{aligned} S &= R + Rln(Q) + RT\left(\frac{\partial lnQ}{\partial T}\right)_v \\ &= Rln(Q\ e) + RT\left(\frac{\partial lnQ}{\partial T}\right)_v \\ &= R\left(ln((q_t q_e q_r q_v)e) + T\left(\frac{\partial lnq}{\partial T}\right)_v\right) \end{aligned} \tag{2.43}$$

Equation 2.43: The entropy calculation in Gaussian 09 (N=1). Q is the total partition function and q_i is a single component of the partition function (t = translation, e = electronic, r = rotation and v = vibration). R is the gas constant and T is the temperature in Kelvin. In the second equation e is substituted into the ln function as it equals 1, hence maintaining the first term in the first equation.[161]

We are concerned with the translational and rotational degrees of freedom gained in the gaseous state compared to the crystalline solid form. We therefore substitute the partition functions for translation and rotation into **Equation 2.43**. The translational partition function is as follows:[161]

$$q_t = \left(\frac{2\pi mKT}{h^2}\right)^{3/2} V \tag{2.44}$$

Equation 2.44: The translational partition function. h is Planck's constant, K is the Boltzmann constant, m is the mass and T is the temperature in Kelvin.[161]

The partial derivative of the natural logarithm of q_t with respect to temperature at constant volume gives a value of $\frac{3}{2T}$ leading to the entropy of translation being:

$$\begin{aligned} S_t &= R\left(ln(q_t\ e) + T\left(\frac{3}{2T}\right)\right) \\ &= R\left(ln(q_t) + ln(e) + \frac{3T}{2T}\right) \\ &= R\left(ln(q_t) + 1 + \frac{3}{2}\right) \end{aligned} \tag{2.45}$$

Equation 2.45: The translational entropy as calculated by Gaussian 09. e remains in this partition function from Stirling's approximation of natural logarithms of factorial quantities.[21,161]

The rotational partition function for a non-linear molecule is as follows:

$$q_r = \frac{\pi^{1/2}}{\sigma_r}\left(\frac{T^{3/2}}{(\theta_{r,x}\ \theta_{r,y}\ \theta_{r,z})^{1/2}}\right) \tag{2.46}$$

Equation 2.46: The rotational partition function. Θ is the characteristic rotational temperature ($\Theta = \frac{h^2}{8\pi^2 IK}$ where I is the moment of inertia. σ is the symmetry number of indistinguishable orientations and T is the temperature in Kelvin.[161]

The partial derivative of the natural logarithm of q_r with respect to temperature at constant volume also gives a value of $\frac{3}{2T}$ leading to the entropy of rotation being:

$$S_r = R\left(ln(q_r) + \frac{3}{2}\right) \tag{2.47}$$

Equation 2.47: The rotational entropy as calculated by Gaussian 09.[161]

2.2.12 Solvation models

In this section, a discussion of the available solvation models including explicit, continuum, and hybrid is presented.

2.2.12.1 Explicit Solvation Models

Explicit solvent models are models in which the solvent molecules are treated explicitly i.e. the coordinates and usually at least some of the molecular degrees of freedom are included. This is a physically realistic picture in which the solvent interacts directly with the solute molecule (contrast to continuum models **Section 2.2.12.2**). These models generally occur in the application of molecular mechanics (MM) and dynamics (MD) or Monte Carlo (MC) simulations, although some quantum chemical calculations do use small solvent clusters. Molecular dynamics simulations allow one to study a system in discrete time intervals and hence to follow a reaction coordinate or system evolution over time. These

simulations employ molecular mechanics forcefields which are generally empirical, parametrised functions designed to efficiently calculate large system properties. These are regularly parametrised to higher level quantum chemical calculation and experimental data.[113]

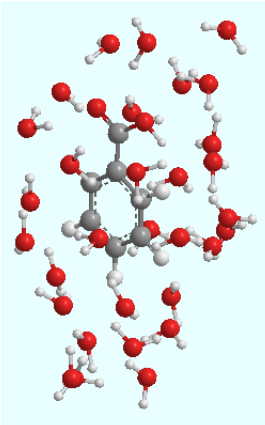


Figure 2.19: Example molecular dynamics simulation box of Benzoic acid in water

In general such forcefield methods are based on an energy evaluation functional containing terms related to bond stretching, angle bending, torsions and terms for repulsion and dispersion, such as the Buckingham potential mentioned previously (**Section 2.2.8, Equation 2.38**). Common solvents often have idealised models generated, reducing the degrees of freedom to evaluate in the energy calculation without great loss of accuracy. Models such as TIPXP (where X is an integer suggesting the number of sites used)[162] and the simple point charge model (SPC)[163] of water have been used extensively. Models of this type typically use a fixed number of sites (often three for water) and place a parametrised point charges and repulsion and dispersion parameter on each. Often aspects of the geometry are fixed such as the bond length or angles.

2.2.12.2 Continuum Solvation Models

Continuum solvation models are models in which no explicit solvent molecules are present and hence their coordinates are not considered. These models usually use a few empirical parameters and represent the solvent as a continuous isotropic field. Continuum models work with a thermally averaged system, hence allowing the solvent to be represented with only a few parameters. The primary parameter is the dielectric constant (ϵ), although others are sometimes used such as surface tensions. The dielectric constant governs to what extent such a medium can be polarised. The solute is encased in a *tessellated* (tiled) cavity which is embedded in the solvent field. The charge distribution of the solute polarises the dielectric medium which hence polarises the solute. This defines a reaction potential, a response to the change in polarisation. This reaction field is iterated to a self consistent solution. The interaction of the solute and solvent is calculated at each of the *tesserae* (tiles) of the solute's cavity. Continuum models have widespread use, including use in forcefield methods. Here we restrict the discussion to continuum models in quantum chemical situations, hence, refer to quantum chemical charge distributions from *ab*

initio methods (HF, Post-HF and DFT). In general these approaches can be thought of in the following way:

$$\hat{H}^{total}(r_{molecule}) = \hat{H}^{molecule}(r_m) + \hat{V}^{molecule+solvent}(r_m) \tag{2.48}$$

Equation 2.48: Separation of the Hamiltonian to include the solute molecule alone and a separate term for the interaction of the solute with the solvent, a perturbation to the gaseous molecules Hamiltonian.[164]

Note that the equation only depends on the solute molecule coordinates (r_m). $\hat{V}^{molecules+solvent}$ is a term made up of interaction operators, generally given the symbol Q. These interaction operators measure the interaction and system changes that occur on going from a gaseous infinitely separated system to one in a continuum solution. Each of these terms directly relates to the free energy of solution with the addition of a fifth term related the thermal averaging.

$$\begin{gathered} Q(m) = Q_{cavity} + Q_{electrostatic} + Q_{repulsion} + Q_{dispersion} \\ G = G_{cavity} + G_{electrostatic} + G_{repulsion} + G_{dispersion} + G_{thermal\ motion} \end{gathered} \tag{2.49}$$

Equation 2.49: Top: Four interaction operators generally considered in the continuum solvation models. Bottom: Five contributing free energy terms from continuum solvation models[164]

Each of the interaction operators has a physical meaning. The first is the cavity creation term. This term accounts for the energy requirement to build a cavity of approximately the shape and size of the solute in the solvent. This can be pictured physically as the energy cost of compressing against the solvents structure. The second term is the electrostatic energy term containing the information pertaining to the solute-solvent polarisation process i.e. the reaction field. The third term is an approximation for the quantum mechanical exchange repulsion. This cannot be explicitly accounted for given the implicit nature of the solvent, but it is approximated based on extensive high level calculations on dimers. The final term is that of the quantum mechanical dispersion energy. This is again an averaging procedure accounting for the solvent charge distribution.[164,165]

There are several flavours of continuum solvation model; generally these differ in how the cavity is constructed, how they account for dispersion/repulsion and cavitation energy, how the charge distribution of the molecule is represented and finally how the solvent is presented. Here we will focus on the methods emanating from the popular *polarisble continuum model* (*PCM*). PCM is the original method but which has been updated to the integral equation formalism PCM (IEFPCM) method in recent years.[166] These models describe the solvent by a single parameter, the dielectric constant. The cavity is constructed by a series of interlocking spheres based on the coordinates of the nuclei.. The spheres define a solvent excluding surface i.e. based on sterics no solvent molecules could reach closer than this point. A second solvent accessible surface is defined by passing a solvent probe over the solvent excluding surface. The former surface is used for calculation of the cavitation energy while

the later is used in the calculation of the repulsion and dispersion, as there is a dependence of solvent radius in the calculation of these terms. The calculation of these non- electrostatic terms comes from scaled particle theory.[167] The electrostatic interactions, leading to the reaction field, are calculated using the *Poisson equation* on an approximate solvent excluding surface (approximate as it is often scaled). The Poisson equation allows the reaction field to be defined.[113,164,168]

$$\nabla\epsilon(r)\nabla\phi(r) = -4\pi\rho(r) \tag{2.50}$$

Equation 2.50: The Poisson equation. ϕ is the electrostatic potential, ϵ is the dielectric constant ρ is the charge distribution and ∇ is the derivative operator.[113]

All methods based in PCM solve the electrostatics using the Poisson equation. Variations on cavity creation have been attempted with SCIPCM, which builds the cavity based on the electron density of the solute. A more recent and very promising addition to this group of methods is the *solvation model based on density* (*SMD*), which comes from the work of Marenich *et al.*[69] This model solves for the electrostatic interactions in the same way as the IEFPCM method. However, it contains a set a specifically parametrised radii for use with the model. These radii should produce accurate electrostatic results, at least for molecules similar to those in the training set, by design. An additional term known as the solvent cavity dispersion solvent structure term is then added to account for non-electrostatics. This empirically parametrised term again is designed for use with the SMD radii. This allows for a variation on the calculation of the repulsion dispersion and cavitation terms compared to traditional PCM models.

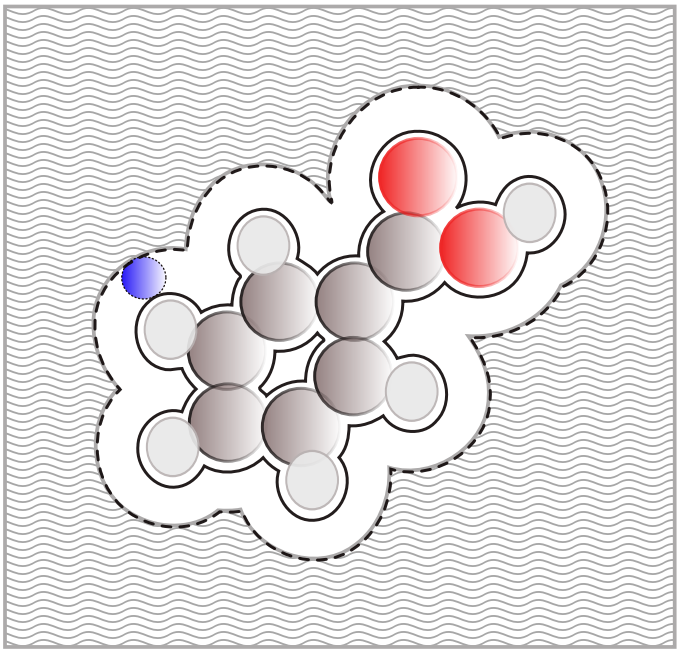


Figure 2.20: The PCM model with solvent excluding surface (solid line) and solvent accessible surface (dash line). The blue ball is a probe sphere of the radius of the average solvent molecule size.

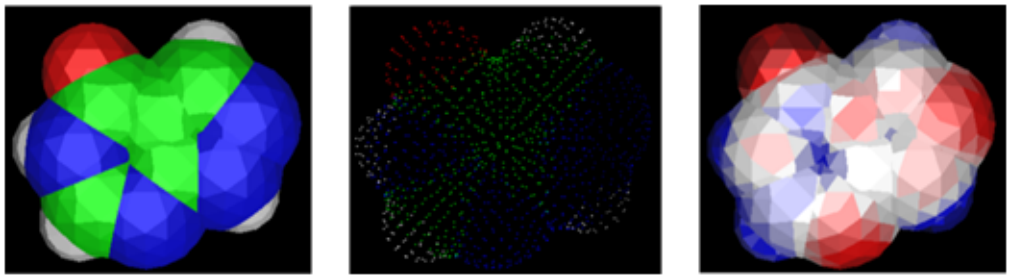


Figure 2.21: The PCM cavity for the molecule allopurinol. Left: The PCM solvent accessible surface. Middle: The central points of the tesserae where the reaction field is evaluated. Right: The polarisation on the surface due to reaction field and solute molecule interactions, Red is negative and blue is positive polarisation.

2.2.12.3 Hybrid Solvation Models

Hybrid models are somewhere in between the explicit and continuum models. These models usually sit slightly closer to one or other of the previous models; for example mixed quantum mechanics and molecular mechanics models, *QM/MM* schemes, sit generally closer to explicit models, having usually a QM core treatment of the solute and perhaps a few explicit water molecules. This is followed by a layer of MM water molecules, generally fading to a continuum description in a third layer. The *reference interaction site model* (RISM) sits on the other side using a continuum representation of solvent density which fluctuates achieving a description of the solvent shell behaviour. Both of these methods use statistical averaging over ensembles.

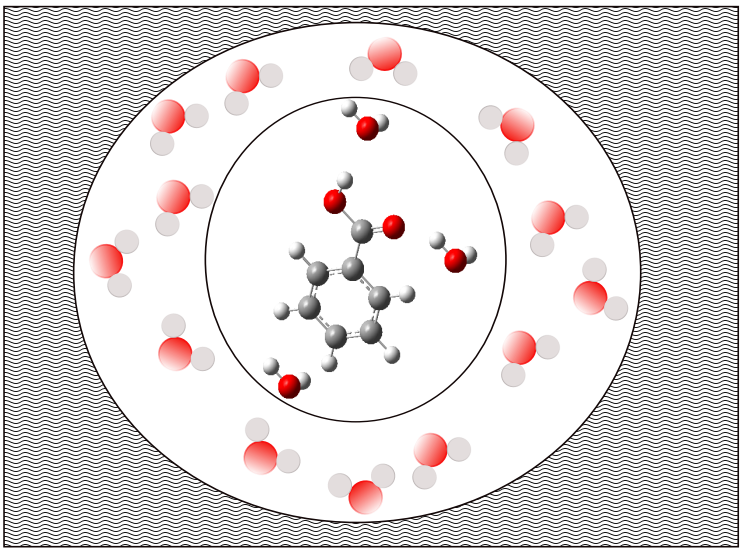


Figure 2.22: A schematic representation of QM/MM. The core is QM, the second layer is MM and the third is the continuum solvent

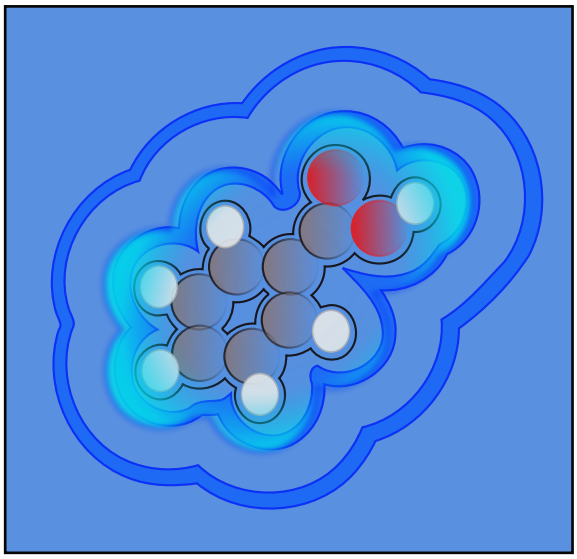


Figure 2.23: A schematic representation of RISM showing solvation shell behaviour around a benzoic acid molecule going to bulk at larger distances.

For the rest of this section we will focus the discussion on the RISM methodology, whilst acknowledging that there are a range of mixed schemes which have been attempted over the years.[112,113,164,169,170]

2.2.12.4 RISM - Background

RISM is a method from classical statistical mechanics based on the *integral equation theory of liquids* (IET). RISM focuses on statistical modelling of the solvent, as in reality a solvent body is a dynamic structure in which an integer number of solvent molecules can only represent a snapshot in time. This is achieved using *Pair Correlation Functions/Radial Distribution Functions* (*PCF/RDF*). These are probabilistic functions representing the chance of finding a solvent atom or molecule at a certain distance from a reference point, in this case often the solute molecule.

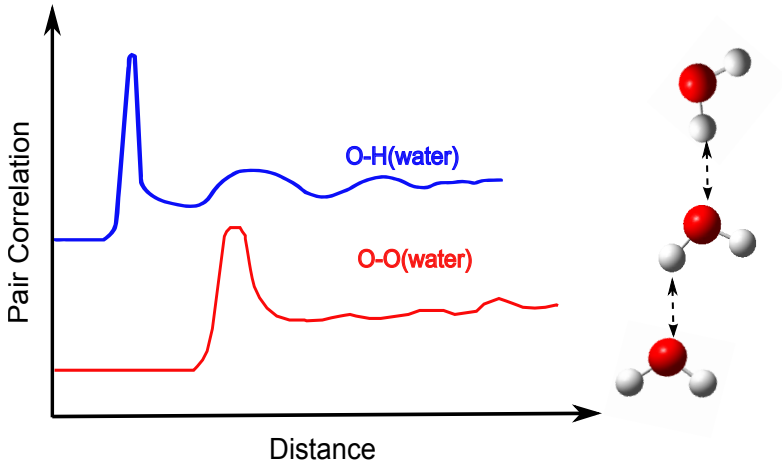


Figure 2.24: A schematic pair correlation function of solvent water. Inspired by section 4.6 in the following reference.[164]

The example pair correlation function above shows the value of such functions when discussing solvent density and structure. We can represent the probability of finding a solvent atom at a certain distance from the reference point of another atom, hence, solvent shell structure is derived naturally from this theory. The calculation method begins with the *Molecular Ornstein-Zernike equation* (*MOZ*).[164] Within this framework a system is defined in a 3D space hence can be defined by three spatial coordinates (r) and three angles (Θ). The MOZ equations utilise relative PCF/RDF's of molecules allowing the total correlation function to be defined as below (**Equation 2.51**). As these equations are of high dimensionality a common approximation is that of spherical symmetry, which removes the consideration of orientational degrees of freedom. **Equation 2.52** shows the MOZ assuming spherical symmetry.[164,165,171]

$$h(r - r^{'}; \Theta - \Theta^{'}) = g(r - r^{'}; \Theta - \Theta^{'}) - 1 \tag{2.51}$$

Equation 2.51: Top: $h(r;\Theta)$ is the total correlation function, $g(r;\Theta)$ is the radial distribution function. This enumerates the effect of a molecule on a second molecule separated by a distance r.[164]

$$h(r) = c(r_{1,2}) + \int dr_3 c(r_{1,3})\rho(r_3)h(r_{2,3}) \tag{2.52}$$

Equation 2.52: The Ornstien-Zernike equation assuming spherical symmetry. ρ is the liquid density, r is the distance separating particles, $h(r)$ is the total correlation function, equivalent to a pair correlation function, $c(r)$ is the direct correlation function.[164]

The MOZ equation splits the total correlation function into two sections: Firstly a direct section, interested in the direct effect of one particle on one other particle separated by a distance r. This is represented by the direct correlation function $c(r)$. The second part is the indirect effect of a third position in a system of three particles. This is represented by the direct correlation function $c(r_{1,3})$ which is the correlation between the first particle and the third particle and the total correlation function $h(r_{2,3})$. This is shown diagrammatically in **Figure 2.25**.

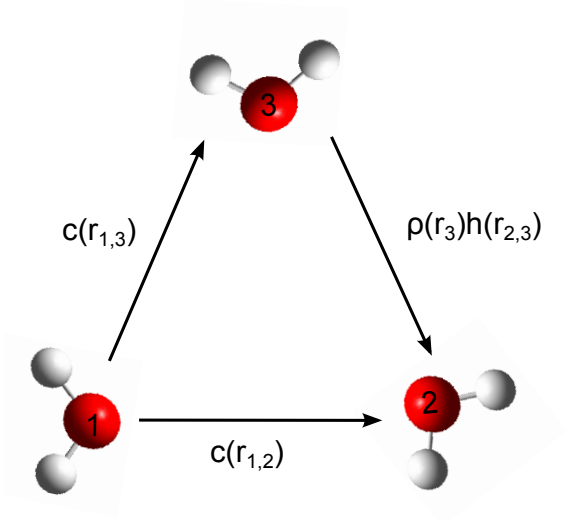


Figure 2.25: A diagram representing the contributions to the total correlation function.

The solutions to this equation are $h(r)$ and $c(r)$, in order to find these two variables a second equation, a closure relation, must be introduced. The exact forms of all terms in this closure relation are not known, hence approximations are made. The most basic of these approximations is the *HyperNetted Chain* (*HNC*) which assumes that those terms whose form is unknown simply do not contribute to the solution (are set to zero). This was initially reasonably successful, but often caused slow convergence and in some cases a divergence.[52] The HNC has been superseded by the *Partially Linearised HyperNetted Chain* (*PLHNC*) also known as the Kovalenko Hirata closure.[172] This closure linearises the exponential function if it exceeds a threshold, in this case if the argument exceeds 0. This leads to a more reliable convergence of the equations.

$$h_\alpha(r) = \begin{cases} e^{-\beta U(r)+T(r)} - 1 & (\text{when} - \beta\upsilon_a(r) + h_a(r) - c_a(r) \leq 0) \\ -\beta U(r) + T(r) & (\text{when} - \beta\upsilon_a(r) + h_a(r) - c_a(r) > 0) \end{cases} \tag{2.53}$$

Equation 2.53: Partially Linearised HyperNetted Chain closure[172] *$\beta = \frac{1}{k_B T}$ $U(r)$ is the interaction potential and $T(r)$ is the indirect correlation function, the difference between the total and the direct correlation function.*

A typical interaction potential used in these calculations is a pairwise additive Lennard-Jones Coulomb potential. An example is provided below in **Equation 2.54**.

$$v(r) = 4\epsilon \left[\left(\frac{\sigma_1}{r_{12}} \right)^{12} - \left(\frac{\sigma_2}{r_{12}} \right)^{6} \right] + \frac{Q_1 Q_2}{r_{12}} \tag{2.54}$$

Equation 2.54: A general solute-solvent interaction potential in the form of a Lennard-Jones potential plus a Colulombic interaction. σ is a parameter that determines the point at which the potential switches and becomes positive, r_{12} is the distance separating the particles, ϵ is the maximum well depth of the potential. Such a potential is similar in form to the Buckingham potential, ***Figure 2.16****. Q_1 and Q_2 are the charges on particles one and two respectively.*[164]

2.2.12.5 3D - RISM

The MOZ equation can be recast into an approximation utilising 3D solute-solvent correlation functions. This approximation is achieved by a partial averaging over the conformational degrees of freedom of the solvent molecules. This allows for the break up of the direct correlation function into partial site contributions, hence the total correlation function is given in **Equation 2.55**. The bulk solvent is modelled by the solvent susceptibility function ($\chi_{\zeta,\alpha}$). This is composed of a term for the intramolecular correlation and a radial intermolecular correlation functions shown diagrammatically in **Figure 2.26**.

$$h(r) = \sum_{\zeta}^{N_{solvent}} \int_{R^3} c_{\zeta}(|r - r_1|)\chi_{\zeta,\alpha}(r_1)dr_1$$
$$\chi_{\zeta,\alpha}(r) = \omega_{\gamma,\zeta}^{solvent} + \rho_{\alpha} h_{\zeta,\alpha}^{solvent} \tag{2.55}$$

Equation 2.55: Top: A re-statement of the Ornstien-Zernike equation; ζ, α label solvent molecule sites. Bottom: The solvent susceptibility function, which defines the bulk solvent response. $\omega_{\gamma,\alpha}^{solvent}$ is the intramolecular correlation function, ρ is the number density and $h_{\zeta,\alpha}^{solvent}$ is the radial total correlation function.

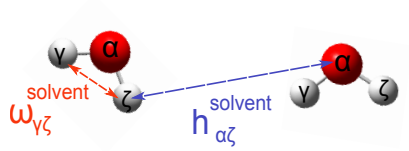


Figure 2.26: A schematic representation of the components of the solvent susceptibility function.[31]

From these equations it is possible to represent the solute-solvent interaction in a way which naturally recovers some information on the solvent structure and hence, provides a more physical picture of the system. 3D-RISM therefore accurately accounts for the spatial correlations of the solvent density which surrounds the solute molecule. Solvent molecules are modelled as a set of atomic sites, with 3D structure described by intramolecular correlation functions. The solute can be of any arbitrary shape and is taken as a single site, generally defined at the origin.

Figure 2.27 shows a plot of an organic molecule surround by solvent, the gradient of the solvent shading provides a diagrammatic interpretation of the solvation shells calculated using 3D - RISM theory.

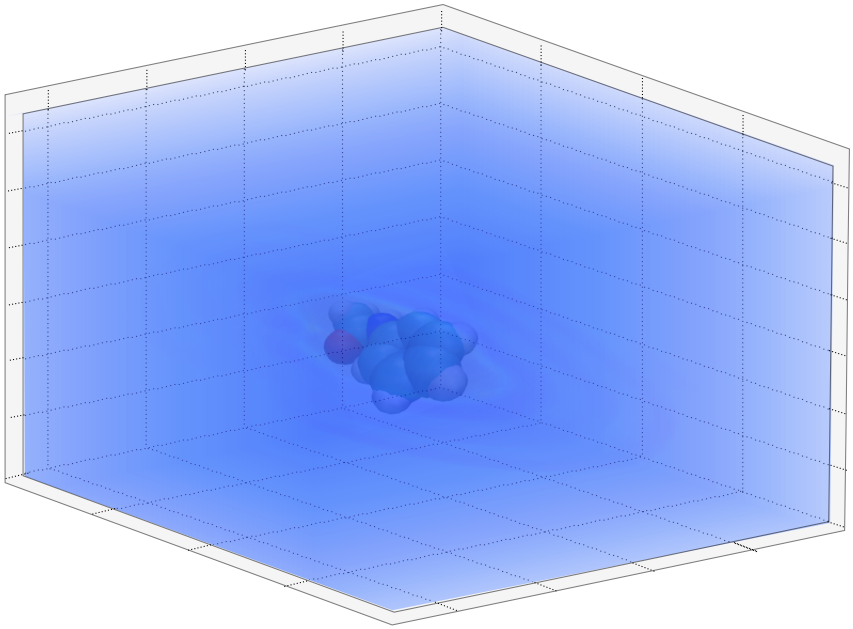


Figure 2.27: A RISM calculated solvent distribution. The gradient determines the density of the solvent around the solute molecule.

2.3 Analysis Statistics

Through this work the following statistics, or some subset of them, are used to analyse the results. There are four in total, the first is the coefficient of determination (R^2), second the Root Mean Square Deviation (RMSD), third the standard deviation (σ), and finally the bias. The definitions of these four statistics are given as follows:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n}(y_{exp}^i - y_{pred}^i)}{\sum_{i=1}^{n}(y_{pred}^i - \bar{y})} \tag{2.56}$$

Equation 2.56: The coefficient of determination. SS_{res} is the residuals sum of squares and SS_{tot} is the total sum of squares.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_{exp}^i - y_{pred}^i)^2}{N}} \tag{2.57}$$

Equation 2.57: Root Mean Square Deviation.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(y_{exp-pred}^i - \bar{y})^2}{N\ -\ 1}} \tag{2.58}$$

Equation 2.58: Standard deviation.

$$Bias = \frac{\sum_{i=1}^{n}(y_{exp}^{i} - y_{pred}^{i})}{N} \tag{2.59}$$

Equation 2.59: Bias.

Within these formulae y_{pred} is the predicted value and y_{exp} is the value from the literature, found experimentally. $\bar{y}$ is the mean of the experimental data. N is the number of input values.

The R^2 value provides a measurement of how well the data fits the model. RMSD provides a measurement of the differences between the predicted values and the actual values. RMSD can be broken down further leading the bias and σ; the bias covers the systematic error of the model and σ estimates the random error of the given model.

Chapter 3

First Principles Predictions of Solubility

"In almost all industries and all biological sciences, we encounter liquid mixtures. There exists an urgent need to understand these systems and to be able to predict their behaviour from the molecular point of view."

Arieh Ben-Naim, 2006

3.1 Sublimation Free Energy Predictions

In this section we will explore the importance of solid state interactions in determining solubility. We have focused our work on predicting the solubility of organic molecules, hence, will focus on organic crystals. Initially data are presented on predicting sublimation thermodynamics, before moving on to hydration thermodynamics and finally solubility. This work was carried out in collaboration with Dr David Palmer.

Sublimation is an endothermic process, which involves taking a molecule from the solid state directly to the gaseous state without a liquid intermediate. Experimentally this occurs in conditions just below the triple point of the substance. The thermodynamic cycle we have chosen to use for solubility prediction goes via the vapour, hence, incorporates predictions of the thermodynamics of sublimation.

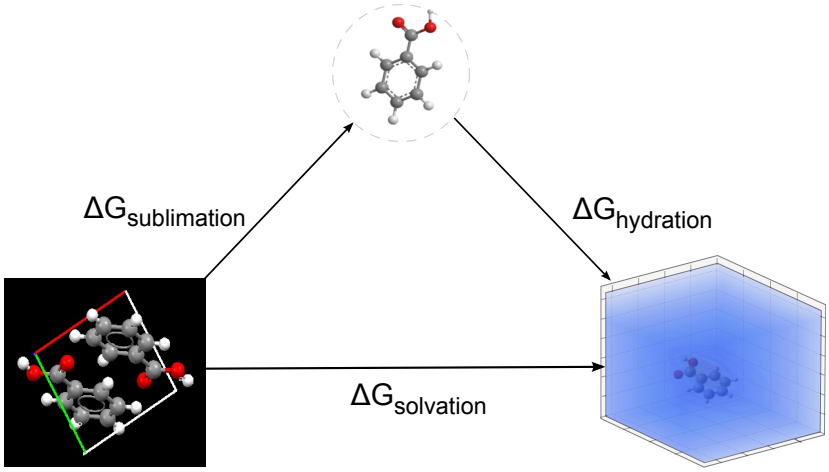


Figure 3.1: Thermodynamic cycle via the gas phase

3.1.1 Testing the Predictability of Sublimation Free energy

We began with a project aiming to produce useful prediction of solubility from 'first principles' methods. The current state of the art in solubility predictions is in the realm of QSPR/QSAR methods. A model free of training set restrictions and physically decomposable into meaningful values and steps would be a valuable tool to anyone interested in solubility prediction. This project is split into three sections over this chapter allowing for a detailed inspection of each prediction step of the thermodynamic cycle. In this section we will focus on the sublimation free energy predictions. These prediction were made using DMACRYS, which is a periodic lattice simulation program, capable of minimising the energy of a crystal unit cell. DMACRYS works on the basis of the theory set out in **Section 2.2.8**, utilising DMA to account for electrostatic interactions and a parametrised Buckingham potential to account for non-electrostatic intermolecular interactions such as dispersion and repulsion.

3.1.2 Calculating ΔG_{sub}

Gibbs free energy of sublimation (ΔG_{sub}) is calculated in the one atmosphere standard state (ΔG^{o}_{sub}), as is consistent with experimental practices. As solubility is calculated in the one mole per litre standard state (Ben - Naim terminology, ΔG^{*}_{sub}) ΔG^{o}_{sub} is later adjusted to meet the Ben - Naim standard state. Within this chapter we will consider all terms in the 1 mole per litre standard state unless explicitly stated otherwise. We calculate ΔG_{sub} using the Gibbs free energy equation

$$\Delta G_{sub} = \Delta H_{sub} - T\Delta S_{sub} \tag{3.1}$$

Equation 3.1: Gibbs free energy equation

3.1.2.1 Calculation of ΔH^o_{sub}

DMACRYS carries out a rigid body minimization of the crystal structure, hence arriving at minimized lattice energies. This lattice energy can be approximately converted into a ΔH_{sub} by the following formula:

$$\Delta H_{sub} = -U_{latt} - 2RT \tag{3.2}$$

Equation 3.2: A. Gavezzotti's approximation to ΔH_{sub}.[173]

This equation is derived by consideration of the packing energy of a crystal, with which it is stated that a half of the energy is approximately equivalent to the lattice energy i.e. the energy difference between a molecule at rest in the solid state and gaseous state. From this point it remains to account for motions of the molecule in the two states and the thermodynamic environment, therefore accounting for the temperature dependence of the ΔH_{sub}. As H is defined as $H = U + (pV)$ in a one molar ideal gas pV = RT hence, the thermodynamic environment can be approximated as an RT correction. From the equipartition theorem, the energy associated with gaseous rotations and translations is 3RT and the energy associated with crystal lattice both vibrational and rotational is equal to $2 \times 6 \times \frac{1}{2}RT$. This gives the following relationship leading to the -2RT term.

$$\Delta H_{sub} = -U_{latt} + RT + \Delta U_{vib} = -U_{latt} + RT + (3RT - 6RT) = -U_{latt} - 2RT \tag{3.3}$$

Equation 3.3: A. Gavezzotti's approximation to ΔH_{sub}.[173]

3.1.2.2 Calculation of ΔS^o_{sub}

The entropy contribution to ΔG_{sub} was calculated as follows:

$$\Delta S_{sub} = (S^{rotation}_{sub} + S^{translation}_{sub}) - S^{cryst,vibrations}_{sub} \tag{3.4}$$

Equation 3.4: Approximate calculation ΔS_{sub}.

In this equation, we sum the contribution to the rotational and translational entropy in the gas phase at 298K. We then subtract the intermolecular vibrational contributions from the crystal, as these will disappear during the sublimation process. We assume that the electronic entropy is consistent across the phase change and that the intermolecular and intramolecular contributions in the crystal are decoupled hence, there is a net zero change in intramolecular entropy over the phase transition. Additional conformational entropy in flexible molecules was ignored, although empirical correction were attempted applying corrections for the number of rotatable bonds.

The calculation of the gaseous rotational and translational entropy was done using statistical thermodynamics in Gaussian 09 assuming an ideal gas.[161] Crystal

contributions to the entropy were calculated by DMACRYS at 298K. From the third law of thermodynamics we know that at 0K the entropy of a perfect crystal is zero. As our system is considered at 298K we must compute the effects of thermal motion. Within a crystalline structure rotational and translational entropies can be assumed to be zero. We therefore require only to account for the vibrational contributions emanating from the crystal lattice. We therefore consider the phonon modes calculated by rigid body lattice dynamics. We calculate the phonon modes at the gamma point (k=0). The free energy expression is given by Day *et al.*[174]

3.1.3 Dataset Generation - DLS-25

We generated a dataset upon which to test our predictive methods. This dataset is called *Drug-Like Solubility - 25* (*DLS-25*), and contains 25 drug-like organic molecules with experimental data taken from the published literature. The diagram below **(Figure 3.2)** shows the chemical structures, names and Cambridge Structural Database (CSD) refcodes. The molecules were selected on the basis of:

- A known solubility value in the literature. Were possible Cheqsol solubilities were used.
- A single crystal X-ray structure of the molecule existed in the CSD.
- Where possible experimentally determined sublimation and hydration free energies were available in the literature.

The lattice energy calculations required a single-crystal structure as input for the energy minimisation, hence the above criteria. As polymorphic characterisation is seldom applied to the crystalline form observed at thermodynamic equilibrium in the solubility experiments, the following algorithm was applied to decide which polymorphic form to use:

- Download all appropriate crystal structures of a compound from the CSD. Suitable here means 3D coordinates are available for a crystal structure which contains one molecular species, hence avoiding salts, solvates and co-crystals.
- The lattice energy of each viable entry was calculated.
- The crystal structure with the lowest calculated lattice energy was selected.

Eight molecules in DLS-25 set had polymorphic information provided with their solubility measurements. For these molecules the same form was used in the calculations as observed in experiment. For 10 of the 25 molecules experimental ΔG_{sub} and Gibbs free energy of hydration (ΔG_{hyd}) values were available in the open literature. These molecules are displayed in **Figure 3.3** and highlighted in **Figure 3.2**. We therefore evaluated the performance of our ΔG_{sub} and ΔG_{hyd} predictions using these ten molecules. The ten molecule subset of the full DLS-25 dataset will be collectively referred to as DLS-10 from this point on.

The difficulty in retrieving experimental values for properties such as ΔG_{sub} and ΔG_{hyd} is a great hindrance to the development of quantitative predictive models. The scarcity of such data is likely due to the difficulties of generating such values (due to low solubility and/or low volatility of some of the compounds in

question), this makes model validation difficult.[51] Of the millions of compounds known within organic chemistry only a few thousand have reported ΔG_{hyd} values in the literature.[51] Such data needs clearly documenting with procedural details and conditions. ΔG_{sub} and ΔG_{hyd} are vital thermodynamic quantities, which are used to calculate other properties of interest such as solubility,[6,31] pKa[175] and protein-ligand binding affinities.[176] For these reasons accurate *in silico* predictions of ΔG_{sub} and ΔG_{hyd} is very important; ideally such methods would allow accurate estimates of ΔG_{sub} and ΔG_{hyd} to be obtained for compounds of low volatility and/or low solubility. The lack of such data is a significant challenge in the development of novel computational models.[31,51,177,178] The ten molecules in our dataset for which experimental ΔG_{sub} and ΔG_{hyd} values were available are shown in the **Tables 3.2** and **3.3** respectively. For future work in this area, it would be of great benefit if more well curated data was made available by experimental colleagues.

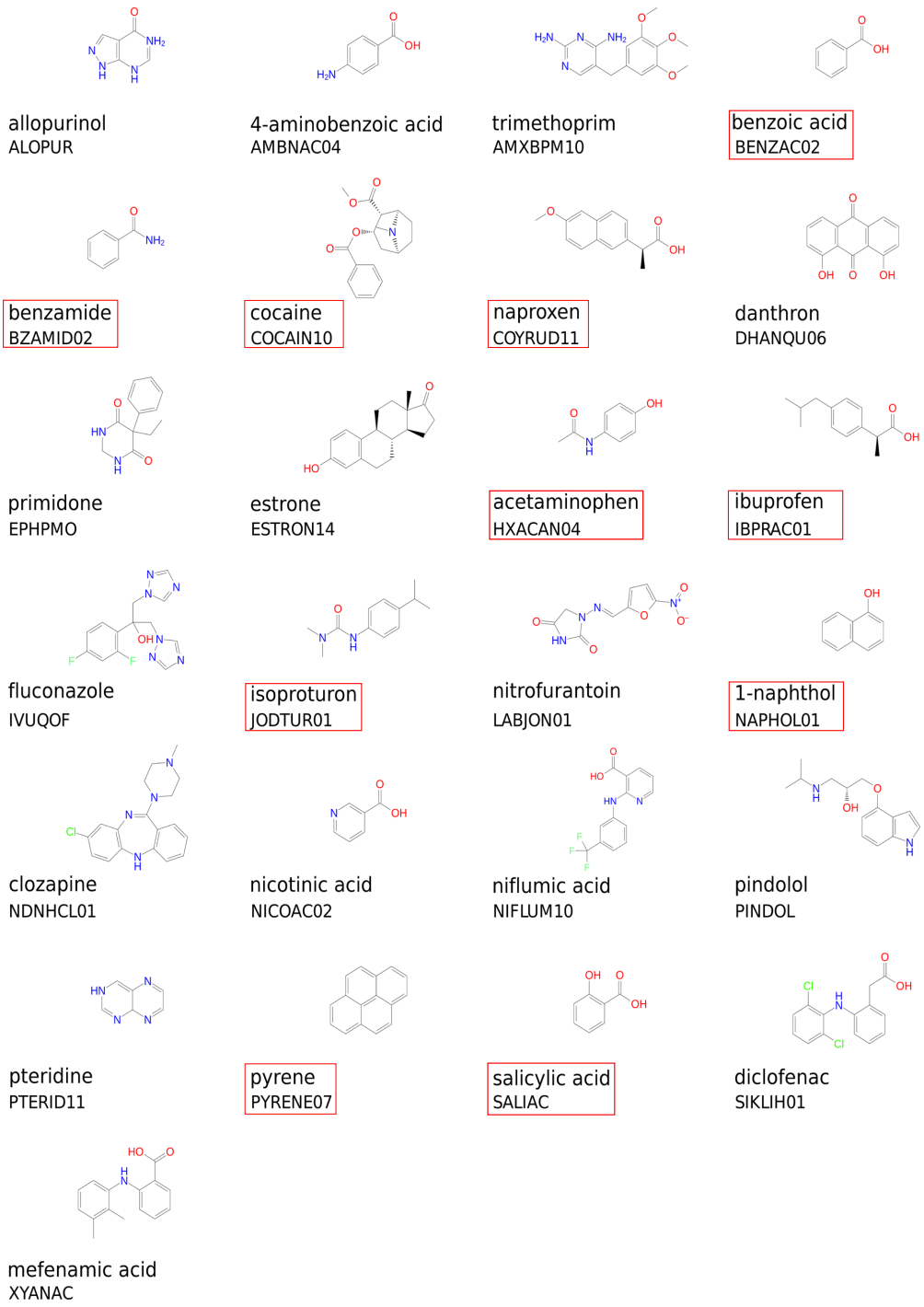


Figure 3.2: The molecular structures, colloquial names and CSD refcodes for the DLS-25 dataset. The refcodes record the polymorph used in our calculations. Image inspired by reference.[31]

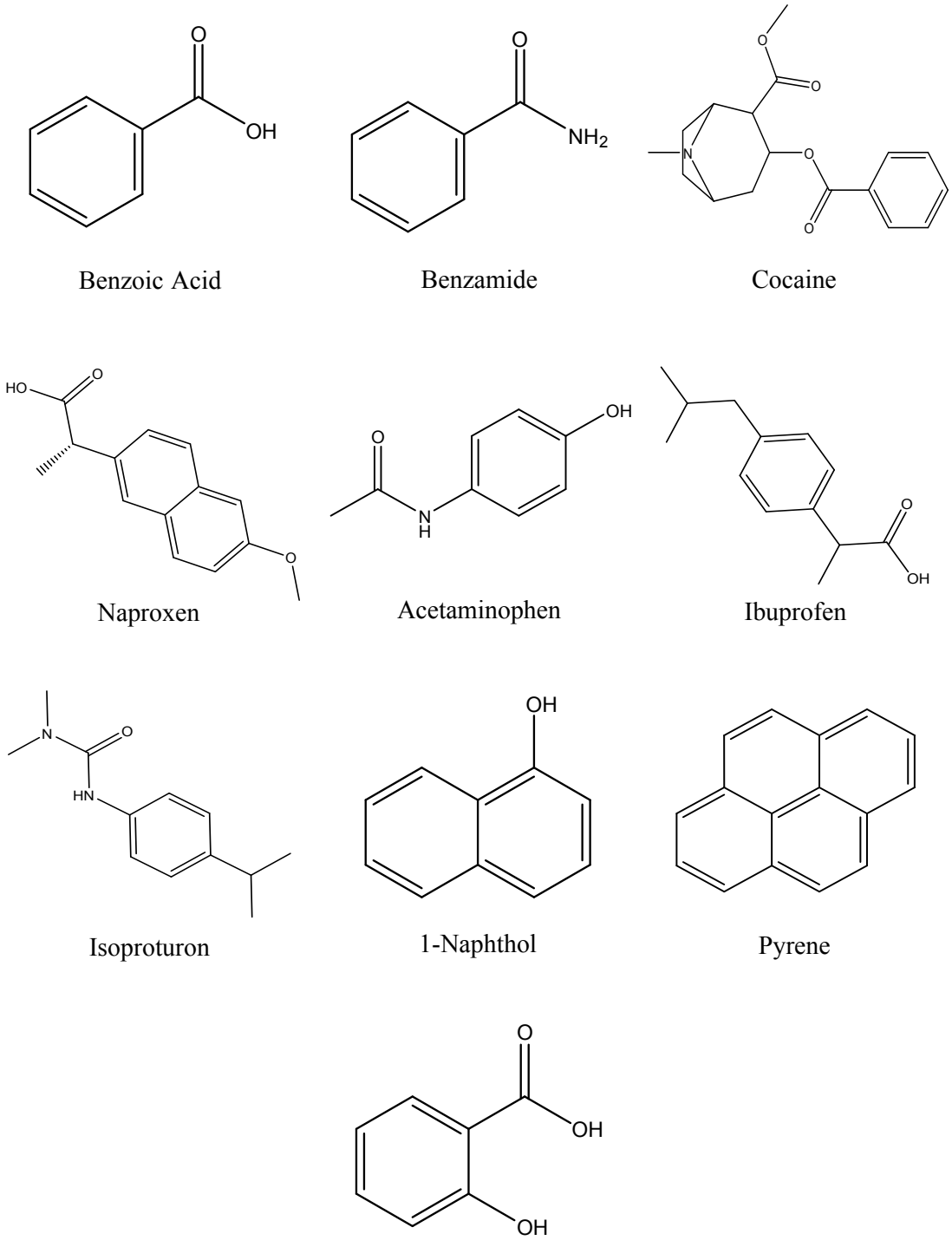


Figure 3.3: DLS-10 a subset of DLS-25 for which experimental values of ΔG_{sub} and ΔG_{hyd} were known.

3.1.4 Predictions From DMACRYS

DMACRYS' model potential encompasses two primary terms: the first, electrostatic and secondly, repulsion - dispersion. The electrostatic contributions were calculated by representing the charge distribution by multipoles upto hexadecapole (rank 4

multipoles). These representations were calculated from three different levels of quantum mechanical theory; HF/6-31G**, MP2/6-31G** and B3LYP/6-31G**. We will refer to each calculated ΔG_{sub} by the following: $\Delta G_{sub}(HF)$, $\Delta G_{sub}(MP2)$ and $\Delta G_{sub}(B3LYP)$. $\Delta G_{sub}(exp)$ will refer to the experimental ΔG_{sub} values.

In the evaluation of the repulsion - dispersion term we selected to use the empirically fitted parameters from the FIT potential. This potential unfortunately lacked hetroatomic parameters for Cl, which was required when evaluating ΔG_{sub} for some of the DLS-25 molecules. Homoatomic parameters where taken from the following reference,[155] we therefore generated the hetroatomic parameters using the mixing rules presented in **Section 2.2.8**. The homoatomatic parameters used in the mixing rules are as follows:

Atom Pair	Description	A_{ik} $kJmol^{-1}$	B_{ik} Å^{-1}	C_{ik} $kJmol^{-1}$Å^6
C-C	Any C atom	369743	3.60	2439.8
H-H	H bonded to C	11971	3.74	136.4
Hp-Hp	H bonded to polar atom	5029.68	4.66	21.50
N-N	Any N atom	254529	3.78	1378.4
O-O	Any O atom	230064.1	3.96	1123.59
F-F	Any F atom	363725	4.16	844
Cl-Cl	Any Cl Atom	924675	3.51	7740.48

Table 3.1: Homoatomic FIT potential parameters for the Buckingham potential

3.1.5 Comparison of Multipoles From Different Levels of Theory

Presented below are the results of the ΔG_{sub} predictions using HF, MP2 and B3LYP. In each case we present: The ΔG_{sub} predictions the DLS-10 molecules compared to experiment and solubility predictions for the DLS-10 molecules, made using the experimental hydration free energies. **Table 3.2** shows the DLS-10 predictions.

Molecule Refcode	ΔG_{sub}^{exp} (kJ/mol)	ΔG_{sub}^{HF} (kJ/mol)	ΔG_{sub}^{MP2} (kJ/mol)	ΔG_{sub}^{B3LYP} (kJ/mol)
BENZAC02	34.23[179,180]	48.91	35.08	34.59
BZAMID02	43.14[29,179]	48.76	36.41	36.98
COCAIN10	54.90[179,180]	61.07	56.23	56.55
COYRUD11	61.06[181]	77.84	65.62	64.84
HXACAN04	59.95[181]	69.13	54.91	53.93
IBPRAC01	42.06[180,181]	65.80	54.89	54.60
JODTUR01	59.45[29,179]	68.22	60.15	59.75
NAPHOL01	35.38[179]	39.73	35.82	33.23
PYRENE07	46.25[29,179]	43.81	41.88	41.77
SALIAC	40.31[179]	42.41	34.04	33.40
R^2		0.67	0.76	0.76
RMSE		11.64	5.63	5.66
σ	10.34	7.00	5.63	5.62
bias		-9.30	0.17	0.71

Table 3.2: ΔG_{sub} predictions using HF, MP2 and B3LYP multipoles with the FIT repulsion and dispersion potential.

We can see from **Table 3.2** that the MP2 and B3LYP models produce very similar results, often differing by less than 1 kJ/mol. This is reflected in the statistics, which show a notable difference only in the bias and a small difference in the RMSE as a result. When applying HF multipoles however, a much greater discrepancy is found compared to the other methods. When compared to experiment, the results clearly show a significantly better correlation between $\Delta G_{sub}(exp)$ and either $\Delta G_{sub}(MP2)$ or $\Delta G_{sub}(B3LYP)$. We can see from the statistics that over the ten molecule dataset either of the models produced with MP2 or B3LYP multipoles provide good predictions, explaining much of the variance in the data. Taking the bias as a measure of the systematic error of the models, we can see that whilst there is an increase in the bias for the use of B3LYP multipoles, in both cases this systematic error remains below 1 kJ/mol over the dataset. For the model applying HF multipoles we see a marginally worse correlation and notable increase in the other statistics showing it to be a poorer model. In order for a models prediction to be useful the RMSE should lye within the standard deviation of the experimental data, otherwise a prediction of the mean of the experimental data will produce a prediction closer to the actual value. Here, the HF model fails this test suggesting it is not making a useful prediction of sublimation. Below **Figures 3.4, 3.5** and **3.6** show plots of the three methods against the experimental sublimation values.

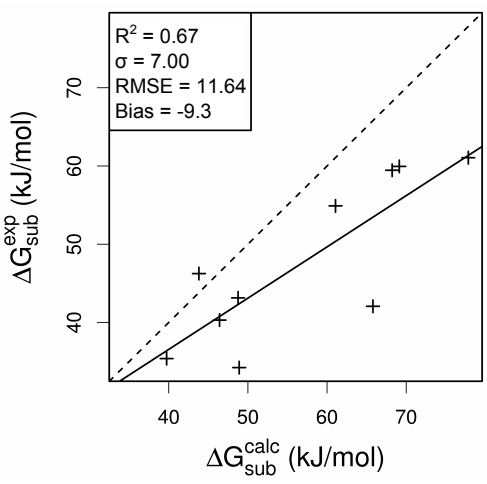


Figure 3.4: Predicted ΔG_{sub} using HF multipoles against experimental ΔG_{sub} .[31]

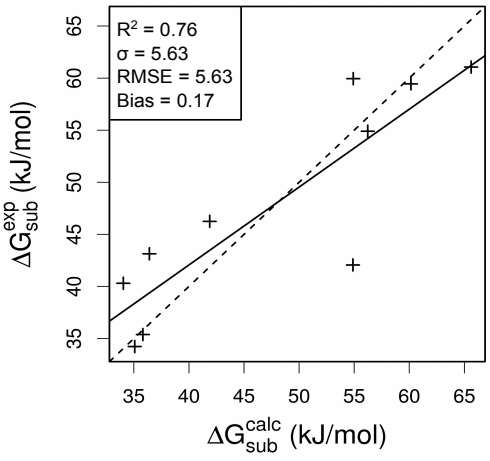


Figure 3.5: Predicted ΔG_{sub} using MP2 multipoles against experimental ΔG_{sub} .[31]

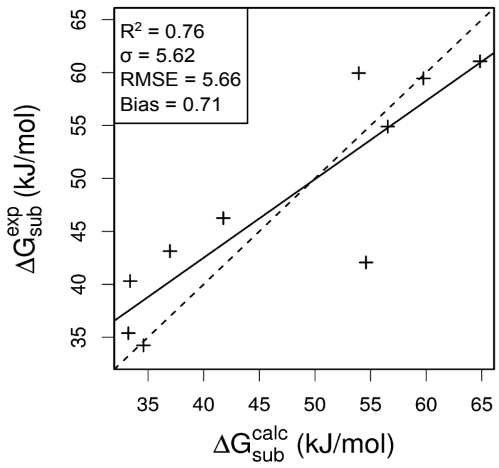


Figure 3.6: Predicted ΔG_{sub} using MP2 multipoles against experimental ΔG_{sub} .[31]

In the following figures we show predictions of solubility using the three sublimation predictions and experimental ΔG_{hyd}. We see in the following results that, as may be suspected from the previous discussion, the use of MP2 and B3LYP multipoles provides a much better agreement with experiment in terms of solubility than the use of HF multipoles. Using either MP2 or B3LYP multipoles provides a very strong correlation coefficient and a low RMSE ($R^2 = 0.81$ and $RMSE = 0.99$ $logS$ $units$). Alternatively, using the HF multipoles we see a much poorer correlation coefficient and RMSE score ($R^2 = 0.61$ and $RMSE = 2.04$ $logS$ $units$).

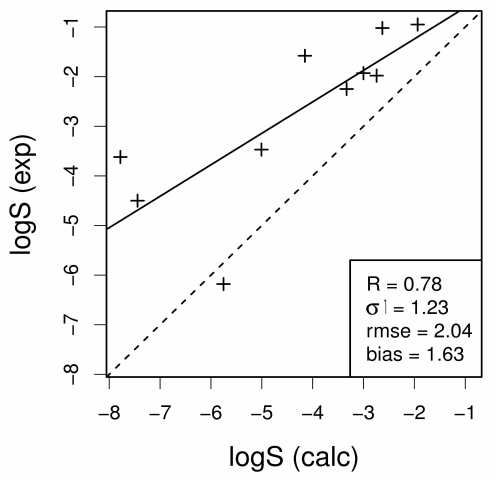


Figure 3.7: Predicted solubility using HF multipoles and experimental hydration free energy against experimental solubility.[31]

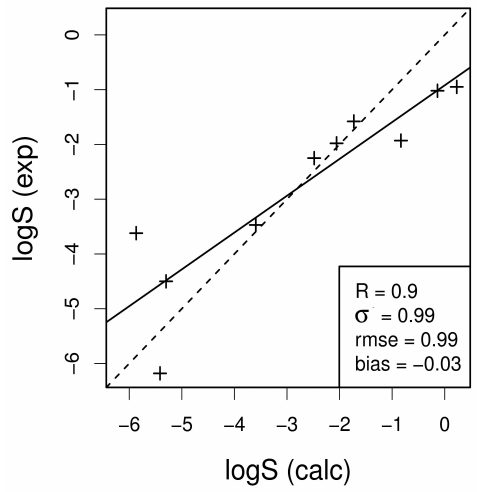


Figure 3.8: Predicted solubility using MP2 multipoles and experimental hydration free energy against experimental solubility.[31]

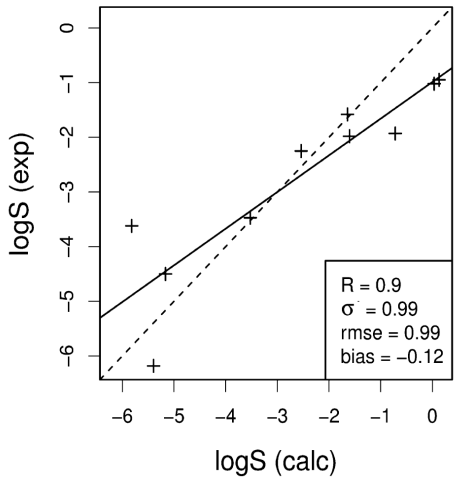


Figure 3.9: Predicted solubility using B3LYP multipoles and experimental hydration free energy against experimental solubility.[31]

Concluding from these results, using either MP2 or B3LYP multipoles can provide good agreement with experiment. In both cases these methods are designed to capture at least some aspect of electron correlation which HF does not. As a result the electrostatic description which is provided should be superior. MP2 is the recommend level of theory, however in view of these results B3LYP offers a comparable but computationally cheaper alternative for molecules such as those of the DLS-10 dataset. We therefore chose to continue this work applying the B3LYP multipoles.

3.2 Hydration Free Energy Predictions

In this section we focus on predicting ΔG_{hyd} by a variety of methods, we define the hydration free energy to be that associated with the direct transfer of a gaseous molecule to the solution. This process involves several changes to the system which are of interest. These can be collected into two physical changes, firstly the generation of a cavity in the solvent, and secondly the solvation of a molecule within the cavity.

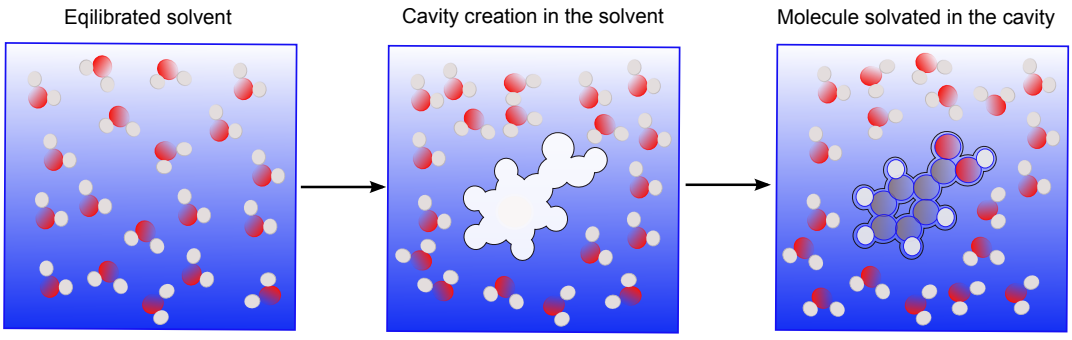


Figure 3.10: A diagram pictorially representing the hydration

There are numerous models in the literature for hydration free energy predictions.[51,52,69,182] For example, the continuum solvent model **(Section 2.2.12.2)** *SMD* is parametrised against hydration free energy data and is the recommended method to calculate such a quantity in Gaussian 09.[183] This kind of parametrisation is not unique to SMD; other models have been parametrised against similar data.[52] Additionally, there are an increasing number of QSAR methodologies to predict such quantities.[52,184]

The calculated ΔG_{hyd} must account for the factors of; cavitation, solvent ordering and solvation of the solute in the cavity. As a result information is required on both the isolated systems (solute and solvent) and the solution. A simple approximation to ΔG_{hyd} is to calculate ΔG_{hyd} using the gas solution partition coefficient equivalent to the Henry's law constant(**Equation 1.4**), hence, allowing a direct computation as follows:[24]

$$\Delta G_{hyd} = -RTlnK_H \tag{3.5}$$

Equation 3.5: A prediction of the hydration free energy from K_H. R is the gas constant and T is the temperature in Kelvin.

In the rest of this chapter, the focus will be on the prediction of ΔG_{hyd} from chemical theory. Methodologically specific details will be introduced and discussed.

3.2.1 Methods and Dataset Selection

The general theory of the methods used in this section can be found in **Section 2.2.12.2** and **Section 2.2.12.3**. Some additional specific details and options that were used in our production calculations are elaborated upon here and the *drug-like-solubility-25* (*DLS-25*) dataset is applied. As with the sublimation free energy predictions, we will here focus on the hydration free energy predictions for the DLS-10 molecules, for which experimental data was available. These ten molecules are shown in **Figure 3.3**. The experimental ΔG_{hyd} data is presented in **Table 3.3**:

Molecule Name	CSD refcode	$\Delta G_{hyd}^{exp}(kJ/mol)$
Benzoic Acid	BENZAC02	-33.14 [179]
Benzamide	BZAMID02	-45.64 [179]
Cocaine	COCAIN10	-49.98 [179]
Naproxen	COYRUD11	-43.30 [181]
Acetaminophen	HXACAN04	-62.05 [181]
Ibuprofen	IBPRAC01	-29.33 [181]
Isoproturon	JODTUR01	-47.58 [179]
1-Naphthol	NAPHOL01	-32.01 [179]
Pyrene	PYRENE07	-18.91 [179]
Salicyclic Acid	SALIAC	-37.21 [179]

Table 3.3: The ten ΔG_{hyd} from literature of the DLS-25 dataset.

We applied several methods when predicting ΔG_{hyd}. These methods are: IEFPCM, SMD and 3DRISM-KH/UC (3D RISM with a Kovalenko-Hirata closure and universal correction). The ΔG_{hyd} values from continuum models were all calculated as the difference between the gaseous isolated system energy and the energy in the continuum field. This is an oversimplification of the actual phenomena accounted for, which includes changes in electrostatic interactions, cavitation, repulsion and dispersion. Helpfully however, the calculation process automatically accounts for this in the software.

$$\Delta G_{hyd} = E_{solution} - E_{gaseous} \tag{3.6}$$

Equation 3.6: Calculating ΔG_{hyd} from continuum models. This is an oversimplification of the actual phenomena accounted for, but is how the quantity is calculated in this work thanks to the software automatically taking these additional factors into account

Optimised gas phase structures were calculated from crystal structures at the same level of theory as was to be applied to the solvated phase. In calculating the solvated phase there were two options.

- Calculate the ΔG_{hyd} by taking the energy difference between an optimised gas phase molecule and same structure in an approximate aqueous environment.
- Calculate the ΔG_{hyd} by taking the energy difference between an optimised gas phase molecule and the re-optimised structure in an approximate aqueous environment.

Although the former may at first seem a crude approximation, it has been shown a number of times to be a reasonable approximation for small molecules,[169,185] this is not always the case for larger multi-functional compounds.[186] We carried out some tests to check how large the difference was for our dataset, if this approximation was applied. The SMD method was applied at the HF and M06-2X levels of theory with a 6-31G* basis set. A summary of these results can be seen in **Figure 3.11**

below. The results for SMD(M062X) are shown by triangular data points, results for SMD(HF) are represented by diamond data points.

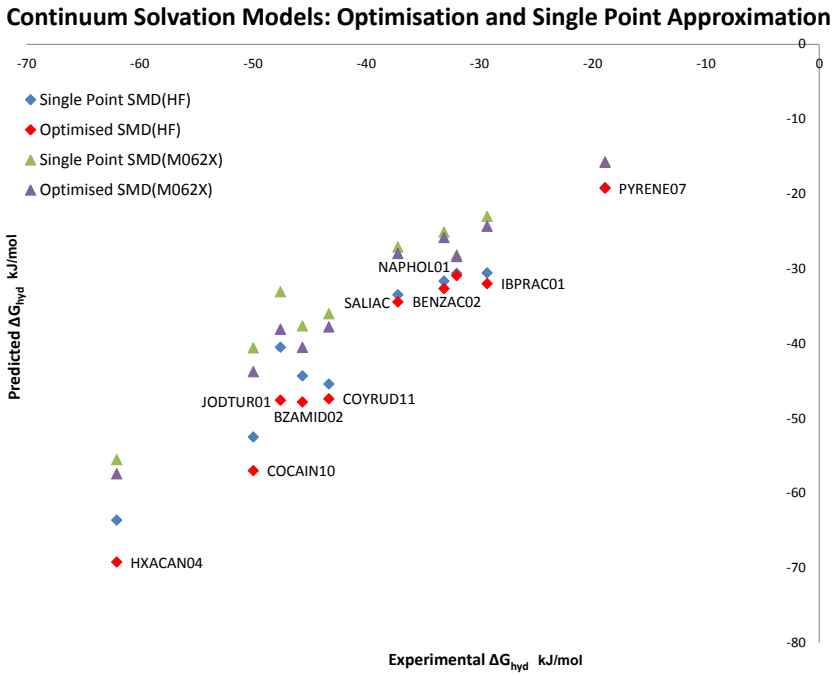


Figure 3.11: Comparison of optimisation and single point hydration free energy calculations.

The biggest energy difference was approximately 7 kJ/mol with the majority being less than 4 kJ/mol. Generally speaking, those molecules which are poly-functional and possess a functionality including nitrogen have the greatest reduction in energy upon optimisation. As the results show only a modest change in energy upon optimisation, and considering the large number of calculations needed to test the variety of methods we were applying in this work, we chose to use the single point approximation to reduce the computational time required. As a result of this, we approximate ΔG_{hyd} to be the energy of step 1 in the following diagram and do not consider step 2 any further.

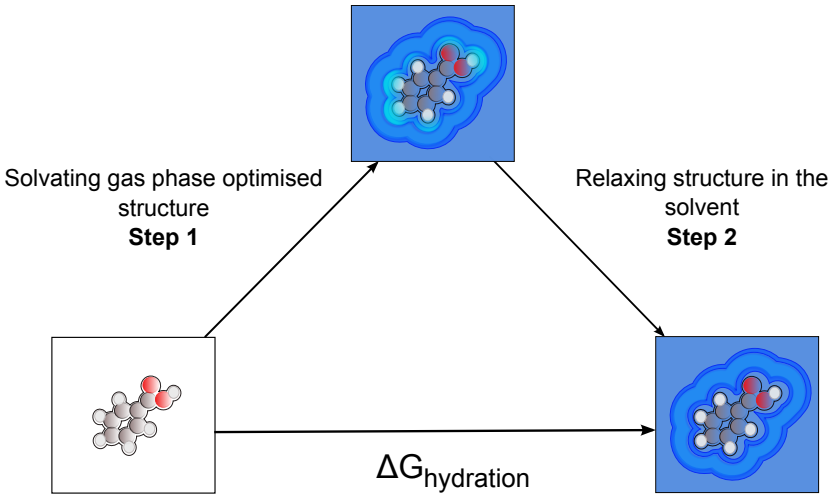


Figure 3.12: Illustration of the steps involved in predicting ΔG_{hyd} .

3.2.1.1 IEFPCM Production Calculation Details

IEFPCM was used within Gaussian 03. We applied the *united atom topological model* radii, optimised for use at the HF/6-31G* level of theory (abbreviated to *UAHF*), to the IEFPCM model. These radii represent H implicitly by including it in the sphere of the heavy atoms it is covalently bonded to. These radii were selected as they are the default radii in Gaussian 03 for making such a calculation using the IEFPCM procedure. We applied this model at the HF/6-31G* level of theory.

3.2.1.2 SMD Production Calculation Details

The SMD model was used again in the Gaussian 09 package. This method, as discussed previously (**Section 2.2.12.2**) solves the same electrostatic equations as IEFPCM but uses its own optimised radii to create a cavity. The SMD model contains several fitted parameters in addition to the radii, which are used to define a functional which calculates the contributions of cavity creation, dispersion and solvent structure to the free energy of hydration. These parameters were optimised on a dataset of $\sim$ 2000 molecules in various solvents and composed of the elements H, C, N, O, F, Si, P, S, Br and Cl.[69] The model is optimised for several levels of electronic structure theory: M05-2X/MIDI!6D, M05-2X/6-31G*, M05-2X/6-31+G**, M05-2X/cc-pVTZ, B3LYP/6-31G*, and HF/6-31G*. We applied the model at two levels of theory; HF/6-31G* and M062X/6-31G* levels. As with IEFPCM, the radii have been optimised for the HF/6-31G* level of theory. We chose the M06-2X functional as it has been shown to perform well for molecules composed of main group elements. Additionally, we were curious how the SMD solvent model would perform for levels for which it was not optimised.

3.2.1.3 3DRISM-KH/UC Production Calculation Details

3DRISM-KH/UC was applied as outlined previously (**Section 2.2.12.3**). The ΔG_{hyd} was calculated by applying the *Gaussian fluctuation* (*GF*) hydration free

energy functional. This functional was originally developed for use with another RISM approximation[187] known as 1D RISM, which is a further approximation to the OZ equation in terms of 1D integrals, but was adapted later for use with the 3DRISM[188] approximation. Unfortunately, the GF functional only provides qualitative results and hence correction schemes have been devised.[31,51] In this work the *Universal Correction* (*UC*) is called upon. Work by Palmer *et al* has shown a strong correlation between the error in the prediction from the GF functional and the partial molar volume calculated by 3D RISM.[31,51] A 2 parameter linear correction was devised by regression now known as the universal correction.

$$\Delta G_{hyd}^{3DRISM-KH/UC} = \Delta G_{hyd}^{GF} + a(\rho V) + b \tag{3.7}$$

Equation 3.7: The universal correction. ΔG_{hyd}^{GF} *is the raw Gaussian fluctuation free energy.* ρV *is the dimensionless partial molar volume. The scalar a is a bias correction of the value -3.312 kcal/mol. The intercept b is valued as 1.152 kcal/mol. Both a and b were calculated from a linear regression of a training set of molecules.*

This correction was applied in all of the 3DRISM calculation presented here. Lennard-Jones parameters were taken from the Amber GAFF forcefield and charges calculated using the semi-empirical Hamiltonian AM1. All RISM calculations were performed using Ambertools with a minimum buffer distance of 30Å.

3.2.2 Calculated Predictions of Hydration Free Energy

The predicted ΔG_{hyd} values of the molecules in the DLS-25 dataset are presented in the rest of this chapter. All four of the analysis statistics described in **Section 2.3** are used in this chapter. The free energies of hydration will be referred to as follows: Those calculated using the 3DRISM-KH/UC method $\Delta G_{hyd}^{3DRISM-UC}$, those using the SMD(M062X) method, $\Delta G_{hyd}^{SMD(M062X)}$, predictions from the SMD(HF) method, $\Delta G_{hyd}^{SMD(HF)}$ and finally values calculated by IEFPCM, ΔG_{hyd}^{PCM}. The predicted hydration free energy values for all 25 molecules are displayed in **Table 3.4**.

Molecule	ΔG_{hyd}^{exp}	$\rho V_{3D-RISM}$	$\Delta G_{hyd}^{3DRISM-UC}$	$\Delta G_{hyd}^{SMD(M06-2X)}$	$\Delta G_{hyd}^{SMD(HF)}$	ΔG_{hyd}^{PCM}
ALOPUR		4.36	-73.88	-63.96	-76.63	-68.32
AMBNAC04		5.21	-56.29	-46.01	-51.70	-44.18
AMXBPM10		11.21	-92.41	-69.63	-79.84	-53.47
BENZAC02	-33.14	4.80	-36.67	-25.10	-31.66	-27.95
BZAMID02	-45.64	5.15	-49.86	-37.66	-44.33	-35.73
COCAIN10	-49.98	11.83	-59.30	-40.59	-52.52	-27.87
COYRUD11	-43.30	9.23	-44.44	-36.04	-45.42	-36.61
DHANQU06		8.18	-39.47	-26.68	-38.26	-21.42
EPHPMO		8.40	-70.40	-63.67	-73.65	-64.39
ESTRON14		11.66	-42.51	-42.78	-52.89	-47.40
HXACAN04	-62.05	5.97	-60.31	-55.53	-63.63	-53.35
IBPRAC01	-29.33	9.83	-31.74	-23.02	-30.56	-18.24
IVUQOF		10.63	-70.75	-78.77	-93.96	-52.13
JODTUR01	-47.58	9.07	-48.01	-33.09	-40.51	-28.79
LABJON01		7.55	-92.77	-75.46	-100.63	-96.48
NAPHOL01	-32.01	5.81	-24.70	-28.15	-30.64	-28.24
NDNHCL01		12.21	-49.05	-53.04	-55.02	-36.99
NICOAC02		4.65	-43.74	-35.98	-44.28	-38.16
NIFLUM10		9.60	-65.73	-25.72	-34.33	-22.43
PINDOL		10.31	-67.78	-52.92	-57.67	-45.73
PTERID11		4.74	-52.06	-55.90	-64.98	-37.49
PYRENE07	-18.91	8.01	-26.18	-15.72	-19.15	-11.30
SALIAC	-37.21	5.02	-36.13	-27.12	-33.49	-29.62
SIKLIH01		10.31	-41.77	-33.80	-42.77	-19.79
XYANAC		9.76	-35.80	-23.63	-26.65	-16.65

Table 3.4: Hydration free energy prediction for the full DLS-25 molecule set. All free energies are quoted in kJ/mol.

The predicted ΔG_{hyd} results for the ten molecules for which we have experimental data are presented in **Table 3.5** and **Figure 3.13**.

Molecule Refcode	ΔG_{hyd}^{exp}	$\Delta G_{hyd}^{3DRISM-UC}$	$\Delta G_{hyd}^{SMD(M062X)}$	$\Delta G_{hyd}^{SMD(HF)}$	ΔG_{hyd}^{PCM}
BENZAC02	-33.14	-36.67	-25.10	-31.66	-27.95
BZAMID02	-45.64	-49.86	-37.66	-44.33	-35.73
COCAIN10	-49.98	-59.30	-40.59	-52.52	-27.87
COYRUD11	-43.30	-44.44	-36.04	-45.42	-36.61
HXACAN04	-62.05	-60.31	-55.53	-63.63	-53.35
IBPRAC01	-29.33	-31.74	-23.02	-30.56	-18.24
JODTUR01	-47.58	-48.01	-33.09	-40.51	-28.79
NAPHOL01	-32.01	-24.70	-28.15	-30.64	-28.24
PYRENE07	-18.91	-26.18	-15.72	-19.15	-11.30
SALIAC	-37.21	-36.13	-27.12	-33.49	-29.62
R		0.93	0.97	0.97	0.88
RMSD		4.85	8.3	2.91	11.58
σ	12.31	4.49	3.06	2.81	5.58
Bias		1.82	-7.71	-0.72	-10.15

Table 3.5: Hydration free energy prediction for the DLS-10 molecule set. All free energies are quoted in kJ/mol.

It is immediately clear that one method outstrips the others in terms of predictive accuracy, SMD(HF). As was mentioned previously, this method is parametrised for use at this level of theory and for making predictions of ΔG_{hyd}, hence, it is not so surprising that the result is good. We can define a criterion for a useful prediction to be one in which *the RMSD of the prediction is within the experimental standard deviation*. If this is not the case, a null prediction of the mean of the experimental data would be a more accurate prediction. In terms of the other three methods we can rank them using this new criterion. 3DRISM-KH/UC is the only other method that produces a low RMSD, bias and high R^2 value. We therefore can interpret this as a useful prediction. This is an interesting result as 3DRISM-KH/UC contains no explicit parametrisation against hydration free energy data. The remaining two methods, SMD(M06-2X) and IEFPCM, perform fairly poorly, with the next nearest RMSD (SMD(M06-2X)) being nearly twice that of 3DRISM-KH/UC. Noting that both of these two methods have a large bias, in fact a bias considerably larger than the corresponding σ, it is a reasonable assertion that the methods contain large systematic errors.

Another interesting point to note is the directionality of the bias. All continuum models are biased in the direction of predicting molecules to be more hydrophilic than experiments would suggest whereas 3DRISM's is biased to predict molecules to be more hydrophobic. This suggests some underlying bias to the continuum models

methods, given the similarities between them. It has been shown previously that RISM methods tend to overestimate the hydrophobicity of molecules,[51] stemming from an overestimation of the energetic cost of cavity creation. This overestimation of the energetic cost of cavity creation is the basis of the UC correction.[51,189] In all cases the methods produce a good correlation coefficient, however given the small size of this dataset this could be a false positive.

It is unreasonable to draw large or sweeping conclusions from such a small dataset. A collection of ten molecules is certainly too small of a dataset to conclusively show the superiority of one method in making these predictions. However, for a small set of data such as this one may expect it to be reasonably easy for these methods to achieve favourable statistical values. Indeed, this is the case for the SMD(HF) method, which does perform better than has been previously reported by other authors.[69,190,191] If a method is performing badly on such a small dataset, it is likely to be amplified when the method is applied to a larger dataset. Presented below is a graphical representation of the data from these methods.

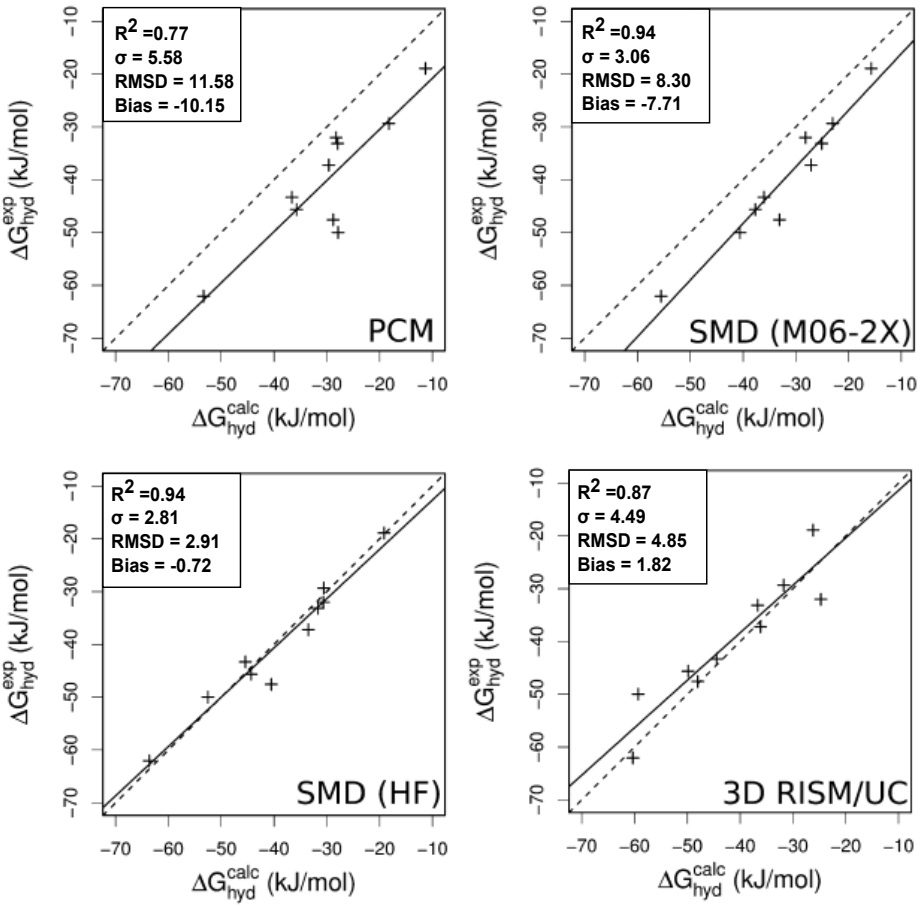


Figure 3.13: Plot of the DLS-10 hydration free energy predictions.[31]

Based on these results, we suggest that continuum models such as SMD can provide a good prediction of the ΔG_{hyd} provided one uses the methods for which the model is optimised. RISM provides a good alternative to such continuum models. The continued development of the RISM methods, including parametrisation, may offer a superior method in the future. For the time being, applications of correction terms

such as the UC allow quantitative values to be computed using 3DRISM.

In summary, a lack of experimental data for drug-like organic molecules hampers continued progress in predicting ΔG_{hyd} from theory. However, utilising the available data (often for small simple or mono-functional organic molecules) several methods have been produced and validated, which are available in the open literature and well known computational chemistry programs. Here we have performed a small scale test of the abilities of some of these methods. We can conclude from our own results and those of others that predictions of ΔG_{hyd} energies are improving. A blind test in 2009, to predict the ΔG_{hyd} of biologically active molecules, found the best predictions had an RMSD ranging from 2.4 - 3.5 kcal/mol (10.46 - 14.64 kJ/mol).[192] This test was based on 63 drug-like molecules. Recent work by others has shown improvement for predicting ΔG_{hyd} of organic molecules.[182] Our results suggest it is now possible to improve upon these results for drug-like molecules.

3.3 First Principles Prediction of Solubility

A first principles prediction of solubility is widely sought as it would enable a universal method to be applied to all systems, and provided such a method was not overly computationally expensive, it would represent a cost saving in the drug development pipeline. This is an industry which currently is suffering from falling numbers of drug candidates making it through clinical trials; current estimates suggest only 5% of compounds in phase I clinical trials will make it to pharmacy shelves.[193] Approximately 40% of lead compounds are estimated to be essentially insoluble in water and thus cannot be developed as a result, solubility is one of the major causes of such attrition.[42] This is whilst investment in drug development continues to increase, hence a universal pre-screening methodology is commercially attractive.[193] The current state of the art predictions come from QSPR methods. These are powerful correlative relationships which enable predictions to be made efficiently for some subspace of the full chemical space. This subspace is dependent upon the training data available. A first principles method would remove this dependence on training data. For clarity we define a first principles method to be one not directly parametrised upon experimental solubility data. The approach we have followed in this chapter relies upon the following thermodynamic cycle (**Figure 3.14**).

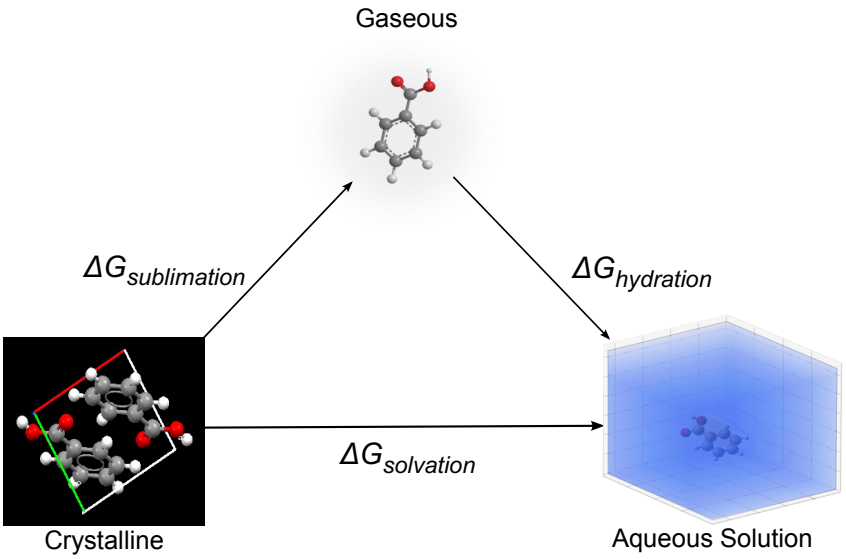


Figure 3.14: Thermodynamic cycle via the vapour

Having made predictions of ΔG_{sub} and ΔG_{hyd} we now wish to predict ΔG_{solv} which can readily be converted in to a prediction of intrinsic solubility(S_0), typically this is quoted as the base 10 logarithm of the intrinsic solubility ($logS$). ΔG_{solv} is predicted as the sum of ΔG_{sub} and ΔG_{hyd}. This is then converted to logS by **Equation 3.8**.

$$\Delta G_{solv} = \Delta G_{sub} + \Delta G_{hyd} \tag{3.8}$$

$$Log_{10}S = log_{10}\left(\frac{P_0}{RT}exp\left(\frac{\Delta G^o_{sub} + \Delta G^*_{hyd}}{-RT}\right)\right)$$

Equation 3.8: Top: The free energy of sublimation plus the free energy of hydration gives a prediction of the free energy of solvation. Bottom: P_0 is the atmospheric pressure, R is the gas constant ($8.314\ Jmol^{-1}K^{-1}$), T is the temperature in Kelvin. ΔG^o_{sub} is the sublimation in free energy in the 1 atmosphere standard state. ΔG^*_{hyd} is the hydration free energy in the 1 $molL^{-1}$ standard state.

3.3.1 A First Principles Prediction of Solubility: Results

This work follows directly from the previous two sections on predicting ΔG_{sub} and ΔG_{hyd}. For the rest of this chapter we will focus of solubility predictions for the full DLS-25 dataset. In total 12 predictions were made utilising variations in methodology (three ΔG_{sub} predictions and four ΔG_{hyd} predictions).

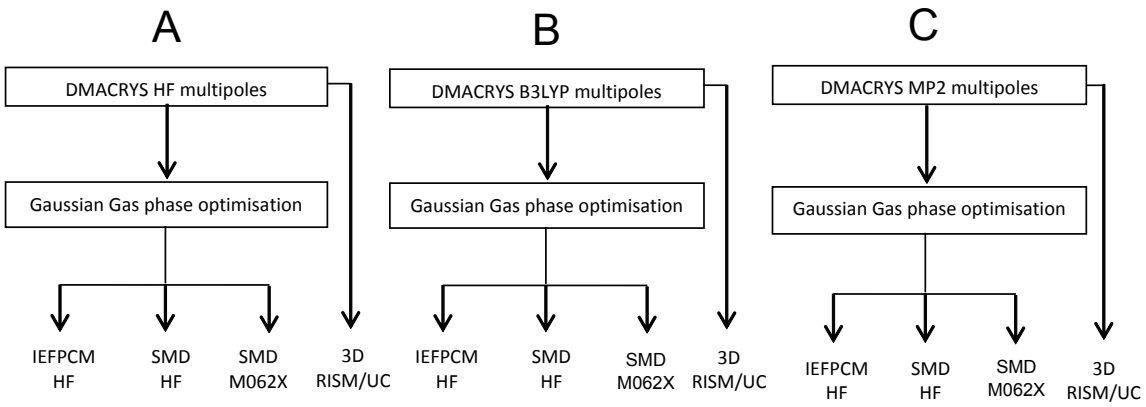


Figure 3.15: Solubility predictions over 12 variations in methodology.

As the intrinsic aqueous solubility can be defined in terms of ΔG_{sub} and ΔG_{hyd}, one could assume that the accuracy of the predicted solubilities would follow implicitly the accuracy of the individual predictions of ΔG_{sub} and ΔG_{hyd}. However, caution must be applied to such conjectures. We must be aware of additional limiting factors in the predictions. Experimental errors in the measurements of ΔG_{sub} and ΔG_{hyd} are substantially greater than those of intrinsic aqueous solubility. This is due to larger errors in the measurement of partial vapour pressures of drug-like molecules, compared to the errors in measurement of concentrations of the same molecules in saturated solutions. This often means that ΔG_{sub} and ΔG_{hyd} data are only available for low molecular weight, non drug-like, molecules as these can be more accurately measured at room temperature. Finally, the ten molecules described previously in **Sections 3.1** are a subset of the full DLS-25 dataset and could be a biased sample, i.e. composed of those molecules in the DLS-25 dataset for which ΔG_{sub} and ΔG_{hyd} can be measured most easily. They could therefore be closer to the training datasets used for the parameterisation of some of the continuum model's, hence giving such a method an unfair advantage. In general here the data follows the trends previously set in the predictions of ΔG_{sub} and ΔG_{hyd}. For example, the predictions of ΔG_{sub} by HF were far worse than those by B3LYP and MP2. This trend holds with all solubility predictions which follow the left hand side flow chart A, with solubility predictions originating from flow chart A being noticeably poorer than for flow chart B or C.

Previously, several parameters have been suggested by Hopfinger *et al* as criteria for the quality of a solubility prediction.[194]

- *Accurate predictions are defined as having an absolute error of* ≤ 0.5 $Log_{10}S$ *absolute error.*
- *Reasonable predictions are defined as having an absolute error of* ≤ 1 $Log_{10}S$ *absolute error*
- *Outliers are defined as molecules with a calculated solubility that is more than two standard deviations outside the experimental data.*

We will use these criteria as a further test of our data. For the rest of this section the focus will be upon the middle flow chart B's predictions, as these were the most accurate. A summary of the results for all methods is provided below.

ΔG_{hyd}^{method}	ΔG_{sub}^{method}	R	RMSE	σ	Bias	|Error|≤0.5	|Error|≤1.0	Outliers
3DRISM/UC	MP2	0.81	1.58	1.58	-0.05	5(20%)	10(40%)	0
	B3LYP	0.85	1.45	1.43	-0.23	5(20%)	12(48%)	0
	HF	0.75	2.51	1.64	1.90	1 (4%)	2(8%)	5
SMD(HF)	MP2	0.81	2.14	2.13	0.14	8(32%)	12(48%)	4
	B3LYP	0.84	2.03	2.03	-0.05	8(32%)	12(28%)	2
	HF	0.75	3.02	2.19	2.08	2(8%)	5(20%)	6
SMD(M06-2X)	MP2	0.84	2.49	1.87	1.65	6(24%)	8(32%)	3
	B3LYP	0.86	2.33	1.82	1.46	6(24%)	10(40%)	2
	HF	0.74	4.20	2.17	3.59	1(4%)	1(4%)	13
PCM(HF)	MP2	0.71	3.57	2.65	2.40	3(12%)	9(36%)	9
	B3LYP	0.74	3.37	2.54	2.21	5(20%)	11(44%)	9
	HF	0.65	5.11	2.69	4.35	0(0%)	2(8%)	13

Table 3.6: Summary of solubility predictions for the DLS-25, using the dataset by 12 methods applied in this work. The final three columns show the quality of our predictions analysed by Hopfinger et al's criteria error ≤ 0.5 logS, error ≤ 1 logS and outliers two σ outside the experimental data respectively.

Below, we present the solubility prediction produced from B3LYP/6-31G* multipoles and all four solvation models. Methods are labelled by the solvation model used as all other aspects are identical. We apply the analysis statistics from **Section 2.3** in addition to Hopfinger *et al*'s criteria, providing a full assessment of the quality of our predictions.

For the complete DLS-25 dataset, the most accurate solubility predictions come from a combination of: ΔG_{hyd} predicted by 3DRISM-UC and ΔG_{sub} calculated using B3LYP/6-31G* multipoles (**Table 3.7** and **Figure 3.16**). This combination results in an $R^2 = 0.72$ and an $RMSE = 1.45$ logS (units referred to mol/L). Hence, this method is the only one with an RMSE within the standard deviation of the experimental data (1.79 logS units). Earlier, we stated that a prediction must have an RMSE within the standard deviation of the experimental data, otherwise the null prediction of the mean would be a more accurate prediction. The 3DRISM method achieved a $\sigma = 1.43$ logS and $Bias = -0.23$ logS (units referred to mol/L). The low bias suggests much of the error is attributable to a random error, not a systematic deviation by the model. Five molecules out of the 25 were predicted with absolute errors <0.5 logS units, whereas 12 more were predicted to within an absolute error of <1 logS unit. The method made no outlying predictions.

A surprising turn of events was that after its promising prediction of ΔG_{hyd}, the method using SMD(HF) did not supply the most accurate solubility prediction. The accuracy of the solubility predictions afforded by ΔG_{hyd} SMD(HF) with ΔG_{sub} B3LYP/6-31G* multipoles was $R^2 = 0.71$ and $RMSE = 2.03$ logS (referred to units of mol/L). By comparison to 3DRISM-UC this is a fairly poor performance. This performance is in part the result of two outliers: NIFLUM10 ($\Delta logS = 4.58$) and PTERID11 ($\Delta logS = -5.09$). NIFLUM10 incorporates a tri-fluorinated methyl moiety, which is an unusual group and may contribute to the prediction errors;

IVUQOF is the only other molecule which contains fluorine, and is also poorly predicted ($\Delta logS = -1.82$ logS) however, falling short of being classified as an outlier. None of NIFLUM10, PTERID11 or IVUQOF appear in the SMD training set.

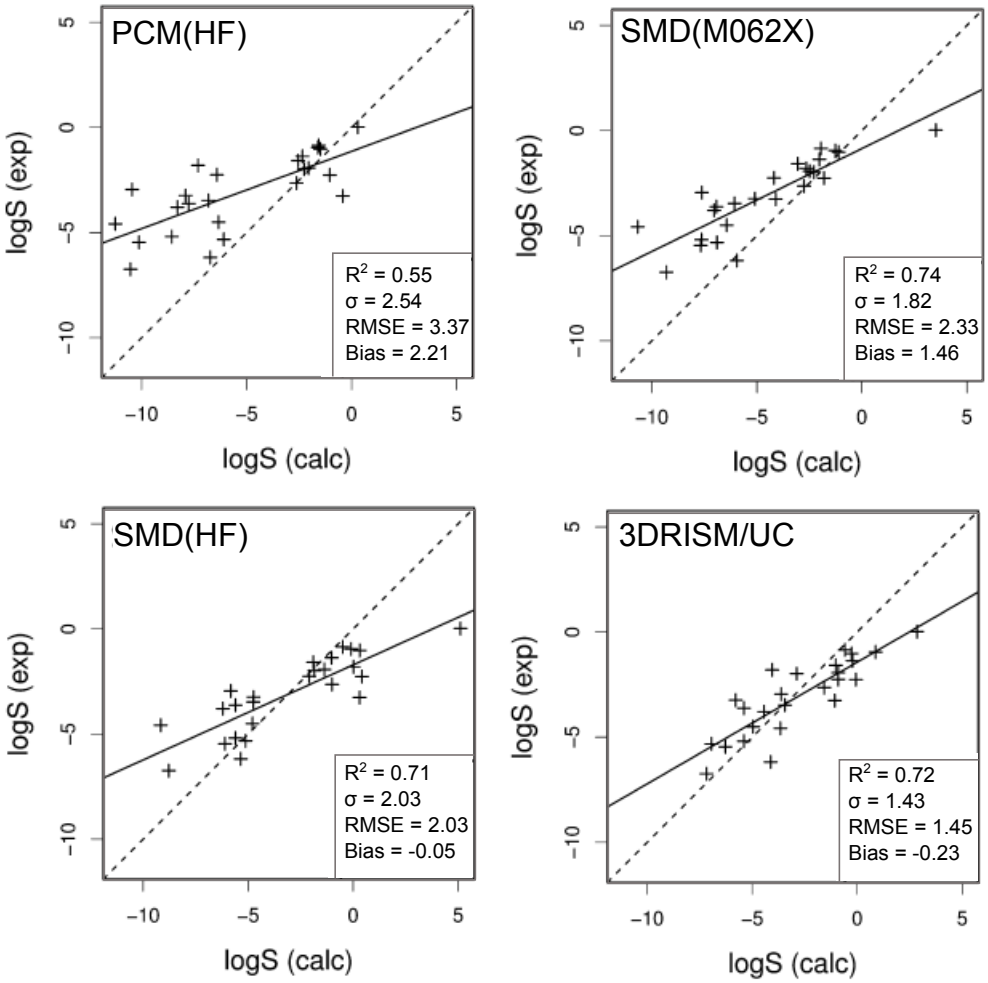


Figure 3.16: Solubility predictions using B3LYP multipoles and a choice of four solvation models PCM(HF), SMD(HF), SMD(M062X) and 3D-RISM.

The SMD training set for solvation free energies in water consist of 274 solutes. 18 of these solutes contain fluorine, 8 of which contain the CF_3 group present in NIFLUM10. One might expect from the occurrence of the CF_3 in the training data that the solubility of molecules with this group predicted using the SMD method would not be outliers. However, in all the solute molecules containing the CF_3 group in the SMD training set are simple organic molecules, often halogenated alkanes, which may not provide a sufficiently robust parametrisation for this group when it is bonded to more complex organic structures such as NIFLUM10. This again is potentially due to the difficulties of making experimental measurements of such systems and/or non-additive behaviour of the group contributions, hence such information is either not available or difficult to use when parameterising models.

The other two methods, using the PCM (HF/6-31G*) or SMD (M06-2X/6-31G*) solvation models, the results are very poor. Both models produce notable *bias* and

standard deviation values, culminating in large RMSE values (RMSE SMD(M06-2X) = 2.33 logS and RMSE PCM(HF) = 3.37 logS). This level of accuracy is not sufficient for a useful prediction of intrinsic aqueous solubility. Given that both of these methods also provided the poorest prediction of hydration free energy, it is perhaps not surprising that they also lead to the poorest predictions of solubility. The SMD(M06-2X) prediction, like its related method SMD(HF), has two outliers. The first is NIFLUM10 (ΔlogS = 6.09). The same molecule is an outlier in the SMD(HF) prediction, adding weight to the argument that the solubility of this molecule is difficult to predict using the SMD method, potentially due to an insufficient training set. The second is AMXBPM10 (ΔlogS = 4.67). Neither of these molecules (NIFLUM10 or AMXBPM10) appear in the SMD training set. The PCM(HF) method has nine outliers, meaning it has the most outliers of any of the four methods for which the results are plotted above (**Figure 3.16** and **Table 3.7**).

Molecule	*logS* Exp	*logS* 3DRISM/UC	Error	*logS* SMD(M06-2X)	Error	*logS* SMD(HF)	Error	*logS* PCM(HF)	Error
ALOPUR	-2.26 [180]	-0.06	-2.20	-1.80	-0.46	0.42	-2.68	-1.04	-1.22
AMBNAC04	-1.37 [29]	-0.22	-1.15	-2.02	0.65	-1.03	-0.34	-2.34	0.97
AMXBPM10	-2.95 [28]	-3.63	0.68	-7.62	4.67	-5.83	2.88	-10.45	7.50
BENZAC02	-1.58 [180]	-1.02	-0.56	-3.05	1.47	-1.90	0.32	-2.55	0.97
BZAMID02	-0.95 [29]	0.87	-1.82	-1.27	0.32	-0.10	-0.85	-1.61	0.66
COCAIN10	-2.25 [180]	-0.91	-1.34	-4.19	1.94	-2.10	-0.15	-6.42	4.17
COYRUD11	-4.50 [28]	-4.96	0.46	-6.44	1.94	-4.79	0.29	-6.34	1.84
DHANQU06	-5.19 [29]	-5.40	0.21	-7.64	2.45	-5.61	0.42	-8.56	3.37
EPHPMO	-2.64 [29]	-1.57	-1.07	-2.75	0.11	-1.00	-1.64	-2.63	-0.01
ESTRON14	-5.32 [195]	-6.94	1.62	-6.90	1.58	-5.13	-0.19	-6.09	0.77
HXACAN04	-1.02 [28]	-0.27	-0.75	-1.11	0.09	0.31	-1.33	-1.49	0.47
IBPRAC01	-3.62 [180]	-5.40	1.78	-6.92	3.30	-5.60	1.98	-7.76	4.14
IVUQOF	-1.80 [180]	-4.05	2.25	-2.65	0.85	0.02	-1.82	-7.32	5.52
JODTUR01	-3.47 [29]	-3.45	-0.02	-6.06	2.59	-4.76	1.29	-6.82	3.35
LABJON01	-3.26 [28]	-1.07	-2.19	-4.11	0.85	0.31	-3.57	-0.42	-2.84
NAPHOL01	-1.98 [28]	-2.88	0.90	-2.28	0.30	-1.84	-0.14	-2.26	0.28
NDNHCL01	-3.24 [194]	-5.79	2.55	-5.09	1.85	-4.75	1.51	-7.91	4.67
NICOAC02	-0.85 [29]	-0.59	-0.26	-1.95	1.10	-0.50	-0.35	-1.57	0.72
NIFLUM10	-4.58 [28]	-3.66	-0.92	-10.67	6.09	-9.16	4.58	-11.25	6.67
PINDOL	-3.79 [28]	-4.43	0.64	-7.04	3.25	-6.21	2.42	-8.30	4.51
PTERID11	0.02 [28]	2.85	-2.83	3.52	-3.50	5.11	-5.09	0.29	-0.27
PYRENE07	-6.18 [29]	-4.12	-2.06	-5.95	-0.23	-5.35	-0.83	-6.73	0.55
SALIAC	-1.93 [28]	-0.91	-1.02	-2.49	0.56	-1.37	-0.56	-2.05	0.12

Continued on next page

Table 3.7 – *Continued from previous page*

Molecule	*logS* Exp	*logS* 3DRISM/UC	Error	*logS* SMD(M062X)	Error	*logS* SMD(HF)	Error	*logS* PCM(HF)	Error
SIKLIH01	-5.46 [28]	-6.28	0.82	-7.67	2.21	-6.10	0.64	-10.13	4.67
XYANAC	-6.74 [28]	-7.17	0.43	-9.31	2.57	-8.78	2.04	-10.53	3.79
R^2		0.72		0.74		0.71		0.55	
σ	1.79	1.43		1.82		2.03		2.54	
Bias		-0.23		1.46		-0.05		2.21	
RMSE		1.45		2.33		2.03		3.37	

Table 3.7: All data relating to predictions of logS using the DMACRYS predictions of ΔG_{sub} and ΔG_{hyd} predictions using IEFPCM, SMD and 3DRISM.

3.4 Summary

The aim of this project has been to determine the accuracy of first principles predictions of intrinsic aqueous solubility of crystalline drug-like molecules utilising widely available computational methods and programs. As it is not currently possible to directly compute ΔG_{solv}, we decomposed ΔG_{solv} into ΔG_{sub} and ΔG_{hyd}, utilising a thermodynamic cycle via the vapour to calculate indirectly ΔG_{solv}. The best predictions of intrinsic aqueous solubility came from the 3D-RISM/UC method (RMSE = 1.45 logS units) in this work. Despite this accuracy being worse than many QSPR/QSAR models exemplified in the blind challenge,[28,194] we believe it does provide a proof of concept that can be expanded enabling more accurate first principles predictions of solubility to be made in the future. In addition, this method gives a full characterization of the thermodynamics involved in the transfer of crystalline solute to a gaseous vapour to an aqueous solution. The solubility of any crystalline solute is in part reliant on the properties of the crystalline precipitate within the solution and not solely on the solution properties. The thermodynamic information can aid in the understanding of why a selected molecule is more soluble than any other selected molecules. This can be of great interest in terms of crystal polymorphs. QSPR methods are unable to provide such data as descriptors focus on describing molecular structures.

Chapter 4

Cheminformatics in Solubility Prediction

"We hope that machines will eventually compete with men in all purely intellectual fields. But which are the best ones to start with?"

Alan Turing, 1950

4.1 Solubility Predictions from combined models

Cheminformatics methods are applied to a wide range of property prediction tasks especially in the pharmaceutical industry. Commonly, cheminformatics is applied to predict aqueous solubility, melting point, boiling point, logP, binding affinities, and toxicology.[196] Informatics methods are generally faster and more amenable to *High Throughput Virtual Screening* (*HTVS*) than methods aiming to solve real physical equations from theoretical chemistry. This has led to the generation of a number of tools as easy-to-use and accessible web-based packages.[197–199] However, whilst informatics methods can produce efficient and generally fairly accurate predictions in a reasonable region of chemical space, they lack the ability to decompose the results into intuitive, physically meaningful quantities,which allow a more intuitive understanding of the actual core physical process. In addition they are constrained to apply to molecules related to their training datasets. To acquire this missing understanding it is necessary to perform atomistic calculations based on real physical equations and/or derive further understanding from experimental data, such as in the use of matched molecular pairs.[40] Generally, neither of these approaches are as amenable to HTVS, although current efforts are now allowing HTVS to be become a reality using matched molecular pairs.[200] It is reasonably common for a combination of data sources to be used within industries to achieve a better understanding. Simulation methods and quantum chemical data are sources of rich deep understanding used to supplement informatics.

On the basis of this we wished to investigate the accuracy of informatics methods presented with information from different sources. We chose to focus on machine

learning within cheminformatics to produce a QSPR approach. In the rest of this section a discussion of the methods and results emanating from the above approach is provided. The results have been published in the following publication.[33]

4.2 A New Dataset: Drug-Like-Solubility-100

Following on from the predictions using the DLS-25 dataset we generated a larger dataset, named DLS-100. This dataset consists of 100 drug-like molecules, 25 of which are the DLS-25 molecules. We generated this data set with the aim to investigate a combined theoretical chemistry and cheminformatics method. The fact that cheminformatics requires in general larger sets of data to acquire statistically robust results meant that it was not appropriate to investigate such combined models using the DLS-25 dataset. The structures and data associated with this dataset are presented in **Appendix H**. The full dataset and all scripts for machine learning are downloadable from the ***Mitchell group website***, the use of which is exemplified in the paper by McDonagh *et al.*[33]

This dataset was built with similar ideas and criteria as those used in the generation of the DLS-25 dataset. The following are the applied criteria:

- *Molecules must be organic and drug-like.*
- *A well documented solubility must be reported in the literature, ideally measured by the CheqSol method.*[201]
- *A usable crystal structure must be available from the CSD.*

We identified 125 CheqSol solubilities from three publications; the solubility challenge,[28] Palmer *et al*[6] and Narasimham *et al.*[202] 41 had usable crystal structures available in the CSD. Where a choice of polymorphs with solubility data existed, we selected the most thermodynamically stable polymorph, hence the one with the lowest solubility. Where information on the solubility of specific polymorphs was not available (which was most often the case), we selected the polymorph giving the lowest lattice energy from our calculations.

4.3 Workflow and Descriptor Generation

Our ultimate goal was to investigate the effect of combining descriptors generated from different sources (i.e. simple counts, graph theory, thermodynamics and quantum chemistry). Our main focus was the relative complementarity, or lack of, between the different descriptors. In analysing these data we focus on the RMSE, aiming to minimise the overall predictive inaccuracies rather than focusing on individual cases.

4.3.1 Chemical Theory Workflow

The initial step involved generating a model based exclusively on theoretical chemistry in an analogous way to that outlined for the first-principles prediction of solubility. We applied DMACRYS to minimise the lattice energy by optimising the unit cell. A single molecule was removed from the unit cell and optimised in the gas phase. The optimised structure was then extracted and placed into a solvation calculation using the SMD solvation model. This was carried out for all 100 molecules. In our previous publication,[31] and presented above in **Table 3.7**, we showed that DMACRYS coupled with 3DRISM could provide a better prediction of solubility than DMACRYS coupled with SMD. We selected the SMD model here for a couple of reasons: Firstly, the SMD model was shown to make good predictions of ΔG_{hyd} in the first-principles prediction of solubility.[31] Despite DMACRYS-3DRISM providing a more accurate absolute solubility prediction than DMACRYS-SMD,[31] the SMD model provided a notably higher R^2 against experimental data for ΔG_{hyd} predictions. As the current model is parametrised in nature, the correlation rather than absolute value was considered to be of higher importance. The two methods produce a near equivalent correlation coefficient R^2 for solubility prediction. Secondly, the SMD model is simpler procedurally to use making it more easily applied by others wishing to follow our work. Finally, the SMD model is available in the commonly used quantum chemistry package G09, hence, anyone wishing to utilise our methodology further is more likely to have access to the SMD model than to the RISM methodology. The overall workflow for this procedure is provided here:

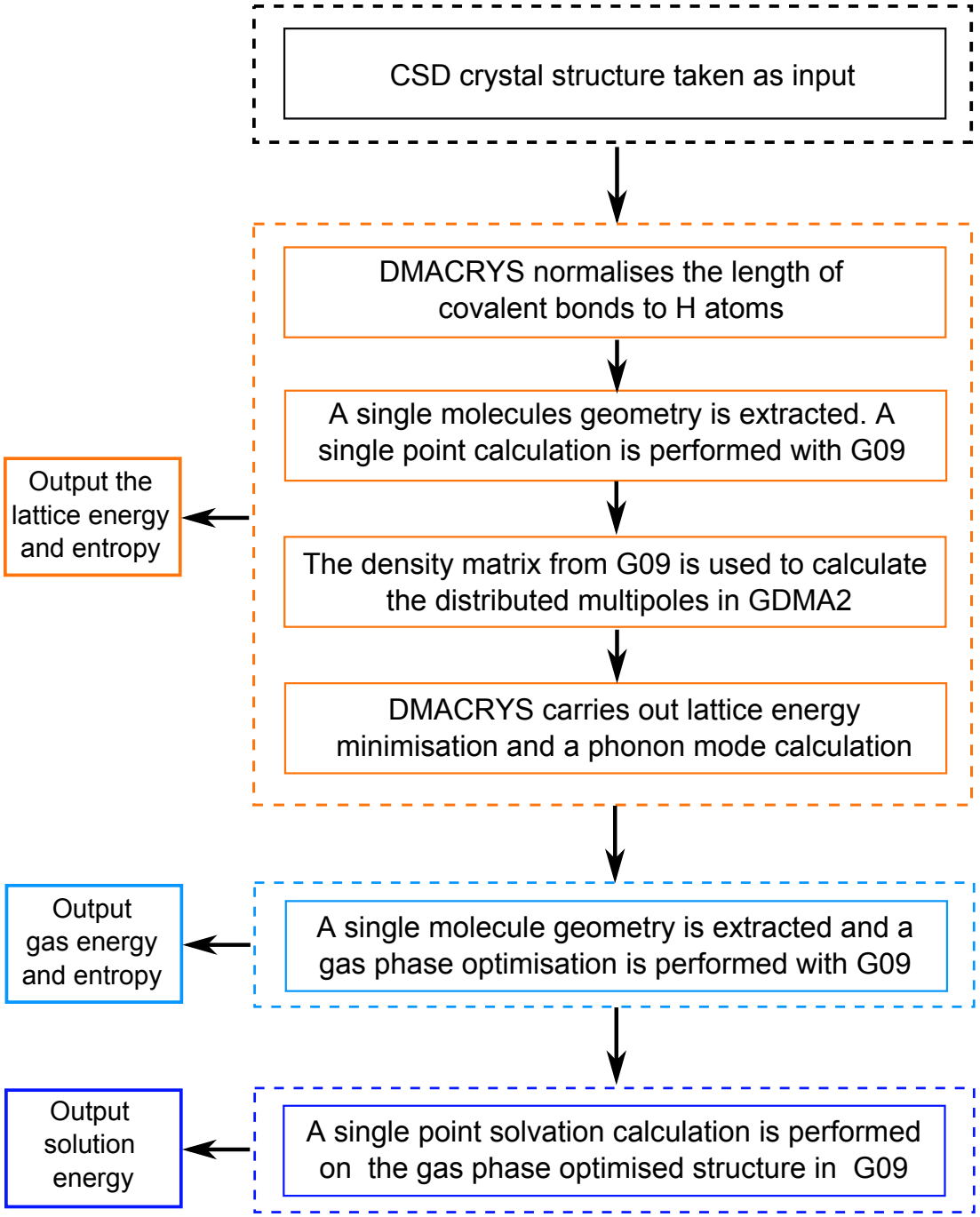


Figure 4.1: Workflow to calculate the thermodynamic parameters required for solubility prediction.

This workflow was run at two different levels of theory: HF/6-31G* and M06-2X/6-31G*. This gave us two benchmarks to which we could compare our cheminformatics predictions of solubility. In total this scheme provides ten chemical parameters usable as descriptors (**Table H.3**).

In addition to this we also calculated 123, two dimensional (calculated from 2D chemical structures) descriptors from the open source java library and toolkit, the *Chemistry Development Kit* (*CDK*).[82] These descriptors were all calculated from SMILES[203] strings. O'Boyle has pointed out the ambiguities within some SMILES strings.[204] In an attempt to minimise this we chose one main SMILES source,

the well-annotated database ChemSpider.[205–208] On a few occasions ambiguities in protonation state were recorded and SMILES strings from another source were used. The sources of the SMILES strings are stated in **Table H.2**. A list of the 2D descriptors used in this work is also provided in **Table H.4**.

4.3.2 Cheminformatics Workflow

Having generated these descriptors, we required a machine learning protocol with which to generate a QSPR model. Here we will apply three machine learning models: Random Forest (RF), Support Vector Machines (SVM) and Partial Least Squares (PLS).As described in **Section 2.1.3**, these machine learning models have several internal parameters which require optimisation. In addition we must produce an unbiased training and test set separation and run this over three machine learning models. To do this efficiently, collaborating with Neetika Nath, a double 10 fold cross validation (CV) approach was chosen. This provides an internal 10 fold CV, in which the machine learning parameters are optimised and a second external 10 fold CV in which an unbiased split of the data is made into a test and training group. A scheme representing the key steps is given below:

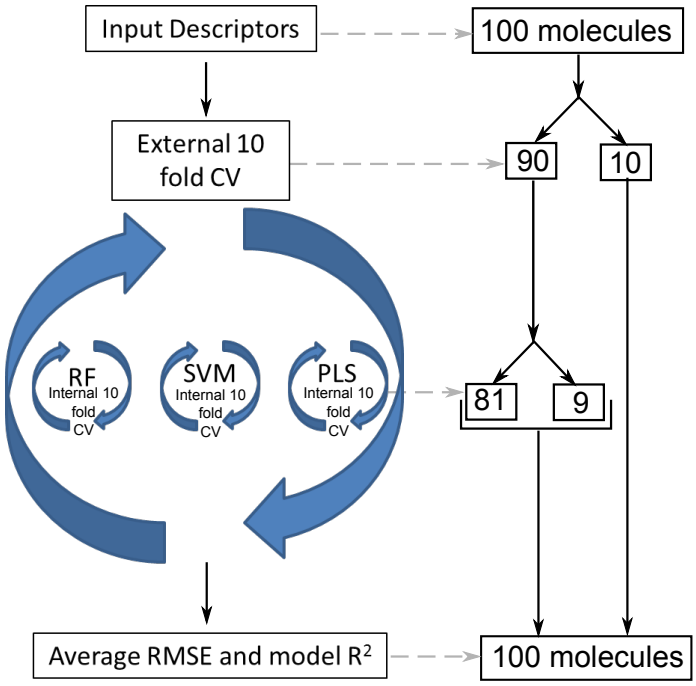


Figure 4.2: Schematic of the key steps involved in the machine learning protocol

Figure 4.2 shows an outline of the cross validation loops, exemplified for 100 molecules as used for the DLS-100 dataset. In the case of 100 molecules the training and test set split is made in an unbiased way from a random selection of 10% of the data reserved as the test set. This is repeated ten times, hence 10 fold cross validation, so that every molecule is used in the test set once. The remaining 90% is used as training data, hence each molecule is used as training data 9 times from the external 10 fold cross validation. The remaining 90% of the original data is then split again. This time 10% is taken randomly to test the parameters being trialled in

the machine learning models. The remaining 90% (81% of the original data) is used to train the machine learning models built using the test parameters. Optimised parameters are considered to be those minimising the RMSE of the internal 10 fold cross validation's test set. This approach is explained in detail in the following complete schematic of the process:

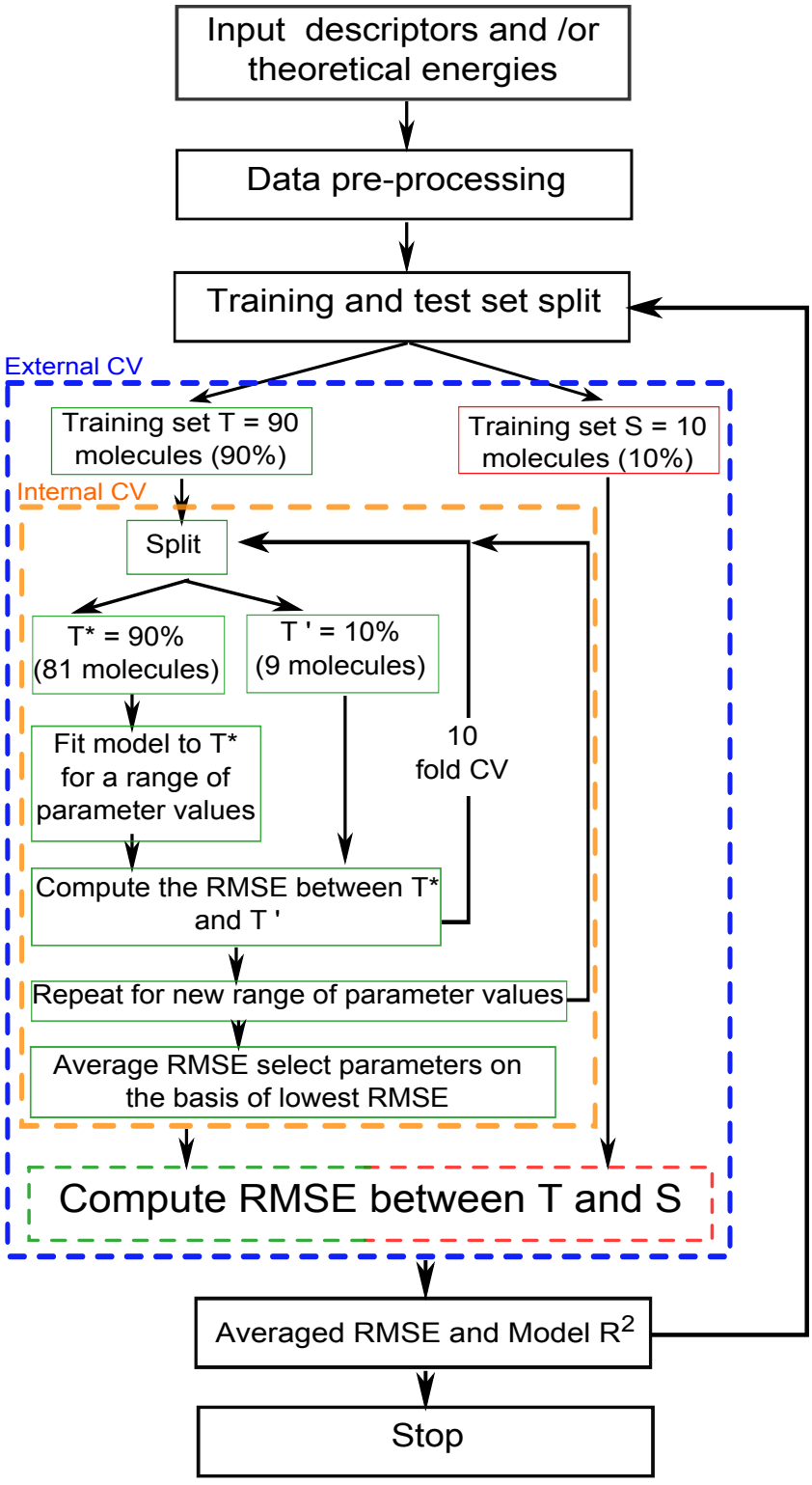


Figure 4.3: Machine learning work flow: Internal 10 fold cross validation optimises the machine learning parameters. External 10 fold cross validation produces an unbiased training and test set predictions.

In addition to the double ten fold cross validation we also have three pre-processing options represented in the overall diagram above in the data pre-processing box. It is important that no single descriptor can overshadow the others. Descriptors which hold the same value for every molecule in the dataset provide no differentiation and

hence are useless to the model. Descriptors holding a wide spread of values can overwhelm the model and hide other descriptors with only a small spread of values. Scaling provides a means of placing descriptors on an equal playing field. The three forms of scaling we use in this work are as follows:

- Scale by the standard deviation and the mean (*auto - scaling*)
- Scale by *principal components analysis* (*PCA*)
- Raw data - No scaling

The first scaling method centres each descriptor on a mean value and normalises each descriptor's standard deviation to one. This is achieved by calculating the mean value of the descriptor by averaging each individual value. The mean value is then taken away from each individual value. The new value is divided by the standard deviation of the descriptor. The new descriptor now has a unitary variance and a centre mean.

$$x_i' = \frac{x_i - \langle x \rangle}{\sigma} \tag{4.1}$$

Equation 4.1: Scaling descriptors by the standard deviation and the mean.[74]

The second scaling method is similar in concept to PLS, in that it reduces the number of descriptors to a set of core components known as principal components. PCA is useful as it allows for the descriptor dimensionality to be reduced hence reducing the possibility of over fitting the model. PCA generates principal components as a linear combination of the descriptors, where the principal components offer a maximal explanation of the variance between the descriptors.

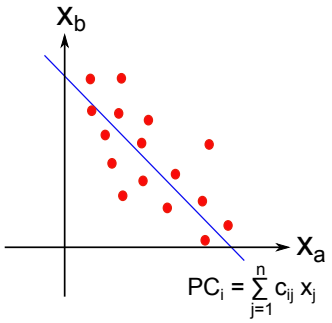


Figure 4.4: A 2D example of principal components analysis.

The final option is to run without any scaling. This enables us to look at the data in a pure form without any effects from scaling.

4.4 Predictions from Chemical Theory

If we apply the same criteria as applied in the first principles predictions of solubility, i.e. *useful predictions are those within the standard deviation of the experimental*

data (*1.71 logS units*), then the two benchmark calculations from theoretical chemistry provided a poor prediction of solubility for the DLS-100 dataset.

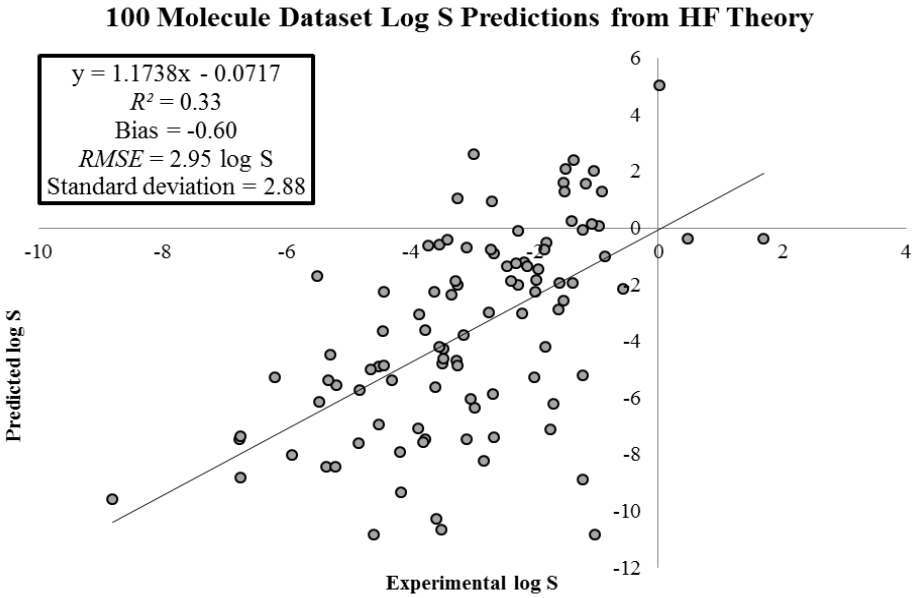


Figure 4.5: A prediction of solubility for DLS-100 using HF

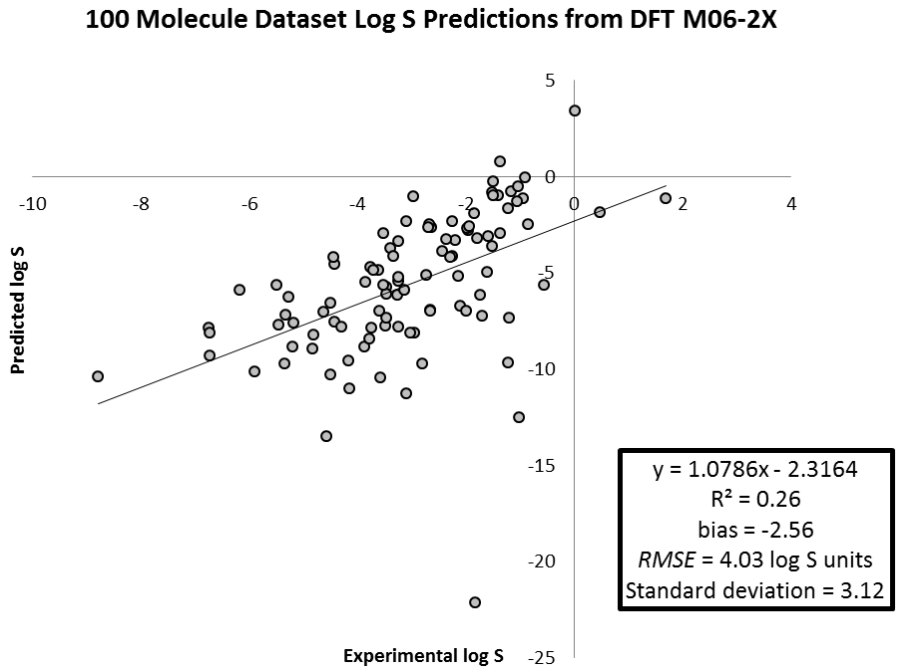


Figure 4.6: A prediction of solubility for DLS-100 using M06-2X

Clearly from the above data both methods fail to make a useful prediction of solubility according to the pre-defined criteria. The predictions made using HF make an improved prediction over predictions from the M06-2X method, although still missing the required predictive accuracy. As the DLS-100 dataset is made up of the DLS-25 dataset and 75 more molecules (DLS-75), it may be instructive to break the

predictions down into these two separate groups. Noting there are slight variations between the calculations in the first principles calculations and these calculations (multipole calculations based on different levels of electronic structure theory and the application of different Buckingham potential parameters), the following results are found when the data are split.

Measure	HF DLS-25	HF DLS-75	HF DLS-100
R^2	0.62	0.25	0.33
σ	2.36	3.02	2.88
Bias	-0.2	-0.73	-0.60
RMSE	2.37	3.11	2.95

Table 4.1: HF; DLS-25, DLS-75 and DLS-100 split

Measure	M06-2X DLS-25	M06-2X DLS-75	M06-2X DLS-100
R^2	0.53	0.19	0.25
σ	2.36	3.32	3.12
Bias	-1.83	-2.78	-2.56
RMSE	2.90	4.33	4.03

Table 4.2: M06-2X; DLS-25, DLS-75 and DLS-100 split

Bearing in mind that the standard deviation of the DLS-25 experimental data is 1.79 logS units and the standard deviation of the DLS-75 and DLS-100 experimental data is 1.71 logS units, we can see that in all cases the RMSE of the predictions exceeds that of the standard deviation of the experimental data. We can also note, that the predictions for the DLS-25 dataset are in both cases improved predictions over the other two datasets. Interestingly, again in both cases, we note that the predictions for the DLS-100 dataset are better, in terms of each statistical measure, than for the DLS-75 dataset. This may suggest that the DLS-75 dataset represents a more difficult dataset than the DLS-25 dataset. Another possible interpretation is that, as the data in the DLS-25 dataset all came from the same experiment(CheqSol) this data is less noisy than the additional data making up the DLS-75 and DLS-100 datasets which comes from a variety of sources. This noise is therefore causing the DLS-75 dataset and hence the DLS-100 dataset to appear as a harder dataset than the DLS-25 dataset due to random errors. Testing these hypotheses is outside the scope of the current work, although other authors have previously considered the themes behind these hypotheses.[36] It was recently suggested by Palmer *et al* that although it is dogma that QSPR and QSAR methods are restricted in accuracy due to the available experimental data, actually the primary reasons for poor predictions from such models is due to failures in the algorithms themselves and incomplete sets of descriptors.[36]

From these results, several conclusions are apparent. Firstly, the present methodologies do not suitably quantify the physical processes occurring during the solvation (transition from solid to solution) of a molecule. Secondly, assuming one

can explain the structure of the underlying data with a general model, using logS values predicted at the current levels of theory as its basis, then the model will be non-linear given the failure of the above linear models to explain the data.

4.5 Predictions from Cheminformatics

4.5.1 Machine learning - Theoretical Chemistry Descriptors

Using energies calculated from theoretical chemistry as descriptors in our machine learning models yields significantly improved results, compared with the result from theoretical chemistry alone. The results produced are useful predictions of solubility with an RMSE within the standard deviation of the experimental data (1.71 logS units). Both the RF and SVM models produce much improved results with PLS producing a slightly poorer result. The charts below show the results overall (in all diagrams the error bars show the standard deviation of the predicted values):

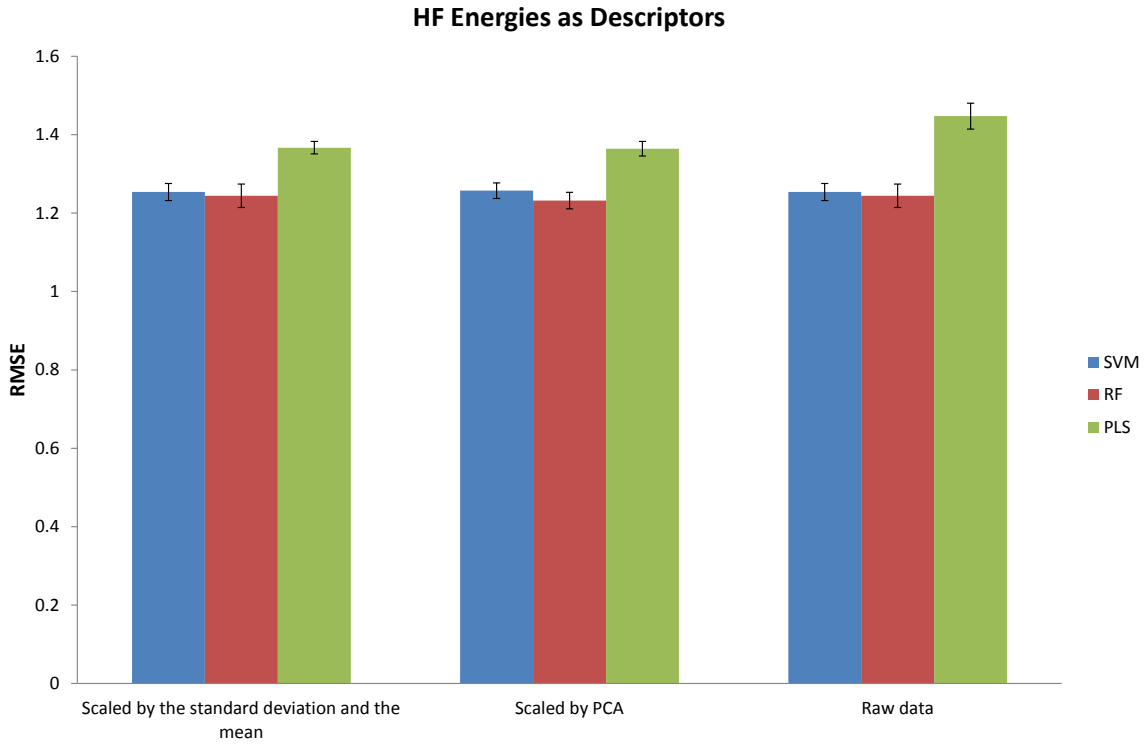


Figure 4.7: A prediction of solubility for DLS-100 using HF energies as descriptors.

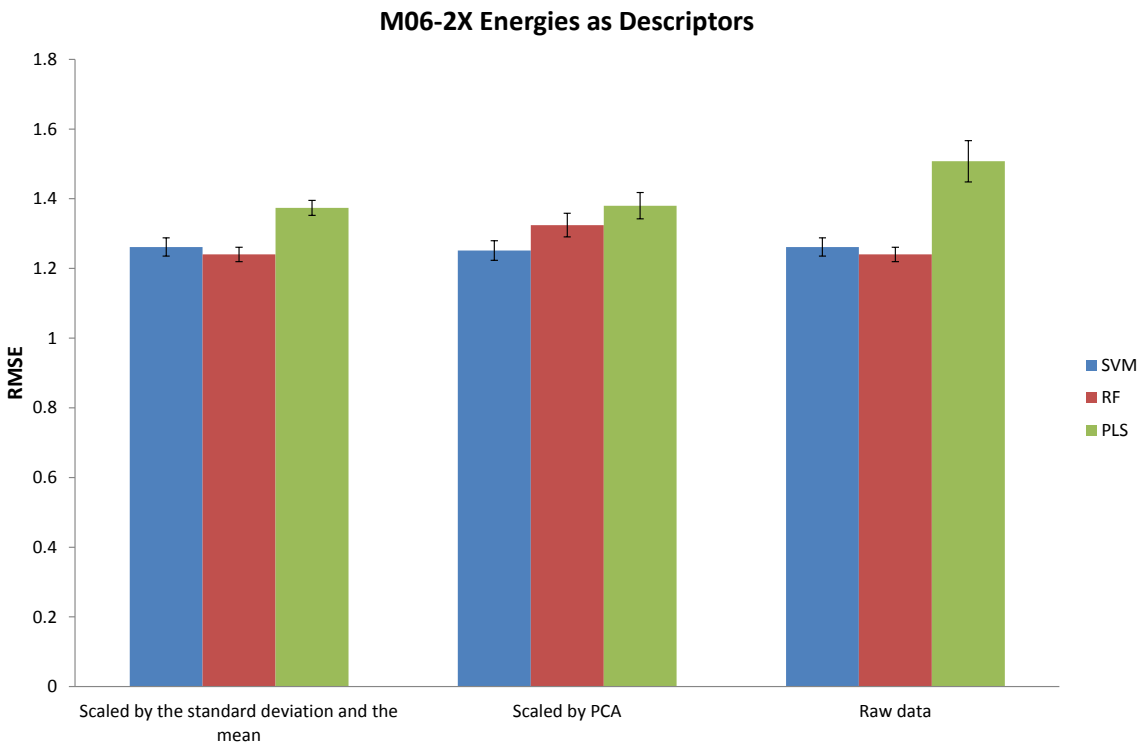


Figure 4.8: A prediction of solubility for DLS-100 using M06-2X energies as descriptors

HF	Scaled by Mean and σ		Scaled by PCA		No Scaling - Raw Data	
Method	RMSE	R^2	RMSE	R^2	RMSE	R^2
SVM	1.25±0.02	0.46±0.02	1.26±0.02	0.46±0.02	1.25±0.02	0.46±0.02
RF	1.24±0.03	0.47±0.03	1.21±0.02	0.5±0.02	1.24±0.03	0.47±0.03
PLS	1.37±0.02	0.36±0.01	1.36±0.02	0.36±0.02	1.45±0.03	0.29±0.03

Table 4.3: DLS-100 results using theoretical chemistry calculated data, at the HF level of theory as descriptors. Results are presented ± the standard deviation in the predictions.

M06-2X	Scaled by Mean and σ		Scaled by PCA		No Scaling - Raw Data	
Method	RMSE	R^2	RMSE	R^2	RMSE	R^2
SVM	1.26±0.03	0.45±0.02	1.25±0.03	0.46±0.02	1.26±0.03	0.45±0.02
RF	1.24±0.02	0.47±0.02	1.32±0.03	0.4±0.03	1.24±0.02	0.47±0.02
PLS	1.37±0.02	0.45±0.02	1.38±0.04	0.45±0.02	1.51±0.06	0.25±0.04

Table 4.4: DLS-100 results using theoretical chemistry calculated data, at the M06-2X level of theory, as descriptors. Results are presented ± the standard deviation in the predictions.

Figure 4.7 and **Figure 4.8** represent the predictive accuracy of each method in terms of the minimizing the RMSE. In this set the best method is RF with HF calculated descriptors scaled with PCA, with the results 1.21 logS units RMSE and R^2=0.5 **(Table 4.4)**. Given that these models are built on only 10 descriptors this is a dramatic improvement and shows that the descriptors are providing useful information to the model which correlates to the experimental data.

4.5.2 Machine learning - CDK Descriptors

In addition to running machine learning using theoretical chemistry descriptors, we ran the machine learning methods using 2D chemical descriptors. It is of interest that when using the 2D chemical descriptors alone as input to the machine learning algorithms a marginally improved prediction of logS is achieved, compared to the equivalent machine learning methods using energies as descriptors. Of particular note is the fact that the RF model can produce a statistically significant improvement on its previous predictions, when presented with data scaled by the mean and standard deviation (**Appendix I**). In all other cases, the changes are not significant. Here, these results suggest that slightly more information, pertinent to the molecule's logS values, is conveyed by cheminformatics descriptors than when the machine learning models are presented with theoretical energies alone.

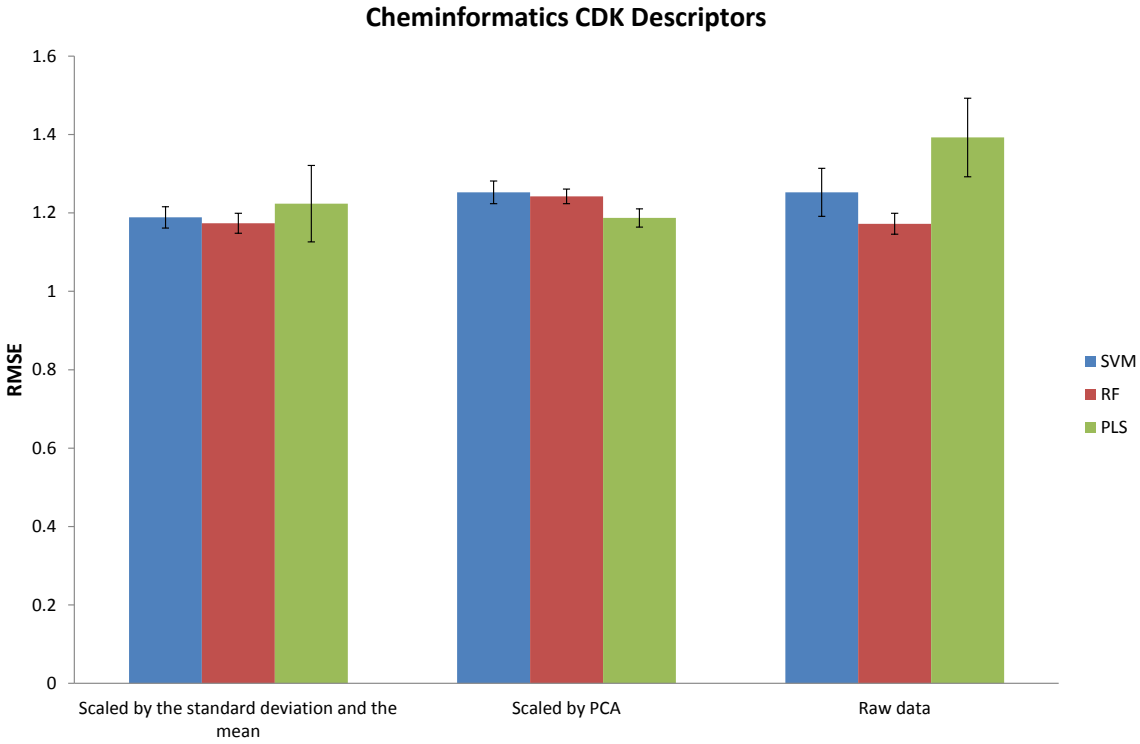


Figure 4.9: A prediction of solubility for DLS-100 using 2D CDK descriptors.

CDK	Scaled by Mean and σ		Scaled by PCA		No Scaling - Raw Data	
Method	RMSE	R^2	RMSE	R^2	RMSE	R^2
SVM	1.19±0.03	0.51±0.02	1.25±0.03	0.46±0.03	1.25±0.06	0.46±0.06
RF	1.17±0.03	0.53±0.02	1.24±0.02	0.48±0.02	1.17±0.03	0.53±0.02
PLS	1.22±0.1	0.52±0.05	1.19±0.02	0.53±0.01	1.39±0.1	0.242±0.06

Table 4.5: DLS-100 results using 2D CDK descriptors. Results are presented ± the standard deviation in the predictions.

Figure 4.9 demonstrates the relative ability of each predictive method. It shows the best method in this case to be a tie between RF either scaled by the standard

deviation and the mean or using the raw data (1.17 logS units RMSE and a R^2=0.53) (**Table 4.5**). For both of the other machine learning methods scaling is required to produce the best result. RF, however, automatically selects descriptors based on importance, i.e. those providing maximal differentiation are selected before those offering less differentiation between the values of the property of interest. For this reason RF in this case is able to use the raw data as effectively as the scaled data.

4.5.3 Machine learning - Mixed Descriptor Sets

We continued this work by combining descriptors and energies, producing a new input dataset containing 133 descriptors. This was input to the machine learning algorithms as was done with the previous data sets. The results generated were generally a moderate improvement over those generated by cheminformatics descriptors alone. The best prediction was provided by PLS using the M06-2X energies and CDK descriptors, scaled by the standard deviation and the mean (1.11 logS units RMSE and R^2=0.59). This implies the theoretical energies provide only a moderate amount of additional, useful, information to the models, above that already present in the CDK descriptors. Some results see a statistically significant improvement on combining the descriptor sets (RF and PLS with descriptors scaled by the mean/standard deviation) compared to using theoretical energies alone. Given this, and the fact that using the descriptors alone provides a small improvement to the results compared to theoretical chemistry descriptors alone, it is a reasonable conclusion that the cheminformatics descriptors contain some modest amount of extra information not present in the theoretical chemistry descriptors. Thus, it is suggested that the cheminformatics descriptors and theoretical chemistry descriptors supply analogous information, with a modest amount of additional information conveyed by combining the descriptors. A reasonable assertion is that the two sets of descriptors are largely non-complementary.

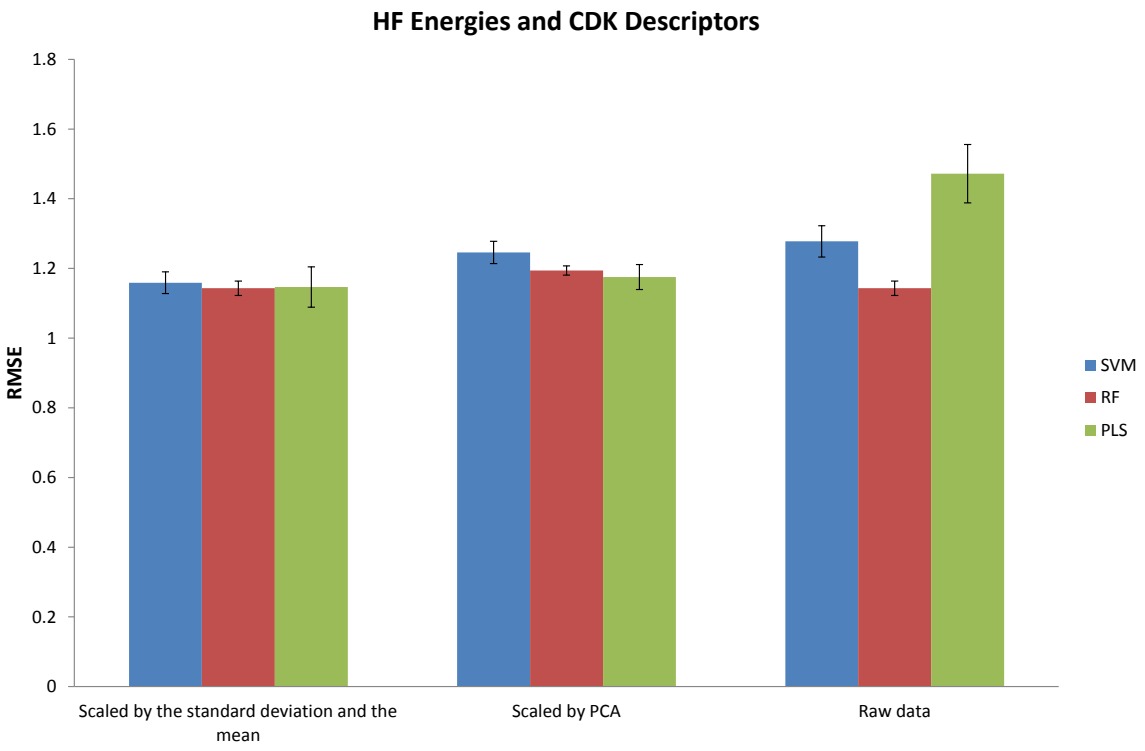


Figure 4.10: A prediction of solubility for DLS-100 using HF energies and CDK descriptors.

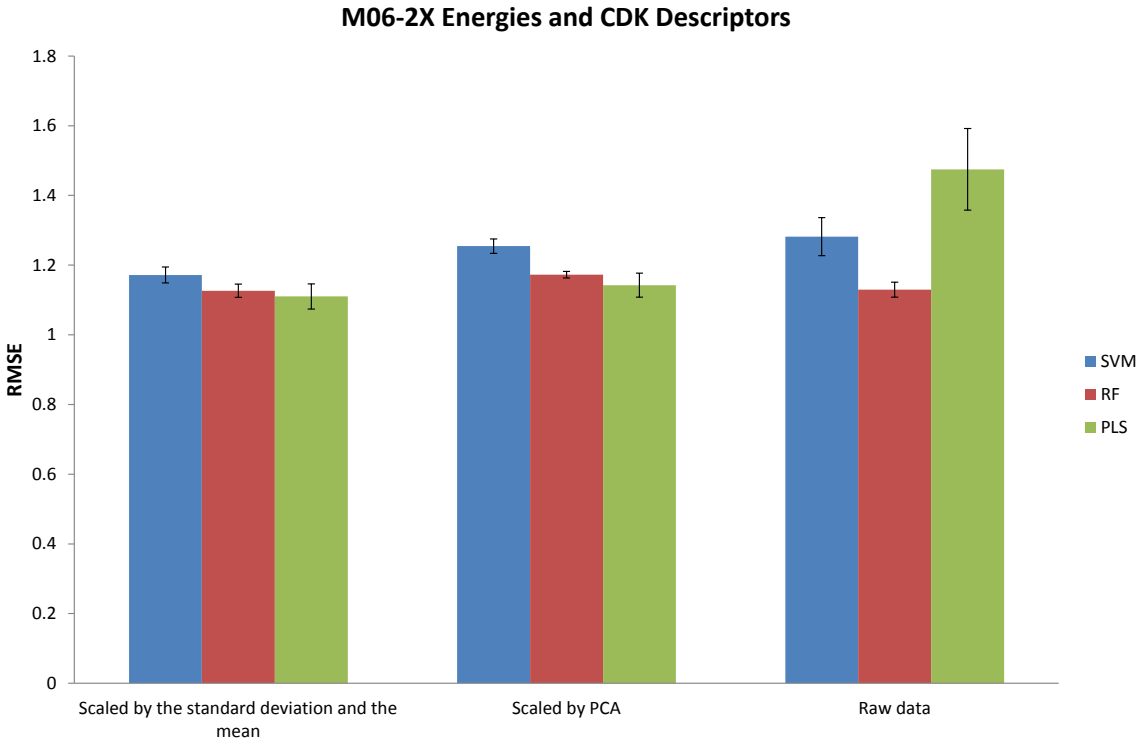


Figure 4.11: A prediction of solubility for DLS-100 using M06-2X energies and CDK descriptors

HF + CDK	Scaled by Mean and σ		Scaled by PCA		No Scaling - Raw Data	
Method	RMSE	R^2	RMSE	R^2	RMSE	R^2
SVM	1.16±0.03	0.54±0.03	1.25±0.03	0.47±0.03	1.28±0.05	0.44±0.04
RF	1.14±0.02	0.56±0.02	1.19±0.01	0.52±0.01	1.14±0.02	0.56±0.02
PLS	1.15±0.06	0.57±0.04	1.18±0.04	0.54±0.03	1.47±0.08	0.35±0.05

Table 4.6: DLS-100 results using theoretical chemistry calculated data, at the HF level of theory as descriptors. Results are presented ± the standard deviation in the predictions.

M062X+CDK	Scaled by Mean and σ		Scaled by PCA		No Scaling - Raw Data	
Method	RMSE	R^2	RMSE	R^2	RMSE	R^2
SVM	1.17±0.02	0.53±0.02	1.25±0.02	0.46±0.02	1.28±0.05	0.43±0.05
RF	1.13±0.02	0.57±0.02	1.17±0.01	0.54±0.01	1.13±0.02	0.57±0.02
PLS	1.11±0.04	0.59±0.02	1.14±0.03	0.56±0.02	1.47±0.12	0.35±0.07

Table 4.7: DLS-100 results using theoretical chemistry calculated data, at the M06-2X level of theory, as descriptors. Results are presented ± the standard deviation in the predictions.

One interesting point, is that the best result comes from the descriptor set combining M06-2X energies with CDK descriptors. The M06-2X energies alone produced the poorest results of the two purely theoretical chemistry predictions (**Figure 4.5** and **Figure 4.6**). This is an unexpected result unlikely to be chemically meaningful but rather the result of a fortuitous correlation with the descriptors providing a correction to the prediction.

4.6 Conclusions from Machine Learning

From these 45 different prediction methods for the solubility of the DLS-100 dataset, we find the RF model performs well with all of the descriptor sets, whether they are scaled or not. The best prediction from the RF method resulted in an RMSE 1.13 logS units. It is generally considered that an accuracy of 1 logS unit constitutes a good prediction. For a difficult dataset such as DLS-100 it appears RF performs well, only narrowly missing the 1 logS unit RMSE criterion in a number of cases. After machine learning all of the methods presented above produce predictions with an RMSE comfortably inside the standard deviation of the experimental data; hence, each prediction is a useful prediction by the criteria previously defined. Note that RF is the only model that makes moderately improved predictions, which are statistically significant in a number of cases, with or without any form of scaling. This is therefore our recommended method for general applicability.[33]

One of the major weaknesses of machine learning models is lack of understanding that can be derived from their results. Whereas theoretical chemistry calculations can be decomposed to reveal further physical information about a system the same is not true for cheminformatics and machine learning. We can, however, analyse which of the descriptors are found to be the most important in certain cases. We

have done this here (**Appendix I**). The XlogP[209] descriptor was consistently found to be the top rated descriptor in all of the data sets it was included in. XlogP is a model which estimates the base ten logarithm of the octanol-water partition coefficient. This estimate is calculated as additive contributions from atoms within the molecule. LogP descriptors have previously been seen as a vital descriptor for QSPR models predicting solubility.[53] It is not surprising that logP is rated as an important descriptor, as it provides information about the solvated phase.[32,196] **Appendix I** shows tables of the top ten important descriptors in the RF models. The χ path and chain indices by Kier and Hall[90,210] were commonly found in the top ten; these descriptors quantify the bonding to heavy atoms over a specific path length or equivalently over a chain length. The Moreau Broto autocorrelation[211] descriptor is also found in the top ten most important descriptors. This descriptor describes how charge and mass are distributed across a particular path length. A final addition to the top ten is Randic's weighted path descriptors,[88,212] which informs the model of the degree of molecular branching present. Adding the theoretical energies and the CDK descriptor sets together finds the following energy terms in the top ten most important descriptors: the free energies of hydration, the free energy of solution and the theoretically predicted logS. The descriptors used in this work are listed in **Appendix H**.[33]

As previously stated we cannot decompose these results to yield deeper chemical meaning as is sometimes possible with theoretical chemistry calculations. For this reason we must be careful when assigning chemical meaning to the descriptor importance information. However, we can see some chemical sense and logic in the selection of most important descriptors. Molecular branching is ranked in the top ten; one can see how information pertinent to the extent and flexibility of the molecule would be important in determining the contribution of entropy, for example. Linking such information with that emanating from the Kier Hall descriptors, allows for the acquisition of knowledge about chain composition, in terms of bonded heavy atoms. The autocorrelation descriptor, describing the distribution of charge and mass, might be considered to impart knowledge of the heavy atom distribution and some limited electronic information from the charges. The degree to which charges are separated, i.e. localised or dispersed, across a molecule is an important factor for determining the enthalpic and entropic contributions. The theoretical energies in the top 10 are all closely related quantities; it comes as no surprise to find the predicted logS value in the top 10, as it is the quantity which we are attempting to predict. One expects the prediction from theoretical chemistry to supply sufficient information to the machine learning methods to be ranked in the top ten most important descriptors. It is again unsurprising that the free energies of solution and hydration are ranked in the top ten most important descriptors, as they provide direct information from quantum chemical calculations on intra- and intermolecular interactions, in a given conformation, within the various physical and chemical environments. Additionally, these predicted values provide information on the energetics of phase transitions.[33]

As a benchmark, we performed the same calculations on a standard dataset from the '*Solubility Challenge*' (*SC*).[28,44,213] The predictions for the SC dataset are made using the 2D cheminformatics descriptors from the CDK. As suitable crystal structures were not available from the CSD for all of the molecules in the SC, it was not possible

to calculate the theoretical energies.[33]

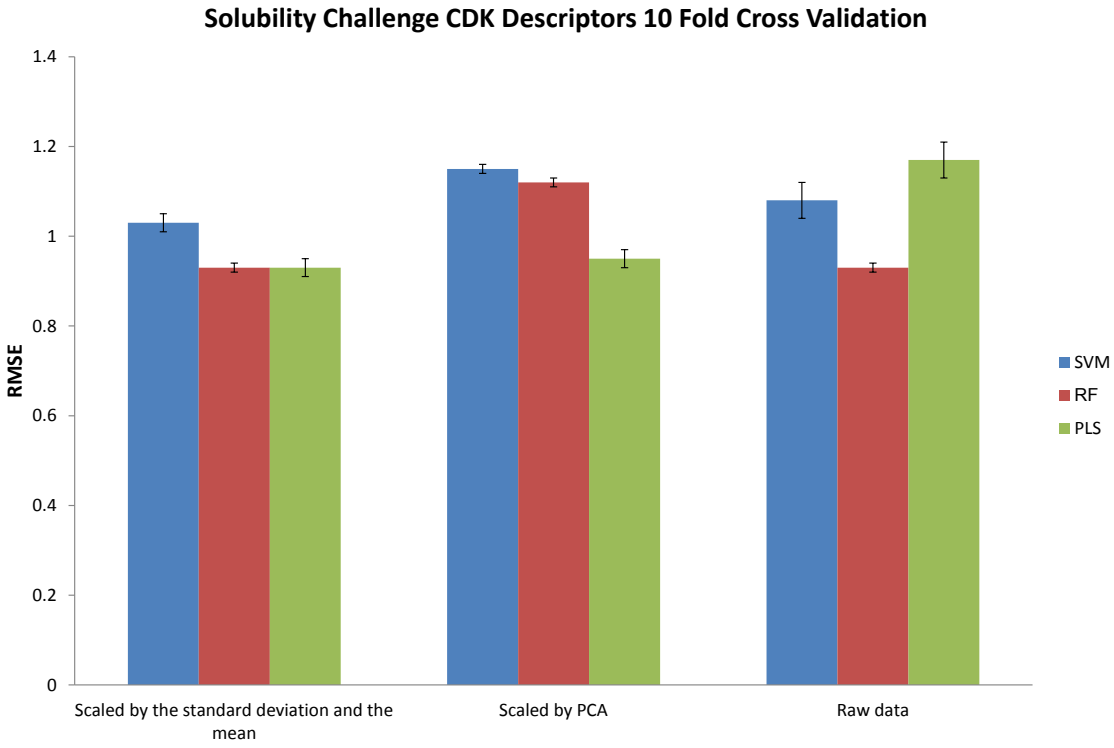


Figure 4.12: A prediction of solubility for solubility challenge dataset using 2D CDK descriptors in our 10 fold cross validation methodology.

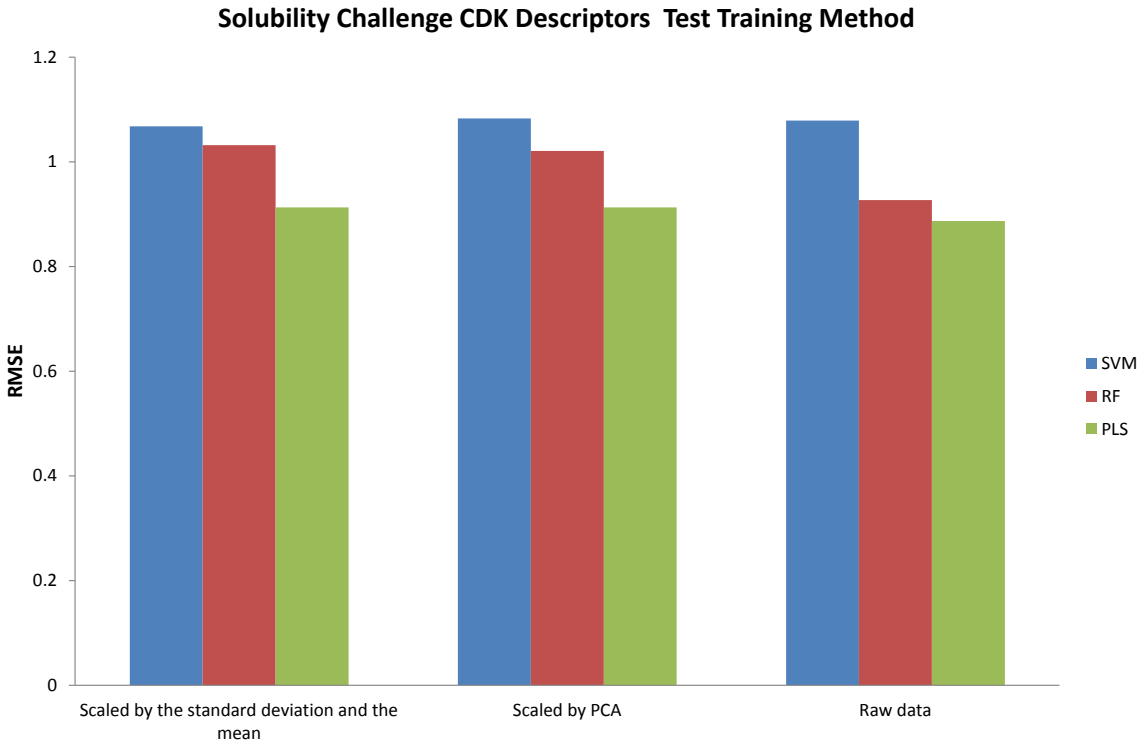


Figure 4.13: A prediction of solubility for solubility challenge dataset using 2D CDK descriptors in the solubility challenge's original test and training data split.

CDK	Scaled by Mean and σ		Scaled by PCA		No Scaling - Raw Data	
Method	RMSE	R^2	RMSE	R^2	RMSE	R^2
SVM	1.03±0.02	0.45±0.02	1.15±0.01	0.31±0.02	1.08±0.04	0.39±0.04
RF	0.93±0.01	0.56±0.01	1.12±0.01	0.36±0.02	0.93±0.01	0.56±0.01
PLS	0.93±0.02	0.55±0.02	0.95±0.02	0.53±0.02	1.17±0.04	0.33±0.03

Table 4.8: A prediction of solubility for solubility challenge dataset using 2D CDK descriptors in our 10 fold cross validation methodology. Results are presented ± the standard deviation in the predictions.

CDK	Scaled by Mean and σ		Scaled by PCA		No Scaling - Raw Data	
Method	RMSE	R^2	RMSE	R^2	RMSE	R^2
SVM	1.07	0.41	1.08	0.39	1.08	0.41
RF	1.03	0.50	1.02	0.50	0.93	0.57
PLS	0.91	0.55	0.91	0.55	0.89	0.58

Table 4.9: A prediction of solubility for solubility challenge dataset using 2D CDK descriptors in the solubility challenges original test and training data split.

Table 4.8 and **Figure 4.12**, validate that our method is performing well by comparison to alternatives by making predictions of the SC dataset within 1 logS unit RMSE. These are levels of predictive accuracy currently in line with those being achieved by some commercial solubility prediction software (MLR-SC62 RMSE 0.95

logS units) and advanced deep-learning methods (RMSE 0.90 logS units).[32] These results are not directly comparable with ours, due to methodological differences and the fact that these results are contingent on correcting eight errors in the original solubility challenge data. Whilst we corrected names and SMILES to ensure consistent structures were used we did not correct the logS values from the original SC. Using the SC data set as a benchmark provides further evidence of the difficult nature of our DLS-100 data set. It is therefore further suggested that DLS-100 should be considered a "difficult data set", given the improvement in predictive accuracy on the SC data set compared to the DLS-100.[33]

Chapter 5

Empirical Models of Solubility Prediction

"At that stage I had barely heard of quantum mechanics, and was only expressing a liking for theory as being able to quantify relationships between measurable properties and make predictions."

Sally Price, 2012

5.1 Empirical Predictions of Solubility

There have been a number of proposed solubility equations containing various parameters and descriptors.[5] These models were generated initially in the 60's, for calculating the differences in logS.[214]. By the 80's expressions capable of accurate predictions of logS had been derived.[46] These relationships were based on the input of a few experimental pieces of data, which could be used effectively as descriptors of physical characteristics related to solubility. In this chapter we describe the application of one of these equations, we attempt to predict the empirical quantities required using cheminformatics and finally apply these quantities to a logS prediction.

5.1.1 The General Solubility Equation: Predicting Melting Points

Predicting crystal structures and their physicochemical properties is an important research area. Predicting melting points is one small region of this research area. Melting points are an attractive property as the well established *General Solubility Equation* (*GSE*, **Equation 5.1**)[46,47,215] links the melting point to solubility with reference to a thermodynamic cycle via a pure melt: this empirically derived relationship has seen wide usage.[54,215,216] The GSE has been proposed as a way to accurately predict solubility using only two pieces of empirical data; the first is the melting point, the second logP. Log P can be reasonably predicted by atom or group

contribution models such as AlogP, XlogP and ClogP which have been discussed previously. Melting points, however, still elude us as predictable quantities. For this reason a good prediction of a crystal's melting point could in principle provide a direct useful prediction of a molecule's solubility.

$$log_{10}S = 0.08 - log_{10}P - 0.01 \times (MP - 25) \quad (5.1)$$
$$log_{10}S = 0.05 - log_{10}P - 0.01 \times (MP - 25)$$

Equation 5.1: The general solubility equation: Top original, bottom revised intercept correction. Log P is the partition coefficient between octanol and water. MP is the melting point in °C.

5.1.2 Melting Point Data

A set of open source melting point data has been made available online.[217] By far the largest of these datasets is the Alfa Aesar dataset of over 8000 molecules and their corresponding SMILES strings. A random subset containing 1100 molecules was taken from this dataset to produce a manageable dataset for melting point prediction. This dataset will be referred to as MP1100. The full dataset is presented with molecular name, predicted logP by the AlogP method[101–103] and the melting point in **Appendix J**. AlogP and ClogP are two of the most popular algorithms to predict the partition coefficient. AlogP was used to predict the partition coefficient as it has been shown to be equivalent to the ClogP algorithm for molecules composed of 21-45 atoms and superior to ClogP for molecules composed of >45 atoms. As 67 of the molecules in the dataset fall in the range of *ge*45 atoms, with many others being between 21 and 45 atoms in size, this method was selected.[218] This dataset covers a diverse range of chemical space and molecules. All molecules are organic and the dataset contains a number of structural isomers. All melting points and SMILES are taken from the Alfa Aesar open source melting point dataset. The dataset is approximately normally distributed as shown below.

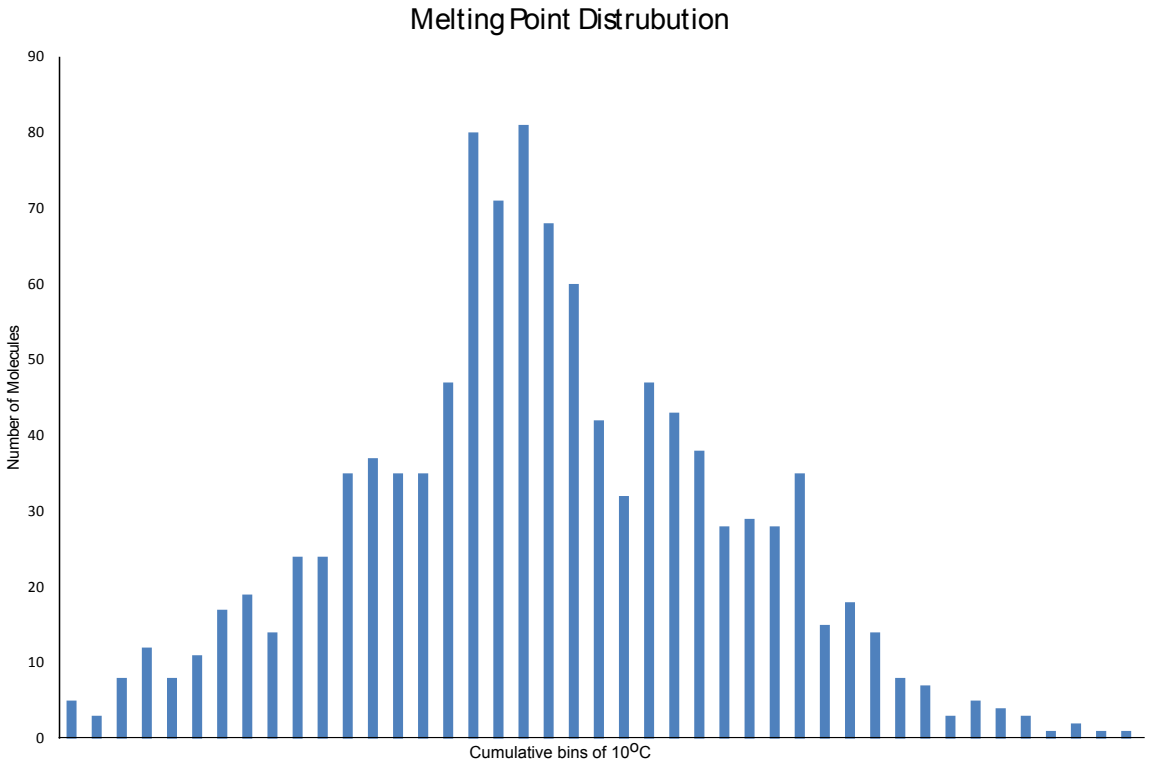


Figure 5.1: Melting point distributions. Each bin is $10°C$*, with each subsequent bin being cumulative over the previous bins.*

5.1.3 Melting Point Predictions

By applying the same machine learning methods as were described in **Section 4.1** the melting points of the MP1100 dataset were predicted. 2D descriptors were again calculated using the CDK for all 1100 molecules; 101 descriptors in total were found to provide information to the models. Given that the number of data points overwhelms the number of descriptors in this dataset, there was little concern of over fitting. A single method of scaling was selected, that of mean and standard deviation scaling. This method was selected as the models created for solubility using this scaling with the 2D CDK descriptors had performed well.

Given the apparent importance of logP, exemplified in the GSE, the melting point predictions were run twice; once containing a logP descriptor and once not containing a logP descriptor. This was done to assess the importance of the descriptor to the final model. It was assumed that this would be negligible and was indeed shown to be so. Given the large number of descriptors it is likely that sufficient information will be provided by other descriptors for a prediction to be made of similar accuracy. Chemically, the melting point will be determined by the solid and molten liquid state properties, ergo, one may imagine a descriptor for the solvated phase having only a minor impact. The results of predictions without the logP descriptor are marginally worse and therefore we focus on the prediction made with the descriptor included. The resultant predictions made without the logP descriptor are shown in **Appendix J**.

The following three figures represent the prediction of each molecule's melting point

in the MP1100 dataset. **Figures 5.2, 5.3** and **5.4** show the prediction by each machine learning method.

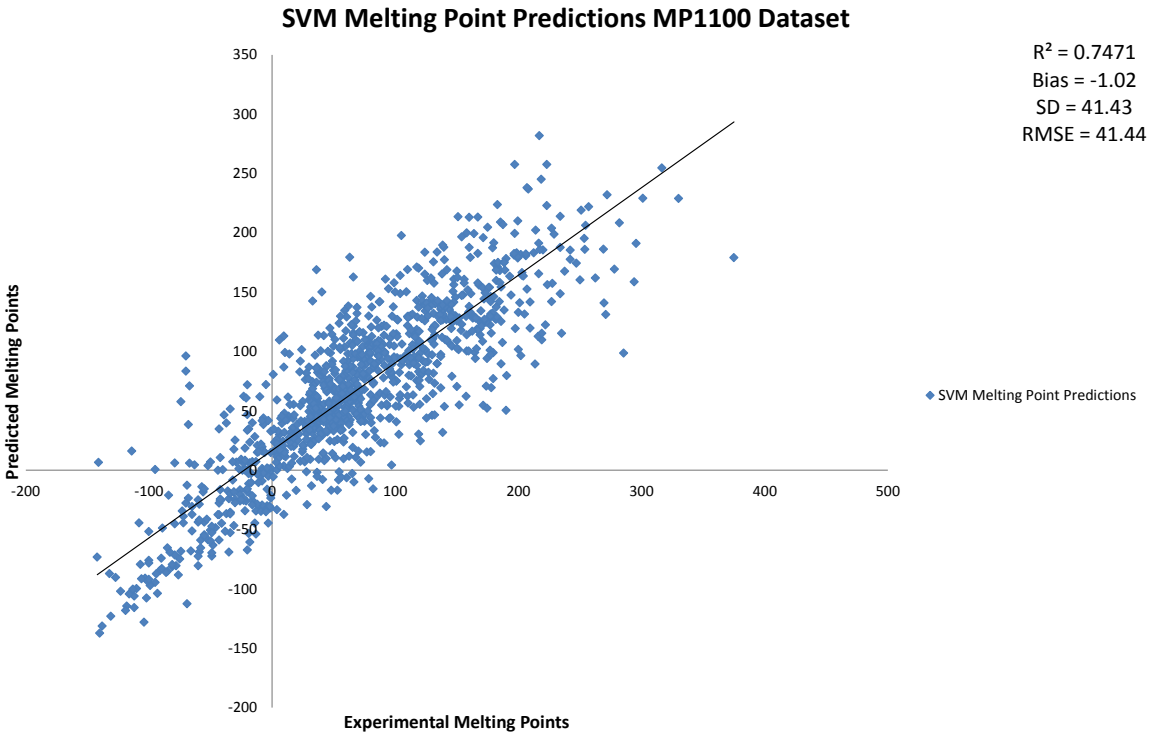


Figure 5.2: A support vector machine prediction of melting points (°C) for the MP1100 dataset using 2D CDK descriptors in our 10 fold cross validation methodology.

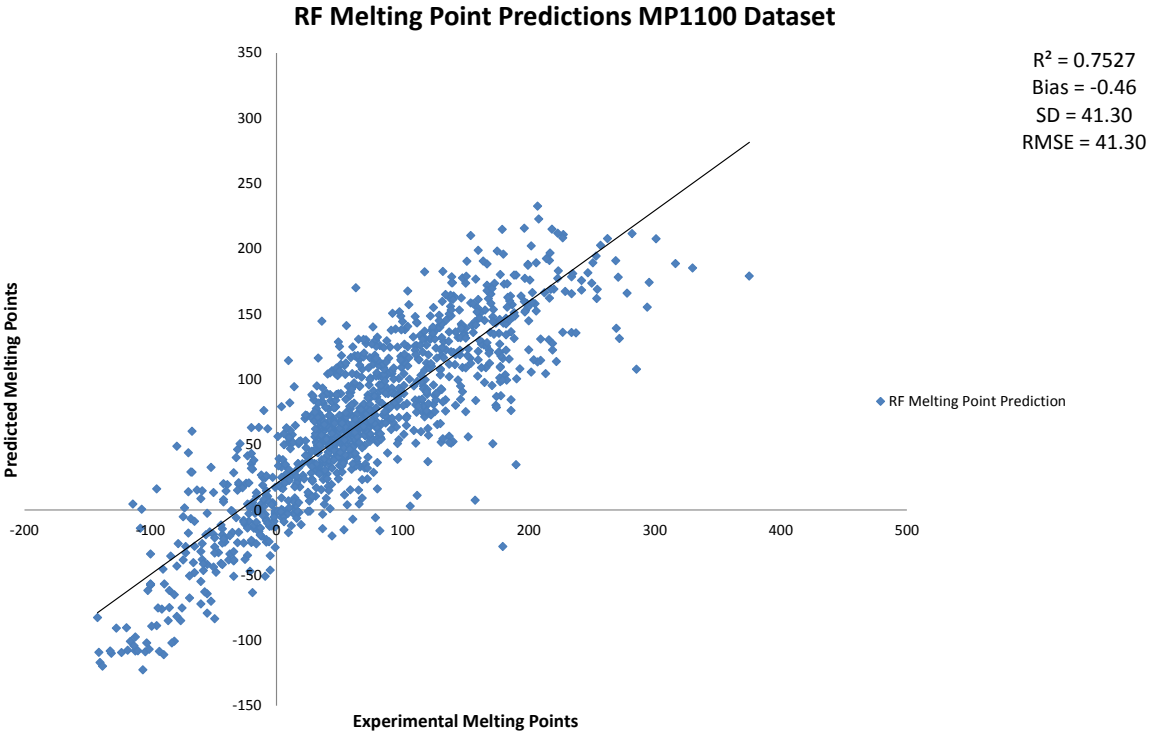


Figure 5.3: A random forest prediction of melting points (°C) for the MP1100 dataset using 2D CDK descriptors in our 10 fold cross validation methodology.

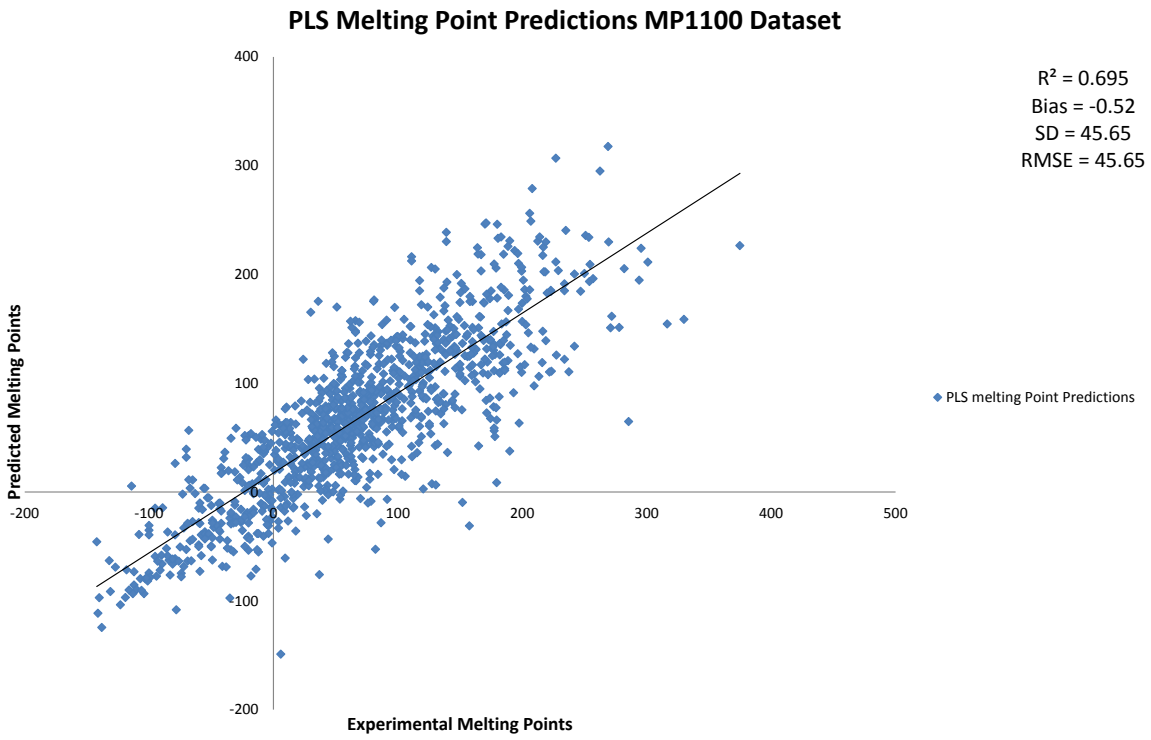


Figure 5.4: A partial least squares prediction of melting points (°C) for the MP1100 dataset using 2D CDK descriptors in our 10 fold cross validation methodology.

We can see that each of the machine learning methods achieves a good correlation coefficient with a low bias, suggesting a low systematic error. The experimental data has a standard deviation of 82.40 $°C$. All of the methods above show an RMSE significantly below this, suggesting the predictions are useful. The RMSE however, in all cases, is above 40 $°C$ meaning that the melting points are at best imprecise predictions. The average absolute predictive errors are: 34.4$°C$ from PLS, 30.7$°C$ from RF and 30.6$°C$ from SVM. Given that these predictions are over a temperature range of 517$°C$, on average the predictive accuracy is reasonably promising. However, at the upper end of the predictive inaccuracy, errors occur upto 220.8$°C$ for PLS, 207.5$°C$ for RF and 195.9$°C$ for SVM. This level of accuracy is not suitable as a quantitative prediction of melting point, although it could be argued that on average these predictions are qualitatively useful. These predictions are in line with existing methodologies using a variety of QSPR/QSAR approaches and datasets containing similar molecules.[196,219–221]

In terms of descriptor importance we find the that the most important descriptor is the topological polar surface area. The next four descriptors making up the top five most important descriptors are; the Zagreb index, weighted paths of length 3 and 4 and finally the hydrogen bond donor count. These descriptors make physical sense as they describe the polarity of a molecule's exposed surface, i.e. the area most available for direct interaction. In addition, information on the molecule's extent in terms of complexity and branching from the Zagreb index and weighted paths is provided by these descriptors. The hydrogen bond donor descriptor is providing at least some information to the model about significant interactions within the solid state not described by the other four descriptors. Interestingly, hydrogen bond acceptor descriptors are found much lower down in the ranking outside of the top ten. The

logP descriptor is also found outside of the top ten. As we stated above it is unlikely that such a descriptor would provide significant amounts of information when trying to predict the solid state property of melting point. The descriptor importance is based on RF's prediction of melting point. The top ten most important descriptors are shown in **Appendix J**.

5.1.4 Solubility from Melting Points

Following from these predictions we studied the overlap between the MP1100 dataset and the existing solubility datasets, DLS-25 and DLS-100. We found an overlap of 30 molecules, which we will refer to as DLS-30. For these 30 molecules we predicted the logS value using the GSE. The GSE was reparametrised by Jain and Yalkowsky[47] and so we present both the results from the original GSE, from Yalkowsky and Valvani,[46] and from the reparametrised version by Jain and Yalkowsky.[47] The three **Figures 5.5, 5.6** and **5.7** show these results.

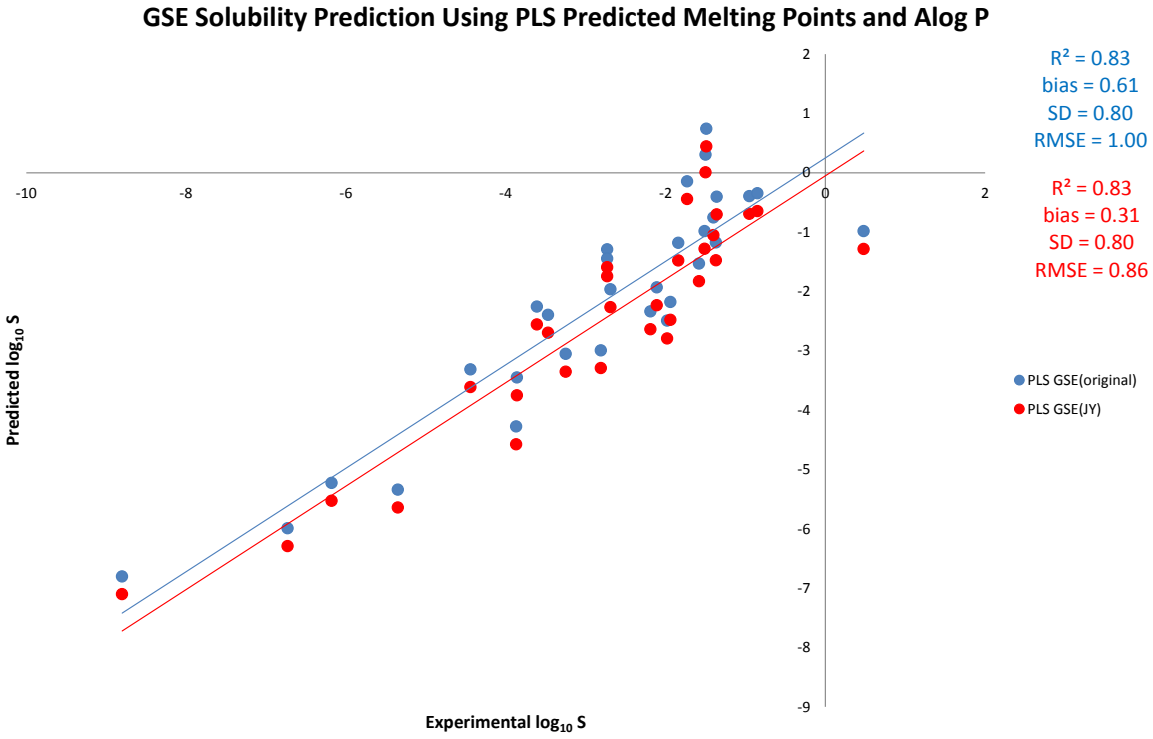


Figure 5.5: A prediction of solubility for the DLS-30 molecules using the general solubility equation with predicted melting points from PLS and predicted logP from AlogP.

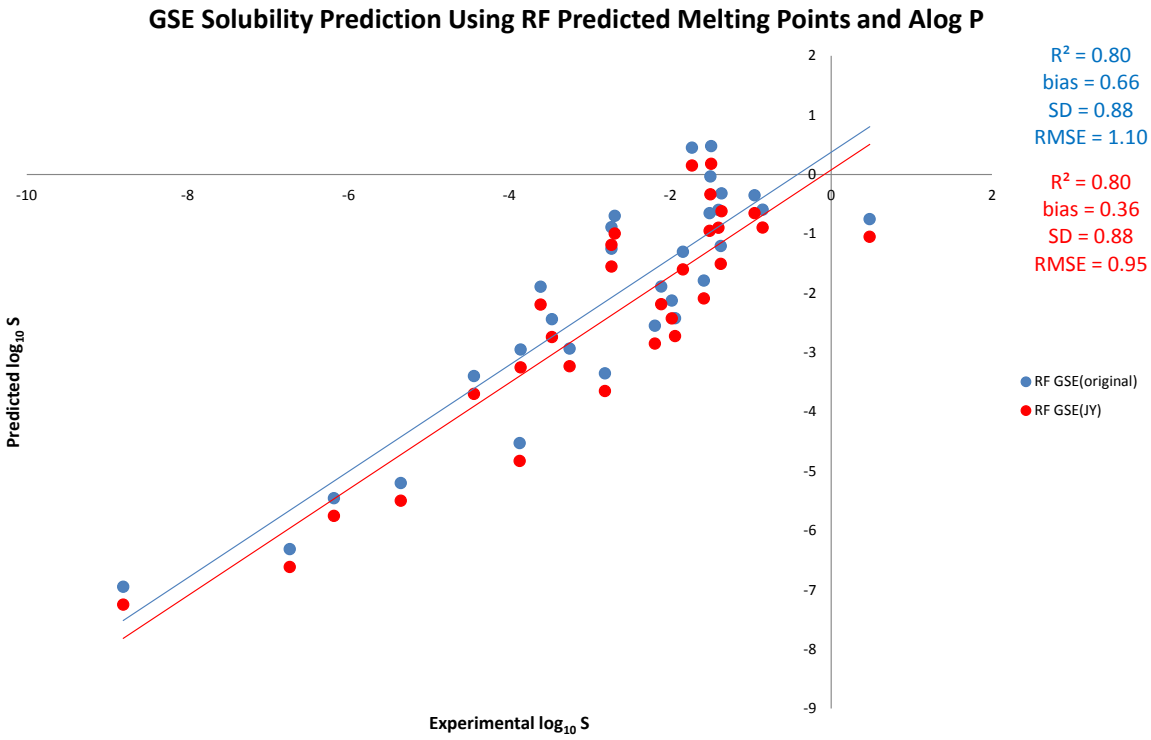


Figure 5.6: A prediction of solubility for the DLS-30 molecules using the general solubility equation with predicted melting points from RF and predicted logP from AlogP.

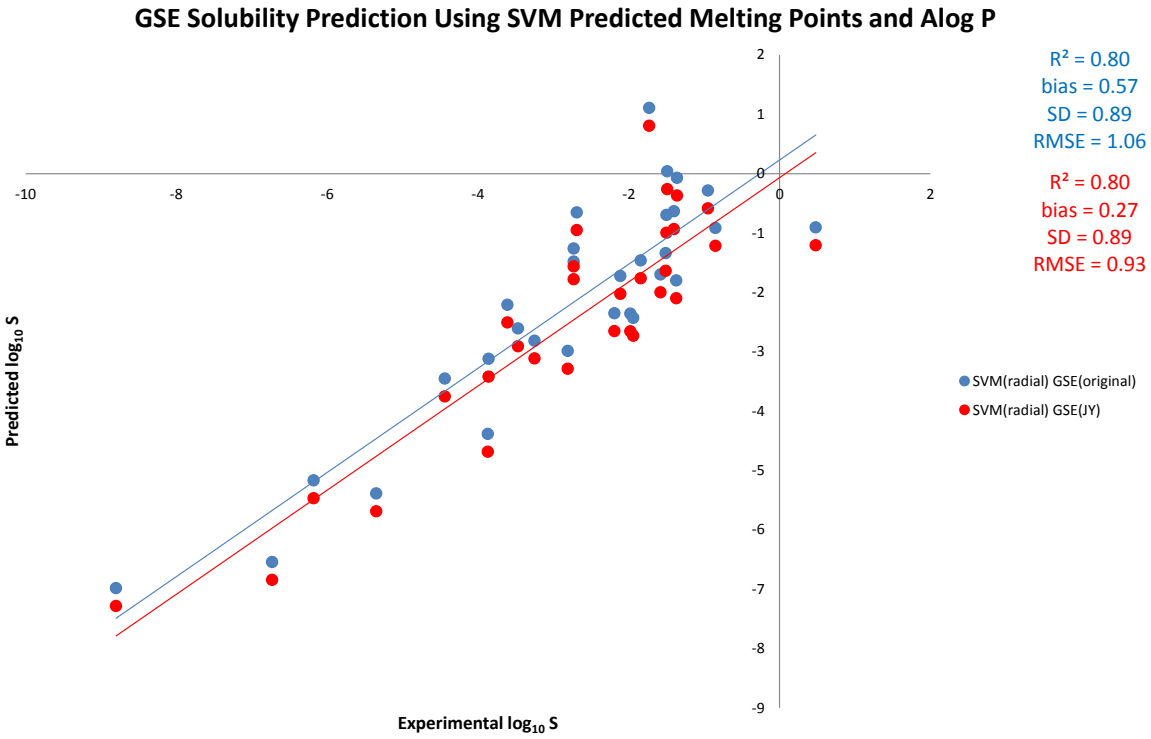


Figure 5.7: A prediction of solubility for the DLS-30 molecules using the general solubility equation with predicted melting points from SVM and predicted logP from AlogP.

The standard deviation of the experimental data is 1.95 logS units. **Table 5.1** summarises the predictive accuracy of each method. The reparametrisation has a

notable effect on the bias where we see consistent reduction in the systematic error compared to the original parametrisation. The predictions of solubility reach the chemical accuracy level, of approximately 1 logS unit RMSE, which is a promising result for a relatively simple model.

These results present the improvement between the two forms of the GSE. The original parametrisation, whilst still performing well in all cases, embodies a much larger systematic error. The reparametrised GSE has a much diminished systematic error leading to an overall improvement in predictive performance. The model utilising PLS predicted melting points performs very well here with low systematic and random errors, leading to an RMSE meeting the chemical accuracy target even with the original form of the GSE.

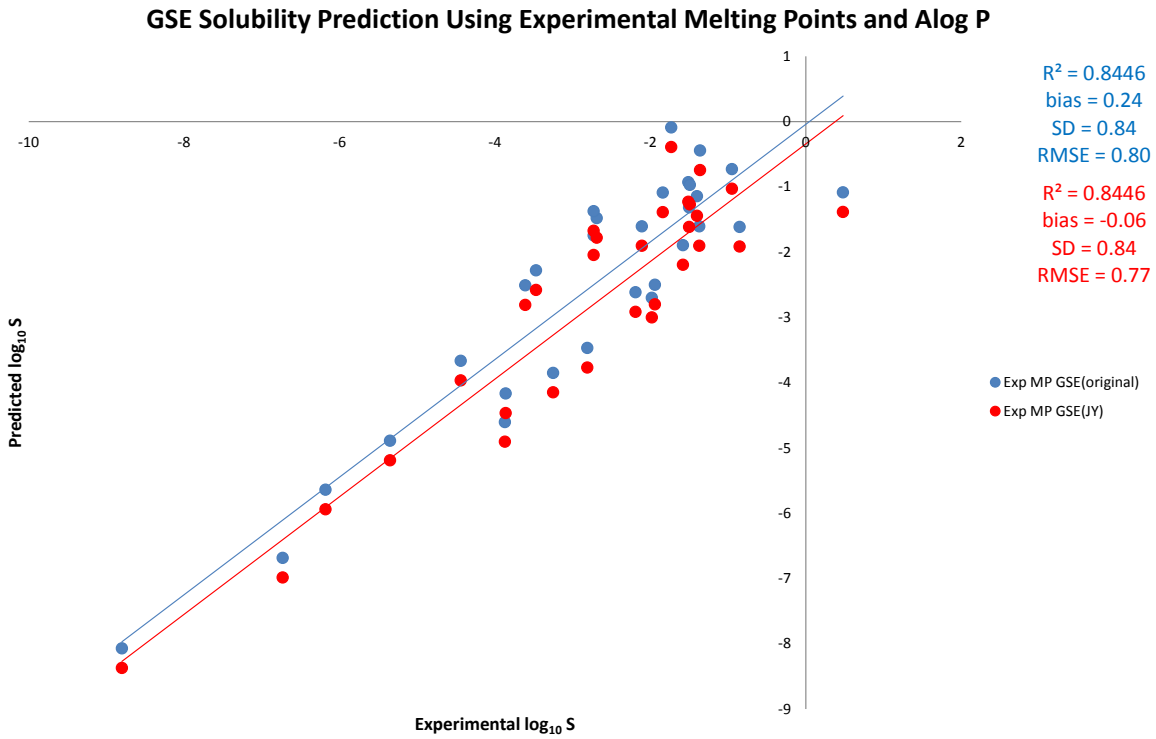


Figure 5.8: A prediction of solubility for the DLS-30 molecules using the general solubility equation with experimental melting points and predicted logP from AlogP.

Figure 5.8 shows the result when the experimental melting points are used in place of the predicted melting points. As would be expected, we see a reduction in the bias, systematic error, and an overall improvement in the RMSE scores. Using either form of the GSE now easily meets the chemical accuracy criteria. This is a good result, providing intuitive confirmation that the method is functioning as expected, showing improvements when presented with more accurate data. When the reparametrised GSE is used we can see the bias reduced to almost zero. This suggests there is little systematic error in the model, providing reassurance that the logP predictions from the AlogP model are also fairly accurate.

Molecule	Exp MP ($^{\circ}C$)	Exp (logS)	AlogP	GSE_{JY}^{Exp} (logS)	PLS MP ($^{\circ}C$)	GSE_{JY}^{PLS} (logS)	RF MP ($^{\circ}C$)	GSE_{JY}^{RF} (logS)	SVM MP ($^{\circ}C$)	GSE_{JY}^{SVM} (logS)
1,3,5-trichlorobenzene	63.50	-4.44	4.08	-3.97	28.30	-3.61	36.96	-3.70	42.39	-3.75
1-Naphthol	96.00	-1.98	2.79	-3.00	75.30	-2.79	38.66	-2.43	61.80	-2.66
4-Aminobenzoic acid	187.50	-1.37	0.78	-1.91	144.60	-1.48	147.66	-1.51	207.01	-2.10
5,5-Diphenylhydantoin	295.50	-3.86	2.26	-4.47	223.95	-3.75	174.20	-3.25	191.13	-3.42
Acetanilide	114.50	-1.40	1.05	-1.45	75.31	-1.05	59.92	-0.90	63.45	-0.93
Adenosine	235.00	-1.73	-1.21	-0.39	240.39	-0.44	181.05	0.15	115.41	0.81
Antipyrine	112.50	0.48	1.01	-1.39	102.40	-1.28	79.32	-1.05	94.70	-1.21
Benzamide	127.00	-0.95	0.51	-1.03	93.34	-0.69	89.24	-0.65	82.50	-0.58
Benzoic acid	122.50	-1.58	1.72	-2.20	85.96	-1.83	112.06	-2.09	102.85	-2.00
Chloramphenicol	150.50	-2.11	1.15	-1.91	183.22	-2.23	178.91	-2.19	162.48	-2.02
Flufenamic acid	134.00	-5.35	4.60	-5.19	178.95	-5.64	165.02	-5.50	183.83	-5.69
Griseofulvin	219.00	-3.25	2.71	-4.15	139.13	-3.35	127.36	-3.23	115.41	-3.11
Hydrochlorothiazide	269.00	-2.69	-0.16	-1.78	317.53	-2.27	190.79	-1.00	186.21	-0.95
Nalidixic acid	229.00	-3.61	1.27	-2.81	203.54	-2.56	167.32	-2.19	198.89	-2.51
Nicotinic acid	237.50	-0.85	0.29	-1.92	110.59	-0.65	135.54	-0.90	167.72	-1.22
Papaverine	146.50	-3.87	4.19	-4.91	113.66	-4.58	138.88	-4.83	124.55	-4.69
Perylene	278.00	-8.80	6.34	-8.37	151.23	-7.10	166.00	-7.25	169.45	-7.28
Pyrene	150.00	-6.18	5.19	-5.94	108.61	-5.53	131.49	-5.75	102.86	-5.47
Quinidine	170.00	-2.81	2.82	-3.77	122.24	-3.29	158.15	-3.65	121.73	-3.29
Salicylamide	140.00	-1.84	0.74	-1.39	148.96	-1.48	161.31	-1.60	177.31	-1.76
Salicylic acid	159.00	-1.94	1.96	-2.80	126.83	-2.48	151.54	-2.73	151.97	-2.73
Sulfacetamide	183.00	-1.51	0.15	-1.23	188.22	-1.28	155.26	-0.95	223.86	-1.64
Sulfadiazine	254.50	-2.73	0.25	-2.05	209.12	-1.59	168.86	-1.19	206.35	-1.56
Sulfamethazine	199.50	-2.73	0.43	-1.68	206.56	-1.75	187.30	-1.55	210.09	-1.78

Continued on next page

Table 5.1 – *Continued from previous page*

Molecule	Exp MP ($^{\circ}C$)	Exp (logS)	AlogP	GSE_{JY}^{Exp} (logS)	PLS MP ($^{\circ}C$)	GSE_{JY}^{PLS} (logS)	RF MP ($^{\circ}C$)	GSE_{JY}^{RF} (logS)	SVM MP ($^{\circ}C$)	GSE_{JY}^{SVM} (logS)
Sulfanilamide	165.50	-1.36	-0.16	-0.75	161.27	-0.70	153.04	-0.62	127.98	-0.37
Thymine	316.50	-1.50	-0.80	-1.62	154.44	0.01	188.64	-0.34	254.67	-1.00
Thymol	50.50	-2.19	3.16	-2.92	22.66	-2.64	44.09	-2.85	24.33	-2.65
Tolbutamide	129.00	-3.47	2.04	-2.58	140.45	-2.69	144.96	-2.74	161.97	-2.91
Triphenylene	196.50	-6.73	5.77	-6.99	127.13	-6.29	159.55	-6.62	182.53	-6.85
Uracil	330.00	-1.49	-1.28	-1.27	158.73	0.44	185.28	0.18	229.04	-0.26
RMSE				0.77		0.86		0.95		0.93
R^2				0.84		0.83		0.80		0.80
σ		1.95		0.84		0.80		0.88		0.89
Bias				-0.06		0.31		0.36		0.27

Table 5.1: All data relating to predictions of melting point, AlogP and logS using the reparameterised version of the GSE.

Molecule name	PLS (|exp-pred|)	RF (|exp-pred|)	SVM (|exp-pred|)
1,3,5-trichlorobenzene	35.2	26.54	21.11
1-Naphthol	20.7	57.34	34.2
4-Aminobenzoic acid	42.9	39.84	19.51
5,5-Diphenylhydantoin	71.55	121.3	104.37
Acetanilide	39.19	54.58	51.05
Adenosine	5.39	53.95	119.59
Antipyrine	10.1	33.18	17.8
Benzamide	33.66	37.76	44.5
Benzoic acid	36.54	10.44	19.65
Chloramphenicol	32.72	28.41	11.98
Flufenamic acid	44.95	31.02	49.83
Griseofulvin	79.87	91.64	103.59
Hydrochlorothiazide	48.53	78.21	82.79
Nalidixic acid	25.46	61.68	30.11
Nicotinic acid	126.91	101.96	69.78
Papaverine	32.84	7.62	21.95
Perylene	126.77	112	108.55
Pyrene	41.39	18.51	47.14
Quinidine	47.76	11.85	48.27
Salicylamide	8.96	21.31	37.31
Salicylic acid	32.17	7.46	7.03
Sulfacetamide	5.22	27.74	40.86
Sulfadiazine	45.38	85.64	48.15
Sulfamethazine	7.06	12.2	10.59
Sulfanilamide	4.23	12.46	37.52
Thymine	162.06	127.86	61.83
Thymol	27.84	6.41	26.17
Tolbutamide	11.45	15.96	32.97
Triphenylene	69.37	36.95	13.97
Uracil	171.27	144.72	100.96
Average	48.25	49.22	47.44

Table 5.2: The absolute differences between the experimental and predicted melting points from PLS, RF and SVM.

From **Tables 5.1, 5.2** and **Figures 5.5 - 5.8**, we can see that predictions of the solubility of the DLS-30 dataset are good. Although this is once again a small dataset, the results agree with previous work showing good solubility prediction using the GSE.[215,216,222] This work further shows that good predictions of solubility are possible even having been given a melting point of fairly low quality. We show above that our predictions of melting point are sufficient, at least on this small dataset, to produce useful predictions of solubility. All of the methods here have

an RMSE within the standard deviation of the experimental data (1.95 logS units), hence, they fulfil our criteria as a useful model of solubility. Further, all models based upon the reparametrised GSE meet the chemical accuracy target of approximately 1 $logS$ unit. In this case the GSE model using the PLS predicted melting points produced the best prediction (RMSE=0.86 R^2=0.83).

It is interesting to note that in **Table 5.2**, the absolute errors in the melting point predictions are in some cases in excess of $100°C$, and yet still the predictions of solubility are reasonably accurate. This suggests the parametrisation of the GSE to be robust against even very poor predictions of the melting point; a useful asset given the difficulty often faced of obtaining reliable experimentally measured melting points. This is likely to be down to the fact that the logP term will dominate many of the logS predictions as only 1% of the predicted melting point value minus $25°C$ enters the GSE equation (**Equation 5.1**).

If we now test the effect and reliance of these models to a change in logP predictions, by replacing the AlogP predictions with XlogP predictions, we obtain the results shown in **Figures 5.9 - 5.12**.

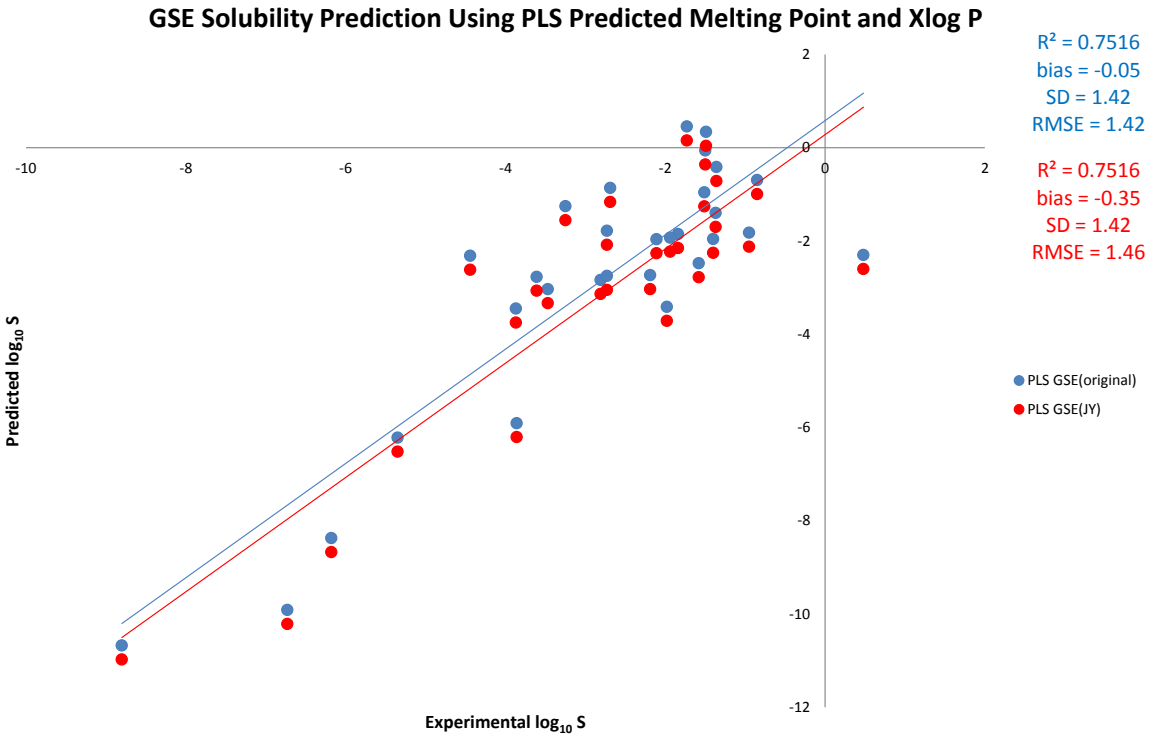


Figure 5.9: A prediction of solubility for the DLS-30 molecules using the general solubility equation with melting points from PLS and predicted logP from XlogP.

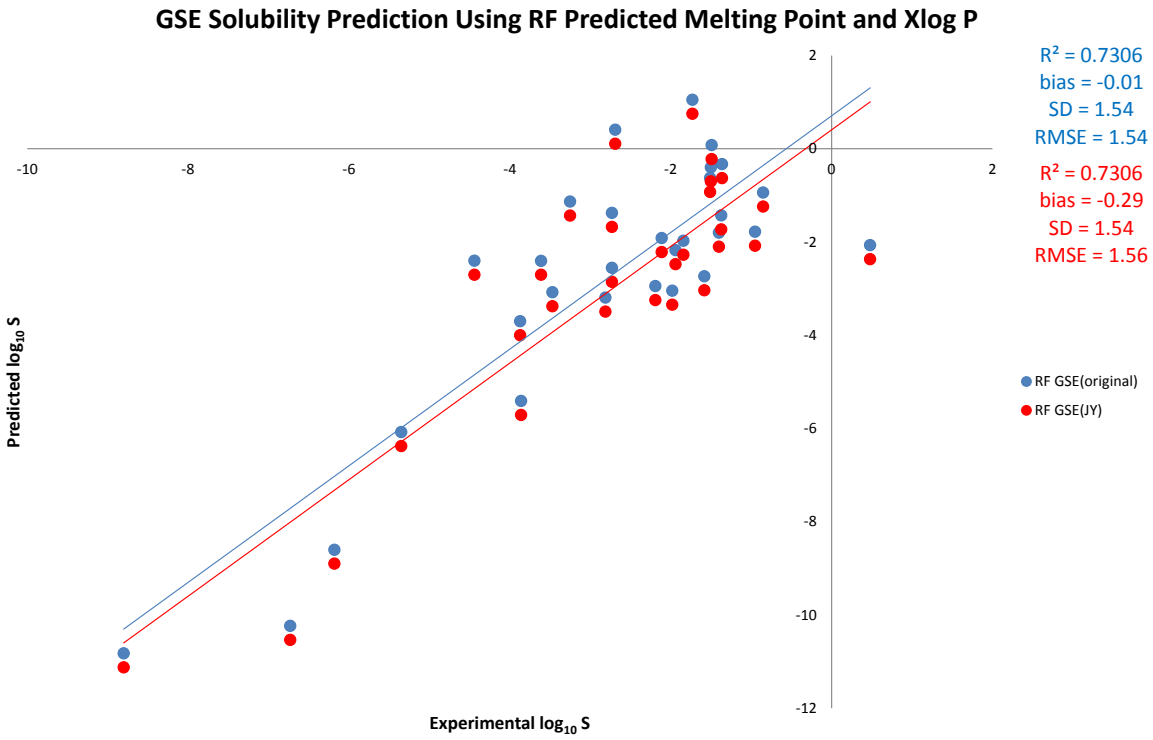


Figure 5.10: A prediction of solubility for the DLS-30 molecules using the general solubility equation with melting points from RF and predicted logP from XlogP.

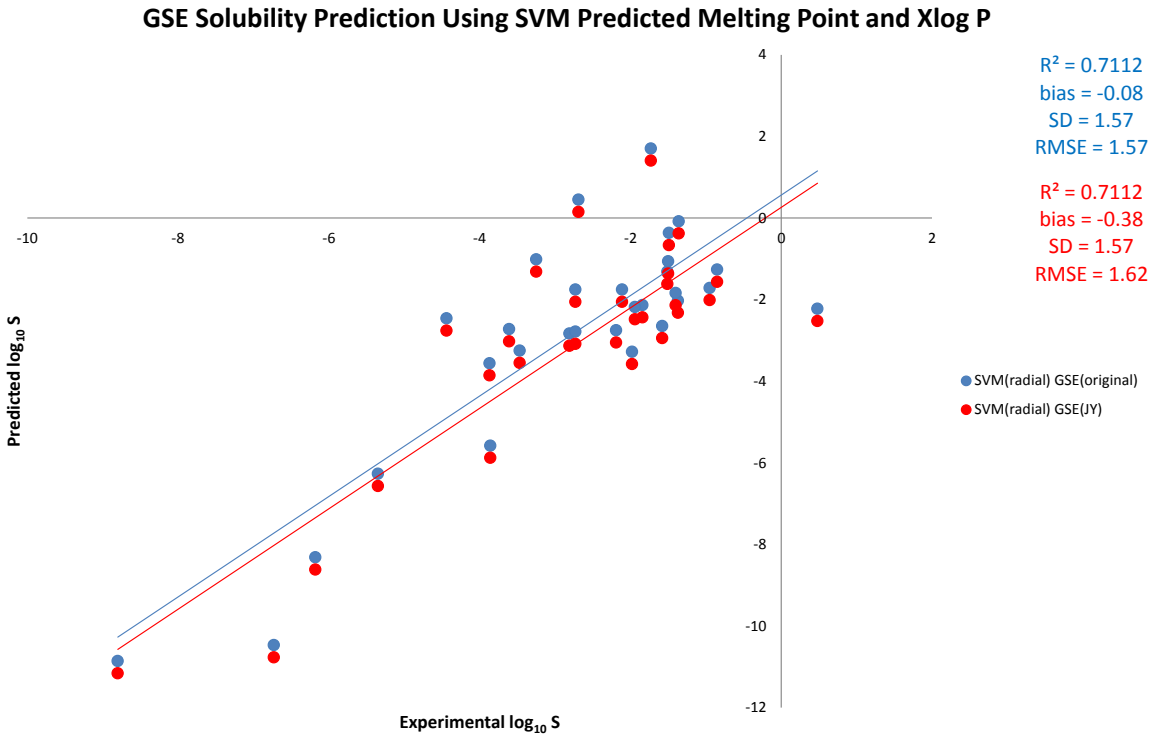


Figure 5.11: A prediction of solubility for the DLS-30 molecules using the general solubility equation with melting points from SVM and predicted logP from XlogP.

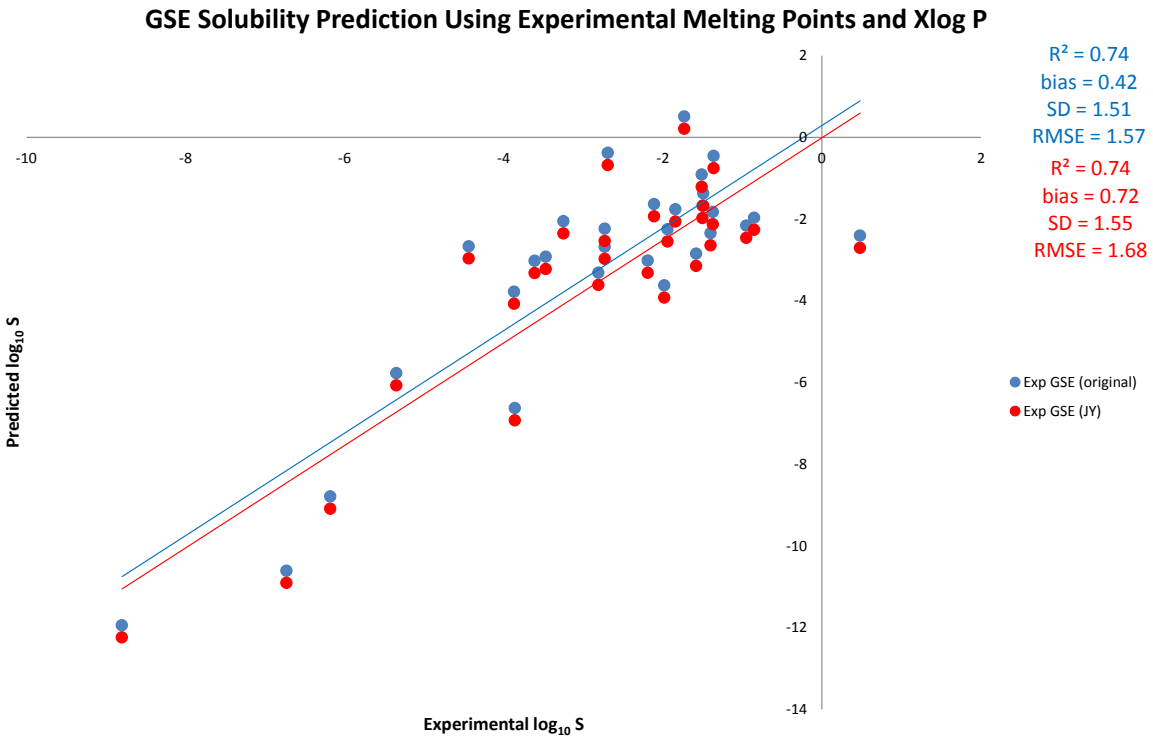


Figure 5.12: A prediction of solubility for the DLS-30 molecules using the general solubility equation with experimental melting points and predicted logP from XlogP.

These plots show a reversal in the optimum parametrisation, so favour the original GSE. We see an increase in the systematic error on going from the original to the reparametrisation and therefore a worsening RMSE score. The RMSE scores are in all cases worse than the RMSE scores calculated when using logP values predicted by the AlogP method, for otherwise equivalent models of solubility. In addition, we find the non-intuitive result of a worse prediction, in terms of RMSE, when the predicted melting points are substituted by the experimentally derived melting points.

From these results we conclude that the model GSE's accuracy is very dependent on the source of the empirical data. If experimental values are provided for both empirical parameters, then previous work has shown that the resultant predictions are generally good. We see here that there is some unpredictability in the use of calculated data in the GSE. We find models that appear for a small test set to make good predictions of solubility from qualitative predictions of melting point and the well known and widely used logP prediction method AlogP. We also find that, as often the logP term dominates the logS prediction, the GSE is sensitive to the choice of logP prediction algorithm. The results of selecting a poor fitting logP model are here shown to be non-intuitive with reasonably poor predictions of solubility for a QSAR/QSPR model.

5.2 Conclusions from the GSE

From this work we find that it is beyond the capabilities of simple machine learning models, utilising a modest number of 2D chemical descriptors, to predict chemically useful and accurate melting points. We can conclude that based on this it would be difficult to make a simple model of solubility, due to inconsistent predictions of melting points. However, on a small dataset, with errors in the melting point prediction on average of approximately 50°C, we can still make a good prediction of solubility although further testing would be required to check its broader applicability. The GSE appears usefully well adapted and parametrised to deal with poor quality melting points, therefore, allowing useful predictions of solubility to be made despite sizeable errors in the predicted melting points.

Chapter 6

Sublimation Thermodynamics

"If someone points out to you that your pet theory of the universe is in disagreement with Maxwell's equations – then so much the worse for Maxwell's equations. If it is found to be contradicted by observation – well, these experimentalists do bungle things sometimes. But if your theory is found to be against the second law of thermodynamics I can give you no hope; there is nothing for it but to collapse in deepest humiliation."

Sir Arthur Eddington, 1915

6.1 Predicting Sublimation Thermodynamics

In this chapter we introduce extended predictions of ΔH_{sub}, ΔS_{sub} and ΔG_{sub} by DMACRYS and machine learning. To conclude the chapter we present predictions of ΔH_{sub} from first principles calculation. These predictions were carried out using a new dataset.

6.1.1 A New Sublimation Dataset

Following from the work presented in **Chapter 3** we decided to investigate predictions of sublimation thermodynamics, using a larger dataset. A dataset of chemically diverse, organic crystals was curated by searching the literature for data which met our criteria:

- Experimental values for ΔH_{sub}, ΔS_{sub} and ΔG_{sub} must be available or calculable from a single literature source.
- Where possible a crystal structure must be available in the CSD.
- Where possible the literature should have a record of the polymorph used in the experiment.

Applying the first of these criteria provided 158 molecules after a search of appropriate literature. Applying the second criteria removed 62 of these molecules, leaving a dataset of 96 molecules. Applying the final criterion would have reduced the dataset to 3. Therefore, we took the approach of minimising the energy of all potential polymorphs of a given molecule, and taking the one with the lowest lattice energy to be the most stable, hence major contributory form to the sublimation thermodynamics. As this process was extremely time consuming we opted to reduce the dataset size to a more manageable 60-molecule set. This set was selected on the basis of the best available crystal structures. Upon minimisation we found repeated convergence failures in several structures. A variety of parameter variations were attempted but were ultimately unsuccessful. The dataset was therefore reduced to 48 molecules. **Table 6.2** shows the reduced dataset, which will be denoted SUB-48.

6.1.2 Sublimation Thermodynamics: Predictions by DMACRYS

Following from our previous work, in **Chapter 3**, we wished to investigate how accurately we could predict sublimation thermodynamics over a larger dataset using computationally efficient methods. We proceeded using the methods found in **Chapter 3**, this time making predictions over the SUB-48 dataset.

We consider here the ΔH_{sub} predictions. These predictions were made applying distributed multipoles, calculated at the B3LYP/6-31G** level of theory, as a model of the electrostatics, and the FIT potential parameters, in the form of a Buckingham potential, to model the repulsion and dispersion.

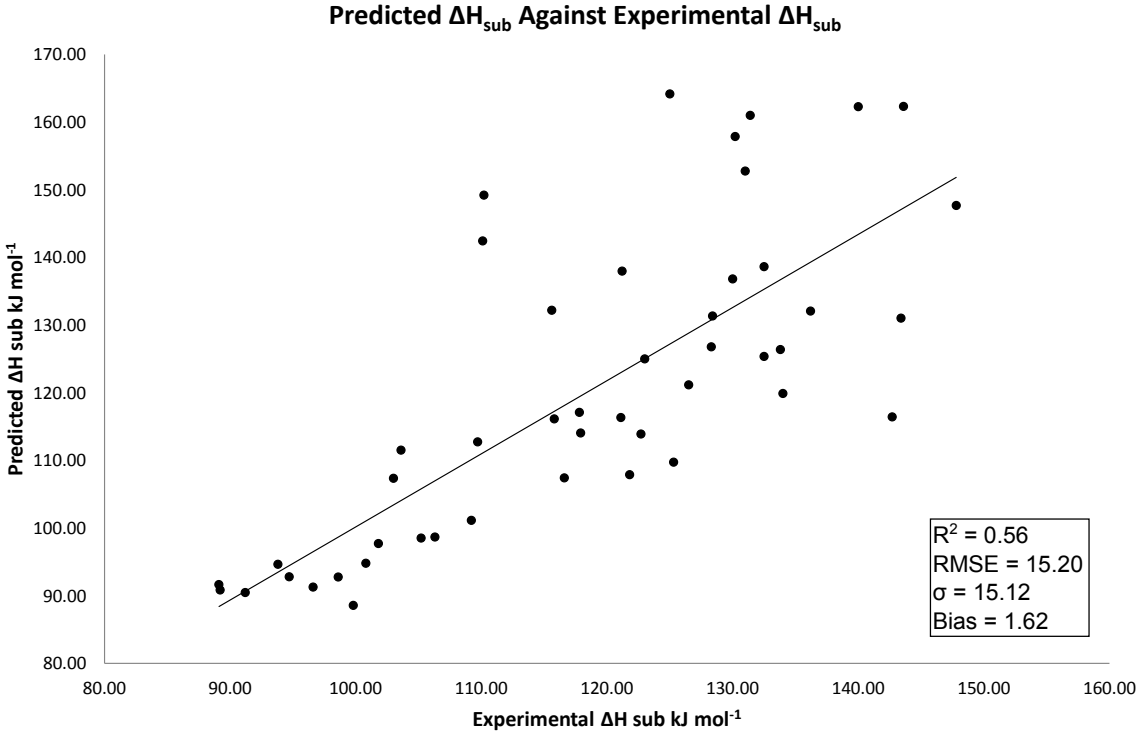


Figure 6.1: Predicted ΔH_{sub} from DMACRYS against experiment.

We can see from **Figure 6.1** that the prediction is fairly poor. The standard deviation of the experimental data is 15.94 kJ/mol, just above the RMSE of the the predicted values. We can therefore classify this as a useful prediction given our previous criteria, however, for practical purposes the predictions are not sufficiently accurate. A reasonable positive correlation exists. We note that the bias and σ suggest that the vast majority of the error is due to random errors which we cannot expect the model to reasonably resolve; only a small amount (approximately 11% of the RMSE) is due to systematic errors. This suggests a large amount of seemingly random variation, which cannot be explained by the model, is incorporated in the data.

Secondly, we made predictions of ΔS_{sub} using the same model as used for ΔH_{sub}. **Figure 6.2** shows a comparison of the calculated and experimental $T\Delta S_{sub}$ data. These predictions are made using the rigid-body approximation, hence ignoring intramolecular contributions.

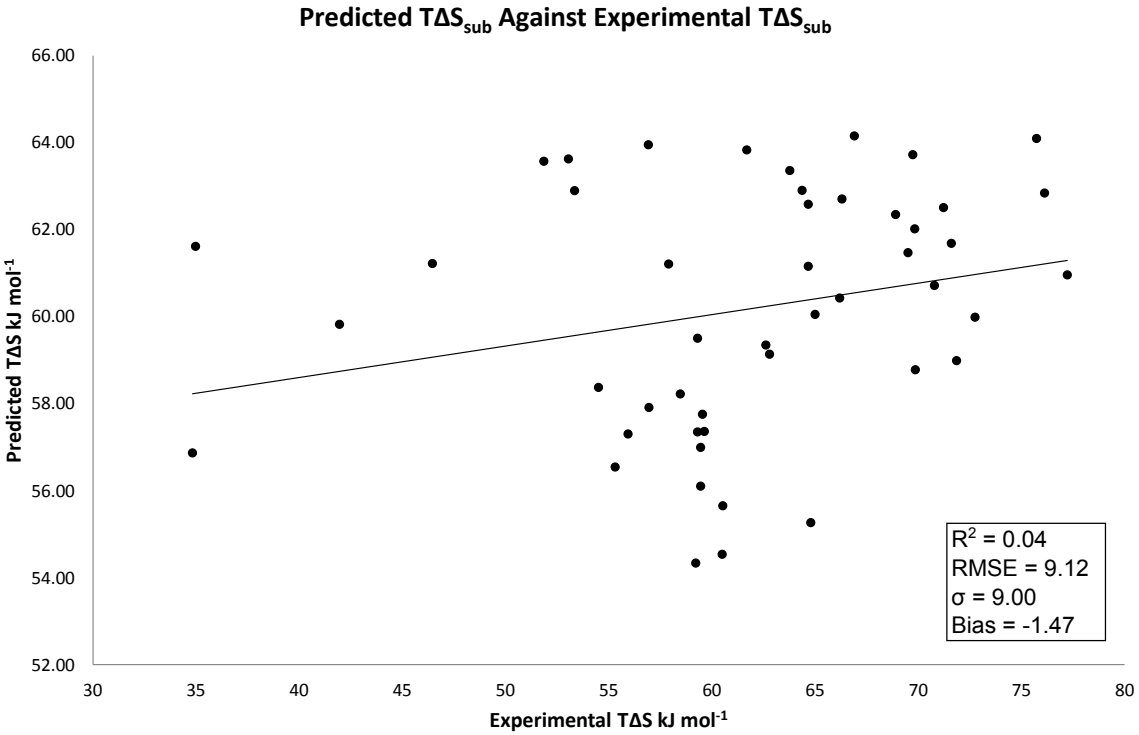


Figure 6.2: Predicted ΔS_{sub} from DMACRYS against experiment.

It is immediately clear that there is no correlation between the predicted and experimental $T\Delta S_{sub}$; we additionally note the similarity with the comparison of the computed and experimental enthalpy data, in that the RMSE is just inside the experimental σ ($9.35 kJmol^{-1}$) and the majority of the RMSE can be attributed to random errors. These results suggest that either some important contribution is being neglected in our model or that there is a far larger error in the experimental data than is quoted in the literature. In either scenario the current level of predictive accuracy is often tens of $kJmol^{-1}$ off.

In an effort to elucidate the origin of the errors, we attempted to correlate the entropy values with other physical properties that one may expect to be correlated with

entropy, including molecular weight and the number of rotatable bonds. **Figures 6.3, 6.4** and **Table 6.1** show the correlations between molecular weight and the number of rotatable bonds with the individual entropy contributions, respectively.

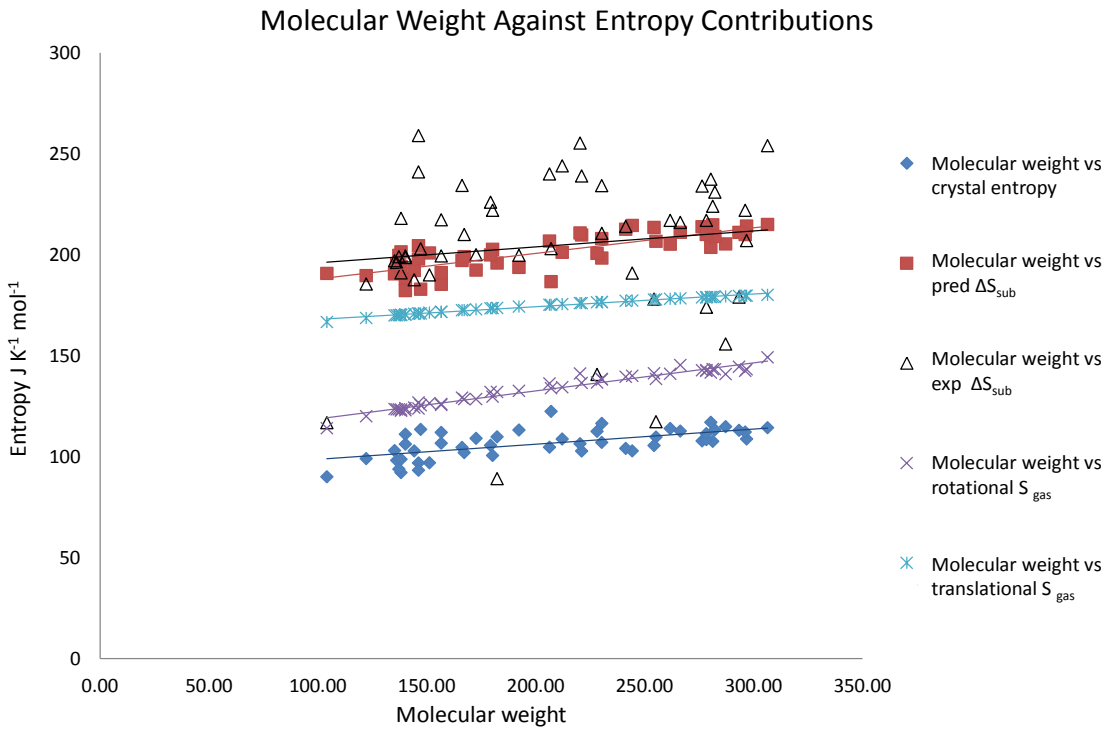


Figure 6.3: Molecular weight against entropy.

Entropy Contribution	R^2
Crystal Entropy	0.45
ΔS_{sub}	0.42
Experimental Entropy	0.01
Rotational Entropy	0.95
Translational Entropy	0.99

Table 6.1: Correlation coefficients for plots of molecular weight against entropy.

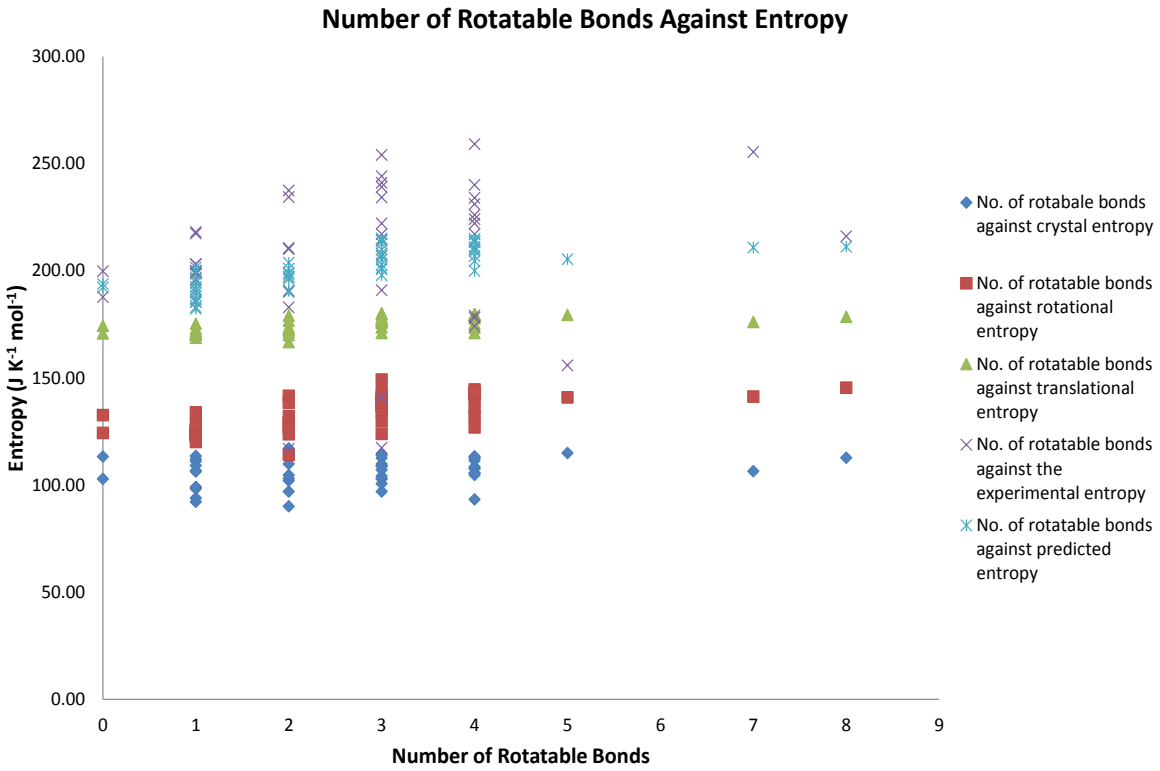


Figure 6.4: Number of rotatable bonds against entropy.

From these plots we can see that the calculated entropy contributions correlate with molecular weight and the number of rotatable bonds, as one would expect, however, the experimental entropy does not. As entropy is a measure of system disorder, one might expect that the molar mass and number of rotatable bonds to correlate with entropy as heavier molecules and molecules with more rotatable bonds generally have an increased number of degrees of freedom.

Finally, the predictions of ΔG_{sub} are presented. These are simply calculated as the difference between ΔH_{sub} and $T\Delta S_{sub}$. **Figure 6.5** below shows a weak correlation and again an RMSE score dominated by random error. In this case though, the RMSE of the prediction is well outside the σ of the experimental data, which is $15.61 kJ/mol$. We therefore find that the predictions of ΔG_{sub} fail our criteria of a useful prediction. Additionally, these values are far too large for practical quantitative use.

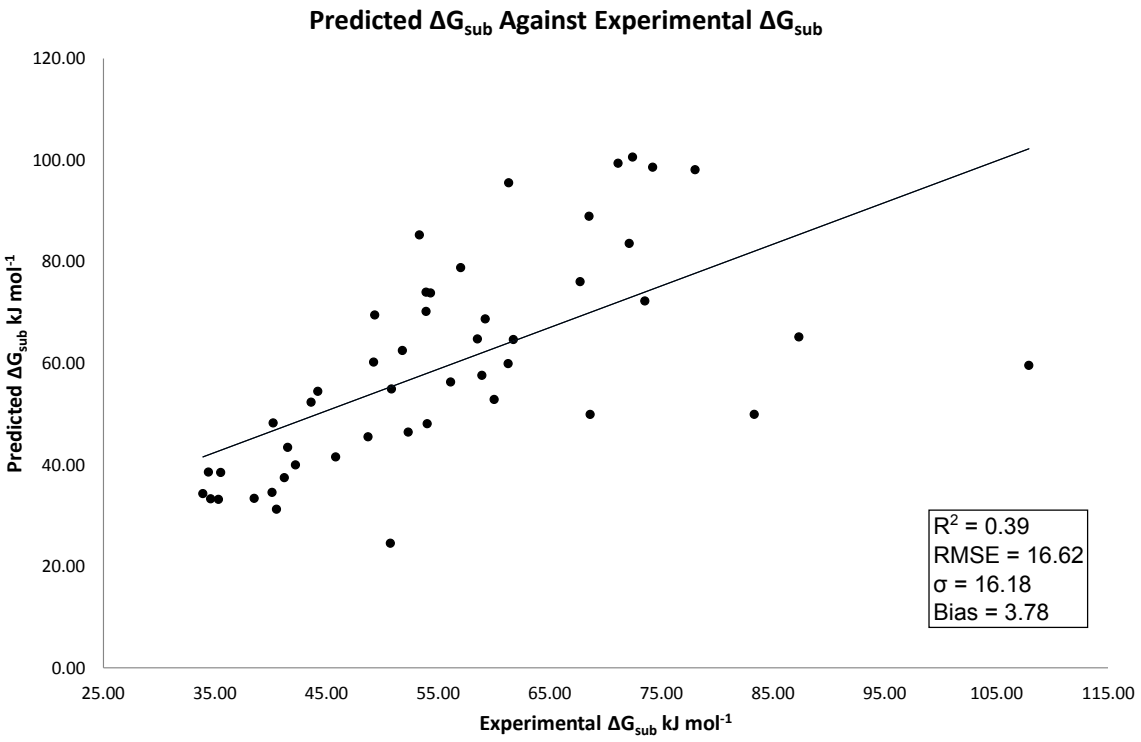


Figure 6.5: Predicted ΔG_{sub} from DMACRYS against experiment.

From these results it is likely that important contributions to ΔG_{sub} are missing in our model. It is also possible that the potential errors in the experimental results are underestimated. Note that these experimental determinations are difficult to carry out. **Table 6.2** shows the raw predicted and experimental data. **Appendix H, Table H.1** contains the full experimental data with the reported error margins quoted.

Chemical Name	Crystal Structure	T(K)	ΔH_{sub}^{exp} (kJ/mol)	ΔH_{sub}^{Pred} (kJ/mol)	$T\Delta S_{sub}^{Exp}$ (kJ/mol)	$T\Delta S_{sub}^{Pred}$ (kJ/mol)	ΔG_{sub}^{Exp} (kJ/mol)	ΔG_{sub}^{Pred} (kJ/mol)
Acetanilide	ACANIL01[223]	301.02	99.8	88.57	59.3	57.35	40.5	31.77
aspirin (acetylsalicylic acid)	ACSALA15[224]	298.15	109.7	112.76	66.19	60.43	43.6	52.33
Benzoic acid	BENZAC01[225]	298.15	89.2	90.85	55.31	56.54	33.9	34.30
isophthalic acid	BENZDC01[226]	298.15	143.4	131.02	69.86	58.78	73.5	72.25
2,4-Dinitrobenzoic acid	BIPJUF[227]	298.15	134	119.89	72.75	59.99	61.25	59.90
Benzophenone	BPHENO03[228]	298	89.1	91.65	54.5	58.37	34.6	33.25
atenolol	CEZVIN[229]	298	140	162.28	64.37	62.9	71.1	99.35
2-Chlorobenzoic acid	CLBZAC01[225]	298.15	106.3	98.67	64.79	55.26	41.5	43.41
4-Chlorobenzoic acid	CLBZAP03[225]	298.15	105.2	98.52	59.45	56.99	45.8	41.53
2,6-Dichloro-4-nitroaniline	CLNOAN[230]	298.15	109.2	101.14	60.52	55.65	48.7	45.49
3-Fluorobenzoic acid	COVJIG[225]	298.15	93.8	94.65	59.45	56.1	34.4	38.54
naproxen	COYRUD11[6]	298.15	128.3	126.79	69.83	62.02	58.5	64.77
2,4-dichlorophenoxy acetic acid	CPXACA[231]	298	123	125	71.22	62.51	51.78	62.46
1,8-diphenylnaphthalene	DPNAPH01[232]	298.15	132.5	125.37	70.78	60.72	61.72	64.65
2-Fluorobenzoic acid	FBENZA02[225]	298.15	94.7	92.8	59.21	54.33	35.5	38.46
flurbiprofen	FLUBIP[233]	298.15	110.2	149.21	56.92	63.95	53.3	85.25
Flufenamic acid	FPAMCA17[234]	298.66	121.2	137.98	66.9	64.16	54.3	73.93
2,4,6-trinitromesitylene (TNM)	HEXTIN01[235]	298.15	103.6	111.52	34.97	61.62	68.6	49.90

Continued on next page

Table 6.2 – *Continued from previous page*

Chemical Name	Crystal Structure	T(K)	ΔH_{sub}^{exp} (kJ/mol)	ΔH_{sub}^{Pred} (kJ/mol)	$T\Delta S_{sub}^{Exp}$ (kJ/mol)	$T\Delta S_{sub}^{Pred}$ (kJ/mol)	ΔG_{sub}^{Exp} (kJ/mol)	ΔG_{sub}^{Pred} (kJ/mol)
4-amino-N-(4-ethylphenyl)-benzenesulfonamide	HUNXIY[236]	298	143.6	162.32	69.73	63.72	74.2	98.56
4-Amino-N-(4-methoxyphenyl)-benzenesulfonamide	HUNXOE[236]	298	125	164.16	51.85	63.57	72.4	100.56
4-Amino-N-(5-chloro-2-methylphenyl)-benzenesulfonamide	HUNXUK[236]	298	131	152.77	61.69	63.83	68.5	88.90
Paracetamol	HXACAN27[237]	304.74	117.9	114.05	57.9	61.21	60	54.17
Ibuprofen	IBPRAC01[224]	298.33	115.8	116.13	71.6	61.69	44.2	54.49
4-hydroxybenzoic acid	JOZZIH[238]	298.15	121.1	116.34	65	60.05	56.1	56.28
Tolfenamic acid	KAXXAI01[239]	298	128.4	131.35	64.67	61.16	53.9	70.16
Ketoprofen	KEMRUP[228]	298	110.1	142.44	53.04	63.63	57	78.78
9-Methylanthracene	MANTHR01[240]	298.15	101.8	97.71	59.54	57.76	42.2	39.96
2,4,6-N-tetranitro-N-methylaniline	MTNANL[235]	298.15	133.8	126.38	46.45	61.22	87.3	65.16
1-naphthol	NAPHOL01[6]	298.15	91.2	90.47	55.93	57.3	35.3	33.17
4-nitrobenzoic acid	NBZOAC01[238]	298.15	116.6	107.42	62.61	59.35	54	48.07
Niflumic acid	NIFLUM10[234]	298.27	130.2	157.87	68.9	62.35	61.3	95.55
nitroguanidine	NTRGUA01[235]	298.15	142.7	116.43	34.82	56.87	107.9	59.56
2,2-Dimethylsuccinic acid	OLENIC[241]	298.15	122.7	113.9	71.85	58.99	50.8	54.91
4-Methylbenzoic acid	PTOLIC01[242]	298.15	98.6	92.76	58.47	58.22	40.1	34.54
Phenacetin	PYRAZB21[223]	307.52	121.8	107.89	69.5	61.47	52.3	48.29

Continued on next page

Table 6.2 – *Continued from previous page*

Chemical Name	Crystal Structure	T(K)	ΔH_{sub}^{exp} (kJ/mol)	ΔH_{sub}^{Pred} (kJ/mol)	$T\Delta S_{sub}^{Exp}$ (kJ/mol)	$T\Delta S_{sub}^{Pred}$ (kJ/mol)	ΔG_{sub}^{Exp} (kJ/mol)	ΔG_{sub}^{Pred} (kJ/mol)
5-chloro-2-nitroaniline	RAPKUP[243]	298.15	100.8	94.8	59.63	57.36	41.2	37.44
salicylic acid	SALIAC15[244]	298.15	96.6	91.28	56.95	57.91	38.5	33.37
Diclofenac	SIKLIH01[245]	298.65	115.6	132.19	66.3	62.71	49.3	69.59
4-cyanobenzoic acid	TAGNAR[238]	298.15	111.2	79.06	60.49	54.54	50.7	24.52
o-Terphenyl	TERPHO02[225]	298.15	103	107.36	62.79	59.14	40.2	48.22
2,4,6-trinitroaniline (TNA)	TNIOAN[235]	298.15	125.3	109.74	41.95	59.82	83.3	49.91
1,3,5-Triphenylbenzene	TPHBEN01[246]	298.15	147.8	147.68	75.73	64.1	72.1	83.58
N-(4-nitrophenyl)-benzene-sulfonamide	UVEMOY[247]	298	132.5	138.63	64.67	62.58	67.7	76.02
4-Amino-N-(4-nitrophenyl)benzenesulfonamide	UVEMUE[247]	298	131.4	161	53.34	62.9	78	98.07
4-Hydroxybenzamide	VIDMAX[248]	298	117.8	117.1	59.3	59.5	58.9	57.57
2-Methylglutaric acid	XIBVIO[241]	298.15	126.5	121.16	77.22	60.96	49.2	60.21
Mefenamic acid	XYANAC[239]	298	136.2	132.07	63.77	63.36	59.2	68.68
4-Heptylbenzoic acid (cr, II)	ZIKWOF[242]	298.15	130	136.82	76.12	62.84	53.9	73.98
R^2				0.56		0.04		0.39
RMSE				15.20		9.12		16.62
σ			15.94	15.12	9.35	9.00	15.61	16.18
Bias				1.62		-1.47		3.78

Table 6.2: Experimental and predicted sublimation data.

In an effort to improve these models we applied machine learning to ΔG_{sub} prediction data. In a methodology analogous to that laid out previously (**Section 4.3.2**), we supply the following descriptor sets independently, to train and test three machine learning models, SVM, RF and PLS. The descriptor sets are: Thermodynamic values, CDK 2D molecular descriptors calculated from SMILES strings or a combination of the two. A single scaling method auto-scaling was applied. **Tables 6.3, 6.4** and **6.5** summarise the results from these methods.

Thermodynamic values	RF	PLS	SVM
R^2	0.29±0.03	0.3±0.02	0.17±0.03
RMSE	13.22±0.32	13.24±0.31	14.48±0.52

Table 6.3: ΔG_{sub} predicted by machine learning using the theoretical chemistry predictions of sublimation thermodynamics as descriptors, RMSE given in (kJ/mol).

CDK 2D Descriptors	RF	PLS	SVM
R^2	0.39±0.04	0.43±0.03	0.49±0.03
RMSE	12.12±0.38	11.65±0.3	11.05±0.31

Table 6.4: ΔG_{sub} predicted by machine learning using the CDK 2D molecular descriptors, RMSE given in (kJ/mol).

Coupled Descriptors	RF	PLS	SVM
R^2	0.47±0.03	0.57±0.03	0.54±0.05
RMSE	11.29±0.25	10.18±0.31	10.58±0.65

Table 6.5: ΔG_{sub} predicted by machine learning applying a combined descriptor set as descriptors, RMSE given in (kJ/mol).

The results show a clear improvement in all cases over the predictions made using exclusively theoretical chemical methods. In all cases we now meet the criteria we set out for a useful prediction; all methods have an RMSE within the standard deviation of the experimental data ($\sigma_{exp} = 15.61 kJ/mol$). As with solubility, the 131 molecular descriptors seem to convey more information to the model than the 9 theoretical chemistry thermodynamic values. However, contrary to the solubility results it seems in this case that the two descriptor sets are complementary; much of the variance is explained once the descriptor sets are combined. We see a sizeable reduction in the RMSE and an improvement in R^2 for these models. These are encouraging results providing a significant improvement in accuracy. However, despite these improvements the predictive accuracy is still low, with sizeable RMSE values still found after coupling the descriptors.

We have also applied analogous methods for ΔS_{sub} and ΔH_{sub} predictions, with promising predictions for ΔH_{sub} and inconclusive results for ΔS_{sub} . These results are presented in tables which can be found in **Appendix K**. These findings support the predictability of ΔH_{sub}. We see better predictions of ΔH_{sub} when the theoretical

energies are used in the descriptor set, with the 2D CDK descriptors offering little information. In terms of ΔS_{sub}, the opposite trend is found. SVM fails to bring the RMSE inside the experimental standard deviation in all cases. The 2D CDK descriptors provide significantly more useful information to the models and RF and PLS produce useful prediction of ΔS_{sub} using these descriptors. None of the machine learning models produce a useful prediction of ΔS_{sub} from the theoretical chemistry values. This suggests our theoretical chemistry values are poorly correlated with experimental values even when non-linear methods are applied. The descriptors consistently rated as the most important in the combined models are those related to molecular branching (weighted path), topological surfaces (TPSA), counts of specific atoms, calculated ΔG_{sub} and Kier Hall smarts (group counting based on molecular fragmentation). The number of rotatable bonds consistently features in the top twenty important descriptors, molecular weight however does not. From the data here it is not possible to find the exact origins of these errors. It is not clear whether the errors are in the experimental or calculated data. It is likely to be a combination of both approximate modelling and experimental errors.

6.1.3 ΔH_{sub} Predictions from First Principles

Having carried out predictions using parametrised minimisation methods and machine learning in an effort to find computationally efficient methods to predict sublimation thermodynamics, we decided to investigate if improvements were accessible by using higher levels of theory. Given the difficulties we found with the entropy data and questions remaining over the quoted errors, we focused on predicting the ΔH_{sub}. We applied the plane wave periodic DFT code CASTEP.[249] Given the expense of these calculations a subset of 24 molecules from the SUB-48 dataset were used. The plane wave basis set requires a cutoff value in terms of energy in order to limit its size. A suitable cutoff energy was found by converging the system energy with increasing cutoff values to a tolerance of 0.01 eV. The plane wave basis cutoff energies were converged with respect to the elements which made up the molecular units, i.e. the data set was split into molecules composed of the same atoms and the basis set was converged for one molecule from each set. This value was used for all molecules composed of the same atoms. The values are shown in **Table 6.6**. A K point grid spacing was used for all production calculations of 0.05 $Å^{-1}$.

Atoms composing the molecules	Cutoff Energy (eV)
CH	1100
CHO	1150
CHNO	1100
CHClO	1100
CHFO	1100
CHClNO	1200

Table 6.6: Plane wave cutoff convergence values

Each structure was then optimised with the PBE functional due to its previous

successful use in similar applications.[59,250] The Tkatchenko Scheffler (TS) dispersion correction was also applied. In order to predict ΔH_{sub} the energy of a single molecular unit of the crystal was required. Therefore, an optimised molecule from the crystal was extracted and manipulated using shell scripts and gaussview. The molecule was placed in a suitably large box, which was 10Å larger than the molecule in all directions, hence minimising any interaction over the periodic boundary. This gas phase molecule was then optimised in the supercell. The lattice energy (E_{latt}) was then calculated as follows:

$$E_{latt} = \left(\frac{E_{crys}}{N_{Molecular\ Units}} \right) - E_{gas} \tag{6.1}$$

Equation 6.1: Calculation of E_{latt} from CASTEP crystal optimisation and gas optimisations.

Table 6.7 presents results of predictions by CASTEP and DMACRYS along with the experimental ΔH_{sub} values. **Figures 6.6** and **6.7** summarise the results from **Table 6.7**.

Molecule chemical name	Refcode	E_{latt}^{CASTEP} (kJ/mol)	ΔH_{sub}^{CASTEP} (kJ/mol)	$U_{latt}^{DMACRYS}$ (kJ/mol)	$\Delta H_{sub}^{DMACRYS}$ (kJ/mol)	ΔH^{Exp} (kJ/mol)
aspirin	ACSALA15	-147.22	142.26	-117.72	112.74	109.70
Benzoic acid	BENZAC01	-115.05	110.09	-95.80	90.85	89.20
isophthalic acid	BENZDC01	-188.40	183.44	-135.98	131.02	143.40
2,4-Dinitrobenzoic acid	BIPJUF	-141.17	136.21	-124.85	119.89	134.00
2,6-Dichloro-4-nitroaniline	CLNOAN	-120.25	115.29	-106.10	101.14	109.20
3-Fluorobenzoic acid	COVJIG	-117.95	112.99	-99.60	94.65	93.80
naproxen	COYRUD11	-172.43	167.47	-131.74	126.79	128.30
2,4-dichlorophenoxy acetic acid	CPXACA	-144.20	139.24	-129.96	125.00	123.00
2-Fluorobenzoic acid	FBENZA02	-122.40	117.44	-97.76	92.80	94.70
flurbiprofen	FLUBIP	-175.64	170.68	-154.16	149.21	110.20
4-amino-N-(4-ethylphenyl) benzenesulfonamide	HUNXIY	-204.98	200.02	-167.28	162.32	143.60
4-Amino-N-(4-methoxyphenyl) benzenesulfonamide	HUNXOE	-196.54	191.58	-169.12	164.16	125.00
9-Methylanthracene	MANTHR01	-135.98	131.02	-102.67	97.71	101.80
2,4,6-N-tetranitro-N-methylaniline (Tetryl)	MTNANL	-129.27	124.31	-131.34	126.38	133.80
1-naphthol	NAPHOL01	-122.95	117.99	-95.43	90.47	91.20
4-Methylbenzoic acid	PTOLIC01	-128.29	123.33	-97.72	92.76	98.60

Continued on next page

Table 6.7 – *Continued from previous page*

Molecule chemical name	Refcode	E_{latt}^{CASTEP} (kJ/mol)	ΔH_{sub}^{CASTEP} (kJ/mol)	$U_{latt}^{DMACRYS}$ (kJ/mol)	$\Delta H_{sub}^{DMACRYS}$ (kJ/mol)	ΔH^{Exp} (kJ/mol)
5-chloro-2-nitroaniline	RAPKUP	-115.82	110.86	-99.76	94.80	100.80
salicylic acid	SALIAC15	-117.68	112.72	-96.23	91.28	96.60
2,4,6-trinitroaniline (TNA)	TNIOAN	-125.25	120.29	-114.69	109.74	125.30
1,3,5-Triphenylbenzene	TPHBEN01	-196.85	191.89	-152.64	147.68	147.80
4-Hydroxybenzamide	VIDMAX	-161.43	156.47	-122.06	117.10	117.80
2-Methylglutaric acid	XIBVIO	-158.33	153.37	-126.12	121.16	126.50
Mefenamic acid	XYANAC	-179.95	174.99	-137.03	132.07	136.20
4-Heptylbenzoic acid (cr, II)	ZIKWOF	-184.83	179.87	-141.78	136.82	130.00
RMSE			34.47		13.61	
R^2		0.84	0.60	0.84	0.64	
σ			19.20		13.54	17.97
Bias			28.63		1.39	

Table 6.7: Predictions of ΔH_{sub} for CASTEP and DMACRYS.

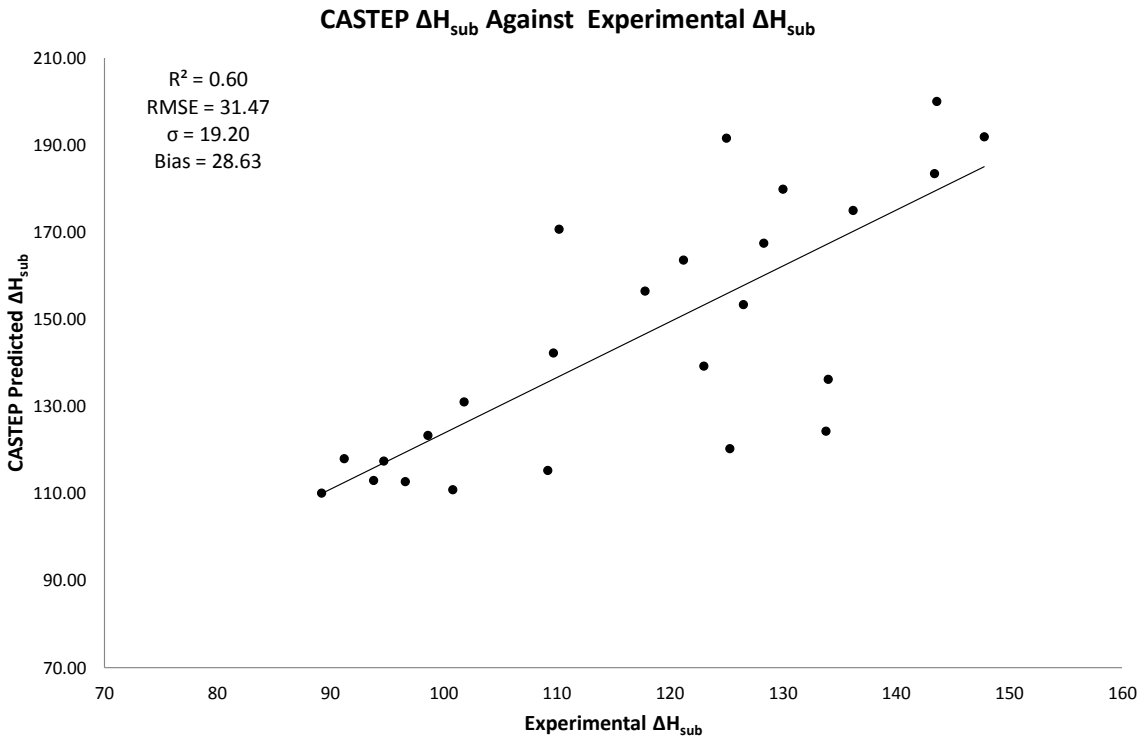


Figure 6.6: Predicted ΔH_{sub} from CASTEP against experiment (kJ/mol).

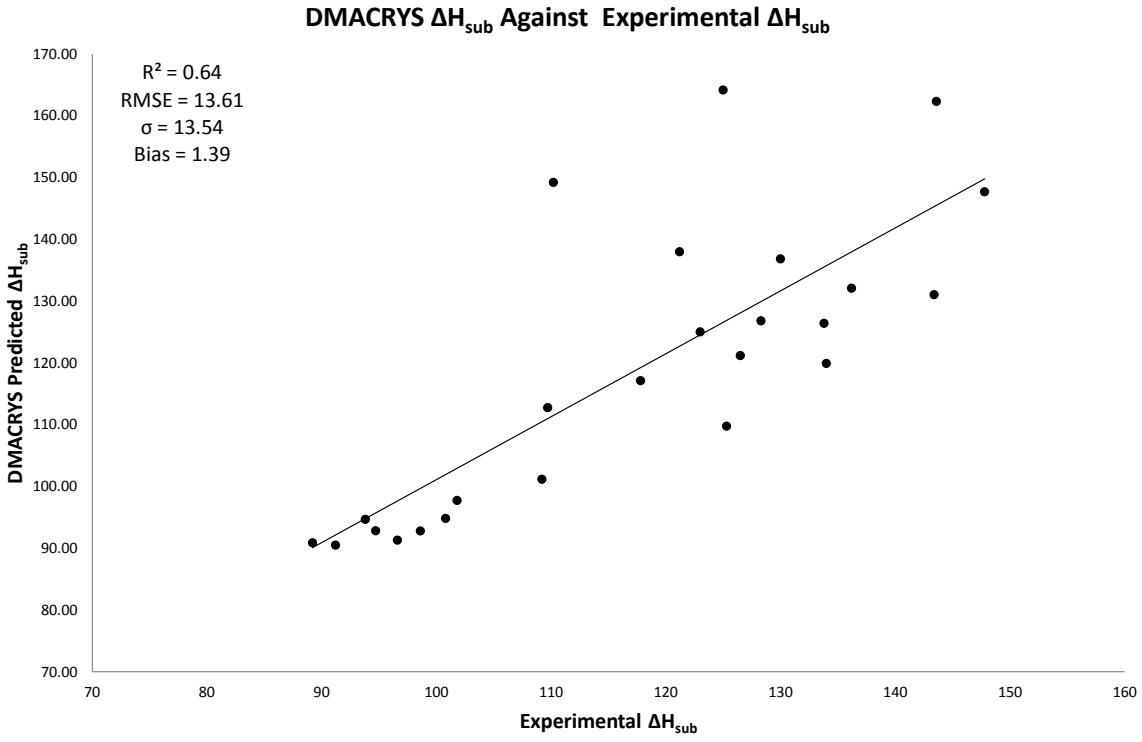


Figure 6.7: Predicted ΔH_{sub} from DMACRYS against experiment (kJ/mol).

It is clear that the use of higher levels of theory does not improve the results, in fact predictions from first principles methods are less accurate. It appears that the first principles methods have a tendency to predict over-bound molecules. It

has been shown previously, that the application of pairwise dispersion corrections schemes, such as the TS dispersion correction scheme, can lead to to large errors and that newer dispersion correction schemes going beyond pairwise additivity are required.[250] It may be that the specific parametrisation for crystal structures of the FIT potential in DMACRYS means that this empirical potential accounts for many body interactions due its derivation from experimental results. As the bias (systematic error) is the predominate error term here, it would in principle be possible to derive an *a priori* correction, however, this would defeat the idea of a first principles prediction. If one correlates the predicted lattice energy from DMACRYS and CASTEP a fairly tight correlation is found (**Figure 6.8**).

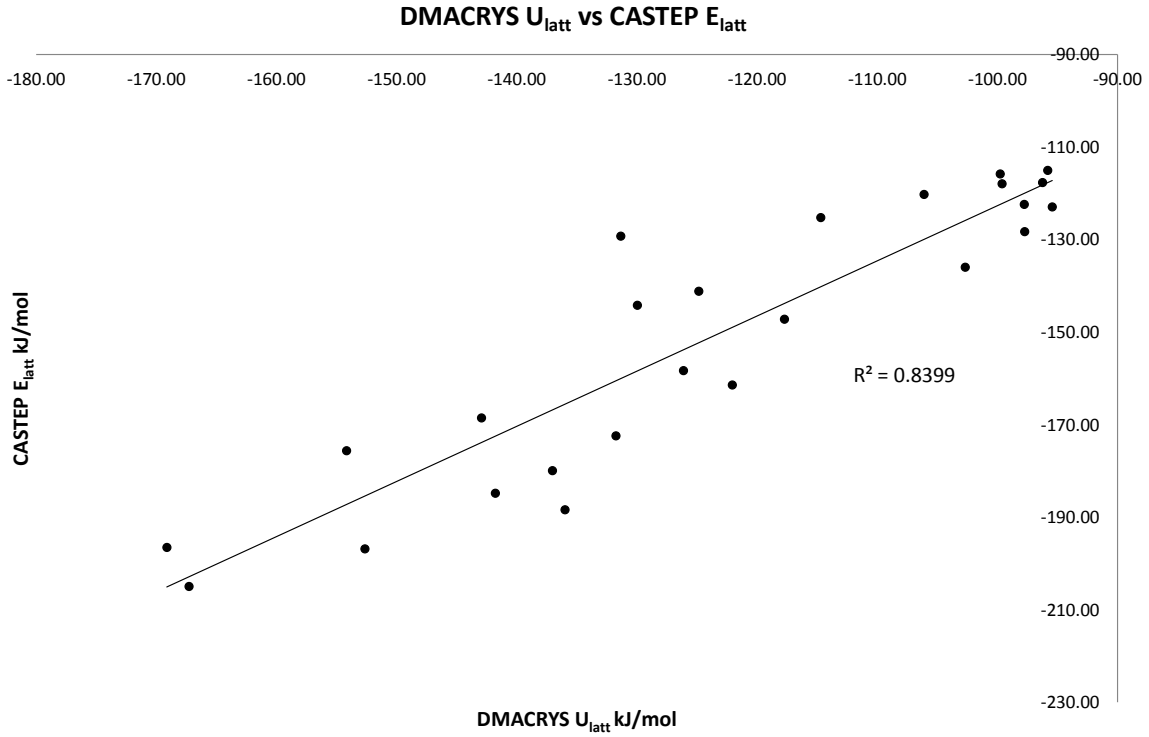


Figure 6.8: Predicted ΔH_{sub} from DMACRYS against CASTEP.

Figure 6.8 shows the lattice energy predicted by CASTEP grows at approximately twice the rate as that predicted by DMACRYS. This is likely due to the systematic error represented by the high bias value. If indeed this is so then a simple correction could be implemented.

6.2 Summary

In the above work we have focused on predictions of sublimation thermodynamics using the periodic lattice simulation program DMACRYS. We have additionally applied machine learning, to predict ΔH_{sub}, ΔS_{sub} and ΔG_{sub} and finally we applied periodic plane wave DFT (using CASTEP), to predict ΔH_{sub}. The results suggest that DMACRYS can provide useful predictions of ΔH_{sub}. The CASTEP PBE-TS methodology offers a qualitative answer, but tends to over-bind the molecules

leading to a large systematic error. ΔG_{sub} is harder to predict. DMACRYS does not make useful predictions of ΔG_{sub}. Machine learning can offer a means by which to achieve reasonable predictions of ΔG_{sub}. ΔS_{sub} remains a stumbling block. Simple approximations do not provide sufficient numerical accuracy and it appears that ΔS_{sub} is quoted with a too small error margin in some experimental reports.

Further work is required for this project to reach a firm conclusion. It is perhaps surprising to find the periodic DFT method did not perform as well as lattice minimisation software. It is possible that the 'off the shelf' value of 0.05 $Å^{-1}$ for the K point spacing is not suitable for all cases. One possibility would be to converge the calculated energy with respect to grid spacing. Additionally the application of newer dispersion corrections could be attempted to test if the over binding we see here is related to the dispersion interactions.

Chapter 7

Conclusion

"The task is, not so much to see what no one has yet seen; but to think what nobody has yet thought, about that which everybody sees."

Erwin Schrödinger, 1952

The primary purpose of this thesis is to investigate methods for predicting intrinsic aqueous solubility, that is the solubility of a neutral molecule in water. In this thesis we present methods ranging from cheminformatics and empirical equations to quantum chemistry. We began with an overview of solubility predictions and experimental considerations. We have discussed industrial uses for such research in the context of the pharmaceutical industry. In **Chapter 2** we discuss a detailed account of the many theoretical methodologies applied in this thesis. An introduction to cheminformatics is provided, followed by detailed discussion of machine readable formats, cheminformatics descriptors and machine learning. Critical aspects of quantum chemistry are discussed before moving on to more detailed explanations of specific theoretical frameworks and methods employed in this thesis.

Following this broad introduction, **Chapter 3** is the first experimental results chapter, covering a proof of concept work for a first principles prediction of solubility.[31] We cover predictions of ΔG_{sub}, ΔG_{hyd} and ΔG_{solv} using a variety of methods. A new dataset, DLS-25, is introduced comprised of 25 drug-like organic molecules. Ten of these molecules are used to test the accuracy of our ΔG_{sub} and ΔG_{hyd} predictions; whilst the full 25 are used for solubility predictions. In total we present 12 first principles predictions of solubility and conclude that only one of these methods produces a useful prediction of solubility; the method coupling DMACRYS with 3DRISM.

Chapter 4 introduces our experimental ideas involving cheminformatics.[33] A larger dataset was curated called DLS-100, which is a dataset comprising 100 drug-like organic molecules. This work extends the previous work by taking a well correlated method from **Chapter 3** and applying it to all 100 molecules. This provided a theoretical benchmark to compare our informatics models against. Three machine learning models are used with three descriptor sets and three scaling methods. The machine learning models were SVM, RF and PLS. The three descriptor sets were 2D

CDK descriptors (123), energy terms from theoretical chemistry (10) and a union of these two sets of descriptors (133). We conclude that these models are able to provide good predictions of solubility with significant improvements in accuracy when 2D CDK descriptors are presented to machine learning models compared to using theoretical chemistry alone. We found there to be little benefit from descriptor scaling, although for some methods auto-scaling shows an improved result. The two sets of descriptors are not complementary therefore, similar information is provided to the models by each set of descriptors.

We followed this work with some calculations using empirical predictions of solubility in **Chapter 5**. Here we applied the GSE to solubility predictions over a small subset of a large dataset from Alfa Aeser for melting point predictions. We applied the machine learning methodology described in **Chapter 4** to melting point prediction. 1100 melting points were predicted with RMSEs of approximately $40°C$. These results were of a similar accuracy to those previously reported. We combined these predicted melting points with two predictions of logP (AlogP and XlogP) using the GSE to predict solubility. For a small dataset of 30 molecules we found excellent agreement with experimentally determined solubility measurements when our predicted melting points were used with AlogP. This was not the case when XlogP was used suggesting the choice of logP calculation is important for solubility predictions by the GSE.

Chapter 6 is the final experimental chapter. This chapter introduces work which has been carried out to test how accurately sublimation thermodynamics can be predicted. A range of methods including cheminformatics, simulation and quantum chemistry are tested. A sublimation dataset is introduced which has been curated from the literature and called SUB-48. We find that cheminformatics provides useful predictions of ΔH_{sub} and ΔG_{sub}. ΔH_{sub} can also be usefully predicted by simulation and qualitatively predicted using quantum chemistry. ΔS_{sub} is poorly predicted and initial efforts to identify the source of the errors have not been successful. The current results suggest it is likely that considerable errors in both *in silico* calculation and experiment exist.

This body of work has examined solubility prediction and its related quantities. We find that, currently, QSPR/QSAR methods remain the current state of the art for solubility prediction, although it is becoming possible for purely theoretical methods to achieve useful predictions of solubility. Theoretical chemistry can offer little useful additional input to informatics models for solubility predictions. However, theoretical chemistry will be crucial for enriching our understanding of the solvation process, and can have a beneficial impact when applied to informatics predictions of properties related to solubility.

7.1 Future Work

Leading on from this project there is much scope for a more thorough investigation of ΔH_{sub}, ΔS_{sub} and ΔG_{sub}. These terms are important in many aspects of chemistry but for solubility, an improved prediction of ΔG_{sub} could give a much improved prediction of solubility. The trade off for this may be computational cost, but a

systematic improvement in the theory would be of benefit as it would be applicable to all systems, and could enable the parametrisation of cheaper more niche methods. Such work may also impact on crystal structure prediction leading to methods capable of property prediction for unknown crystal structures.

There is also further scope to test and extend the first principles method to much larger datasets, although this is currently largely hampered by the accessibility of sufficient, good quality experimental data. Development of RISM methods is another expanding field with applications to solubility being one of the major aims. Recent developments have seen attempts to produce theoretically justified correction schemes for RISM.[251,252] The UC correction applied in this work is born from a knowledge of RISM short falls and a pragmatic solution. Attempts are currently in progress to apply RISM to understand/predict hydrated crystal structures and expand this to solubility prediction.

In industry many advancements occur, and are being currently developed, that take advantage of new computer architecture, database infrastructure and high throughput screening data. Combinatorial libraries and data mining are leading to new information being generated, often from old data. Large databases are being used to apply techniques such as matched molecular pairs[253,254] which allows new empirical rule sets to be generated. This kind of development may lead to new empirical models capable of fast, accurate predictions using existing applications.

Without deviation from the norm, progress is not possible.
Frank Zappa, 1971

Appendix A

Reading and Writing SMILES and InChI Strings

SMILES is a H suppressed string representation. All other atom types are represented by their element symbol. The simplest SMILES string we can write therefore is C, which would represent methane, as H is suppressed these need not be quoted.[74] We can therefore represent butane as CCCC. SMILES are created so as if one were to draw the molecule each atom would be visited only once. As a result of this, rings must have one bond broken in them; this is defined by appending an integer to each element symbol of the pair between which the bond is broken. C1CCCCC1 is cyclohexane. Listed below are some simple rules which allow us to extend this simple idea to represent more complex structures.

- Capital letters represent aliphatic atoms, whereas lower letters represent aromatic atoms.
- Double bonds are represented as =.
- Triple bonds are represented as #.
- Rings are represented by element symbols as normal but an integer is appended to the element symbols of the pair of atoms the bond is broken between.
- Branches are enclosed in brackets and can be nested as required. Once the brackets are closed the structure returns to the branch point.
- A chiral centre can be represented by adding @ to the element symbol. In some cases it is necessary to explicitly define H atoms on chiral centres.
- Cis and trans isomers are represented as C/C=C\C and C/C=C/C respectively.

A simple and instructive example is benzene whose SMILES string is as follows c1ccccc1. We can now define more complicated SMILES as below[74]

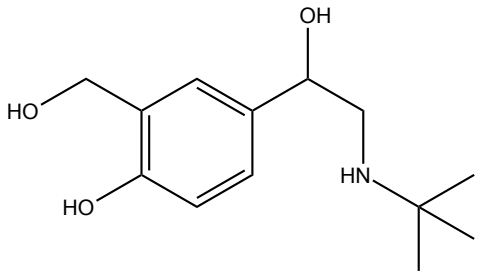


Figure A.1: Salbutamol from the SMILEs string; CC(C)(C)NCC(C1=CC(=C(C=C1)O)CO)O[255]

InChI is a more complex system of sections containing specific information. The information is separated by "/" and the string always begins with "InChI=" this is followed by the version number of the InChI software used and often followed by an "S" to indicate that the standard InChI settings were used in the InChI software when the string was created. The following rules are then applied.[74,76,256]

- The main layer beginning with "/" is a copy of the molecular formula.
- This is followed by a section beginning with "/c" defining the connectivity.
- The next section beginning with "/h" defines terminal H positions and H attachment points.
- There are then a number of possible flags; depending on the chemical structure all relevant flags are applied. Some of the common flags are: "/q" (charge), "/p" (proton balance) and "/t" (tetrahedral parity).
- The next section defines the stereochemistry flagged by "/s" and takes a value 1 (absolute), 2 (relative) and finally 3 is (racemic).
- Following this section is the flag "/f" for the fixed H structure.
- If it is necessary a connectivity flag 2 "/h" can be invoked at this point to define the location of fixed and mobile H's.

The InChI string can be compressed to an InChIkey, which is a 27 character string, whilst the InChIkey cannot be reconverted into the structure due to the data compression, it will not be cut up by search engines, making it a useful tool for database searching and digital curation. The InChI string for the salbutamol (Figure A.1) molecule shown above is presented below along with its InChIkey.[76,255]
InChI=1S/C13H21NO3/c1-13(2,3)14-7-12(17)9-4-5-11(16)10(6-9)8-15/h4-6,12,14-17H,7-8H2,1-3H3[255]
InChIkey NDAUXUAQIAJITI-UHFFFAOYSA-N[255]

Appendix B

Atomic Units

Atomic units allow a convenient set of units to be used in the quantum mechanical calculations.

Symbol	Meaning	S.I. unit	Atomic unit (A.U.)
m_e	Electron mass	$9.1094 \times 10^{-31} Kg$	1
a_0	Bohr radius	$0.52918 \times 10^{-10} m$	1
e	Charge of an electron	$1.6022 \times 10^{-19} C$	1
$\hbar = \frac{h}{2\pi}$	Reduced Planck's constant	$1.0546 \times 10^{-34} Js$	1
$4\pi\epsilon_0$	Vacuum permittivity	$1.113 \times 10^{-10} C^2/Jm$	1
$E_H = \frac{\hbar^2}{m_0 a_0^2}$	Energy (Hartrees)	$27.2114eV$	1

Table B.1: Atomic units and SI units[257]

This allows the simplification of the Hamiltonian by setting many of the physical constants to unity (1). Take the Hamiltonian of the hydrogen atom for example:

$$\hat{H} = -\frac{\hbar^2}{2m_e}\nabla^2 - \frac{e^2}{4\pi\epsilon_0 r} \tag{B.1}$$

Equation B.1: The Hamiltonian of the hydrogen atom in S.I. units

$$\hat{H} = -\frac{1}{2}\nabla^2 - \frac{1}{r} \tag{B.2}$$

Equation B.2: The Hamiltonian of the hydrogen atom in atomic units (A.U.) units

Appendix C

The Slater Determinant

If we take a generalised 2 x 2 determinant of molecular orbitals we can solve it for a generalised wave function;

$$\Psi(1,2) = \frac{1}{\sqrt{n!}} \begin{vmatrix} \phi_1(1) & \phi_2(1) \\ \phi_1(2) & \phi_2(2) \end{vmatrix} \tag{C.1}$$

$$\Psi(1,2) = \phi_1(1)\phi_2(2) - \phi_2(1)\phi_1(2) \tag{C.2}$$

If we now exchange the two electron's coordinates (i.e. change the orbital they are in in the determinant) we get the result of the negative of the wavefunction above;

$$\Psi(1,2) = \frac{1}{\sqrt{n!}} \begin{vmatrix} \phi_2(1) & \phi_1(1) \\ \phi_2(2) & \phi_1(2) \end{vmatrix} \tag{C.3}$$

$$\Psi(1,2) = \phi_2(1)\phi_1(2) - \phi_1(1)\phi_2(2) \tag{C.4}$$

This generalisation holds for larger determinants, below an example calculation is given for a 3x3 determinant.

$$\Psi(1,2,3) = \frac{1}{\sqrt{n!}} \begin{vmatrix} \phi_1(1) & \phi_2(1) & \phi_3(1) \\ \phi_1(2) & \phi_2(2) & \phi_3(2) \\ \phi_1(3) & \phi_2(3) & \phi_3(3) \end{vmatrix} \tag{C.5}$$

One method to solve a determinant greater than 2x2 is to break it down into smaller determinants. This sub-division has a accompanied sign convention. This convention is shown below. We proceed to divide the 3x3 determinant into 2x2 determinants.

$$\Psi(1,2,3) = \frac{1}{\sqrt{n!}} \begin{vmatrix} & + & - & + \\ + & \phi_1(1) & \phi_2(1) & \phi_3(1) \\ - & \phi_1(2) & \phi_2(2) & \phi_3(2) \\ + & \phi_1(3) & \phi_2(3) & \phi_3(3) \end{vmatrix} \tag{C.6}$$

By selecting each of the values in the top row and eliminating that row and corresponding column we can make three 2x2 matrices, these need to be multiplied by the selected value in the top row.

To begin with if we select $\phi_1(1)$ and eliminate the row 1 and column 1 we create the following 2x2 matrix which is then multiplied by $\phi_1(1)$;

$$+\phi_1(1) \begin{vmatrix} \phi_2(2) & \phi_3(2) \\ \phi_2(3) & \phi_3(3) \end{vmatrix} \tag{C.7}$$

Continuing this process we arrive at two further 2x2 matrices for the two remaining values in the first row.

$$-\phi_2(1) \begin{vmatrix} \phi_1(2) & \phi_3(2) \\ \phi_1(3) & \phi_3(3) \end{vmatrix} \tag{C.8}$$

$$+\phi_3(1) \begin{vmatrix} \phi_1(2) & \phi_2(2) \\ \phi_1(3) & \phi_2(3) \end{vmatrix} \tag{C.9}$$

Moving to solve each of these matrices as we did as the first gives the final expression as follows for a 3x3 matrix, with the appropriate sign applied to the separate solutions.

$$\Psi(1,2,3) = \frac{1}{\sqrt{n!}}(\phi_1(1)[\phi_2(2)\phi_3(3) - \phi_3(2)\phi_2(3)] - \phi_2(1)[\phi_1(2)\phi_3(3) - \phi_3(2)\phi_1(3)] + \phi_3(1)[\phi_1(2)\phi_2(3) - \phi_2(2)\phi_1(3)]) \tag{C.10}$$

Appendix D

Orthonormality

Orthonormality is an important concept in much of physics. It is a combination of two important mathematical constructs; orthogonality and normality. If we take a exemplary case of two vectors, we can define then as orthogonal if the following relationship holds:[258,259]

$$A \cdot B = \sum_i a_i b_i^* = 0 \tag{D.1}$$

Equation D.1: Orthogonality definition. i is the component label of vectors A and B i.e. if A and B are 3D vectors $a_1 b_1^* + a_2 b_2^* + a_3 b_3^*$.

This follows from the definition of the dot product:

$$Re(A \cdot B) = \|A\|\|B\|Cos(\phi) \tag{D.2}$$

Equation D.2: Re refers to the real portion of the vectors A and B. Where $\phi = 90°$ *the cos function is zero, hence the dot product is zero.* ϕ *represents the angle between the vectors, so if this is equal to 90 degrees the vectors are by definition orthogonal.*

We can also define a vector as unit vector (normal) if the following is true:

$$\sum_i a_i a_i^* = 1 \tag{D.3}$$

Equation D.3: Vector normality definition.

Vectors are considered orthonormal if **Equation D.1** and **Equation D.3** are true. This is demonstrated here and is generalisable to infinite dimensional vectors. Functions can be thought of as infinite dimensional vectors with x coordinate in place of the i subscript (i.e. each function is one component) and the y coordinate is the magnitude of function or output of the function.[258] The orthogonality definition then can be restated:

$$\int a(z)b^*(z)dz = 0 \tag{D.4}$$

Equation D.4: Function version of the orthogonality definition. $a(z)$ and $b^*(z)$ are functions.

The normality definition can then also be restated for a normalised function a(z):

$$\int a(z)a^*(z)dz = 1 \tag{D.5}$$

Equation D.5: Function version of the normality definition.

The property of being orthonormal is generalised using the delta function, and notationally simplified using the Dirac Bra-Ket notation:

$$\begin{array}{ll} \int \Psi_j^* \Psi_i dr & = \delta_{ij} \\ \langle \Psi_i | \Psi_j \rangle & = \delta_{ij} \end{array} \tag{D.6}$$

Equation D.6: Function orthonormality definition. $\delta = 0$ when $i \neq j$ and $\delta = 1$ when $i = j$.

Appendix E

The Variational Principle

The variational principle is a cornerstone, a crucial component to quantum chemistry. Assuming we know the exact solutions of the Schrödinger equation, where i indexes the solution number. There are infinitely many solutions with infinitely many energies, one corresponding to each solution, ϵ_0 is the lowest ground state energy. The Hamiltonian is a *Hermitian operator* or *self-adjoint* operator (the adjoint is obtained by taking the complex conjugate and then transposing the matrix), hence the solutions form a complete basis. We may also work with these equations so that the solutions are orthonormal **(Appendix D)**.[113]

$$\hat{H}\Psi_i = \epsilon_i \Psi_i \tag{E.1}$$

Equation E.1: Schrödinger equation with infinite solutions.

$$\frac{\int \Psi^*_{trial} H \Psi_{trial} dr}{\int \Psi^*_{trial} \Psi_{trial}} = E_{trial} \geq E_0 \tag{E.2}$$

Equation E.2: The variational principle: A method to assess the quality of a trial wavefunction. Ψ^*_{trial} is the complex conjugate of Ψ_{trial}.

We can expand an approximate wavefunction (θ) in the basis of the exact solutions as they are a complete set. We can calculate the energy of the approximate wave function by calculating its expectation value **Equation E.2**. These two equations can be combined to calculate the expectation value of the approximate function.[113]

$$\theta = \sum_i a_i \Psi_i$$
$$E = \frac{\langle\theta|\hat{H}|\theta\rangle}{\langle\theta|\theta\rangle} \tag{E.3}$$
$$E = \frac{\sum_i \sum_j a_i a_j \langle\Psi_i|\hat{H}|\Psi_j\rangle}{\sum_i \sum_j a_i a_j \langle\Psi_i|\Psi_j\rangle}$$

Equation E.3: Top: Approximate wave function expanded in the basis of the exact solution. Middle: Expectation value. Bottom: Expectation value of the constructed wavefunction.

Applying the orthonormality relation **(Appendix D)** we can simplify the bottom equation in **E.3**. We can then prove the variational principle by taking away E_0 and proving the result is always positive or zero.

$$E = \frac{\sum_i a_i^2 \epsilon_i}{\sum_i a_i^2} \tag{E.4}$$
$$E - E_0 = \frac{\sum_i a_i^2 (\epsilon_i - E_0)}{\sum_i a_i^2} \geq 0$$

Equation E.4: As a_i^2 is always positive and E_0 is the lowest energy by definition the answer is always ≥ 0

This finally arrives at the stated variational principle.[113]

$$E_0 \leq E_{trial} = \frac{\langle\Psi_{trial}|\hat{H}|\Psi_{trial}\rangle}{\langle\Psi_{trial}|\Psi_{trial}\rangle} \tag{E.5}$$

Equation E.5: The variational principle

Appendix F

Eigenvalues and Eigenvectors

Taking an operator in matrix form known as Z we can state it in an eigenvalue problem:[260]

$$Z\Psi = z\Psi \tag{F.1}$$

In order to solve this equation we need to rewrite it in to the following:

$$\begin{aligned} Z\Psi - z\Psi &= Z\Psi - z\Psi \\ Z\Psi - z\Psi &= 0 \\ (Z - zI)\Psi &= 0 \end{aligned} \tag{F.2}$$

I here is the identity matrix. Solutions to this equation occur when the determinant of the matrix $(Z - zI)$ is equal to zero.

$$det(Z - zI) = 0 \tag{F.3}$$

Calculating the eigenvalue (Z is defined from an example in the following reference):[260]

$$Z = \begin{pmatrix} -1 & -1 \\ 2 & -4 \end{pmatrix} \tag{F.4}$$

$$(Z - zI) = \begin{pmatrix} -1 & -1 \\ 2 & -4 \end{pmatrix} - z \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{F.5}$$

$$(Z - zI) = \begin{pmatrix} -1 & -1 \\ 2 & -4 \end{pmatrix} - \begin{pmatrix} z & 0 \\ 0 & z \end{pmatrix} \tag{F.6}$$

$$(Z - zI) = \begin{pmatrix} -1 - z & -1 - 0 \\ 2 - 0 & -4 - z \end{pmatrix} \tag{F.7}$$

$$(Z - zI) = \begin{pmatrix} -1 - z & -1 \\ 2 & -4 - z \end{pmatrix} \tag{F.8}$$

$$\begin{gathered} det(Z - zI) = (-1 - z)(-4 - z) - (-1)(2) \\ det(Z - zI) = (z^2 + 5z + 4) + 2 \\ det(Z - zI) = (z^2 + 5z + 6) \end{gathered} \tag{F.9}$$

$$\begin{gathered} factorise\ z^2 + 5z + 6 = 0 \\ (z + 2)(z + 3) \\ z = -2, -3 \end{gathered} \tag{F.10}$$

Calculating the eigenvector:[260] By placing the values for z in the following matrix one at a time we can find the eigenvectors.

$$(Z - zI) = \begin{pmatrix} -1 - z & -1 \\ 2 & -4 - z \end{pmatrix} \tag{F.11}$$

Substitute in z=-2

$$(Z - zI) = \begin{pmatrix} -1 - (-2) & -1 \\ 2 & -4 - (-2) \end{pmatrix} \tag{F.12}$$

$$(Z - zI) = \begin{pmatrix} 1 & -1 \\ 2 & -2 \end{pmatrix} \tag{F.13}$$

Given our original expression we can now solve for the eigenvector.

$$(Z - zI)(\Psi_n) = 0 \tag{F.14}$$

$$(Z - zI)(\Psi_n) = \begin{pmatrix} 1 & -1 \\ 2 & -2 \end{pmatrix} \begin{pmatrix} \Psi_1 \\ \Psi_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{F.15}$$

We can use this to generate relations between Ψ_1 and Ψ_2. Initially following the conventions of matrix multiplication we derive the following equations:

$$\begin{aligned} 1\Psi_1 - 1\Psi_2 &= 0 \\ 2\Psi_1 - 2\Psi_2 &= 0 \end{aligned} \tag{F.16}$$

This leads to the following relations between Ψ_1 and Ψ_2 by rearrangement of the equations above:

$$\begin{aligned} 1\Psi_1 &= 1\Psi_2 \\ 2\Psi_1 &= 2\Psi_2 \\ &\therefore \\ \Psi_1 &= \Psi_2 \end{aligned} \tag{F.17}$$

Therefore :

$$\Psi_i = 1 and \begin{pmatrix} \Psi_1 \\ \Psi_2 \end{pmatrix} = C \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{F.18}$$

Equation F.18: C is a multiplicative constant.

The same procedure can be carried out for z=-3.

Appendix G

DFT Existence and Variational Theorem

The proof for the *existence theorem* goes via *reductio ad absurdum*, meaning an impossible result is returned from an assumption to the contrary. Assume that two different external potentials (ν_a, ν_b) produce the same non-degenerate, ground state, electron density (ρ_0) in a molecule. We can define two Hamiltonians, one with ν_a and the other ν_b ($\hat{H}_a$, $\hat{H}_b$). The ground state wave functions ($\Psi_{0,a}$, $\Psi_{0,b}$) for these systems can then be defined and hence the eigenvalues (ϵ_a, ϵ_b). Wave function theory tells us that if we apply $\hat{H}_a$ to $\Psi_{0,b}$ then the eigenvalue (ϵ_x) is greater than ϵ_a.

$$\epsilon_{0,a} \leq \langle \Psi_{0,b} | \hat{H}_a | \Psi_{0,b} \rangle \tag{G.1}$$

This can be rewritten. Substituting in the expression for ϵ_b and representing the difference between $\epsilon_{0,a}$ and ϵ_x as the difference in the two Hamiltonians $\hat{H}_a$, $\hat{H}_b$ applied to the ground state wave function of b.

$$\begin{aligned} \epsilon_{0,a} &\leq \langle \Psi_{0,b} | \hat{H}_a | \Psi_{0,b} \rangle \\ \epsilon_{0,a} &\leq \langle \Psi_{0,b} | \hat{H}_a - \hat{H}_b | \Psi_{0,b} \rangle + \langle \Psi_{0,b} | \hat{H}_b | \Psi_{0,b} \rangle \\ \epsilon_{0,a} &\leq \langle \Psi_{0,b} | \nu_a - \nu_b | \Psi_{0,b} \rangle + \epsilon_{0,b} \end{aligned} \tag{G.2}$$

Equation G.2: expansion of the variational equation utilising different external potentials

We can now explicitly write out the last integral of **Equation G.2**. In this step we integrate the difference in the one-electron potential operators over the ground state density.

$$\epsilon_{0,a} \leq \int [\nu_a(r) - \nu_b(r)]\rho_0(r)dr + \epsilon_{0,b} \tag{G.3}$$

Equation G.3: Energy eigenvalue of the mixed external potentials. ρ_0 is the ground state density.

An alternative expression can be reached, which is equally valid, just by swapping the a and b subscripts in **Equation G.3**. No difference between a and b has been defined to stop us from being able to do this. If we sum the inequalities we reach the following statements:

$$\begin{aligned} \epsilon_{0,a} + \epsilon_{0,b} &\leq \int [\nu_a(r) - \nu_b(r)]\rho_0(r)dr + \int [\nu_b(r) - \nu_a(r)]\rho_0(r)dr + \epsilon_{0,b} + \epsilon_{0,a} \\ \epsilon_{0,a} + \epsilon_{0,b} &\leq \int [\nu_a(r) - \nu_b(r) + \nu_b(r) - \nu_a(r)]\rho_0(r)dr + \epsilon_{0,b} + \epsilon_{0,a} \\ \epsilon_{0,a} + \epsilon_{0,b} &\leq \epsilon_{0,a} + \epsilon_{0,b} \end{aligned} \tag{G.4}$$

Clearly this result is not possible and hence this is proof that the ground state density uniquely defines the external potential.[112]

A second theorem then provides means to optimise the density. This is a variational theorem analogous to the one used in the wave function methods **(Appendix E)**. Assuming we have a well behaved density that integrates to the correct number of electrons, we can proceed as follows:[112,113]

$$\epsilon_0(\rho) \leq \epsilon_0(\rho') \tag{G.5}$$

Appendix H

DLS-100 and SUB-48

Shown below, the table provides the data for the SUB-48 dataset with all references to the origin of the sublimation data.

Molecule name	Refcode and data reference	ΔH_{sub}^{exp} (kJ/mol)	ΔS_{sub}^{exp} $(JK^{-1}\,mol^{-1})$	ΔG_{sub}^{exp} (kJ/mol)
Acetanilide	ACANIL01[223]	99.8	197	40.5
aspirin (acetylsalicylic acid)	ACSALA15[224]	109.7	222	43.6
Benzoic acid	BENZAC01[225]	89.2±0.8	185.5±2.7	33.9±1.1
isophthalic acid	BENZDC01[226]	143.4±1.7	234.3±4.5	73.5±2.2
2,4-Dinitrobenzoic acid	BIPJUF[227]	134±3	244±8	61.2514
Benzophenone	BPHENO03[228]	89.1±0.3	182.89	34.6
atenolol	CEZVIN[229]	140.0±3.7	215±9	71.1
2-Chlorobenzoic acid	CLBZAC01[225]	106.3±0.5	217.3±1.6	41.5±0.7
4-Chlorobenzoic acid	CLBZAP03[225]	105.2±0.7	199.4±2.1	45.8±1.0
2,6-Dichloro-4-nitroaniline	CLNOAN[230]	109.2±0.9	203.0±2.6	48.7±1.2
3-Fluorobenzoic acid	COVJIG[225]	93.8±0.5	199.4±1.6	34.4±0.7
naproxen	COYRUD11[6]	128.3	234.2281879	58.5
2,4-dichlorophenoxy acetic acid	CPXACA[231]	122± 5	239±4	51.778
1,8-diphenylnaphthalene	DPNAPH01[232]	132.5±0.6	237.4±1.6	61.71919
2-Fluorobenzoic acid	FBENZA02[225]	94.7±0.5	198.6±1.6	35.5±0.7
flurbiprofen	FLUBIP[233]	110.2	190.94	53.3
Flufenamic acid	FPAMCA17[234]	121.2	224	54.3
2,4,6-trinitromesitylene (TNM)	HEXTIN01[235]	103.6±1.2	117.3±1.6	68.6±2.8
4-amino-N-(4-ethylphenyl)benzenesulfonamide	HUNXIY[236]	143.6±0.9	233±2	74.2
4-Amino-N-(4-methoxyphenyl)benzenesulfonamide	HUNXOE[236]	124±1	173±2	72.4
4-Amino-N-(5-chloro-2-methylphenyl)benzenesulfonamide	HUNXUK[236]	130±1	206±3	68.5
Paracetamol	HXACAN27[237]	117.9±0.7	190±2	60

Continued on next page

Table H.1 – *Continued from previous page*

Molecule name	Refcode and data reference	ΔH_{sub}^{exp} (kJ/mol)	ΔS_{sub}^{exp} $(JK^{-1}\,mol^{-1})$	ΔG_{sub}^{exp} (kJ/mol)
Ibuprofen	IBPRAC01[224]	115.8	240	44.2
4-hydroxybenzoic acid	JOZZIH[238]	121.1±0.4	218.0±1.4	56.1±0.1
Tolfenamic acid	KAXXAI01[239]	128.4±0.8	216±4	53.9±0.4
Ketoprofen	KEMRUP[228]	110.1±0.5	178±1	57
9-Methylanthracene	MANTHR01[240]	101.8±1.0	199.7±3.0	42.2±1.3
2,4,6-N-tetranitro-N-methylaniline (Tetryl)	MTNANL[235]	133.8±1.6	155.8±2.0	87.3±3.2
1-naphthol	NAPHOL01[6]	91.2	187.58	35.3
4-nitrobenzoic acid	NBZOAC01[238]	116.6±0.6	210.0±2.0	54.0±0.1
Niflumic acid	NIFLUM10[234]	130.2±0.8	231 ±2	61.3
nitroguanidine	NTRGUA01[235]	142.7±2.0	116.8±3.4	107.9±4.6
2,2-Dimethylsuccinic acid	OLENIC[241]	122.7±2.7	241±8	50.8±3.6
4-Methylbenzoic acid	PTOLIC01[242]	98.6±0.7	196.1±1.8	40.1±0.8
Phenacetin	PYRAZB21[223]	121.8±0.7	226±2	52.3
5-chloro-2-nitroaniline	RAPKUP[243]	100.8±0.3	200.0±0.9	41.2±0.4
salicylic acid	SALIAC15[244]	96.6	191	38.5
Diclofenac	SIKLIH01[245]	115.6	222	49.3
4-cyanobenzoic acid	TAGNAR[238]	111.2±0.4	202.9±1.4	50.7±0.1
o-Terphenyl	TERPHO02[225]	103.0±0.4	210.6±1.3	40.2±0.6
2,4,6-trinitroaniline (TNA)	TNIOAN[235]	125.3±0.8	140.7±0.3	83.3±1.2
1,3,5-Triphenylbenzene	TPHBEN01[246]	147.8±0.7	254±2	72.1±0.9
N-(4-nitrophenyl)-benzene-sulfonamide	UVEMOY[247]	132.5±1.6	217±7	67.7
4-Amino-N-(4-nitrophenyl)benzenesulfonamide	UVEMUE[247]	131.4±2.6	179±7	78

Continued on next page

Table H.1 – *Continued from previous page*

Molecule name	Refcode and data reference	ΔH_{sub}^{exp} (kJ/mol)	ΔS_{sub}^{exp} $(JK^{-1}\,mol^{-1})$	ΔG_{sub}^{exp} (kJ/mol)
4-Hydroxybenzamide	VIDMAX[248]	117.8±0.6	198±2	58.9
2-Methylglutaric acid	XIBVIO[241]	126.5±2.1	259±8	49.2±3.2
Mefenamic acid	XYANAC[239]	136.2±0.8	213±3	59.2±0.1
4-Heptylbenzoic acid (cr, II)	ZIKWOF[242]	130.0±0.9	255.3±2.6	53.9±1.2

Table H.1: Experimental sublimation data and references.

Below is a table with the molecule name, CSD refcode, Log S and SMILES source listed. Following from this are several images of the structures of the molecules determined from their SMILES.

Chemical name	CSD refcode	Log S exp (mol/L)	Smiles Source
Acetanilide	ACANIL01	-1.4	ChemSpider
Adenosine	ADENOS10	-1.73	ChemSpider
Allopurinol	ALOPUR	-2.26	ChemSpider
4-Aminobenzoic acid	AMBNAC04	-1.37	ChemSpider
4-Aminosalicylic acid	AMSALA01	-1.96	ChemSpider
Trimethoprim	AMXBPM10	-2.95	ChemSpider
Antipyrine	ANTPYR10	0.48	ChemSpider
Acetazolamide	ATDZSA	-2.44	ChemSpider
Benzoic acid	BENZAC02	-1.58	ChemSpider
Salbutamol	BHHPHE	-1.22	ChemSpider
Quinidine	BOMDUC	-2.81	ChemSpider
Benzamide	BZAMID02	-0.95	ChemSpider
Thiamphenicol	CABCIR01	-2.15	ChemSpider
(RS)-Atenolol	CEZVIN	-1.3	ChemSpider
Chloral Hydrate	CHORLH01	1.7	ChemSpider
Cimetidine	CIMETD	-3.6	Solubility Challenge
Chloramphenicol	CLMPCL02	-2.11	ChemSpider
Diuron	CLPHUR02	-3.76	ChemSpider
Chlorprothixene	CMAPTX	-6.75	ChemSpider
Cocaine	COCAIN10	-2.25	ChemSpider
Corticosterone	CORTIC	-3.24	ChemSpider
Naproxen	COYRUD11	-4.5	ChemSpider
Sertraline	CUTPEN	-4.83	ChemSpider
Cytosine	CYTSIN01	-1.16	ChemSpider
Dapsone	DAPSUO03	-3.09	ChemSpider
1,8-Dihydroxyanthraquinone	DHANQU06	-5.19	ChemSpider

Continued on next page

Table H.2 – *Continued from previous page*

Chemical name	CSD refcode	Log S exp (mol/L)	Smiles Source
Cortisone	DHPRTO02	-3.27	ChemSpider
Diazepam	DIZPAM10	-3.75	ChemSpider
Sulindac	DOHREX	-5	ChemSpider
Primidone	EPHPMO	-2.64	ChemSpider
Estrone	ESTRON14	-3.95	ChemSpider
Hydroflumethiazide	EWUHAF01	-2.97	ChemSpider
Alclofenac	FICJAC	-3.13	ChemSpider
Flurbiprofen	FLUBIP	-4.15	ChemSpider
Famotidine	FOGVIG02	-2.65	ChemSpider
Flufenamic acid	FPAMCA	-5.35	ChemSpider
5-Fluorouracil	FURACL02	-1.03	ChemSpider
Equilin	GODTIC	-5.28	ChemSpider
Griseofulvin	GRISFL	-3.25	Wikipedia
Hydrochlorothiazide	HCSBTZ04	-2.69	ChemSpider
Fluometuron	HODHIS	-3.46	ChemSpider
Paracetamol	HXACAN04	-1.02	ChemSpider
Ibuprofen	IBPRAC01	-3.6	ChemSpider
Propranolol	IMITON	-3.49	ChemSpider
Thymol	IPMEPL	-2.19	ChemSpider
Fluconazole	IVUQOF	-1.8	ChemSpider
Pentoxifylline	JAKGEH	-0.56	ChemSpider
Isoproturon	JODTUR01	-3.47	ChemSpider
Guanine	KEMDOW	-3.56	ChemSpider
Khellin	KHELIN	-3.02	ChemSpider
Nitrofurantoin	LABJON01	-3.24	ChemSpider
L-DOPA (Levodopa)	LDOPAS03	-1.12	ChemSpider

Continued on next page

Table H.2 – *Continued from previous page*

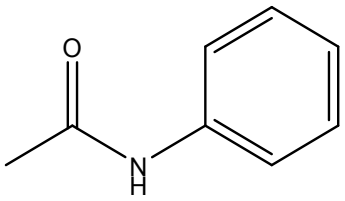
Chemical name	CSD refcode	Log S exp (mol/L)	Smiles Source
Indoprofen	LEKMET	-4.82	ChemSpider
Metoclopramide	METPRA	-3.57	ChemSpider
Metronidazole	MNIMET	-1.22	ChemSpider
Papaverine	MVERIQ01	-3.87	ChemSpider
Nalidixic acid	NALIDX01	-3.61	ChemSpider
1-Naphthol	NAPHOL01	-1.98	ChemSpider
Clozapine	NDNHCL01	-3.24	ChemSpider
Norethisterone	NETIND01	-4.63	ChemSpider
Nicotinic acid	NICOAC02	-0.85	ChemSpider
Niflumic acid	NIFLUM10	-4.59	ChemSpider
Oxytetracycline	OXYTET	-2.95	ChemSpider
Perylene	PERLEN05	-8.8	ChemSpider
Perphenazine	PERPAZ	-4.16	ChemSpider
Phenobarbital	PHBARB09	-2.29	Solubility Challenge
Phthalic acid	PHTHAC01	-1.49	ChemSpider
5,5-Diphenylhydantoin	PHYDAN01	-3.86	ChemSpider
Pindolol	PINDOL	-3.79	Solubility Challenge
Progesterone	PROGST12	-4.42	ChemSpider
Pteridine	PTERID11	0.02	ChemSpider
Phenacetin	PYRAZB21	-2.37	ChemSpider
Pyrene	PYRENE07	-6.18	ChemSpider
Pyrazinamide	PYRZIN	-0.91	ChemSpider
Salicylic acid	SALIAC	-1.94	ChemSpider
Salicylamide	SALMID07	-1.84	ChemSpider
Glipizide	SAXFED	-5.46	Wikipedia
Diclofenac	SIKLIH01	-5.49	ChemSpider

Continued on next page

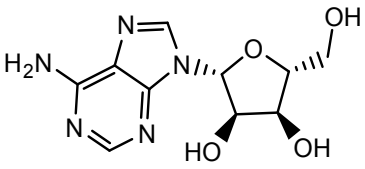
Table H.2 – *Continued from previous page*

Chemical name	CSD refcode	Log S exp (mol/L)	Smiles Source
Sulfamethoxazole	SLFNMB01	-2.7	ChemSpider
Sulfamethazine	SLFNMD01	-2.73	ChemSpider
Sulfacetamide	SLFNMG01	-1.51	ChemSpider
Sulfanilamide	SULAMD01	-1.36	ChemSpider
Sulfadiazine	SULDAZ01	-2.73	ChemSpider
Gliclazide	SUVGUL	-4.07	ChemSpider
Nadolol	TAYGAC	-1	ChemSpider
1,3,5-trichlorobenzene	TCHLBZ	-4.44	ChemSpider
Thalidomide	THALID03	-3.7	ChemSpider
Trihexyphenidyl	THEXPL	-5.2	ChemSpider
Thymine	THYMIN01	-1.5	ChemSpider
Thebaine	TICTUU	-2.66	ChemSpider
Triphenylene	TRIPHE11	-6.73	ChemSpider
Uracil	URACIL	-1.49	ChemSpider
Uric acid	URICAC	-3.4	ChemSpider
Atropine	WALPIJ	-2	ChemSpider
Linuron	WAMXUD	-3.52	ChemSpider
Mefenamic acid	XYANAC	-6.74	ChemSpider
Mifepristone	ZIDLED	-5.75	ChemSpider
Tolbutamide	ZZZPUS02	-3.47	ChemSpider
Codeine	ZZZTSE03	-1.56	ChemSpider
Strychnine	ZZZUEE04	-3.33	ChemSpider

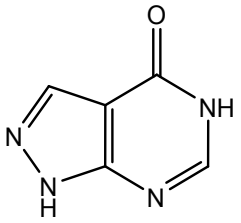
Table H.2: DLS-100 dataset



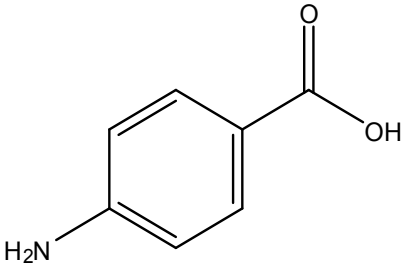
Acetanilide



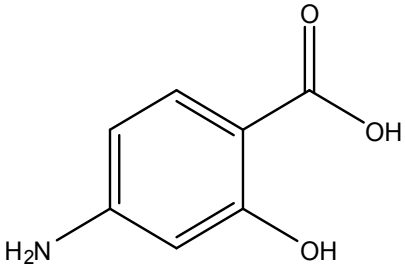
Adenosine



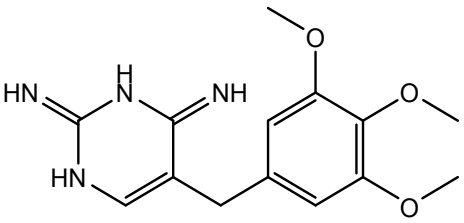
Allopurinol



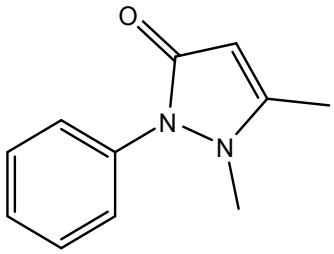
4-Aminobenzoic acid



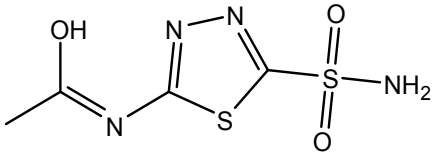
4-Aminosalicylic acid



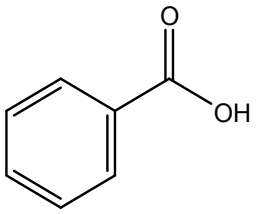
Trimethoprim



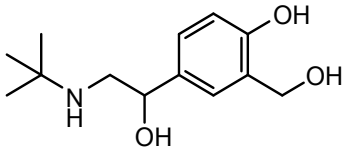
Antipyrine



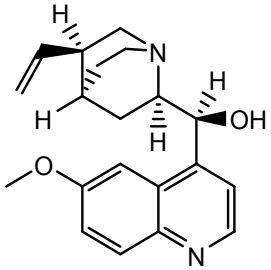
Acetazolamide



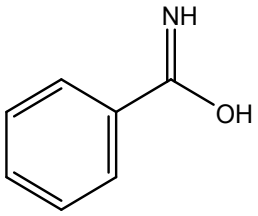
Benzoic acid



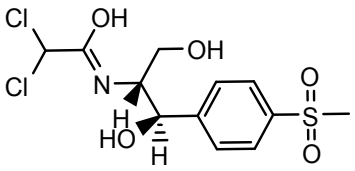
Salbutamol



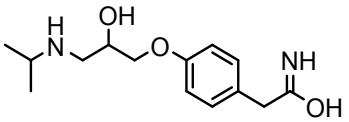
Quinidine



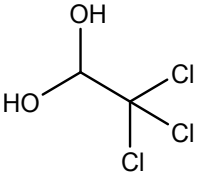
Benzamide



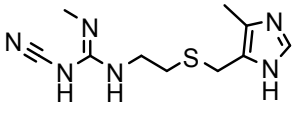
Thiamphenicol



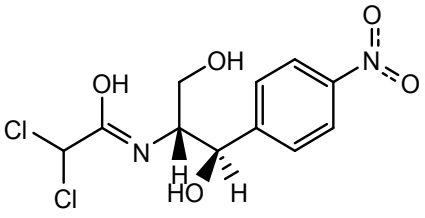
(RS)-Atenolol



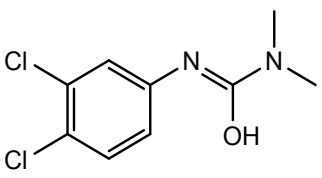
Chloral Hydrate



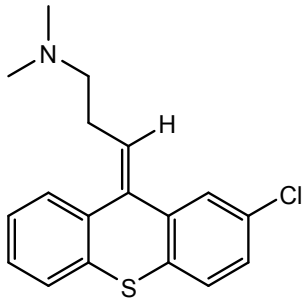
Cimetidine



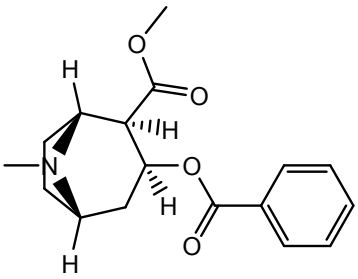
Chloramphenicol



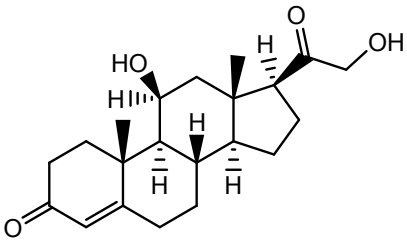
Diuron



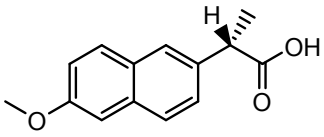
Chlorprothixene



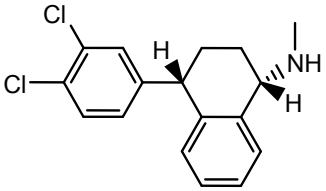
Cocaine



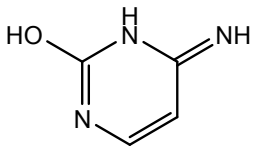
Corticosterone



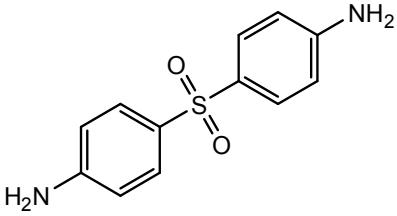
Naproxen



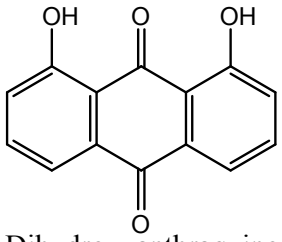
Sertraline



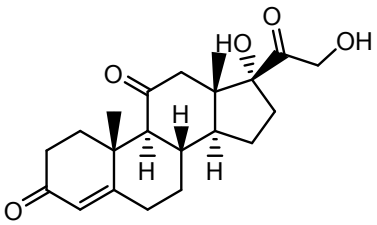
Cytosine



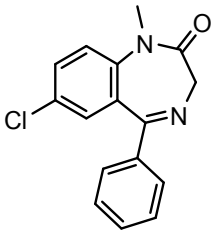
Dapsone



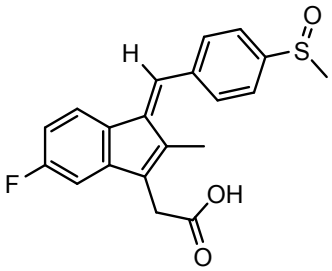
1,8-Dihydroxyanthraquinone



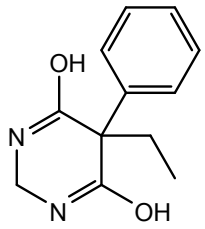
Cortisone



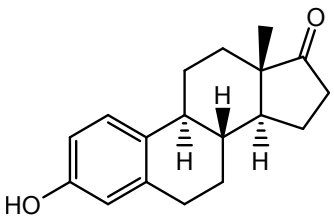
Diazepam



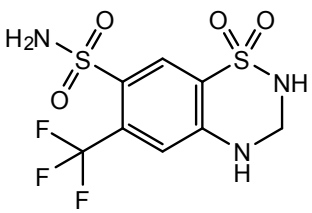
Sulindac



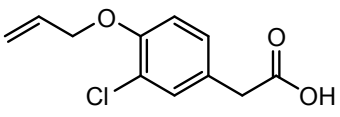
Primidone



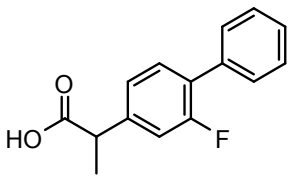
Estrone



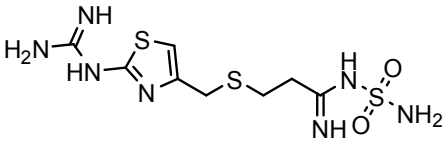
Hydroflumethiazide



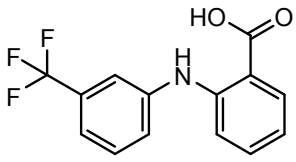
Alclofenac



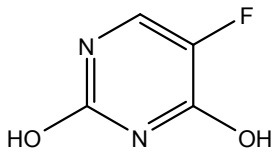
Flurbiprofen



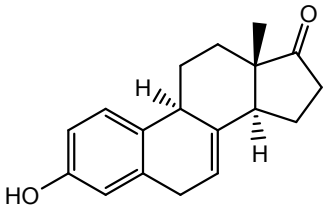
Famotidine



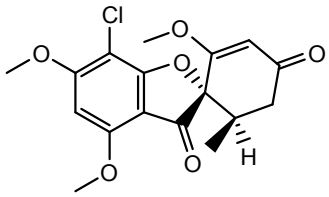
Flufenamic acid



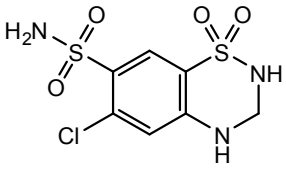
5-Fluorouracil



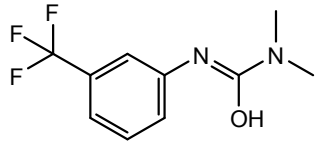
Equilin



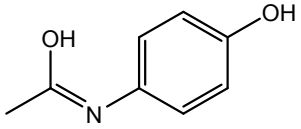
Griseofulvin



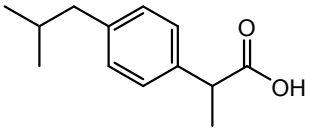
Hydrochlorothiazide



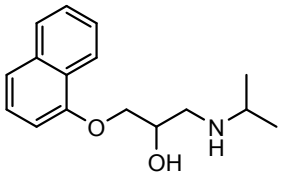
Fluometuron



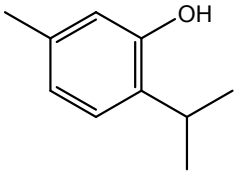
Paracetamol



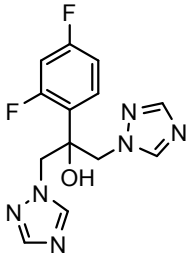
Ibuprofen



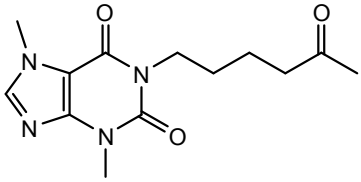
Propranolol



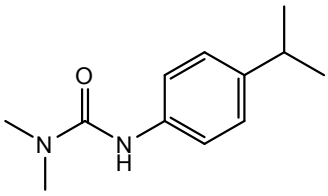
Thymol



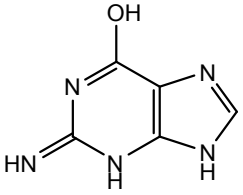
Fluconazole



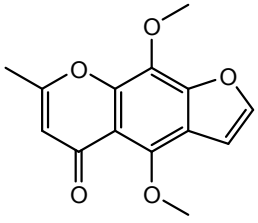
Pentoxifylline



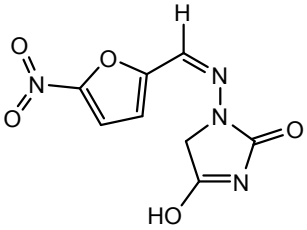
Isoproturon



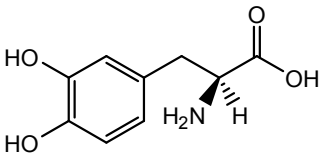
Guanine



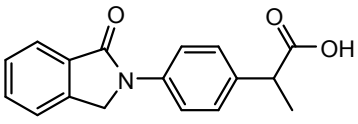
Khellin



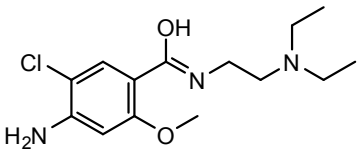
Nitrofurantoin



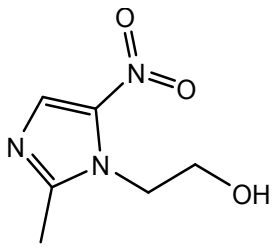
L-DOPA (Levodopa)



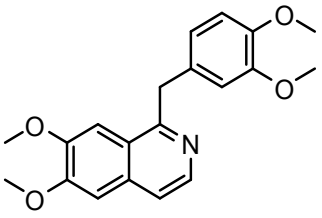
Indoprofen



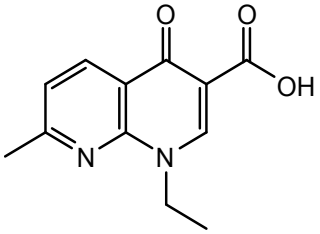
Metoclopramide



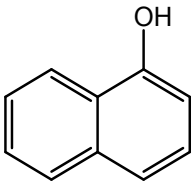
Metronidazole



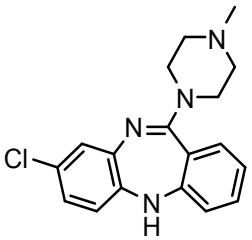
Papaverine



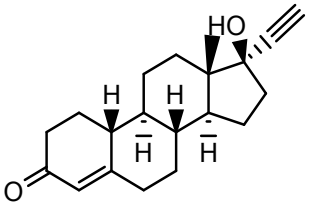
Nalidixic acid



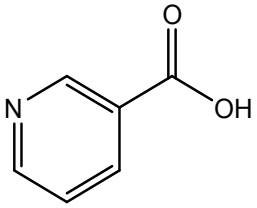
1-Naphthol



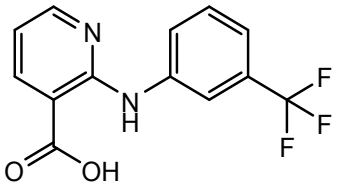
Clozapine



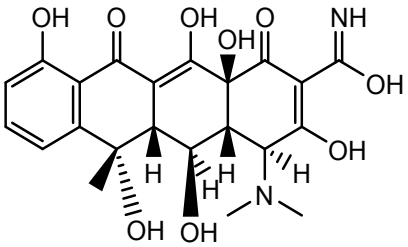
Norethisterone



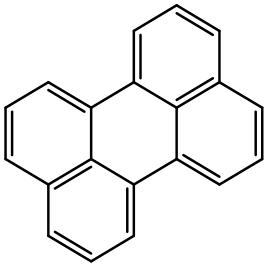
Nicotinic acid



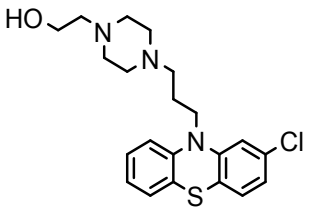
Niflumic acid



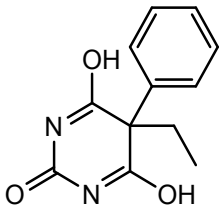
Oxytetracycline



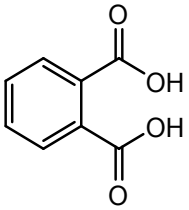
Perylene



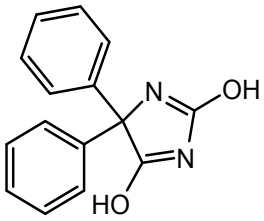
Perphenazine



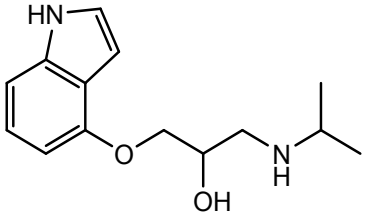
Phenobarbital



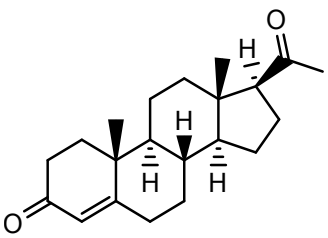
Phthalic acid



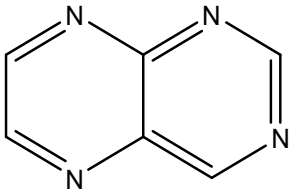
5,5-Diphenylhydantoin



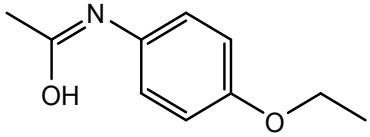
Pindolol



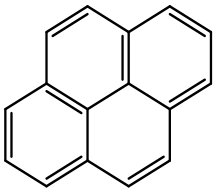
Progesterone



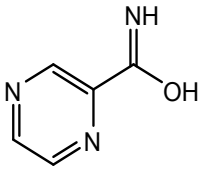
Pteridine



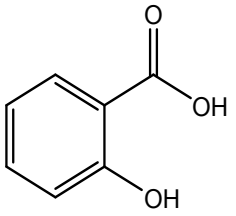
Phenacetin



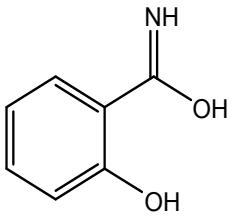
Pyrene



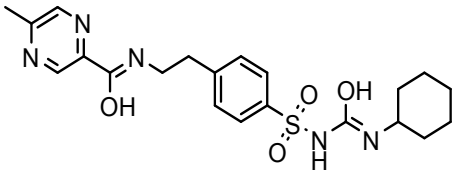
Pyrazinamide



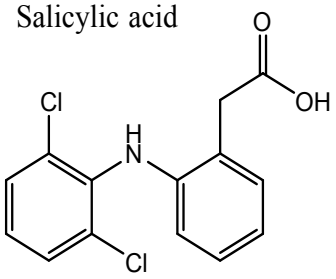
Salicylic acid



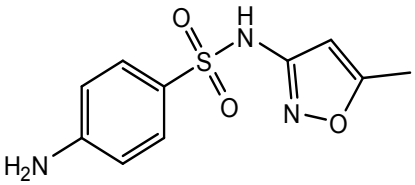
Salicylamide



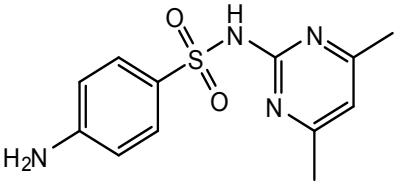
Glipizide



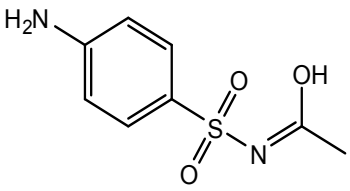
Diclofenac



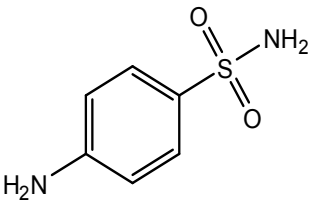
Sulfamethoxazole



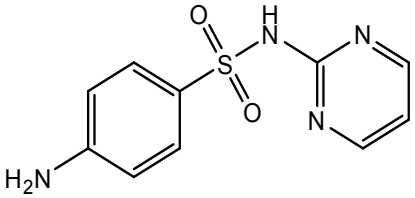
Sulfamethazine



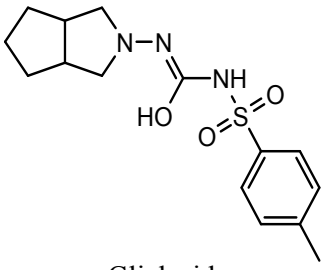
Sulfacetamide



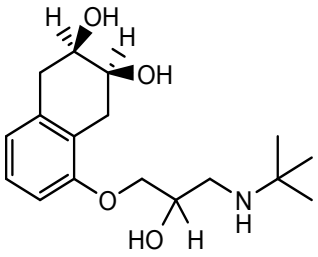
Sulfanilamide



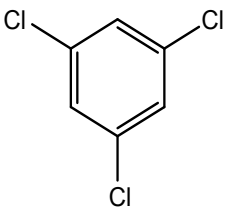
Sulfadiazine



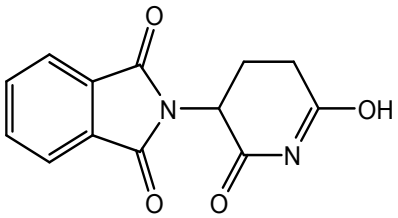
Gliclazide



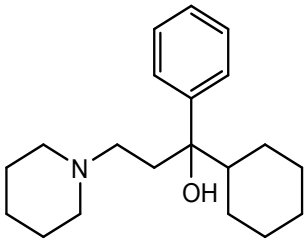
Nadolol



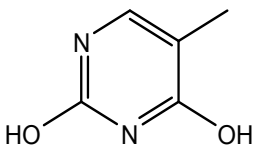
1,3,5-trichlorobenzene



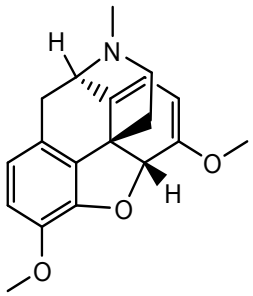
Thalidomide



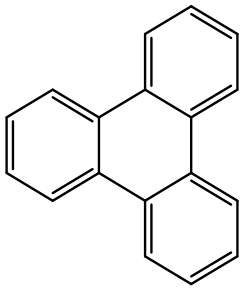
Trihexyphenidyl



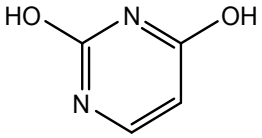
Thymine



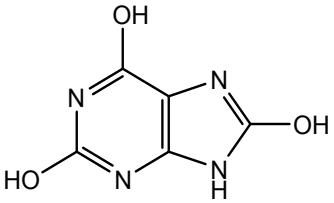
Thebaine



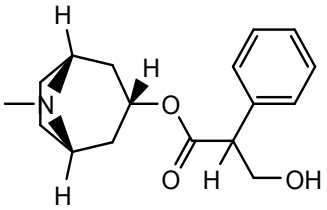
Triphenylene



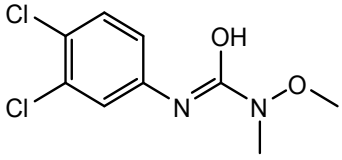
Uracil



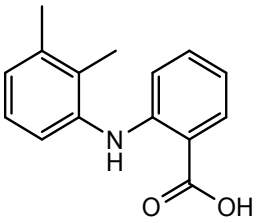
Uric acid



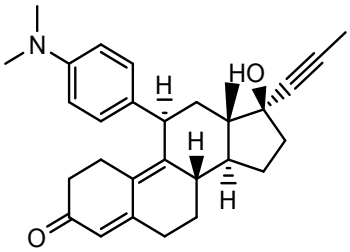
Atropine



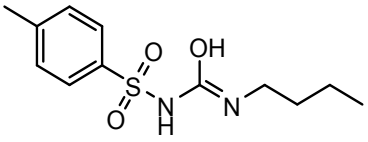
Linuron



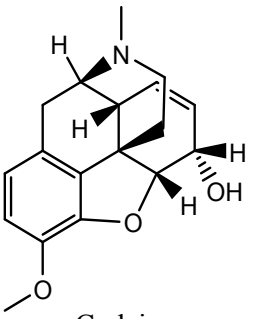
Mefenamic acid



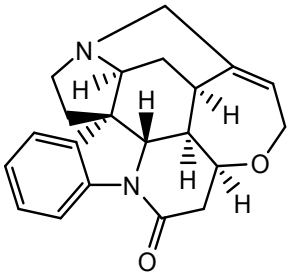
Mifepristone



Tolbutamide



Codeine



Strychnine

Figure H.1: DLS-100 dataset, structures converted from InChI strings to 2D structures

Descriptor	No. of descriptors	Meaning
U_{latt}	1	The lattice energy of the crystal as calculated by DMACRYS.
S_{crys}	1	The entropy of the crystal calculated by DMACRYS.
S_{rot}	1	The rotational entropy of a single molecule in the gas phase, as calculated by Gaussian 09.
S_{trans}	1	The translational entropy of a single molecule in the gas phase, as calculated by Gaussian 09.
ΔG_{sub}	1	Sublimation free energy predicted by thermochemical calculations.
Gaseous energy	1	The gaseous absolute energy as calculated by Gaussian 09.
solution energy	1	The solution absolute energy as calculated by Gaussian 09.
ΔG_{hyd}	1	The free energy of hydration calculated by thermochemical calculations.
ΔG_{solv}	1	The solvation free energy calculated by thermochemical calculations.

Table H.3: Calculated physical chemical values used as descriptors.

Descriptors Explanations

Descriptor	No. of descriptors	Meaning
ALOGP	3	Stands for additive logP. It is calculated using molar refractivity values defined by Ghose and Crippen.
BCUT	6	Used for defining chemical diversity by encoding intermolecular interactions.
Fragment complexity	1	Represents the complexity of a fragment in terms of bond number number of hydrogen atoms and hetroatoms.
Aromatic atoms count	1	Count of the number of atoms in an aromatic system.
Aromatic bonds count	1	Count of the number of aromatic bonds.

Continued on next page

Table H.4 – *Continued from previous page*

Descriptor	No. of descriptors	Meaning
Atom count	14	Count of the number of time a particular element is present in a molecule.
Autocorrelation (charge)	5	Describes how charge varies along a molecular structure.
Autocorrelation (mass)	5	Describes how mass varies along a molecular structure.
Autocorrelation (polarizability)	5	Describes how polarizability varies along a molecular structure.
Number of bonds	1	Counts the number of binds.
BPol	1	The sum of squared difference between atomic polarizabilities.
Carbon types	9	Topologically describes the connectivity between Carbon atoms.
Chi chain	10	Descriptors providing information on molecular branching and size.
Chi cluster	8	Descriptors providing information on molecular branching and size.
Chi path cluster	6	Descriptors providing information on molecular branching and size.
Chi path	16	Descriptors providing information on molecular branching and size.
Eccentric connectivity index	1	Topological descriptor for distance and information on which atoms are adjacent to others.
Hbond donor count	2	A count of the number H bond donors.
Kier Hall Smarts	79	A count of E-state fragments.
Kappa shape indices	3	Compares the extent of molecular shape complexity against linear and completely bonded.
Largest chain	1	The number of atoms in the longest chain of atoms.
Largest Pi system	1	The number of atoms in the largest pi system.
Longest aliphatic chain	1	The number of atoms in the longest aliphatic group.
Petitjean number	1	A measure of the maximum separating distance between the two most distant vertices.
Topological	4	Provides information on the topological shape.

Continued on next page

Table H.4 – *Continued from previous page*

Descriptor	No. of descriptors	Meaning
Number of rotatable bonds	1	Count the number of rotatable bonds.
TPSA	1	Estimate of the topological polar surface area
Vertex adjacency information	1	Provides informations on the adjacency of vertices.
Molecular weight	1	Calculation of molecular weight.
Weighted path	5	Branching descriptor.
Wiener numbers	2	Information is encoded on the size of paths and polarity.
X log P	1	Atom type calculation estimating Log P.
Zagreb index	1	Complexity descriptor.

Table H.4: Descriptors used from the CDK two dimensional descriptors.

Appendix I

Statistical Significance and Descriptor Importance

This appendix presents the statistical significance data from the cheminformatics solubility prediction project in **Chapter 5 Section 4.1**. This Appendix also contains tables showing the most important descriptors in the Random Forest models scaled by: the mean and standard deviation, PCA and using the raw data.

Scaled by the mean and standard deviation

Partial Least Square

	mxd	hfd	dd	hf	mx
mxd	x				
hfd	0.14	x			
dd	0.19	0.06	x		
hf	0.00	0.02	0.04	x	
mx	0.00	0.02	0.09	0.20	x

Support Vector Regression

	mxd	hfd	dd	hf	mx
mxd	x				
hfd	0.29	x			
dd	0.36	0.07	x		
hf	0.03	0.04	0.13	x	
mx	0.05	0.06	0.09	0.37	x

Random Forest Regression

	mxd	hfd	dd	hf	mx
mxd	x				
hfd	0.26	x			
dd	0.02	0.11	x		
hf	0.00	0.01	0.01	x	
mx	0.02	0.02	0.07	0.25	x

mxd = M06-2X + CHEMOINFORMATICS DESCRIPTORS
hfd = HF + CHEMOINFORMATIC DESCRIPTORS
dd = CHEMOINFORMATICS DESCRIPTORS
hf = HF
mx = MX06-2X

Chemoinformatics descriptors

	SVR	RF	PLS
SVR	x		
RF	0.13	x	
PLS	0.18	0.23	x

HF + Chemoinformatics Descriptors

	SVR	RF	PLS
SVR	x		
RF	0.12	x	
PLS	0.06	0.22	x

MX06-2X + Chemoinformatics Descriptors

	SVR	RF	PLS
SVR	x		
RF	0.03	x	
PLS	0.16	0.28	x

HF

	SVR	RF	PLS
SVR	x		
RF	0.25	x	
PLS	0.03	0.01	x

MX06-2X

	SVR	RF	PLS
SVR	x		
RF	0.20	x	
PLS	0.03	0.01	x

SVR = SUPPORT VECTOR REGRESSION
RF = RANDOM FOREST
PLS = PARTIAL LEAST SQUARE

Principal components

Partial Least Square

	mxd	hfd	dd	hf	mx
mxd	x				
hfd	0.18	x			
dd	0.11	0.15	x		
hf	0.00	0.01	0.01	x	
mx	0.00	0.02	0.01	0.11	x

Support Vector Regression

	mxd	hfd	dd	hf	mx
mxd	x				
hfd	0.31	x			
dd	0.23	0.08	x		
hf	0.09	0.19	0.12	x	
mx	0.05	0.16	0.19	0.23	x

Random Forest Regression

	mxd	hfd	dd	hf	mx
mxd	x				
hfd	0.10	x			
dd	0.01	0.08	x		
hf	0.19	0.20	0.38	x	
mx	0.01	0.01	0.10	0.07	x

mxd = M06-2X + CHEMOINFORMATICS DESCRIPTORS
hfd = HF + CHEMOINFORMATIC DESCRIPTORS
dd = CHEMOINFORMATICS DESCRIPTORS
hf = HF
mx = MX06-2X

Chemoinformatics descriptors

	SVR	RF	PLS
SVR	x		
RF	0.41	x	
PLS	0.20	0.23	x

HF + Chemoinformatics Descriptors

	SVR	RF	PLS
SVR	x		
RF	0.15	x	
PLS	0.13	0.25	x

MX06-2X + Chemoinformatics Descriptors

	SVR	RF	PLS
SVR	x		
RF	0.02	x	
PLS	0.06	0.08	x

HF

	SVR	RF	PLS
SVR	x		
RF	0.11	x	
PLS	0.01	0.00	x

MX60-2X

	SVR	RF	PLS
SVR	x		
RF	0.15	x	
PLS	0.04	0.26	x

SVR = SUPPORT VECTOR REGRESSION
RF = RANDOM FOREST
PLS = PARTIAL LEAST SQUARE

Raw data set

Partial Least Square

	mxd	hfd	dd	hf	mx
mxd	x				
hfd	0.03	x			
dd	0.19	0.17	x		
hf	0.11	0.13	0.21	x	
mx	0.17	0.24	0.23	0.27	x

Support Vector Regression

	mxd	hfd	dd	hf	mx
mxd	x				
hfd	0.28	x			
dd	0.24	0.29	x		
hf	0.06	0.22	0.11	x	
mx	0.09	0.14	0.20	0.37	x

Random Forest Regression

	mxd	hfd	dd	hf	mx
mxd	x				
hfd	0.23	x			
dd	0.02	0.16	x		
hf	0.01	0.01	0.01	x	
mx	0.02	0.02	0.07	0.25	x

mxd = M06-2X + CHEMOINFORMATICS DESCRIPTORS
hfd = HF + CHEMOINFORMATIC DESCRIPTORS
dd = CHEMOINFORMATICS DESCRIPTORS
hf = HF
mx = MX06-2X

Chemoinformatics descriptors

	SVR	RF	PLS
SVR	x		
RF	0.10	x	
PLS	0.10	0.05	x

HF + Chemoinformatics Descriptors

	SVR	RF	PLS
SVR	x		
RF	0.06	x	
PLS	0.07	0.01	x

MX06-2X + Chemoinformatics Descriptors

	SVR	RF	PLS
SVR	x		
RF	0.07	x	
PLS	0.17	0.02	x

HF

	SVR	RF	PLS
SVR	x		
RF	0.25	x	
PLS	0.01	0.00	x

MX60-2X

	SVR	RF	PLS
SVR	x		
RF	0.20	x	
PLS	0.01	0.01	x

SVR = SUPPORT VECTOR REGRESSION
RF = RANDOM FOREST
PLS = PARTIAL LEAST SQUARE

Descriptor Importance

Descriptor only	Descriptor and HF	Descriptor and M06-2X	HF	M06-2X
X log P	X log P	X log P	ΔG solvation	ΔG solvation
WTPT 3	WTPT 3	DFT log S	HF log S	ΔG solution
VCH 7	DFT log S	ΔG solution	ΔG solution	DFT log S
ATSc2	ΔG solution	WTPT 3	ΔG sublimation	$S_{rotations}$
SP 6	VCH 7	VCH 7	U_{latt}	S_{trans}
ATSc1	ΔG solvation	ΔG solvation	$S_{crystal}$	Solution energy
SP 5	ATSc1	ATSc1	$S_{rotations}$	U_{latt}
SP 7	SP 6	ATSc2	$S_{translation}$	$S_{crystal}$
ATSm4	ATSc2	WTPT 2	Soln energy	Gas energy
ATSm1	WTPT 2	SP 6	Gas energy	ΔG sublimation

Table I.1: Top 10 variables ranking of variable importance in Random Forest scaled by mean/σ.

Descriptor only	Descriptor and HF	Descriptor and M06-2X	HF	M06-2X
X log P	X log P	X log P	ΔG solvation	ΔG solvation
WTPT 3	WTPT 3	ΔG solution	HF log S	ΔG solution
VCH 7	DFT log S	DFT log S	ΔG solution	DFT log S
ATSc2	ΔG solution	WTPT 3	ΔG sublimation	$S_{rotations}$
ATSc1	VCH 7	ΔG solvation	U_{latt}	$S_{translation}$
SP 6	ΔG solvation	VCH 7	$S_{crystal}$	Solution energy
SP 5	ATSc1	ATSc1	$S_{rotations}$	U_{latt}
ATSm5	ATSc2	ATSc2	$S_{translations}$	$S_{crystal}$
ATSm4	SP 6	WTPT 2	Solution energy	Gas energy
SP 7	SP 5	SP 6	Gas energy	ΔG sublimation

Table I.2: Top 10 variables ranking of variable importance in Random Forest raw data.

CDK	Descriptors Definitions
XLogP	Predicted logP based on the atom-type.
WTPT.3	A set of weighted path descriptors from Randic. They describe the degree of molecular branching.
VCH.7	Kier and Hall's χ indices (orders 3 -6).
ATSc2	The Moreau-Broto autocorrelation partial charge model.
SP.6	Kier and Hall's χ path indices (orders 0-7).
ATSc1	The Moreau-Broto autocorrelation partial charge model.
SP.5	Kier and Hall's χ path indices orders (0 - 7).
SP.7	Kier and Hall's χ path indices (orders 0 - 7).
ATSm4	The Moreau-Broto autocorrelation calculated by the application of partial charges.
WTPT.2	A set of weighted path descriptors from Randic. They describe the degree of molecular branching.
ATSm5	The Moreau-Broto autocorrelation calculated by the application of partial charges.

Table I.3: Descriptor names and meaning.

Appendix J

MP1100

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
(1,2-dibromoethyl)benzene	4.01	72
(1-bromoethyl)benzene	3.61	-65
(1R)-camphor	2.85	177.5
(1R)-camphor-10-sulfonic acid	-0.53	200
(1R)-camphor-10-sulfonyl chloride	2.06	66
(1R)-camphorquinone	2.57	199.5
(1R)-endo-fenchyl alcohol	2.3	40.5
(1R,2s)-10,2-camphorsultam	1	180.5
(1R,3S)-camphoric acid	1.51	185.5
(1S)-camphanic acid	1.77	200
(1S)-camphanic chloride	2.15	66
(1S)-camphor	2.85	177.5
(1S)-camphor-10-sulfonyl chloride	2.06	66
(1S)-camphorquinone	2.57	198.5
(1S)-camphorsulfonylimine	1.55	227
(1S,2R)-10,2-camphorsultam	1	183
(1S,2R)-1-phenyl-2-(1-pyrrolidinyl)-1-propanol	2.04	45
(1S,4S)-2-boc-2,5-diazabicyclo(2.2.1)heptane	0.72	75
(2-aminoethoxy)acetic acid	-3.18	178
(2-bromoethyl)benzene	3.28	-56
(2-carboxyphenyl)iminodiacetic acid	0.81	217
(2h)1,4-benzothiazin-3(4h)-one	1.65	177

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
(2-hydroxyethyl)hydrazine	-1.48	-70
(2S,5s)-2,5-hexanediol	0.31	53
(3,4-dimethoxyphenylthio)acetic acid	2	102
(3-bromo-2,4,6-trimethylphenylcarbamoyl)methyliminodiacetic acid	-0.39	194
(3-chloropropyl)trimethoxysilane	1.98	-50
(3R-cis)-tetrahydro-3-trichloromethyl-1h,3h-pyrrolo(1,2-c)oxazol-1-one	2.06	110
(3S,4R)-4-(4-fluorophenyl)-1-methyl-3-piperidinemethanol	2.06	95.5
(4-bromobutoxy)benzene	3.73	41.5
(4-chlorophenoxy)acetyl chloride	2.65	19
(4-chlorophenylsulfonyl)acetonitrile	1	171
(4-chlorophenylthio)acetic acid	2.71	105.5
(4-chlorophenylthio)acetonitrile	2.91	80
(4-fluorophenylthio)acetic acid	2.05	77.5
(4-fluorophenylthio)acetonitrile	2.31	33.5
(4-imidazolyl)acetonitrile	0.43	137
(4-tert-butylphenoxy)acetonitrile	3.37	68.5
(5-mercapto-1,3,4-thiadiazol-2-ylthio)acetic acid	0.31	167
(benzylthio)acetic acid	2.2	61.5
(chloromethyl)cyclopropane	2.12	-91
(e)-3-dimethylamino-1-(2-pyridyl)-2-propen-1-one	0.49	131
(e)-alpha-(4-chlorophenyl)cinnamonitrile	4.44	115.5
(e)-alpha-methylstilbene	4.84	80.5
(methoxymethyl)diphenylphosphine oxide	1.63	117

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
(methylamino)acetaldehyde dimethyl acetal	-0.3	-73
(methylthio)acetic acid	-0.15	13.5
(pentafluorophenyl)diphenylphosphine	5.02	71
(phenylsulfonyl)acetamide	-0.36	156
(phenylsulfonyl)acetic acid	0.18	111.5
(phenylthio)acetic acid	2.28	63.5
(R)-1-boc-3-hydroxypiperidine	1.51	46.5
(R)-3-(boc-amino)-3-(4-bromophenyl)propionic acid	3.01	144
(R)-3-boc-thiazolidine-2-carboxylic acid	1.15	91.5
(R)-limonene	4.5	-74
(S)-1-boc-3-hydroxypiperidine	1.51	37
(S)-2-(6-methoxy-2-naphthyl)propionic acid	3.29	155
(S)-2-(boc-amino)-4-phenylbutyric acid	2.92	78
(S)-2-pyrrolidinone-5-carboxylic acid	-1.01	160
(S)-3-(boc-amino)-4-(4-pyridyl)butyric acid	1.43	141.5
(S)-3-(boc-amino)-4-phenylbutyric acid	2.82	104
(S)-3-(boc-amino)-5-methylhexanoic acid	2.23	54
(S)-3-(boc-amino)piperidine	1.17	124.5
(S)-3-boc-thiazolidine-2-carboxylic acid	1.15	91.5
(S)-4-sec-butyloxazolidine-2,5-dione	0.73	65.5

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
(S)-alpha-methoxy-alpha-(trifluoromethyl)phenylacetic acid	2.01	44
(S)-nicotine	0.87	-79
(S,s)-hydrobenzoin	2.05	149
(trimethylsilyl)acetic acid	1.33	41
(2.2)paracyclophane	5.23	285.5
(4-(trifluoromethyl)phenylsulfonyl)acetonitrile	1.49	143.5
(4-(trifluoromethyl)phenylthio)acetic acid	2.98	116
1-(1-methyl-4-piperidinyl)piperazine	0.1	30.5
1-(1-propynyl)cyclohexanol	2.49	48
1-(2,4,6-triisopropylphenylsulfonyl)-1,2,4-triazole	3.03	111
1-(2,4-difluorophenyl)piperazine	1.43	75
1-(2,5-dimethylphenyl)piperazine	2.18	44.5
1-(2-aminoethyl)piperazine	-1.4	-19
1-(2-aminophenyl)pyrrole	2.06	93.5
1-(2-bromoethyl)-4-nitrobenzene	3.12	69
1-(2-fluorophenyl)piperazine	1.35	46
1-(2-furoyl)piperazine	-0.13	68
1-(2-furyl)-2-nitroethylene	1.39	73.5
1-(2-hydroxyethyl)-2-imidazolidinone	-1.57	48.5
1-(2-hydroxyethyl)piperazine	-1.67	-39
1-(2-hydroxyethyl)piperidine	0.56	16
1-(2-methoxyethyl)homopiperazine	-0.26	66.5
1-(2-methoxyphenyl)piperazine	1.42	37
1-(2-naphthoyl)-3,3,3-trifluoroacetone	3.29	74
1-(2-naphthyl)ethanol	2.77	73.5
1-(2-nitrophenyl)piperidine	3.29	77
1-(2-nitrophenyl)pyrrole	2.81	59

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
1-(2-phenylethyl)-4-(phenylethynyl)benzene	6.07	105
1-(2-phenylethyl)-4-piperidone	1.81	58
1-(2-pyridylazo)-2-naphthol	3.87	139.5
1-(2-thenoyl)-3,3,3-trifluoroacetone	2.37	42.5
1-(2-thiazolylazo)-2-naphthol	4.17	139
1-(2-trifluoromethylphenyl)imidazole	2.28	51.5
1-(3,4-dichlorophenyl)piperazine	2.72	63
1-(3,4-dimethylphenyl)piperazine	2.18	62.5
1-(3,5-dichlorophenyl)-2,5-dimethyl-1h-pyrrole	4.83	79
1-(3-aminophenyl)ethanol	0.49	67.5
1-(3-aminopropyl)imidazole	-0.71	-68
1-(3-chlorophenoxy)-3-butyn-2-ol	2.04	38
1-(3-hydroxypropyl)piperazine	-1.22	50
1-(3-methoxybenzoyl)-2-(1-naphthoyl)hydrazine	2.83	190
1-(3-trifluoromethylphenoxy)-3-butyn-2-ol	1.87	31.5
1-(4-aminophenyl)ethanol	0.5	69.5
1-(4-biphenylyl)ethanol	3.42	97
1-(4-bromophenyl)ethanol	2.45	37
1-(4-chlorophenyl)cyclohexane-1-carboxylic acid	3.73	152
1-(4-chlorophenyl)cyclopropanecarboxylic acid	2.76	153.5
1-(4-chlorophenylsulfonyl)-3,3-dimethyl-2-butanone	2.69	97
1-(4-ethoxyphenyl)ethanol	2.28	46.5
1-(4-ethoxyphenyl)ethynyl-4-n-pentylbenzene	6.42	62
1-(4-fluorophenyl)piperazine	1.49	31.5

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
1-(4-fluorophenyl)pyrrole	2.77	51
1-(4-hydroxyphenyl)-5-mercaptotetrazole	1.31	170
1-(4-iodophenyl)pyrrole	3.57	130.5
1-(4-methoxy-4-biphenylylsulfonyl)proline	2.18	154
1-(4-methoxybenzoyl)-2-(1-naphthoyl)hydrazine	2.86	188.5
1-(4-methoxyphenyl)-1-cyclohexanecarbonitrile	3.81	43
1-(4-methoxyphenyl)-1h-1,2,4-triazole	1.16	97
1-(4-methoxyphenyl)ethynyl-4-n-pentylbenzene	6.13	47
1-(4-methoxyphenyl)ethynyl-4-n-propylbenzene	5.3	61
1-(4-methoxyphenyl)imidazole	1.62	66
1-(4-nitrophenyl)-3-(2-thienyl)-2-propen-1-one	3.53	169
1-(4-nitrophenyl)-5-(trifluoromethyl)-1h-pyrazole-4-carboxylic acid	2.68	201
1-(4-nitrophenyl)glycerol	0.74	96
1-(4-nitrophenyl)piperazine	1.27	131
1-(4-nitrophenyl)piperidine	3.37	103.5
1-(4-nitrophenylazo)-2-naphthol	5.04	250
1-(4-pyridyl)piperazine	0.56	138.5
1-(6-methoxy-2-naphthyl)ethanol	2.9	111.5
1-(boc-amino)cyclopentanecarboxylic acid	2.12	132.5
1-(chloromethyl)naphthalene	3.91	32
1-(cyanoacetyl)piperidine	0.49	87.5
1-(cyanoacetyl)pyrrolidine	-0.14	72.5
1-(heptafluorobutyryl)imidazole	2.4	10.5
1-(methylsulfonyl)imidazole	-1.01	88.5
1-(o-tolyl)piperazine	1.7	46.5

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
1-(pentafluorobenzoyl)imidazole	2.51	55.5
1-(pentafluorophenyl)ethanol	1.94	33
1-(phenylethynyl)cyclohexanol	3.34	61
1-(phenylsulfonyl)indole	2.56	78
1-(phenylsulfonyl)pyrrole	1.53	88.5
1-(p-toluenesulfonyl)indole	2.73	85
1-(p-toluenesulfonyl)pyrrole	1.66	101
1-(trans-cinnamoyl)imidazole	2.14	130
1-(trifluoromethyl)cyclohexanecarboxylic acid	1.71	69.5
1-(trifluoromethyl)cyclopentanecarboxylic acid	1.39	37
1-(trimethylsilyl)imidazole	0.78	-42
1,1-(azodicarbonyl)dipiperidine	2.09	134.5
1,1,1,2-tetrafluoro-2-iodo-2-(trifluoromethoxy)ethane	2.55	43.5
1,1,1,3,3,3-hexafluoro-2-propanol	2.58	-3
1,1,1-tris(chloromethyl)ethane	2.86	18
1,1,1-tris(hydroxymethyl)ethane	-1.34	190
1,1,2,2-tetrachloroethane	2.57	-43
1,1,2,2-tetrafluoroethyl methyl ether	1.83	-107
1,1,2-trichloro-3,3,3-trifluoro-1-propene	3.19	-114
1,1,2-trimethyl-1h-benzo(e)indole	4.38	116.5
1,1,3,3,5,5-hexamethyltrisiloxane	2.71	-67
1,1,3,3-tetramethyldisiloxane	1.66	-78
1,1,3-triphenylpropargyl alcohol	4.6	81
1,1,4,4-tetraphenyl-1,3-butadiene	7.24	197.5

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
1,1:3,1-terphenyl-5-boronic acid	4.34	294
1,10-decanedicarbonitrile	2.86	17.5
1,10-decanediol	2.7	72.5
1,10-decanedithiol	4.77	17
1,10-dibromodecane	3.6	28
1,10-diiododecane	5.55	30
1,10-phenanthroline, anhydrous	2.31	117.5
1,12-diaminododecane	3.47	70
1,12-dibromododecane	5.61	40
1,12-dodecanediol	3.74	82.5
1,18-octadecanedicarboxylic acid	6.79	127
1,1-bi(2-naphthol)	4.6	216.5
1,1-bis(methylthio)-2-nitroethylene	0.87	125.5
1,1-carbonyldiimidazole	-0.33	118
1,1-carbonyldipiperidine	1.61	44.5
1,1-cyclohexanediacetic acid	1.31	180
1,1-cyclopropanedicarboxylic acid	-0.21	131.5
1,1-cyclopropanedicarboxylic acid monomethyl ester	0.28	49
1,1-dioxobenzo(b)thiophen-2-ylmethyl chloroformate	1.69	76.5
1,1-diphenyl-2-propyn-1-ol	3.04	46
1,1-diphenylacetone	3.57	61
1,1-diphenylethanol	3.3	80
1,1-diphenylethylene	4.53	6
1,1-thiocarbonyldi-2(1h)-pyridone	1.59	164.5
1,2,3,4-tetrafluorobenzene	2.34	-42
1,2,3,4-tetrahydrocarbazole	3.77	119
1,2,3,4-tetrahydroisoquinoline	1.31	-30
1,2,3,4-tetrahydronaphthalene	3.79	-36
1,2,3,4-tetrahydroquinoline	2.27	14
1,2,3-benzotriazin-4(3h)one	0.25	223

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
1,2,3-hexanetriol	-0.85	65.5
1,2,3-thiadiazole-4-carboxaldehyde	0.25	85
1,2,3-thiadiazole-4-carboxylic acid	-0.66	227.5
1,2,3-triacetoxybenzene	1.58	166
1,2,3-tribromopropane	3.24	16.5
1,2,3-trichloro-4-nitrobenzene	3.65	54.5
1,2,3-trichloro-5-nitrobenzene	3.69	69.5
1,2,3-trichlorobenzene	4.07	54
1,2,3-trichloropropane	2.29	-14
1,2,3-trimethoxybenzene	2.02	44
1,2,4,5-tetrabromobenzene	5.03	177.5
1,2,4,5-tetrachlorobenzene	4.61	140
1,2,4,5-tetrafluorobenzene	2.43	4
1,2,4,5-tetrakis(isopropylthio)benzene	7.01	77
1,2,4,5-tetramethylbenzene	4.05	78.5
1,2,4-benzenetricarboxylic anhydride	0.98	167
1,2,4-butanetriol	-1.59	-20
1,2,4-triacetoxybenzene	1.63	99
1,2,4-triazole	-0.8	120
1,2,4-triazolo(4,3-a)pyridin-3(2h)-one	0.19	234
1,2,4-triazolo(4,3-a)pyridine-3-thiol	1.2	209.5
1,2,4-tribromobenzene	4.42	42
1,2,4-trichloro-5-iodobenzene	4.48	103.5
1,2,4-trichlorobenzene	4.08	17
1,2,4-trifluoro-5-nitrobenzene	1.97	-11
1,2,4-trimethylbenzene	3.62	-44
1,2,4-triphenyl-1,4-butanedione	4.81	128
1,2,5-trichloro-3-iodobenzene	4.47	52
1,2,6-hexanetriol	-0.8	-20
1,2-benzenedimethanol	0.33	61.5
1,2-benzenedithiol	1.8	23
1,2-benzisoxazol-3(2h)-one	0.79	138.5

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
1,2-bis(1-naphthyl)ethane	6.63	161.5
1,2-bis(2-chloroethoxy)ethane	1.26	-32
1,2-bis(2-nitrophenoxy)ethane	3.29	168.5
1,2-bis(carboxymethylthio)ethane	0.11	110
1,2-bis(chlorodimethylsilyl)ethane	3.58	37
1,2-bis(dimethoxyphosphoryl)benzene	0.78	81
1,2-bis(dimethylphosphino)ethane	2.77	179.5
1,2-bis(diphenylphosphino)benzene	7.92	187
1,2-bis(diphenylphosphino)ethane	6.88	140.5
1,2-bis(methanesulfonamido)benzene	0.14	212.5
1,2-bis(phenylsulfonyl)ethane	1.6	179.5
1,2-bis(phenylthio)ethane	4.98	69
1,2-bis(2-(trifluoromethyl)phenyl)ethane	5.19	75
1,2-cyclohexanedione	0.77	36.5
1,2-di(3-indenyl)ethane	5.88	123
1,2-di(p-tolyl)ethane	5.79	81
1,2-diaminopropane	-1.36	-37
1,2-dianilinoethane	3.34	66
1,2-dibenzoylbenzene	4.51	146.5
1,2-dibenzoylethane	3.2	146
1,2-dibromo-2,4-dicyanobutane	2.14	50
1,2-dibromo-3,5-difluorobenzene	3.62	37
1,2-dibromo-4,5-difluorobenzene	3.67	33
1,2-dibromobenzene	3.77	5
1,2-dibromobutane	3.38	-65
1,2-dibromoethane	2.08	9.5
1,2-dibromopropane	2.65	-55
1,2-dichloro-3-iodobenzene	4.07	35
1,2-dichloro-3-nitrobenzene	3.05	62
1,2-dichloro-4-fluorobenzene	3.46	-1

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
1,2-dichloro-4-iodobenzene	4.07	30
1,2-dichloro-4-nitrobenzene	3.11	41
1,2-dichlorobenzene	3.45	-18
1,2-dichloroethane	1.48	-35
1,2-dichloropropane	2.13	-100
1,2-diethoxybenzene	3.09	44
1,2-diethoxyethane	0.78	-74
1,2-diethylbenzene	4.55	-31
1,2-difluoro-4,5-dimethoxybenzene	2.2	40
1,2-difluorobenzene	2.24	-34
1,2-dihydronaphthalene	3.53	-8
1,2-diiodoethane	2.72	82
1,2-dimethoxy-4,5-dinitrobenzene	1.58	132.5
1,2-dimethoxyethane	0.03	-69
1,2-dimethyl-5-nitroimidazole	-0.01	137
1,2-dimethylimidazole	0.37	36.5
1,2-diphenoxyethane	3.42	95
1,2-diphenylethane	4.74	51.5
1,2-epoxyoctadecane	8.31	32.5
1,2-ethanedithiol	0.84	-41
1,2-ethylenediphosphonic acid	-0.89	219
1,2-octanediol	1.66	31
1,2-o-isopropylidene-alpha-D-glucofuranose	-0.94	159.5
1,2-phenylene phosphorochloridite	2.53	30
1,2-phenylenediacetic acid	1.2	151
1,2-phenylenediacetonitrile	0.69	59
1,2-propanediol	-1.1	-60
1,2-propanediol diacetate	0.77	-31
1,3,3-trimethyl-2-methyleneindoline	3.38	-10
1,3,5,7-cyclooctatetraene	3.1	-5
1,3,5-benzenetricarbonyl chloride	3.12	36
1,3,5-benzenetricarboxylic acid	0.87	375
1,3,5-tribenzoylbenzene	5.45	117.5
1,3,5-tribenzylhexahydro-1,3,5-triazine	3.42	50
1,3,5-tribromobenzene	4.42	122.5

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
1,3,5-trichloro-2,4,6-trifluorobenzene	4.06	63
1,3,5-trichloro-2-iodobenzene	4.48	54.5
1,3,5-trichloro-2-nitrobenzene	3.6	70.5
1,3,5-trichlorobenzene	4.08	63.5
1,3,5-triethynylbenzene	2.03	106
1,3,5-trifluoro-2-nitrobenzene	1.89	3.5
1,3,5-trifluorobenzene	2.36	-5.5
1,3,5-triisopropylbenzene	5.52	-7
1,3,5-trimethoxy-2-nitrobenzene	1.65	151.5
1,3,5-trimethoxybenzene	1.96	52
1,3,5-trimethyl-1h-pyrazole	0.73	35.5
1,3,5-trimethyl-1h-pyrazole-4-carboxaldehyde	0.58	81.5
1,3,5-trioxane	-0.95	60.5
1,3,5-triphenylbenzene	7.3	173
1,3,5-tris(2-hydroxyethyl)cyanuric acid	-1.57	138
1,3,5-tri-tert-butylbenzene	7.17	71
1,3-benzenedimethanol	0.28	58
1,3-benzenedisulfonyl chloride	1.66	59
1,3-benzodioxole	1.71	-18
1,3-bis(2-hydroxyhexafluoroisopropyl)benzene	3.66	9.5
1,3-bis(4-piperidinyl)propane	2.73	65.5
1,3-bis(diphenylphosphino)propane	7.23	60.5
1,3-bis(hydroxymethyl)urea	-2.19	126
1,3-bis(trifluoromethyl)benzene	3.7	-35
1,3-bis((trimethylsilyl)ethynyl)benzene	4.73	58
1,3-bis(tris(hydroxymethyl)methylamino)propane	-2.14	169
1,3-cyclohexadiene	2.3	-89
1,3-cyclohexanediol, cis trans	0.02	30
1,3-cyclohexanedione	0.34	103

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
1,3-cyclooctadiene	3.56	-53
1,3-cyclopentanedione	-0.14	152
1,3-di(4-pyridyl)propane	2.39	54.5
1,3-di-2-thienyl-2-propen-1-one	3.6	99
1,3-diacetylbenzene	1.68	36
1,3-diacetylindole	2.29	145.5
1,3-diamino-2-propanol	-1.99	42.5
1,3-diaminopropane	-1.41	-12
1,3-dibenzoylbenzene	4.6	98.5
1,3-dibenzoyloxybenzene	4.42	117.5
1,3-dibenzyl-5-cyanohexahydropyrimidine	2.62	79.5
1,3-dibromo-2,2-diethylpropane	4.39	39.5
1,3-dibromo-2,2-dimethoxypropane	1.88	65.5
1,3-dibromo-5-fluoro-2-iodobenzene	4.08	134
1,3-dibromobenzene	3.73	-7
1,3-dibromopropane	2.55	-34
1,3-dichloro-2-fluorobenzene	3.48	38
1,3-dichloro-2-nitrosobenzene	3.07	171.5
1,3-dichloro-2-propanol	0.71	-4
1,3-dichloro-4-fluorobenzene	3.45	-23
1,3-dichloro-5,5-dimethylhydantoin	-1.48	132
1,3-dichloro-5-iodobenzene	4.07	57
1,3-dichloro-5-nitrobenzene	3.08	63.5
1,3-dichloroacetone	0.72	44
1,3-dichlorobenzene	3.45	-25
1,3-dichloropropene, cis trans	2.07	-84
1,3-diethoxybenzene	2.78	11
1,3-difluorobenzene	2.25	-59
1,3-dihydroxyacetone dimer	-1.91	78.5
1,3-dihydroxynaphthalene	2.02	124.5
1,3-diiodobenzene	3.8	36
1,3-diiodopropane	3.21	-20
1,3-diisopropylbenzene	4.71	-63
1,3-dimethoxybenzene	2	-52

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
1,3-dimethyl-2,4-dioxo-7-n-propyl-2,3,4,7-tetrahydropyrrolo(2,3-d)pyrimidine-6-carboxylic acid	0.7	139
1,3-dimethyl-2-imidazolidinone	-0.53	8
1,3-dimethyl-3,4,5,6-tetrahydro-2(1h)-pyrimidinone	-0.29	-23
1,3-dimethyl-6-methylamino-2,4-dioxo-1,2,3,4-tetrahydropyrimidine-5-carboxaldehyde	-1.11	203
1,3-dimethylbarbituric acid	-0.67	122
1,3-dimethyluracil	-0.98	120.5
1,3-di-n-butyl-2-thiobarbituric acid	2.7	62
1,3-dinitrobenzene	1.7	89
1,3-di-o-tolylguanidine	2.9	177
1,3-dioxolane	-0.61	-95
1,3-diphenoxy-2-propanol	2.7	82
1,3-diphenoxybenzene	4.96	60
1,3-diphenyl-1,3-propanedione	3.14	78
1,3-diphenyl-1-butanone	4.04	70
1,3-diphenylacetone	3.24	34
1,3-diphenylguanidine	2.67	149
1,3-diphenylisobenzofuran	6.15	134.5
1,3-di-tert-butylbenzene	5.58	9.5
1,3-dithiane	1.58	53.5
1,3-indanedione	1.54	130
1,3-phenylene diisocyanate	1.21	48
1,3-phenylenediacetic acid	1.3	175
1,3-phenylenediacetonitrile	0.72	33.5
1,3-propanediol	-1.18	-26
1,3-propanediol di-p-toluenesulfonate	1.68	91
1,3-propanedithiol	1.13	-79
1,3-propanesultone	-0.86	31.5
1,4,10,13-tetrathia-7,16-diazacyclooctadecane	1.75	130

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
1,4,5,6-tetrahydro-6-oxopyridazine-3-carboxylic acid	-1.49	197
1,4,6,7-tetramethylnaphthalene	5.28	63.5
1,4,8,11-tetraazacyclotetradecane	-0.66	186
1,4,8,11-tetraazatricyclo(9.3.1.1(4,8))hexadecane	0.34	84
1,4,8,11-tetrakis(ethoxycarbonylmethyl)-1,4,8,11-tetraazacyclotetradecane	1.49	88
1,4,8,11-tetramethyl-1,4,8,11-tetraazacyclotetradecane	1.25	35
1,4,8-tri-boc-1,4,8,11-tetraazacyclotetradecane	2.57	51
1,4-benzenedimethanol	0.17	119
1,4-benzenedithiol	1.81	97
1,4-benzodioxan-6-amine	1.17	28
1,4-benzodioxane-2-carboxylic acid	0.99	124
1,4-benzodioxane-2-thiocarboxamide	1.36	174
1,4-benzodioxane-6-carboxaldehyde	1.12	50
1,4-benzodioxane-6-sulfonyl chloride	1.76	66
1,4-bipiperidine	1.48	68
1,4-bis(1-hydroxycyclohexyl)-1,3-butadiyne	3.08	174.5
1,4-bis(2,2,2-trifluoroethoxy)benzene	3.49	76
1,4-bis(2-hydroxyisopropyl)benzene	2.13	144
1,4-bis(2-methylstyryl)benzene	6.9	181.5
1,4-bis(2-phenylethyl)benzene	6.82	89
1,4-bis(3-aminopropyl)piperazine	-0.67	14

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
1,4-bis(4-methyl-5-phenyloxazol-2-yl)benzene	6.45	234
1,4-bis(5-phenyloxazol-2-yl)benzene	5.76	242
1,4-bis(diphenylphosphino)butane	7.61	134
1,4-bis(glycidyloxy)benzene	1.28	111
1,4-bis(phenylethynyl)benzene	5.98	177
1,4-bis(trifluoromethyl)benzene	3.66	-1
1,4-bis(trimethylsilyl)-1,3-butadiyne	3.67	111.5
1,4-bis((trimethylsilyl)ethynyl)benzene	4.73	120
1,4-butanediol	-0.63	19.5
1,4-cyclohexadiene	2.31	-49
1,4-cyclohexanedione	0.1	77.5
1,4-diacetoxy-2-butyne	1.31	27.5
1,4-diacetoxybenzene	1.59	120.5
1,4-diacetoxybutane	1.06	12.5
1,4-diacryloylpiperazine	-0.27	93
1,4-diaminoanthraquinone	3	262.5
1,4-diaminobutane	-0.98	26.5
1,4-diazabicyclo(2.2.2)octane	-0.53	157.5
1,4-dibenzoylbenzene	4.64	163
1,4-dibenzyloxybenzene	4.99	126.5
1,4-dibromo-2,3-butanedione	0.78	117.5
1,4-dibromo-2,5-difluorobenzene	3.57	62
1,4-dibromo-2,5-dimethoxybenzene	3.72	146
1,4-dibromo-2-fluorobenzene	3.65	34
1,4-dibromo-2-nitrobenzene	3.31	83.5
1,4-dibromobenzene	3.71	88
1,4-dibromobutane	2.98	-20
1,4-dibromonaphthalene	4.84	81
1,4-dibromopentane	3.54	-34
1,4-dichloro-2-fluorobenzene	3.44	4
1,4-dichloro-2-iodobenzene	4.07	20.5
1,4-dichloro-2-nitrobenzene	3.04	55

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
1,4-dichloro-5,6,7,8-tetrahydro-5,8-ethanophthalazine	3.92	209.5
1,4-dichlorobenzene	3.46	54
1,4-dichlorobutane	2.51	-38
1,4-dicyclohexylbenzene	7.45	104
1,4-diethoxybenzene	2.8	70
1,4-diethylbenzene	4.36	-43
1,4-difluoro-2,5-dimethoxybenzene	2.16	120.5
1,4-difluoro-2-nitrobenzene	1.9	-12
1,4-difluorobenzene	2.26	-13
1,4-diformylpiperazine	-1.48	127.5
1,4-dihydroxyanthraquinone	2.98	196.5
1,4-diiodobenzene	3.8	130
1,4-diiodobutane	3.56	6
1,4-diisopropylbenzene	4.7	-17
1,4-dimethoxy-2-fluorobenzene	2.21	24.5
1,4-dimethoxybenzene	2.05	56
1,4-dimethoxynaphthalene	3.2	84.5
1,4-dimethylpiperazine	-0.01	-1
1,4-dinitrobenzene	1.7	174
1,4-di-o-tosyl-2,3-o-isopropylidene-l-threitol	1.99	90
1,4-dioxane	-0.23	11.8
1,4-dioxane-2,3-diol	-1.47	101
1,4-dioxane-2,5-dione	-0.67	82
1,4-diphenoxybenzene	4.97	72.5
1,4-diphenyl-1-butanone	3.85	55
1,4-diphenylbutadiyne	4.55	86.5
1,4-dipropionyloxybenzene	2.46	113
1,4-di-tert-butylbenzene	5.68	77
1,4-dithio-dl-threitol	0.18	41.5
1,4-dithioerythritol	0.18	83
1,4-naphthoquinone	1.61	122
1,4-oxathiane 4,4-dioxide	-1.36	132
1,4-oxazepan-5-one	-1.06	81
1,4-phenylene diisothiocyanate	3.96	130.5
1,4-phenylenediacetic acid	1.34	254
1,4-phenylenediacetonitrile	0.75	96
1,4-piperazinedipropionitrile	0.24	64.5

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
1,5,5-trimethylhydantoin	0.03	162.5
1,5-bis(diphenylphosphino)pentane	8.17	47
1,5-diaminonaphthalene	1.49	190.5
1,5-diaminopentane	-0.27	9
1,5-dibromopentane	3.51	-34
1,5-dichloro-2,4-dinitrobenzene	2.8	100.5
1,5-dichloroanthraquinone	4.14	247
1,5-dichloropentane	3.11	-72
1,5-difluoro-2,4-dinitrobenzene	1.56	74
1,5-dimethylnaphthalene	4.37	80
1,5-dinitronaphthalene	2.68	216
1,5-hexadiene	3.05	-141
1,5-pentamethylene-1h-tetrazole	0.56	60
1,5-pentanediol	-0.1	-16
1,5-pentanedithiol	2.01	-72
1,6-anhydro-beta-D-glucopyranose	-2.18	182
1,6-diaminohexane	0.27	41
1,6-dibromo-2-hydroxynaphthalene-3-carboxylic acid	4.42	251
1,6-dibromohexane	4.16	-2
1,6-dichlorohexane	3.6	-13
1,6-dicyanohexane	0.75	-3
1,6-dihydroxynaphthalene	1.99	138
1,6-diisocyanatohexane	1.88	-67
1,6-dimethoxynaphthalene	3.21	58.5
1,6-diphenoxy-2,4-hexadiyne	4.46	81.5
1,6-hexanediol	0.59	41.5
1,6-hexanedithiol	2.53	-21
1,7-diaminoheptane	0.79	28
1,7-dihydroxynaphthalene	2	182
1,7-heptanediol	1.16	17.5
1,7-phenanthroline	2.04	79
1,8,9-trihydroxyanthracene	2.73	179
1,8-bis(dimethylamino)naphthalene	3.85	49
1,8-cineole	3.36	1
1,8-diaminonaphthalene	1.22	63

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
1,8-diaminooctane	1.29	52
1,8-diazabicyclo(5.4.0)undec-7-ene	1.6	-70
1,8-dibenzyl-1,4,8,11-tetraazacyclotetradecane	2.28	67.5
1,8-dibromooctane	5.18	15.5
1,8-dichloroanthraquinone	4.14	202
1,8-dichlorooctane	4.63	-8
1,8-naphthalic anhydride	2.42	271
1,8-naphthalimide	2.06	301
1,8-nonadiyne	3.09	-21
1,8-octanediol	1.63	59.5
1,9-diaminononane	1.76	38
1,9-diphenyl-1,3,6,8-nonatetraen-5-one	5.52	142
1,9-nonanediol	2.11	46.5
1-(3,5-bis(trifluoromethyl)phenyl)ethanol	3.1	72
1-(3,5-bis(trifluoromethyl)phenyl)pyrrole	4.1	42
10,11-dihydrocarbamazepine	2.36	205
10,12-docosadiynedioic acid	5.37	111
10,12-pentacosadiynoic acid	8.19	63.5
10,12-tricosadiynoic acid	7.64	57
10-hydroxybenzo(h)quinoline	3.16	104
10-methylphenothiazine	4.13	101.5
10-phenyl-1-decanol	5.82	36
10-undecen-1-ol	4.58	-3
10-undecenoic acid	3.84	23.5
10-undecynoic acid	3.13	41.5
11-heneicosanol	9.06	72
12-aminododecanoic acid	0.25	186
12-hydroxystearic acid	6.61	74.5
12-tricosanone	9.5	67
14-heptacosanone	10.22	77.5
15-hydroxypentadecanoic acid	5.2	86
16-hydroxyhexadecanoic acid	5.77	99
18-crown-6	-0.39	39
18-pentatriacontanone	10.95	86.5
1-acetamidoadamantane	2.61	147.5

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
1-acetyl-2-naphthol	2.91	65
1-acetyl-3-thiosemicarbazide	-1.22	166.5
1-acetyl-4-(4-hydroxyphenyl)piperazine	1.17	182.5
1-acetyl-5-bromo-7-nitroindoline	2.34	197
1-acetyl-5-bromoindoline	2.13	119.5
1-acetyl-5-nitroindoline	1.31	176.5
1-acetylimidazole	-0.33	101
1-acetylisatin	0.72	142
1-acetylnaphthalene	2.97	10
1-acetylpiperazine	-1.03	32
1-acetylpiperidine-4-carbonyl chloride	0.72	132
1-acetylpiperidine-4-carboxylic acid	-0.3	182
1-acetylpyrene	4.99	87.5
1-adamantaneacetic acid	2.77	137
1-adamantaneethanol	3.34	74
1-adamantanemethanol	2.7	116.5
1-amino-2,4-dibromoanthraquinone	3.74	227
1-amino-4-hydroxyanthraquinone	2.97	208
1-amino-5-chloroanthraquinone	4	207
1-aminoanthraquinone	3.21	253.5
1-aminoindane	1.5	2
1-aminopyrene	4.26	117.5
1-aza-18-crown-6	-0.76	47.5
1-benzhydrylpiperazine	2.53	91.5
1-benzoyl-4-piperidone	1.03	54.5
1-benzoylnaphthalene	4.4	75.5
1-benzoylpiperidine	2.18	49
1-benzyl-1,2,3-triazole-4,5-dicarboxylic acid	0.16	180
1-benzyl-1,4,7,10-tetraazacyclododecane	0.09	85
1-benzyl-3-hydroxy-1h-indazole	3.35	165
1-benzyl-4-boc-piperazine	2.92	72

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
1-benzyl-4-cyano-4-hydroxypiperidine	1.18	97
1-benzyl-4-hydroxypiperidine	1.45	62
1-benzyl-5-phenylbarbituric acid	1.97	164
1-benzylimidazole	1.58	70.5
1-benzyloxy-3-iodobenzene	4.48	50.5
1-benzyloxy-4-bromobenzene	4.53	62
1-benzyloxy-4-iodobenzene	4.49	62
1-benzyloxycarbonyl-4-piperidone	1.12	39.5
1-boc-2-(hydroxydimethylsilyl)pyrrole	2.7	55
1-boc-2-piperidone	1.79	35.5
1-boc-3-azetidinone	0.75	50.5
1-boc-3-cyanoazetidine	1.24	69.5
1-boc-3-hydroxyazetidine	0.69	40
1-boc-3-hydroxypiperidine	1.51	68
1-boc-3-oxopiperazine	0.48	158
1-boc-3-piperidone	1.19	38
1-boc-3-pyrrolidinone	0.83	36
1-boc-4-cyanopiperidine	1.94	46.5
1-boc-4-hydroxypiperidine	1.45	63
1-boc-4-piperidinemethanol	1.83	80
1-boc-4-piperidone	1.08	74
1-boc-6-amino-1h-indazole	2.33	171.5
1-boc-azetidine-3-carboxylic acid	1.11	101.5
1-boc-imidazole	1.58	46
1-boc-indoline	3.01	47
1-boc-isonipecotic acid	1.66	150
1-boc-nipecotic acid ethyl ester	2.13	33
1-boc-piperazine	0.58	46
1-boc-pyrrole-2-carboxaldehyde	2.23	50.5
1-bromo-2,3,5,6-tetramethylbenzene	4.55	60
1-bromo-2,3,5-trichlorobenzene	4.68	59
1-bromo-2,3-dichlorobenzene	4.21	58

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
1-bromo-2,4,5-trifluorobenzene	2.99	-19
1-bromo-2,4-dichlorobenzene	4.2	26.5
1-bromo-2,4-difluorobenzene	2.98	-4
1-bromo-2,4-dimethoxybenzene	2.94	25
1-bromo-2,4-dinitrobenzene	2.32	71.5
1-bromo-2,5-difluoro-4-nitrobenzene	2.51	56
1-bromo-2-chloro-4-nitrobenzene	3.26	60.5
1-bromo-2-chlorobenzene	3.61	-13
1-bromo-2-chloroethane	1.65	-18
1-bromo-2-ethylbenzene	4.04	-68
1-bromo-2-fluoro-4-iodobenzene	3.46	36
1-bromo-2-fluorobenzene	2.91	-8
1-bromo-2-hexadecanone	7.41	56
1-bromo-2-iodobenzene	3.7	9.5
1-bromo-2-methoxynaphthalene	4.21	79.5
1-bromo-2-methylpropane	2.57	-118
1-bromo-2-naphthol	3.64	78.5
1-bromo-2-nitrobenzene	2.59	42
1-bromo-3,3-diphenylpropane	5.32	40.5
1-bromo-3,5-bis(trifluoromethyl)benzene	4.18	-16
1-bromo-3,5-dichlorobenzene	4.22	75
1-bromo-3,5-difluorobenzene	3	-27
1-bromo-3,5-di-tert-butylbenzene	6.95	64
1-bromo-3-chlorobenzene	3.59	-22
1-bromo-3-chloropropane	2.12	-59
1-bromo-3-fluorobenzene	3.02	-8
1-bromo-3-iodobenzene	3.69	-9
1-bromo-3-methylbutane	3	-112
1-bromo-3-phenoxypropane	3.12	9.5
1-bromo-4-chloro-2-nitrobenzene	3.2	70
1-bromo-4-chlorobenzene	3.63	65

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
1-bromo-4-fluoro-2-nitrobenzene	2.59	38
1-bromo-4-fluorobenzene	2.98	-16
1-bromo-4-fluoronaphthalene	4.24	35
1-bromo-4-iodobenzene	3.69	91
1-bromo-4-isopropylbenzene	4.24	-12
1-bromo-4-nitrobenzene	2.66	125.5
1-bromo-4-tert-butylbenzene	4.94	15.5
1-bromoadamantane	4.7	119
1-bromobutane	2.73	-112
1-bromodecane	5.92	-30
1-bromododecane	6.8	-10
1-bromoheptane	4.4	-58
1-bromohexadecane	7.61	17
1-bromohexane	3.88	-85
1-bromonaphthalene	3.99	-1
1-bromooctadecane	8.22	27.5
1-bromooctane	4.91	-55
1-bromopentadecane	7.36	18.5
1-bromopentane	3.27	-95
1-bromoperfluorooctane	4.74	6
1-bromopropane	2.18	-110
1-bromopyrene	6.04	94.5
1-bromotetradecane	7.31	5
1-bromoundecane	6.45	-9
1-butanesulfonyl chloride	1.43	-29
1-butanethiol	2.51	-116
1-butanol	0.84	-89.5
1-chloro-2,4-bis(trifluoromethyl)benzene	4.31	-59
1-chloro-2,4-difluoro-3-nitrobenzene	2.44	47
1-chloro-2,4-difluorobenzene	2.93	-26
1-chloro-2,4-dinitrobenzene	2.29	50
1-chloro-2-fluorobenzene	2.81	-43
1-chloro-2-iodobenzene	3.63	1
1-chloro-2-methylpropane	2.29	-131
1-chloro-2-nitrobenzene	2.48	32.5
1-chloro-3,4-dinitrobenzene	2.24	22.5
1-chloro-3,5-dibromobenzene	4.36	92.5

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
1-chloro-3,5-dimethoxybenzene	2.75	34
1-chloro-3-methylbutane	2.81	-104
1-chloro-3-nitrobenzene	2.49	45
1-chloro-4-fluoro-2-nitrobenzene	2.52	38
1-chloro-4-fluorobenzene	2.85	-22
1-chloro-4-iodobenzene	3.66	54.5
1-chloro-4-nitrobenzene	2.56	84
1-chloroacetyl-3-pyrazolidinone	-1.1	145.5
1-chloroadamantane	4.35	165
1-chloroanthraquinone	3.53	160
1-chlorobutane	2.37	-123
1-chloroethyl chloroformate	1.4	-65
1-chlorohexadecane	8.81	9
1-chlorohexane	3.63	-94
1-chloroisoquinoline	2.73	34
1-chloronaphthalene	3.95	-8
1-chlorooctane	4.82	-58
1-chloropentane	3.12	-99
1-chlorophthalazin-4-one	0.53	272
1-chlorophthalazine	1.65	110.5
1-cyano-1-cyclopropanecarboxylic acid	0.04	146
1-cyanoacetyl-3,5-dimethyl-1h-pyrazole	1.01	120
1-cyanomethylpiperidine	1	25
1-cyclohexene-1-acetic acid	1.85	33
1-cyclohexene-1-carboxylic acid	1.65	32
1-cyclohexyl-2-pyrrolidinone	1.86	12.5
1-cyclopentene-1-carboxylic acid	1.09	119
1-cyclopentyl-2,2-dimethyl-1-propanol	3.15	49
1-decanesulfonyl chloride	4.45	32
1-decanethiol	6.24	-26
1-decanol	4.24	6
1-decene	5.63	-66
1-decyne	5	-44
1-difluoromethoxy-4-nitrobenzene	2.13	34

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
1-dimethylamino-2-nitroethylene	-0.48	103
1-dimethylamino-2-propanol	-0.06	-85
1-dodecanesulfonyl chloride	5.4	40.5
1-dodecanethiol	7.13	-7
1-dodecanol	5.36	25.5
1-dodecene	6.53	-37
1-dodecyne	6.03	-19
1-eicosanol	8.9	64
1-ethoxy-2-propanol	0.14	-100
1-ethyl-2-phenylindole	5.28	85.5
1-ethyl-3-methyl-1h-pyrazole-5-carboxylic acid	0.66	138.5
1-ethyl-4-((4-methoxyphenyl)ethynyl)benzene	4.99	35
1-ethyl-4-((4-n-hexylphenyl)ethynyl)benzene	7.14	10
1-ethyl-4-((4-n-propylphenyl)ethynyl)benzene	5.89	50
1-ethyl-4-((p-tolyl)ethynyl)benzene	5.14	72.5
1-ethyl-4-iodobenzene	3.93	-17
1-ethylpiperazine-2,3-dione	-0.87	110
1-ethylpiperidine	2.11	-20
1-ethynylcyclohexanol	1.09	31.5
1-ethynylcyclopentanol	1.05	25
1-ethynylpyrene	5.13	114
1-fluoro-2,4-dinitrobenzene	1.66	27
1-fluoro-2-iodobenzene	2.8	-41.5
1-fluoro-2-nitrobenzene	1.84	-8
1-fluoro-3,5-dimethyl-2-nitrobenzene	2.51	55.5
1-fluoro-3-iodo-5-nitrobenzene	2.71	77.75
1-fluoro-3-iodobenzene	2.88	-42
1-fluoro-3-nitrobenzene	1.88	2
1-fluoro-4-iodobenzene	2.93	-20
1-fluoro-4-nitrobenzene	1.97	23
1-fluoronaphthalene	3.37	-10
1-formylpiperidine	0.21	-31
1h-1,2,3-triazole	-0.73	24
1h-1,2,4-triazole-1-acetic acid	-1.25	203

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
1h-benzotriazole	1.19	97.5
1-heptadecanol	7.82	53.5
1-heptanethiol	4.28	-43
1-heptanol	2.53	-34
1-heptene	4	-119
1-heptylamine	2.57	-23
1-heptyne	3.26	-81
1-hexadecanesulfonyl chloride	7.18	57
1-hexadecanethiol	8.44	21
1-hexadecanol	7.17	49
1-hexadecene	8.11	4
1-hexadecyne	7.47	14.5
1-hexanethiol	3.65	-81
1-hexanol	2.03	-52
1-hexene	3.38	-140
1-hexylamine	1.98	-19
1-hexyne	2.63	-132
1h-indazole	1.61	147.5
1h-indene	3.04	-2
1h-pyrazole	0.03	68
1h-pyrazole-4-carboxylic acid	-0.25	282
1-hydroxycyclohexyl phenyl ketone	2.04	47.5
1-hydroxyisoquinoline	2.02	213.5
1-hydroxymethyl-5,5-dimethylhydantoin	-0.53	105
1-indanol	1.59	52
1-indanone	1.77	40
1-iodo-2,3,4,5-tetramethylbenzene	4.29	31
1-iodo-2,4-dinitrobenzene	2.43	88.5
1-iodo-2-methylpropane	3.3	-93
1-iodo-2-nitrobenzene	2.83	50
1-iodo-3,5-dinitrobenzene	2.46	101
1-iodo-3-nitrobenzene	2.82	36.5
1-iodo-4-nitrobenzene	2.92	174.5
1-iodobutane	3.11	-103
1-iododecane	6.03	-16
1-iodododecane	6.78	-3
1-iodoheptane	4.69	-48
1-iodohexadecane	8.25	22
1-iodohexane	4.19	-75
1-iodonaphthalene	4.06	4

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
1-iodooctane	5.17	-46
1-iodopentane	3.64	-86
1-iodopropane	2.65	-101
1-methoxy-2-propanol	-0.44	-100
1-methoxynaphthalene	3.34	5
1-methyl-1h-benzotriazole	0.99	64.5
1-methyl-1h-pyrazole-3-carboxylic acid	0.08	154
1-methyl-1h-pyrazole-5-carboxylic acid	-0.13	223.5
1-methyl-2-phenylindole	4.81	98.5
1-methyl-2-pyridone	-0.06	31
1-methyl-2-pyrrolidinone	-0.72	-24
1-methyl-2-quinolinone	1.27	75
1-methyl-3-n-propyl-2-pyrazolin-5-one	0.27	115
1-methyl-3-phenylpiperazine	0.94	58
1-methyl-3-trifluoromethyl-2-pyrazolin-5-one	0.7	179
1-methyl-4-(4-piperidinyl)piperazine	0.08	54.5
1-methyl-5-nitro-1h-indazole	1.68	160
1-methyl-6-nitro-1h-indazole	1.41	125.5
1-methylbenzimidazole	1.55	60.5
1-methylbenzimidazole-2-carboxaldehyde	1.4	119.5
1-methylcyclohexanecarboxylic acid	2.14	37.5
1-methylcyclohexanol	1.81	26
1-methylcyclopentanol	1.3	36.5
1-methylcyclopentene	2.89	-142
1-methylfluorene	4.56	85
1-methylhydantoin	-1.2	157.5
1-methylimidazole	-0.18	-6
1-methylimidazole-4,5-dicarbonitrile	-0.07	84.5
1-methylimidazole-4-carboxylic acid	-0.22	242

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
1-methylimidazole-5-carboxaldehyde	-0.68	54.5
1-methylindole-3-carboxaldehyde	1.66	69
1-methylisatin	0.68	131
1-methylisoquinoline	2.55	10
1-methylnaphthalene	3.84	-22
1-methylpiperazine	-0.92	-6
1-methylpiperidine	1.31	-50
1-methylpyrrole	1.31	-57
1-methylpyrrole-2-carboxylic acid	0.74	136.5
1-methylpyrrolidine	0.54	-90
1-methylsulfonyl-1h-benzotriazole	0.48	110
1-methylsulfonyl-4-nitrobenzene	0.62	137
1-methylthio-1-methylamino-2-nitroethylene	0	112.5
1-naphthaldehyde	2.96	1.5
1-naphthaleneacethydrazide	1.71	168
1-naphthaleneboronic acid	1.88	207
1-naphthalenemethanol	2.17	60.5
1-naphthoic acid	2.79	161
1-naphthoic hydrazide	1.48	167
1-naphthol	2.79	96
1-naphthoyl chloride	3.35	17.5
1-naphthyl acetate	2.9	44.5
1-naphthyl isocyanate	2.42	4
1-naphthyl isothiocyanate	4.33	55.5
1-naphthylacetamide	2.07	177
1-naphthylacetic acid	2.97	129.5
1-naphthylacetonitrile	2.83	34
1-n-butyl-4-((4-butylphenyl)ethynyl)benzene	7.04	41
1-n-butyl-4-((4-ethoxyphenyl)ethynyl)benzene	5.96	54
1-n-butyl-4-((4-methoxyphenyl)ethynyl)benzene	5.61	40
1-n-hexyl-4-((p-tolyl)ethynyl)benzene	6.6	42
1-n-hexyltheobromine	1.8	80.5

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
1-nitro-4-(trifluoromethoxy)benzene	2.76	15
1-nitro-4-n-propylbenzene	3.5	-14
1-nitronaphthalene	3.2	57
1-nitropropane	0.91	-108
1-nonanol	3.76	-7
1-nonylamine	3.69	-1
1-nonyne	4.45	-50
1-o-acetyl-2,3,5-tri-o-benzoyl-beta-D-ribofuranose	4.18	129
1-octadecanesulfonyl chloride	8.31	58
1-octadecanethiol	8.98	31.5
1-octadecanol	8.27	58
1-octadecene	9.03	16
1-octanesulfonyl chloride	3.43	14
1-octanethiol	4.95	-49
1-octanol	3.21	-16
1-octen-3-ol	2.43	-49
1-octene	4.61	-102
1-octylamine	3.24	-1
1-octyne	3.86	-60
1-pentadecanol	6.6	45.5
1-pentadecene	7.7	-4
1-pentadecyne	7.16	10
1-pentanethiol	3.02	-76
1-pentanol	1.47	-79
1-pentene	2.84	-138
1-pentenylboronic acid	1.48	80
1-pentylamine	1.39	-55
1-pentyne	2.13	-106
1-phenyl-1,2,3-butanetrione 2-oxime	1.21	129
1-phenyl-1,2-ethanediol	0.46	67
1-phenyl-2-propyn-1-ol	1.25	28
1-phenyl-2-pyrrolidinone	1.05	65.5
1-phenyl-3-pyrazolidinone	0.46	121.5
1-phenyl-3-trifluoromethyl-2-pyrazolin-5-one	2	193
1-phenylcyclohexanol	2.97	60
1-phenylcyclohexene	4.53	-11
1-phenylcyclopentanecarboxylic acid	3.04	159.5

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
1-phenylcyclopropanecarboxylic acid	2.08	87
1-phenylimidazole	1.65	13
1-phenylisatin	1.95	140.5
1-phenylpiperazine	1.54	18
1-phenylpyrrole	2.9	58.5
1-phenylsemicarbazide	0.12	172
1-piperonylpiperazine	0.42	42.5
1-propanethiol	1.72	-113
1-propanol	0.21	-127
1-propylamine	0.31	-83
1-pyrenebutyric acid	5.15	185
1-pyrenecarboxaldehyde	4.62	126.5
1-tert-butyl-2-imidazolidinone	0.5	137
1-tetradecanethiol	7.81	6.5
1-tetradecanol	6.21	38.5
1-tetradecene	7.27	-12
1-tetralone	2.29	7
1-thio-beta-D-glucose tetraacetate	1.67	116
1-tridecene	6.89	-23
1-trifluoromethylcyclopropane-1-carboxylic acid	1.09	88
1-trimethylsilyl-1-propyne	2.8	-69
1-tritylimidazole	4.72	222
1-undecanol	4.83	14
1-undecene	6.11	-49
1-undecyne	5.55	-25
2-(1,3-benzodioxol-5-yl)piperazine	-0.26	122
2-(1-boc-4-piperidinyloxy)-n,n-dimethylacetamide	1.31	58
2-(1-boc-4-piperidinyloxy)-n-cyclopropylacetamide	1.66	86
2-(1-boc-4-piperidinyloxy)-n-methylacetamide	1.05	95
2-(1-cyclohexenyl)ethylamine	1.76	-55
2-(1-naphthyl)ethanol	2.66	62
2-(1-piperazinyl)aniline	0.84	115
2-(1-piperazinyl)phenol	0.78	124.5

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
2-(1-piperazinyl)pyrimidine	0.32	33
2-(1-piperidinyl)aniline	2.76	45.5
2-(1-piperidinyl)benzonitrile	3.06	43
2-(1-piperidinyl)phenol	3	74.5
2-(1-piperidinyl)thiazole-5-carboxaldehyde	2.02	48.5
2-(1-pyrrolidinyl)phenol	2.23	110.5
2-(1-pyrrolyl)benzoic acid	1.86	104
2-(2,2,2-trimethylacetamido)benzeneboronic acid	1.59	269.5
2-(2,4,5-trichlorophenoxy)propionic acid	3.88	178
2-(2,4-dichlorophenoxy)propionic acid	3.13	114
2-(2,4-dinitrobenzyl)pyridine	2.5	92
2-(2,6-dimethoxyphenyl)-4,4-dimethyl-2-oxazoline	2.62	69
2-(2-aminoethoxy)ethanol	-1.41	-11
2-(2-aminophenyl)benzimidazole	2.72	214
2-(2-carboxyvinyl)benzeneboronic acid	1.06	171
2-(2-chloro-4-methoxyphenyl)-3-oxobutyronitrile	2.1	79
2-(2-chlorophenoxy)ethylamine	1.46	39.5
2-(2-chlorophenyl)benzimidazole	4.14	234
2-(2-ethoxyphenoxy)ethyl bromide	3.26	42
2-(2-furyl)-1,3-diphenylimidazolidine	4.2	130
2-(2-furyl)piperazine	-0.21	86.5
2-(2-hydroxyethyl)pyridine	0.45	-8
2-(2-hydroxyphenyl)benzothiazole	4.04	132
2-(2-methoxyphenyl)-5-phenyl-1,3,4-oxadiazole	3.34	100

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
2-(2-methoxyphenyl)piperazine	0.35	79.5
2-(2-methyl-1,3-dioxolan-2-yl)benzeneboronic acid	0.82	110.5
2-(2-naphthoxy)ethanol	2.46	74
2-(2-naphthyl)piperazine	1.54	101.5
2-(2-n-butoxyethoxy)ethyl acetate	1.42	-32
2-(2-pyridyl)benzimidazole	2.55	223
2-(2-thienyl)piperazine	0.24	83
2-(2-thienyl)pyridine	2.79	62.5
2-(3,4-dimethoxyphenyl)ethanol	1.63	46
2-(3,4-dimethoxyphenyl)ethylamine	0.9	11
2-(3-chloro-4-fluorophenyl)indole	4.8	171.5
2-(3-chlorophenoxy)propionic acid	2.56	114.5
2-(3-pyridyl)benzimidazole	2.3	257
2-(3-thienyl)piperazine	0.24	97.5
2-(4-aminophenyl)-1,1,1,3,3,3-hexafluoro-2-propanol	2.72	150
2-(4-aminophenyl)ethanol	0.32	109
2-(4-aminophenyl)ethylamine	0.02	29.5
2-(4-benzyloxyphenyl)ethanol	3.29	86
2-(4-biphenylyl)-2-propanol	3.89	91.5
2-(4-biphenylyl)-5-(4-tert-butylphenyl)-1,3,4-oxadiazole	6.16	137
2-(4-biphenylyl)ethylamine	3.23	52
2-(4-bromophenyl)-5-(1-naphthyl)-1,3,4-oxadiazole	5.08	147.5
2-(4-bromophenyl)-5-phenyl-1,3,4-oxadiazole	3.93	169.5
2-(4-chloro-2-methylphenoxy)acetic acid hydrazide	1.36	150

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
2-(4-chloro-3-nitrobenzoyl)benzoic acid	3.19	199.5
2-(4-chlorobenzoyl)benzoic acid	3.3	150
2-(4-chlorobenzyl)pyridine	3.56	8
2-(4-chlorophenoxy)ethanol	1.91	30
2-(4-chlorophenoxy)isobutyric acid	2.81	119.5
2-(4-chlorophenoxy)nicotinic acid	2.69	158
2-(4-chlorophenyl)indole	4.7	206
2-(4-chlorophenylthio)-6-fluorobenzonitrile	4.8	69.5
2-(4-chlorophenylthio)benzaldehyde	4.75	72.5
2-(4-chlorophenylthio)nicotinic acid	3.3	220
2-(4-cyanophenyl)-5-n-pentyl-1,3-dioxane	3.89	57
2-(4-cyanophenyl)-5-n-propyl-1,3-dioxane	3.25	57.5
2-(4-ethoxyphenyl)ethanol	2.16	42
2-(4-ethylphenyl)-5-n-propylpyrimidine	3.98	42
2-(4-fluorophenoxy)nicotinic acid	2.32	183.5
2-(4-fluorophenyl)indole	4.14	188.5
2-(4-hydroxyphenyl)ethanol	0.85	90.5
2-(4-hydroxyphenyl)propionic acid	1.52	131
2-(4-methoxybenzoyl)thiophene	3.3	73
2-(4-methoxyphenyl)ethanol	1.61	28
2-(4-morpholinyl)-5-(trifluoromethyl)aniline	1.73	126.5
2-(4-morpholinyl)aniline	1.02	96.5
2-(4-n-hexyloxyphenyl)-5-n-octylpyrimidine	7.35	59.5

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
2-(4-pyridyl)benzimidazole	2.31	218.5
2-(4-tert-butylphenyl)ethanol	3.72	33
2-(4-toluoyl)benzoic acid	2.95	138
2(5h)-furanone	0.12	3.5
2-(5-isoxazolyl)-4-methylphenol	2.37	176
2-(5-isoxazolyl)phenol	2.07	184.5
2-(5-nitro-2-pyridyloxy)ethanol	0.67	113
2-(8-chloro-1-naphthylthio)acetic acid	3.76	156
2-(allylthio)nicotinic acid	1.44	146.5
2-(aminomethyl)pyridine	-0.19	-20
2-(boc-amino)-5-cyanopyridine	2.05	172.5
2-(boc-amino)benzeneboronic acid	1.51	124
2-(boc-amino)pyridine	2.29	93
2-(bromoacetyl)naphthalene	3.37	82.5
2-(bromoacetyl)thiophene	2.01	31
2-(bromomethyl)benzonitrile	2.75	72
2-(bromomethyl)benzothiazole	3.06	48
2-(bromomethyl)naphthalene	4.01	53
2-(carboxymethoxy)benzoic acid	1.03	189
2-(carboxymethylthio)benzoic acid	1.42	217.5
2-(diethylamino)ethanol	0.56	-70
2-(difluoromethoxy)benzoic acid	1.71	99
2-(dimethylamino)ethanol	-0.46	-60
2-(dimethylamino)ethyl acrylate	0.58	-60
2-(dimethylamino)ethyl methacrylate	0.87	-30

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
2-(diphenylphosphino)benzoic acid	4.94	177.5
2-(di-tert-butylphosphino)-2-(n,n-dimethylamino)biphenyl	7.4	115.5
2-(di-tert-butylphosphino)-2-methylbiphenyl	7.76	91.5
2-(ethoxycarbonyl)benzeneboronic acid	0.81	131.5
2-(ethoxycarbonyl)phenyl isocyanate	1.82	29
2-(ethylamino)ethanol	-0.49	-9
2-(ethylsulfonyl)ethanol	-0.64	38
2-(ethylthio)ethanol	0.61	-100
2-(ethylthio)nicotinic acid	1.51	185
2-(ethylthio)nicotinoyl chloride	2.29	51
2-(methacryloyloxy)ethyl 3,5-diaminobenzoate	1.09	91
2-(methacryloyloxy)ethyl 3,5-dinitrobenzoate	2.3	71
2-(methoxycarbonyl)benzeneboronic acid	0.34	69
2-(methoxycarbonyl)benzenesulfonamide	0.62	124
2-(methoxycarbonyl)phenyl isocyanate	1.38	47
2-(methylamino)ethanol	-1.05	-5
2-(methylamino)pyridine	1.13	14.5
2-(methylsulfonyl)acetanilide	0.59	145
2-(methylsulfonyl)benzoic acid	0.41	138.5
2-(methylsulfonyl)ethanol	-1.65	28.5
2-(methylsulfonyl)thiophene	0.74	47.5
2-(methylthio)benzeneboronic acid	1.25	160
2-(methylthio)benzimidazole	2.47	204
2-(methylthio)benzoic acid	2.24	167

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
2-(methylthio)benzonitrile	1.95	35.5
2-(methylthio)naphthalene	4.02	62.5
2-(methylthio)nicotinic acid	0.83	216.5
2-(methylthio)nicotinoyl chloride	1.66	94
2-(methylthio)oxazolo(5,4-c)pyridine	1.33	80
2-(methylthio)pyrazine	0.99	44
2-(n-boc-methylamino)-5-iodo-3-methylpyridine	3.21	98.5
2-(n-hexyloxy)ethanol	1.82	-42
2-(nitromethylene)thiazolidine	-0.36	142
2-(n-propylthio)nicotinamide	1.57	147
2-(n-propylthio)nicotinic acid	1.96	161
2-(phenylsulfonyl)thiophene	1.82	123
2-(phthalimido)ethanesulfonyl chloride	1.01	158
2-(p-hydroxyphenylazo)benzoic acid	3.67	206
2-(p-toluenesulfonyl)ethanol	0.33	51.5
2-(p-tolyloxy)benzaldehyde	3.65	55
2-(trifluoromethoxy)benzamide	1.52	154
2-(trifluoromethoxy)benzeneboronic acid	1.89	119
2-(trifluoromethoxy)benzenesulfonamide	2.42	188.5
2-(trifluoromethoxy)benzenesulfonyl chloride	3.12	30.5
2-(trifluoromethoxy)benzoic acid	2.95	79
2-(trifluoromethoxy)phenylacetic acid	2.8	55

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
2-(trifluoromethyl)acetanilide	2.26	95.5
2-(trifluoromethyl)acetophenone	2.55	16
2-(trifluoromethyl)acrylic acid	1.59	50
2-(trifluoromethyl)aniline	2.24	-34
2-(trifluoromethyl)benzaldehyde	2.35	-40
2-(trifluoromethyl)benzamide	1.6	162.5
2-(trifluoromethyl)benzeneboronic acid	1.73	109
2-(trifluoromethyl)benzenesulfonamide	1.94	182
2-(trifluoromethyl)benzenesulfonyl chloride	2.92	23
2-(trifluoromethyl)benzoic acid	2.76	108.5
2-(trifluoromethyl)benzonitrile	2.47	17.5
2-(trifluoromethyl)benzophenone	3.76	59.5
2-(trifluoromethyl)benzoyl chloride	2.82	-22
2-(trifluoromethyl)benzyl alcohol	2.08	4
2-(trifluoromethyl)benzyl bromide	3.62	28.5
2-(trifluoromethyl)cinnamic acid	3.23	202
2-(trifluoromethyl)nicotinic acid	1.63	186
4-aminobenzoic acid	0.78	187.5
5,5-diphenylhydantoin	2.26	295.5
acetanilide	1.05	114.5
adenosine	-1.21	235
antipyrine	1.01	112.5
benzamide	0.51	127
benzoic acid	1.72	122.5
chloramphenicol	1.15	150.5
flufenamic acid	4.6	134
griseofulvin	2.71	219

Continued on next page

Table J.1 – *Continued from previous page*

Molecule Name	Predicted log P (AlogP)	Melting Point (°C)
hydrochlorothiazide	-0.16	269
nalidixic acid	1.27	229
nicotinic acid	0.29	237.5
papaverine	4.19	146.5
perylene	6.34	278
pyrene	5.19	150
quinidine	2.82	170
salicylamide	0.74	140
salicylic acid	1.96	159
sulfacetamide	0.15	183
sulfadiazine	0.25	254.5
sulfamethazine	0.43	199.5
sulfanilamide	-0.16	165.5
thymine	-0.8	316.5
thymol	3.16	50.5
tolbutamide	2.04	129
triphenylene	5.77	196.5
uracil	-1.28	330

Table J.1: MP-1100 dataset

Melting Point Predictions Without the log P Descriptor

Figures J.1, J.2 and **J.3** show predictions made with out the XlogP descriptor. With the exception that the XlogP descriptors removal, the analysis is identical to those presented in **Section 5.1.3**. This test does not prove or disprove the importance of a log P descriptor in other applications, rather it assess the effect of this one potentially important descriptor in this particular context. We must be careful not to over interpret the results, and hence condemn or commend the log P descriptors validity in other situations.

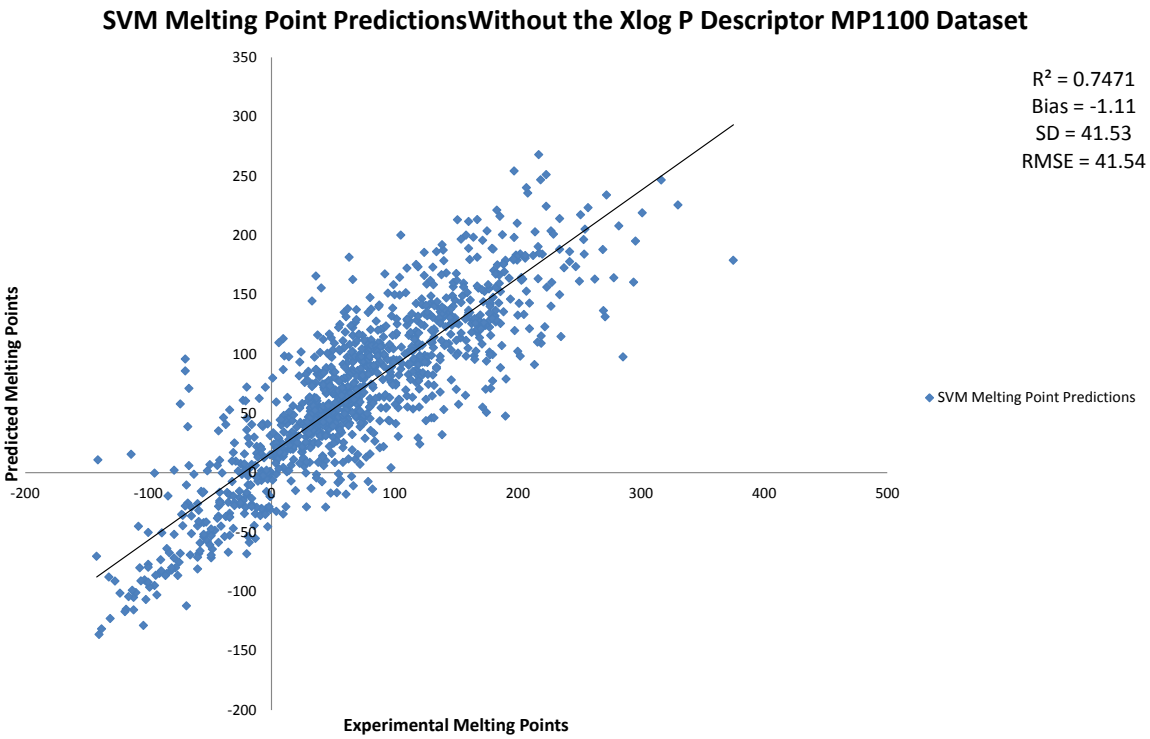


Figure J.1: A prediction of melting points for the MP1100 dataset using 2D CDK descriptors, not including log P descriptors, in our 10 fold cross validation methodology.

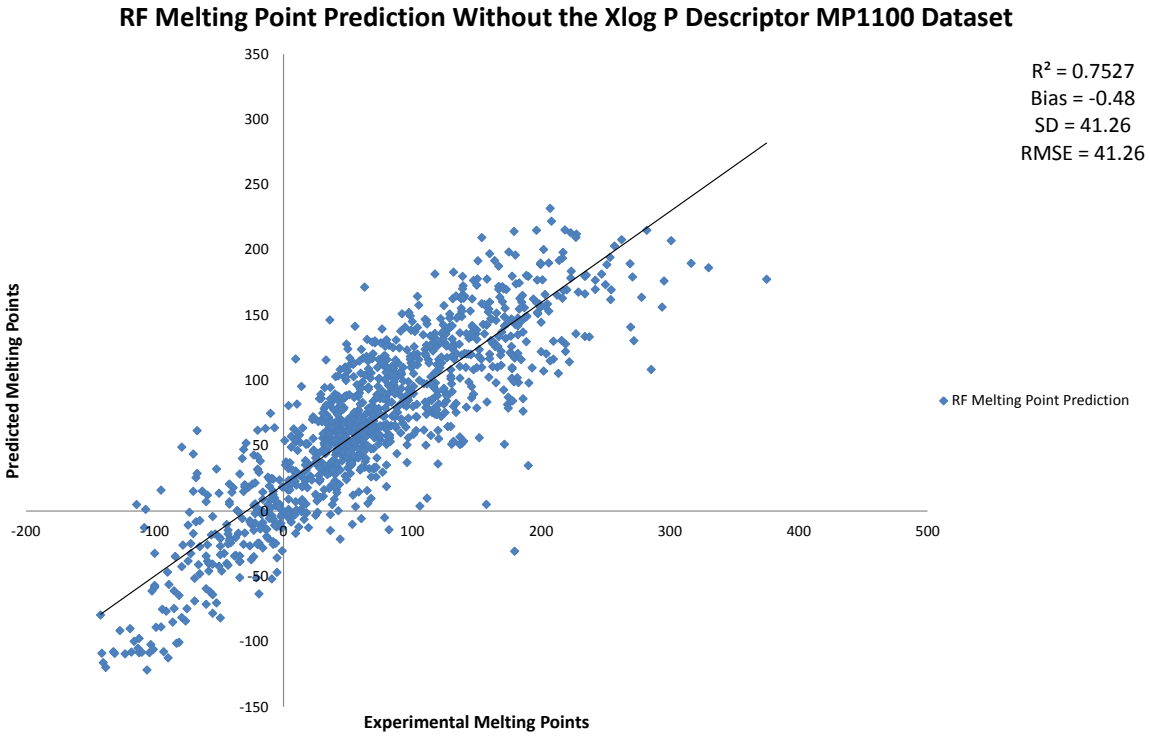


Figure J.2: A prediction of melting points for the MP1100 dataset using 2D CDK descriptors, not including log P descriptors, in our 10 fold cross validation methodology.

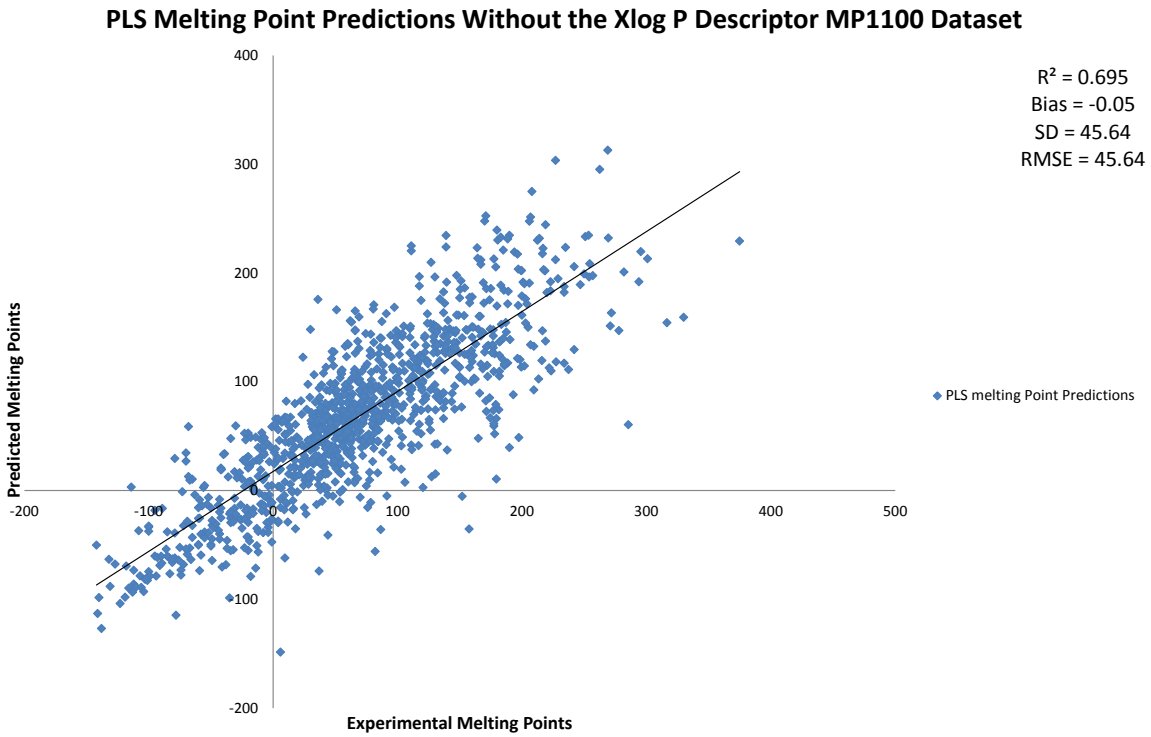


Figure J.3: A prediction of melting points for the MP1100 dataset using 2D CDK descriptors, not including log P descriptors, in our 10 fold cross validation methodology.

Melting Point Descriptor Importance

Descriptor only	Rank
TPSA	1
Zagreb Index	2
WTPT.3	3
WTPT.4	4
Hydrogen bond donor count	5
MDEO.12	6
Wiener numbers	7
Molecular weight	8
MW (Group weights)	9
MDEN.11	10

Table J.2: Top 10 variables ranking in Random Forest scaled by mean/σ.

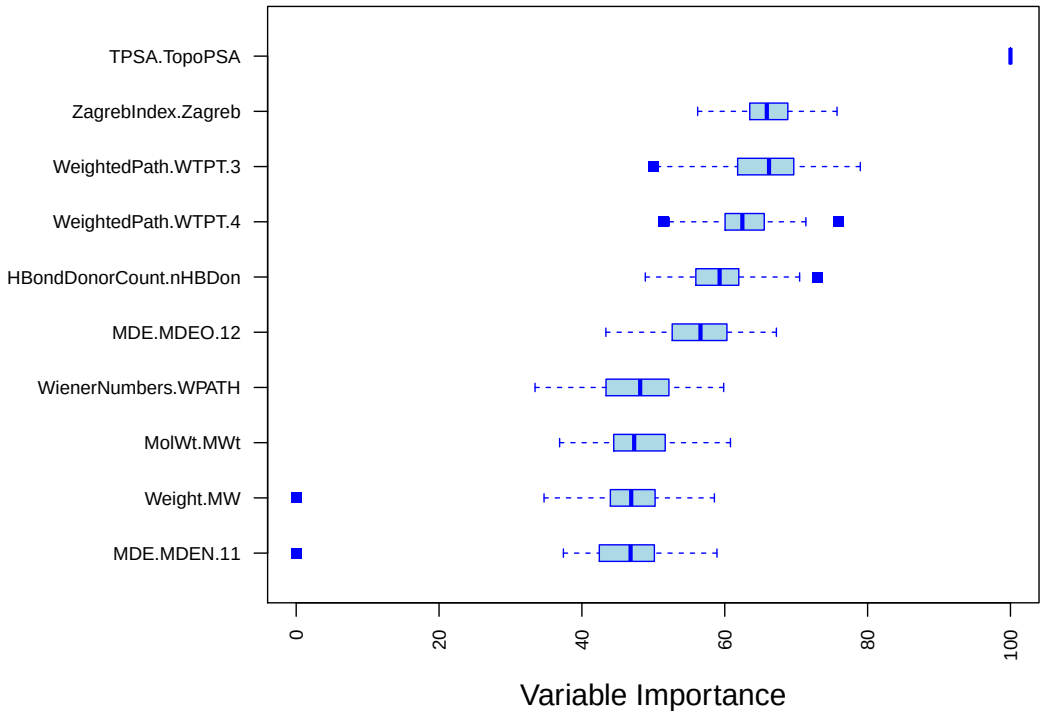


Figure J.4: Variable importance

Appendix K

Machine learning ΔS_{sub} and ΔH_{sub}

Here we present the inconclusive results of machine learning ΔS_{sub} and ΔH_{sub}. The SVM, RF and PLS algorithms have been applied here over three datasets: using theoretical chemistry, 2D CDK descriptors and a combination of the two.

ΔH_{sub} Machine learning Predictions

CDK Descriptors	SVM	RF	PLS
R^2	0.5±0.03	0.37±0.03	0.35±0.05
RMSE	11.22±0.36	12.52±0.34	13.45±0.88

Table K.1: ΔH_{sub} predicted by machine learning applying CDK 2D descriptor set.

Theoretical Chemistry	SVM	RF	PLS
R^2	0.54±0.03	0.58±0.02	0.39±0.03
RMSE	10.7±0.35	10.33±0.31	12.43±0.36

Table K.2: ΔH_{sub} predicted by machine learning applying theoretical chemistry descriptors set.

Combined descriptors	SVM	RF	PLS
R^2	0.56±0.03	0.59±0.02	0.35±0.07
RMSE	10.5±0.3	10.17±0.28	13.52±1.01

Table K.3: ΔH_{sub} predicted by machine learning applying the combination of descriptors.

ΔS_{sub} Machine learning Predictions

CDK Descriptors	SVM	RF	PLS
R^2	0.02±0.03	0.33±0.07	0.25±0.08
RMSE	9.39±0.29	7.7±0.32	8.07±0.47

Table K.4: ΔS_{sub} predicted by machine learning applying CDK 2D descriptor set.

Theoretical Chemistry	SVM	RF	PLS
R^2	0.04±0.04	0.01±0.01	0.01±0.02
RMSE	10.33±1.05	9.78±0.42	9.72±0.4

Table K.5: ΔS_{sub} predicted by machine learning applying theoretical chemistry descriptors set.

Combined descriptors	SVM	RF	PLS
R^2	0.03±0.03	0.33±0.08	0.25±0.07
RMSE	9.34±0.31	7.66±0.35	8.07±0.43

Table K.6: ΔS_{sub} predicted by machine learning applying the combination of descriptors.

Appendix L

Scripts and Code

Listing L.1: A Bash script to automate DMACRYS

```
#!/bin/bash
#SBATCH -J NaMe_DMACRYS
#SBATCH -N 1
#SBATCH -n 12
#SBATCH -p exclusive
#SBATCH -t 24:00:00

source/opt/intel/composerxe/bin/compilervars.sh intel64
source /opt/intel/impi/4.0.1/bin64/mpivars.sh
export I_MPI_FABRICS=shm:dapl


sleep 10
for writelogfile in 1;
do
#############################################################################
# Originally by David S. Palmer regenerated and adapted by James L. McDonagh
# Simple script to set up and run DMACRYS calculations
#
#
# Current working directory must contain:
#         - ${mol}.fdat      ... input file
#         - cutoff           ... dmacrys 'cutoff' file
#         - ${mol}.dsp
... file giving the numbers of three atoms defining line and plane, e.g. "C1 C2 C3"
#         - fitcrys.pots     ... potentials file for the Buckingham potential

# Notes   - ${mol}.dsp must be created by the user
#                   run setup_for_dmacrys.sh
#
use the *.nnl and *.nem files created as a result to find the atoms define
#           the axis
#                   write the atoms in the a fil $mol.dsp
#
#Leipzig#####################################################################
#               #         >> cp ${mol}.fdat ${mol}.dat
#
#               #         >> rpluto ${mol}
#
#               #         >> (click "labl")
#
#
#         >> write to ${mol}.dsp the atom labels of three atoms that define #
#                   a line and a plane.                                     #
#
# ############################################################################
# Usage:
#         - qsub dmacrys_setup.sh | tee dmacrys_setup.jobid
#
#############################################################################
```

```
#User-Defined Variables
dmacrysdir=/home/st-andrews/jm222/Scripts/DMACRYS_Scripts
#Directory containing scripts and templates/ directory
formchk=/opt/app/gaussian/g03/formchk
#formchk executable
neighcrys=/usr/local/bin/neighcrys.out
#neighcrys executable
dmacrys=/usr/local/bin/dmacrys.out
#dmacrys executable
gdma=/usr/local/bin/gdma
#gdma executable
fvibsetup=/home/st-andrews/jm222/Scripts/DMACRYS_Scripts/FvibSetup.pl
#Graeme Day's perl script for setting up lattice entropy calculations
elastphon=/home/st-andrews/jm222/Scripts/DMACRYS_Scripts/elastphon.out
#Graeme Day's perl script for calculating lattice entropies
ext=dat
#either "fdat" or "res".
g03method='METH/BAS'
#Level of QM theory for DMA.  Must match names of variables in
gaussian_setup_dmacrys.py
cutoff=cutoff
#dmacrys cutoff file
mol=MOLE
#name or ID of molecule, e.g. $mol.fdat or $mol.res



#Other variables (do not edit)
home=`pwd`

###################################
#
# Step 1: Find the atom labels
#
###################################

echo "Step 1: Finding atom labels ..."

if [ ! -d neighcrys1 ]; then mkdir neighcrys1; fi
cp cutoff $mol.$ext ${mol}.dsp neighcrys1
cd neighcrys1
#########################################################################
#sed "s/DSPMOL/$mol/g" $dmacrysdir/templates/fort_template1.22 > fort.22#
#########################################################################
sed "s/DSPMOL/$mol/g" $dmacrysdir/fort_template1.22 > fort.22
sed -i "s/DSPEXT/$ext/g" fort.22

#Run neighcrys using non-interactive mode (commands are read from fort.22)
$neighcrys << EOF
n
EOF

# neighcrys (which prints either ${mol}.res.dmain or ${mol}.dmain, depending
on whether the input file is in .res or .fdat format)
for i in `ls *.$ext.*`; do j=`echo $i | sed 's/\.'$ext'//g'`; mv $i $j; done

#Run dmacrys
$dmacrys < $mol.dmain > $mol.dmaout

echo "Step 1: FINISHED!"

###################################
#
# Step 2: Set up axes file
#
###################################

echo "Step 2: Setting up axes file ..."

a1=`head -1 ${mol}.dsp | awk '{print $1}'`
a2=`head -1 ${mol}.dsp | awk '{print $2}'`
a3=`head -1 ${mol}.dsp | awk '{print $3}'`
```

```
#HACK - the "-a" option causes grep to treat binary files as text. 'Tis a hack.
For some reason, bash sometimes thinks fort.XX files are binary.
l1=`grep -a -A 50 "Inequivalent" fort.21 | grep " $a1 " | head -1 | awk '{print $3}' |
sed 's/.\{10\}/&\n/g' | head -1`
l2=`grep -a -A 50 "Inequivalent" fort.21 | grep " $a2 " | head -1 | awk '{print $3}' |
sed 's/.\{10\}/&\n/g' | head -1`
l3=`grep -a -A 50 "Inequivalent" fort.21 | grep " $a3 " | head -1 | awk '{print $3}' |
sed 's/.\{10\}/&\n/g' | head -1`



echo "Input Atom Label | DMACRYS Atom Label"
echo "-----------------------------------"
echo "$a1                | $l1"
echo "$a2                | $l2"
echo "$a3                | $l3"

echo "MOLX 1" > axis
echo "X LINE   $l1 $l2 1" >> axis
echo "Y PLANE  $l1 $l2 1 $l3 2" >> axis
echo "ENDS" >> axis

cd $home

echo "Step 3: Finished!"

###################################
#
# Step 3: Get Coordinates
#
# ... by re-running neighcrys (with axes file) to generate .dmain file
#(which contains coordinates)
# (it is also possible to standardize the lengths of bonds to
#  hydrogens in this step)
#
###################################

echo "Step 3: Getting coordinates ..."

if [ ! -d neighcrys2 ]; then mkdir neighcrys2; fi
cp neighcrys1/axis $mol.$ext cutoff neighcrys2
cd neighcrys2
sed "s/DSPMOL/$mol/g" $dmacrysdir/fort_template2.22 > fort.22
sed -i "s/DSPEXT/$ext/g" fort.22

#Run neighcrys using non-interactive mode (commands are read from fort.22)
$neighcrys << EOF
n
EOF

#neighcrys (which prints either ${mol}.res.dmain or ${mol}.dmain, depending
on whether the input file is in .res or .fdat format !?)
for i in `ls *.$ext.*`; do j=`echo $i | sed 's/\.'$ext'//g'`; mv $i $j; done

cd $home

echo "Step 3: Finished!"

###################################
#
# Step 4: Set up Gaussian calculation
#
###################################

echo "Step 4: Setting up Gaussian calculation ..."

if [ ! -d gaussian ]; then mkdir gaussian; fi
cd gaussian

cp ../neighcrys2/fort.21 ./
```

```
#Extract coordinates to .geom and .xyz files
python $dmacrysdir/dmacrys_coord2gaussian.py fort.21 $mol

#Backup xyz file
mkdir XYZ
cp *.xyz XYZ/

#Create gaussian .com input file from .xyz file
#$dmacrysdir/xyz2com_jm.sh

for fname in `ls *.xyz`

do

echo  %mem=4gb >> $fname.com
echo  %NProcshared=12 >> $fname.com
echo "#" $g03method SP SCF='(tight)' nosymm Formcheck >> $fname.com
#Density=MP2 Formcheck=all
echo >> $fname.com
echo $fname >>$fname.com
echo >> $fname.com
echo 0 1 >> $fname.com
sed -n 3,500p $fname >> $fname.com
echo >> $fname.com

done

for i in `ls *xyz.com | sed 's/\.xyz\.com// g'`;
do
echo $i
  mv $i.xyz.com $i.com

done


#Run Gaussian
#$dmacrysdir/write_jobfile.sh

echo waiting for Gaussian PID to end.
date
g03run ${mol}.com

cd $home
echo "Step 4: Finished!"

###################################
#
# Step 5: Do GDMA
#
###################################

echo "Step 5: Doing GDMA ..."
if [ ! -d gdma ]; then mkdir gdma; fi
cd gdma
sed "s/DSPMOL/$mol/g" $dmacrysdir/data_template > data
cp ../gaussian/Test.FChk ./
echo waiting for GDMA PID to end.
gdma < data

cp $mol.old $mol.punch
python /home/st-andrews/jm222/Scripts/DMACRYS_Scripts/
dmacrys_punchlabels.py $mol.old ../gaussian/$mol.geom
cp  new.punch $mol.punch
cd $home
echo "Step 5: Finished!"

###################################
#
# Step 6: Do neighcrys
#
###################################

echo "Step 6: Doing neighcrys ..."
```

```
if [ ! -d neighcrys3 ]; then mkdir neighcrys3; fi
cp cutoff $mol.$ext ./neighcrys2/axis ./gdma/new.punch neighcrys3
cp ~/Scripts/DMACRYS_Scripts/fitcrys.pots neighcrys3/
cd neighcrys3
sed "s/DSPMOL/$mol/g" $dmacrysdir/fort_template3.22 > fort.22
sed -i "s/DSPEXT/$ext/g" fort.22

#Run neighcrys using non-interactive mode (commands are read from fort.22)
$neighcrys << EOF
n
EOF

# neighcrys (which prints either ${mol}.res.dmain or ${mol}.dmain, depending
on whether the input file is in .res or .fdat format)
for i in `ls *.$ext.*`; do j=`echo $i | sed 's/\.'$ext'//g'`; mv $i $j; done

cd $home
echo "Step 6: Finished!"


###################################
#
# Step 7: Calculate lattice energy
#
###################################

echo "Step 7: Calculating lattice energy ..."
if [ ! -d dmacrys ]; then mkdir dmacrys; fi
cd dmacrys
cp ../neighcrys3/$mol.dmain ../neighcrys3/fort.20 .

echo waiting for DMACRYS plut lattice minimisation PID.
#Run dmacrys
$dmacrys <$mol.dmain> $mol.dmaout

cd $home
echo "Step 7: Finished!"


###################################
#
# Step 8: Set up properties calculation
#
###################################
echo "Step 8: Setting up properties calculation ..."
if [ ! -d neighcrys_props ]; then mkdir neighcrys_props; fi

#cp ./dmacrys/*.dmain neighcrys_props
#cp ./dmacrys/fort.20 neighcrys_props
cd neighcrys_props
cp ../dmacrys/*.dmain ./
cp ../dmacrys/fort.20 ./

#Change options
$dmacrysdir/dmacrys_prop.sh
cp $mol.dmain ${mol}_min.dmain

cd $home
echo "Step 8: Finished!"


###################################
#
# Step 9: Calculate properties
#
###################################

echo "Step 9: Calculating properties ..."
if [ ! -d dmacrys_props ]; then mkdir dmacrys_props; fi
cd dmacrys_props
cp ../neighcrys_props/${mol}_min.dmain ../neighcrys_props/fort.20 .
sed -i 's/STAR PLUT/STAR PROP/g' ${mol}_min.dmain
```

```
#Run dmacrys
echo waiting for DMACRYS prop lattice mimisation PID.
$dmacrys <${mol}_min.dmain> ${mol}_min.dmaout

calculate entropies from phonon modes
perl $fvibsetup < ${mol}_min.dmaout > fort.10
$elastphon
$dmacrysdir/done_dmacrys.sh
cp dmacrys.end ../


cd $home
echo "Step 9: Finished!"

###################################
#
# End
#
###################################

#Write log file
done 2>&1 | tee dmacrys_setup.log
```

Listing L.2: A Bash script to automate the addition of new potential parameters to the Buckingham potential parameter file

```
#!/bin/bash
# add a potential parameter one at a time. Created by James McDonagh 2011

home=`pwd`

echo enter filename to act on.
read file

echo enter the atom label of the atom type to be added.
read new_atom

v=1
while [ $v == "1" ]
do

        if [ $v == "2" ]
        then
        echo Exiting loop
        break
        fi


        echo do you want to add a new potential parameter to a file ? '(y/n)'
        read ans

        if [ $ans == "n" ]
        then
                v=2
                echo Exiting

        elif [ $ans == "y" ]
        then

                echo enter the atom label of the existing atom to be mixed with
                read exist_atom

                echo ' BUCK    '$exist_atom'    '$new_atom'' >> $file
                echo please enter the parameters
                echo enter a
                read a
                echo enter b
                read b
                echo enter c
                read c
                echo '   ' $a'       '$b'       '$c'         0.00    70.00' >> $file
                echo '  ENDS' >> $file
```

```
                echo potential parameters entered for $exist_atom    $new_atom
        fi
done
echo all paramters added.
exit
```

Listing L.3: A Bash script to automatically set up CASTEP calculations from cif files

```
#!/bin/bash

##############################################################################
# This Script reads in files names as csd refcodes from csd cif files and setup #
# CASTEP calculations making a directory for each csd refcode.                 #
#                                                                              #
# Input should be a sub directory called cif containing the cif's you wish to  #
# act on. A template directory should also be set up containing a job submission  #
# script and template parameter file for CASTEP (please edit the               #
# Castep_templates line accordingly).                                          #
#                                                                              #
# Output will be a directory for each cif file named castep_<csd_refcode>      #
# containing the cif file Castep cell file and Castep parameter file.          #
# Script originally created in the University of St Andrews by James McDonagh  #
# 2013                                                                         #
##############################################################################

for writelogfile in 1;
do

# Fixed variable
home=`pwd`
castpath=/usr/local/progs/CASTEP-5.5.2
#CASTEP installation location on the current machine
castscript=/usr/local/progs/CASTEP-5.5.2/scripts
#CASTEP's scripts directory in the installation
script=/home/st-andrews/jm222/Scripts
#My scripts directory
Castep_templates=/home/st-andrews/jm222/Scripts/CASTEP/templates
#My CASTEP templates directory

echo This script sets up CASTEP-5.5.2 calculations on Wardlaw. You must
have cif s of all of the molecules you wish to run, preferably in a
subdirectory of the the directory where this script is being run.

echo do you wish to continue to do this? '(y/n)'
read ans

if [ $ans == "y" ]
then
        echo Continuing
else
        echo Exiting
        exit
fi

# Find the cif files

dir=./cif
if [ -d $dir ]
then
        echo cif directory exists as a sub directory of $home
        for i in `ls cif/*cif | awk 'BEGIN{FS="/"}{print $2}' | sed 's/\.cif//g'`
        do
                if [ -d castep_$i ]
                then
                        cp cif/$i.cif castep_$i
                        cp $Castep_templates/job-castep castep_$i
                        sed -i 's/MOLECULE/'$i'/g' castep_$i/job-castep
                else
                        mkdir castep_$i
                        cp cif/$i.cif castep_$i
                        cp $Castep_templates/job-castep castep_$i
                        sed -i 's/MOLECULE/'$i'/g' castep_$i/job-castep
```

```
                fi
        done
else
        echo cif directory does not exist as a sub directory of $home.
        Please enter the directory location:
        read cif_dir
        echo $cif_dir>> file_path.tmp
        sed 's(/(&\n(g' file_path.tmp
        col_no=`grep -c '/' file_path.tmp`
        for j in `ls $cif_dir/*cif | awk 'BEGIN{FS="/"}{print '$col_no'}' | sed 's/\.cif//g'`
        do
                if [ -d castep_$j ]
                then
                        cp cif/$j.cif castep_$j
                else
                        mkdir castep_$j
                        cp cif/$j.cif castep_$j
                        cp $Castep_templates/job-castep castep_$i
                fi
        done
        rm file_path.tmp
fi

# Make Castep cell files

for cast_dirs in `ls -d castep_*`
do
echo $cast_dirs
        cd $cast_dirs
        cif2cell *.cif -p castep -o cell.cell
        echo perl script returned $?

        cp cell.cell cell_1
        sed -i '/ENDBLOCK SPECIES_POT/ a\kpoints_mp_spacing 0.05' cell_1
        sed -i '/kpoints_mp_spacing 0.05/{x;p;x;}' cell_1
        sed -i '/kpoints_mp_spacing 0.05/ a\#symmetry_generate' cell_1
        sed -i '/symmetry_generate/{x;p;x;}' cell_1

######### From Previous pdb version ####################
#for line_no in `grep -n '%BLOCK kpoints_list' cell_1 | #
sed 's/%//g;s/[a-z]//g;s/[A-Z]//g;s/://g;s/_//g'`       #
#do
#
#       x=$(($line_no+2))
#
#       y=$(($line_no+1))
#
#       sed -i ''$x'd;'$y'd;'$line_no'd' cell_1                          #
#   done
############################################################

                for nam in `ls *.cif | sed 's/'.'cif//g'`
                do
                        mv cell_1 $nam.cell
                        echo $nam
                done



        cd $home
done

# Make Castep param file

echo A template parameter file '(<filename>.param)' is stored and will
be printed to the screen.

echo This should be adequate for most jobs, if you are are happy
with it allow it to be copied to all directories if not then make
the changes 1 at a time in this script by following the instructions.

echo Alternatively if you want to make a number of bigger changes
```

```
exit the script'(ctrl c)' and change the template which can be found in
echo $Castep_templates.
echo
echo Do you wish to exit before this step? '(y/n)'
read exi
if [ $exi == "y" ]
then
        echo Exiting script
        exit
else
        echo Continuing

fi
v=1
cp $Castep_templates/MOLECULE.param $Castep_templates/MOLECULE_1.param
while [ $v == "1" ]
do
        if [ $v == "2" ]
        then
        echo Exiting loop
        break
        fi

        echo the template is currently as follows in : '(program note v = '$v')'
        cat $Castep_templates/MOLECULE_1.param

        echo Do you want to use the template as is '(y)' or make changes '(n)'
        read aw
        if [ $aw == "y" ]
        then
                v=2
                echo Continuing and copying the template file to the respective directories:
                 '(program note v = '$v')'
                for castep_dir in `ls -d castep_*`
                do
                        cd $castep_dir
                        cp $Castep_templates/MOLECULE_1.param ./
                        cp $Castep_templates/MOLECULE_phonon.param ./

                        for mol in `ls *.cif | sed 's/'.'cif//g'`
                        do
                                mv MOLECULE_1.param $mol.param
                                sed -i 's/MOLECULE/'$mol'/g' $mol.param
                                mv MOLECULE_phonon.param $mol'_phonon'.param
                sed -i 's/MOLECULE/'$mol'/g' $mol'_phonon'.param
                        done
                        cd $home
                done

        else
                echo Do you wish to substitute '(s)' or delete '(d)' lines in the template ?
                 '(lines that are hashed are comments and do not need deleting)'
                read a
                if [ $a == "s" ]
                then
                        echo Enter the text you would like to be replaced
                        '(i.e. the text that is currently in the template you want to replace)'
                        read from
                        echo Enter the text you like to replace the old text
                        '(i.e. the text that want in the finalised Castep file)'
                        read too
                        sed -i "s/$from/$too/g" $Castep_templates/MOLECULE_1.param
                        cp $Castep_templates/MOLECULE_1.param $Castep_templates/
                        MOLECULE_altered.param
                elif [ $a == "d" ]
                then
                        echo Enter the text line you would like to be removed
                        read rem
                        sed -i "/$rem/d" $Castep_templates/MOLECULE_1.param
                        cp $Castep_templates/MOLECULE_1.param
                        $Castep_templates/MOLECULE_altered.param
                fi
        fi
```

```
done

echo Done !
echo All cif files in the specified cif directory have a directory made
called Castep_CSD_code the cif files have been copied to the directory
and converted to cell file and a parameters file has been created
according to the specification entered. A job-castep file has also
been place in all directories and is ready to run each molecules
specified calculation.

done 2>&1 | tee Castep.log
exit
```

Listing L.4: An R script to run a neural network.

```
run_nn <- function(input, path){

  ################################
  ##### Libraries
  ################################

  library(neuralnet)
  library(caret)
  library(dismo)
  library(kernlab)
  library(lattice)
  library(plyr)
  library(pROC)
  library(randomForest)
  library(raster)
  library(reshape)
  library(rJava)
  library(sp)

  Debug <- F

  ########################################
  ##### Resolve the path
  ########################################

  setwd(path)
  message("Working directory: ", path)
  # INPUT DATA SHOULD BE A .csv FILE

  ########################################
  ##### Scale Data - Auto Scale
  ########################################

  data <- input
  columnWithRowNames <- row.names(data)# unique row names column (molecule-name column)

  # Scale the data Note all data is scaled by the same values
  message("Scaling the input data to have a standard mean
  and standard deviation - Auto scaling ")
  all_scaled_data <- scale(data[-c(15)])

  # Define a training and test set from the scaled data 75 training 25 test
  scaled_training_data <- all_scaled_data[1:75, ]
  scaled_test_data <- all_scaled_data[76:100, ]

  # Prepare the training data
  Train <- scaled_training_data[, -(ncol(data))] # Removes the last column i.e.
                                                 # the experimental values
  Test <- scaled_test_data
  # If Debug is T (true) tyou can check the data
  if(Debug == T) message("Raw Data: ", input_1)
  if(Debug == T) message("scaled Data", x)

  ########################################
  ##### Begin Neural Network Training
  ########################################
```

```
  #Train the neural network
  #Going to have 10 hidden layers
  #Threshold is a numeric value specifying the threshold for the partial
  #derivatives of the error function as stopping criteria.
  nn <- neuralnet(logS~ALogP+ALogp2+AMR+XLogP
  +WTPT_3+VCH_7+ATSc2+SP_6+ATSc1+SP_5+SP_7
  +ATSm4+ATSm1,scaled_training_data,
  hidden=(20), threshold=0.01, rep=10,
  algorithm="rprop+")

  print(nn)

  #Plot the neural network
  plot(nn)

  ######################################
  ##### Test the Neural Network
  ######################################

  #Test the neural network on some training data
  #testdata <- as.data.frame((1:10)^2) #Generate some squared numbers
  net.results <- compute(nn,scaled_test_data[, 2:14]) #Run them through the neural network

  #Lets see what properties net.sqrt has
  #ls(net.results)

  #Lets see the results
  print(net.results$net.result)

  #Lets display a better version of the results
  #cleanoutput <- cbind(testdata,sqrt(testdata),
   #                    as.data.frame(net.results$net.result))
  #colnames(cleanoutput) <- c("Input","Expected Output","Neural Net Output")
  #print(cleanoutput)
}
```

Bibliography

1. Google Scholar. *Search solubility models in 2013 - 2014* <http://scholar.google.co.uk/scholar?q=+%22solubility+models%22&btnG=&hl=en&as_sdt=0,5&as_ylo=2013&as_yhi=2014> (2014).
2. Young, H. K. Antimicrobial resistance spread in aquatic environments. *Journal of Antimicrobial Chemotherapy* **31,** 627–635. ISSN: 0305-7453 (1993).
3. Fromme, H. *et al.* Occurrence of phthalates and bisphenol A and F in the environment. *Water Research* **36,** 1429–1438. ISSN: 0043-1354 (2002).
4. Kitano, H. Computational systems biology. *Nature* **420,** 206–210 (2002).
5. Jorgensen, W. L. & Duffy, E. M. Prediction of drug solubility from structure. *Advanced Drug Delivery Reviews* **54,** 355–366. ISSN: 0169-409X (2002).
6. Palmer, D. S. *et al.* Predicting intrinsic aqueous solubility by a thermodynamic cycle. *Molecular Pharmaceutics* **5,** 266–279. ISSN: 1543-8384 (Mar. 2008).
7. Michel, J. Current and emerging opportunities for molecular simulations in structure-based drug design. *Physical Chemistry Chemical Physics* **16,** 4465–4477 (10 2014).
8. Speck-Planche, A. *et al.* Rational design of new agrochemical fungicides using substructural descriptors. *Pest Management Science* **67,** 438–445. ISSN: 1526-4998 (2011).
9. Michel, J., Tirado-Rives, J. & Jorgensen, W. L. Energetics of Displacing Water Molecules from Protein Binding Sites: Consequences for Ligand Optimization. *Journal of the American Chemical Society* **131.** PMID: 19778066, 15403–15411 (2009).
10. Fagerberg, J., Karlsson, E., Ulander, J., Hanisch, G. & Bergström, C. Computational Prediction of Drug Solubility in Fasted Simulated and Aspirated Human Intestinal Fluid. English. *Pharmaceutical Research,* 1–12. ISSN: 0724-8741 (2014).
11. Papadatos, G. *et al.* Lead optimization using matched molecular pairs: inclusion of contextual information for enhanced prediction of HERG inhibition, solubility, and lipophilicity. *Journal of chemical information and modeling* **50,** 1872–1886 (2010).
12. sirius-analytical. *Solubility Definitions* <http://www.sirius-analytical.com/science/solubility/solubility-definitions> (2014).
13. Sell, C. S. in *Chemistry and the Sense of Smell* 32–187 (John Wiley & Sons, Inc., 2014). ISBN: 9781118522981. doi:10.1002/9781118522981.ch2. <http://dx.doi.org/10.1002/9781118522981.ch2>.

14. Jouyban, A. & Fakhree, M. A. Experimental and computational methods pertaining to drug solubility. *Toxicity and Drug Testing* **1** (2012).

15. Noyes, A. A. & Whitney, W. R. The rate of solution of solid substances in their own solutions. *Journal of the American Chemical Society* **19,** 930–934 (1897).

16. Bauer, J. *et al.* Ritonavir: an extraordinary example of conformational polymorphism. *Pharmaceutical research* **18,** 859–866 (2001).

17. Chemburkar, S. R. *et al.* Dealing with the impact of ritonavir polymorphs on the late stages of bulk drug process development. *Organic Process Research & Development* **4,** 413–417 (2000).

18. Hassan, M. A., Salem, M. S., Sueliman, M. S. & Najib, N. M. Characterization of famotidine polymorphic forms. *International journal of pharmaceutics* **149,** 227–232 (1997).

19. Phadnis, N. V. & Suryanarayanan, R. Polymorphism in anhydrous theophylline—implications on the dissolution rate of theophylline tablets. *Journal of pharmaceutical sciences* **86,** 1256–1263 (1997).

20. Henck, J. O. & Kuhnert-Brandstatter, M. Demonstration of the terms enantiotropy and monotropy in polymorphism research exemplified by flurbiprofen. *Journal of pharmaceutical sciences* **88,** 103–108 (1999).

21. Atkins, P. *Physical Chemistry* (Oxford University Press, 1978).

22. Bergström, C. A. *et al.* Absorption classification of oral drugs based on molecular surface properties. *Journal of medicinal chemistry* **46,** 558–570 (2003).

23. Le, J. *Drug Absorption* <`http://www.merckmanuals.com/professional/clinical_pharmacology/pharmacokinetics/drug_absorption.html`> (2014).

24. Palmer, D. *Computational studies on the Aqueous Solubility of Pharmaceutical Molecules* PhD Thesis (Univeristy of Camebridge Churchill College, 2008).

25. Budavari, S. Merck index (1989).

26. Howard, P. & Meylan, W. *Physical/chemical property database (PHYSPROP)* 1999.

27. Yalkowsky, S. H., He, Y. & Jain, P. *Handbook of aqueous solubility data* (CRC press, 2010).

28. Llinàs, A., Glen, R. C. & Goodman, J. M. Solubility Challenge: Can You Predict Solubilities of 32 Molecules Using a Database of 100 Reliable Measurements? *Journal of Chemical Information and Modeling* **48,** 1289–1303. ISSN: 1549-9596 (2008).

29. Rytting, E., Lentz, K., Qing, C. X., Qian, F. & Venkatesh, S. Aqueous and cosolvent solubility data for drug-like organic compounds. *The AAPS Journal* **7,** E78–E105 (2005).

30. Pontolillo, J. & Eganhouse, R. P. *The search for reliable aqueous solubility (Sw) and octanol-water partition coefficient (Kow) data for hydrophobic organic compounds: DDT and DDE as a case study* (US Department of the Interior, US Geological Survey Reston, Virginia, 2001).

31. Palmer, D. S., McDonagh, J. L., Mitchell, J. B. O., van Mourik, T. & Fedorov, M. V. First-Principles Calculation of the Intrinsic Aqueous Solubility of Crystalline Druglike Molecules. *Journal of Chemical Theory and Computation,* 3322–3337. ISSN: 1549-9618 (2012).

32. Lusci, A., Pollastri, G. & Baldi, P. Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug-Like Molecules. *Journal of Chemical Information and Modeling* **53,** 1563–1575. ISSN: 1549-9596 (2013).

33. McDonagh, J. L., Nath, N., De Ferrari, L., van Mourik, T. & Mitchell, J. B. O. Uniting Cheminformatics and Chemical Theory To Predict the Intrinsic Aqueous Solubility of Crystalline Druglike Molecules. *Journal of Chemical Information and Modeling* **54,** 844–856. ISSN: 1549-9596 (2014/03/24 2014).

34. Ren, z. Bao, Chong, g. Hai, Li, r. Wei & Wang, a. Fu. Solubility of Potassium p-Chlorophenoxyacetate in Ethanol+ Water from (295.61 to 358.16) K. *Journal of Chemical & Engineering Data* **50,** 907–909 (2005).

35. sirius-analytical. *Solubility validations - CheqSol* <http://www.sirius-analytical.com/science/solubility/solubility-validations> (2014).

36. Palmer, D. S. & Mitchell, J. B. O. Is Experimental Data Quality the Limiting Factor in Predicting the Aqueous Solubility of Druglike Molecules? *Molecular Pharmaceutics,* null (2014).

37. Le, J. *Drug Excretion* <http://www.merckmanuals.com/professional/clinical_pharmacology/pharmacokinetics/drug_excretion.html> (2014).

38. Novartis. *Drug-discovery an the development process* <http://www.novartis.com/innovation/research-development/drug-discovery-development-process/index.shtml> (2014).

39. Hughes, J., Rees, S., Kalindjian, S. & Philpott, K. Principles of early drug discovery. *British Journal of Pharmacology* **162,** 1239–1249. ISSN: 1476-5381 (2011).

40. Leach, A. G. *et al.* Matched Molecular Pairs as a Guide in the Optimization of Pharmaceutical Properties; a Study of Aqueous Solubility, Plasma Protein Binding and Oral Exposure. *Journal of Medicinal Chemistry* **49,** 6672–6682 (2006).

41. Food & (FDA), D. A. *The Biopharmaceutics Classification System (BCS) Guidance* <http://www.fda.gov/AboutFDA/CentersOffices/OfficeofMedicalProductsandTobacco/CDER/ucm128219.htm> (2014).

42. Savjani, K. T., Gajjar, A. K. & Savjani, J. K. Drug solubility: importance and enhancement techniques. *ISRN pharmaceutics* **2012** (2012).

43. Goodman, J. *The Goodman Group - Solubility Challenge* 2008. <http://www-jmg.ch.cam.ac.uk/data/solubility/>.

44. Hopfinger, A. J., Esposito, E. X., Llinàs, A., Glen, R. C. & Goodman, J. M. Findings of the Challenge To Predict Aqueous Solubility. *Journal of Chemical Information and Modeling* **49,** 1–5. ISSN: 1549-9596 (2009/01/26 2008).

45. Klamt, A., Eckert, F., Hornig, M., Beck, M. E. & Bürger, T. Prediction of aqueous solubility of drugs and pesticides with COSMO–RS. *Journal of Computational Chemistry* **23,** 275–281. ISSN: 1096-987X (2002).

46. Yalkowsky, S. H. & Valvani, S. C. Solubility and partitioning I: solubility of nonelectrolytes in water. *Journal of pharmaceutical sciences* **69,** 912–922 (1980).

47. Jain, N. & Yalkowsky, S. H. Estimation of the aqueous solubility I: application to organic nonelectrolytes. *Journal of pharmaceutical sciences* **90,** 234–252 (2001).

48. Ben-Naim, A. Standard thermodynamics of transfer. Uses and misuses. *The Journal of Physical Chemistry* **82,** 792–803. ISSN: 0022-3654 (1978).

49. Ben-Naim, A. & Marcus, Y. Solvation thermodynamics of nonionic solutes. *The Journal of Chemical Physics* **81,** 2016–2027 (1984).

50. Ben-Naim, A. *Molecular theory of solutions* (Oxford Univ. Press, 2006).

51. Palmer, D. S., Frolov, A. I., Ratkova, E. L. & Fedorov, M. V. Towards a universal method for calculating hydration free energies: a 3D reference interaction site model with partial molar volume correction. *Journal of Physics: Condensed Matter* **22,** 492101. ISSN: 0953-8984 (2010).

52. Ratkova, E. L. & Fedorov, M. V. Combination of RISM and Cheminformatics for Efficient Predictions of Hydration Free Energy of Polyfragment Molecules: Application to a Set of Organic Pollutants. *Journal of Chemical Theory and Computation* **7,** 1450–1457. ISSN: 1549-9618 (2011).

53. Salahinejad, M., Le, T. C. & Winkler, D. A. Aqueous Solubility Prediction: Do Crystal Lattice Interactions Help? *Molecular Pharmaceutics* **10,** 2757–2766. ISSN: 1543-8384 (2013/07/01 2013).

54. Sanghvi, T., Jain, N., Yang, G. & Yalkowsky, S. H. Estimation of Aqueous Solubility By The General Solubility Equation (GSE) The Easy Way. *QSAR & Combinatorial Science* **22,** 258–262. ISSN: 1611-0218 (2003).

55. Karamertzanis, P. G. *et al.* Can the Formation of Pharmaceutical Cocrystals Be Computationally Predicted? 2. Crystal Structure Prediction. *Journal of Chemical Theory and Computation* **5,** 1432–1448. ISSN: 1549-9618 (2009).

56. Karamertzanis, P. G. & Pantelides, C. C. Ab initio crystal structure prediction-I. Rigid molecules. *Journal of Computational Chemistry* **26,** 304–324. ISSN: 1096-987X (2005).

57. Lehmann, C. W. Crystal Structure Prediction - Dawn of a New Era. *Angewandte Chemie International Edition* **50,** 5616–5617. ISSN: 1521-3773 (2011).

58. Pisani, C. *et al.* Cryscor: a program for the post-Hartree-Fock treatment of periodic systems. *Phys. Chem. Chem. Phys.* **14,** 7615–7628 (21 2012).

59. Marom, N. *et al.* Many-Body Dispersion Interactions in Molecular Crystal Polymorphism. *Angewandte Chemie International Edition* **52,** 6629–6632. ISSN: 1521-3773 (2013).

60. Lommerse, J. P. M. *et al.* A test of crystal structure prediction of small organic molecules. *Acta Crystallographica Section B* **56,** 697–714. ISSN: 0108-7681 (2000).

61. Motherwell, W. D. S. *et al.* Crystal structure prediction of small organic molecules: a second blind test. *Acta Crystallographica Section B* **58,** 647–661. ISSN: 0108-7681 (2002).

62. Day, G. M. *et al.* A third blind test of crystal structure prediction. *Acta Crystallographica Section B* **61,** 511–527. ISSN: 0108-7681 (2005).

63. Day, G. M. *et al.* Significant progress in predicting the crystal structures of small organic molecules - a report on the fourth blind test. *Acta Crystallographica Section B* **65,** 107–125. ISSN: 0108-7681 (2009).

64. Neumann, M. A. & Leusen, J. Frank J. J.and Kendrick. A Major Advance in Crystal Structure Prediction. *Angewandte Chemie International Edition* **47,** 2427–2430. ISSN: 1521-3773 (2008).

65. Price, S. L. *et al.* Modelling organic crystal structures using distributed multipole and polarizability-based model intermolecular potentials. *Physical Chemistry Chemical Physics* **12,** 8478–8490. ISSN: 1463-9076 (2010).

66. Chaplot, S. L. & Rao, K. R. Crystal structure prediction - evolutionary or revolutionary crystallography? *Current Science* **91,** 1448–1450 (2006).

67. Clark, S. J. *et al.* First principles methods using CASTEP. *Zeitschrift für Kristallographie* **5-6,** 567–570 (2005).

68. Gerogiokas, G. *et al.* Prediction of Small Molecule Hydration Thermodynamics with Grid Cell Theory. *Journal of Chemical Theory and Computation* **10,** 35–48 (2013).

69. Marenich, A. V., Cramer, C. J. & Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *The Journal of Physical Chemistry B* **113,** 6378–6396. ISSN: 1520-6106 (2009).

70. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **28,** 31–36. ISSN: 0095-2338 (1988).

71. Weininger, D. SMILES. 3. DEPICT. Graphical depiction of chemical structures. *Journal of Chemical Information and Computer Sciences* **30,** 237–243. ISSN: 0095-2338 (1990).

72. Weininger, D., Weininger, A. & Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of Chemical Information and Computer Sciences* **29,** 97–101. ISSN: 0095-2338 (1989).

73. Stein, S. E., Heller, S. R. & Tchekhovskoi, D. An open standard for chemical structure representation: The IUPAC chemical identifier (2003).

74. Leach, A. R. & Gillet, V. J. *An introduction to chemoinformatics* Revised Edition (Springer, 2007).

75. Dalby, A. *et al.* Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *Journal of Chemical Information and Computer Sciences* **32,** 244–255. ISSN: 0095-2338 (1992).

76. Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D. & Pletnev, I. InChI - the worldwide chemical structure identifier standard. *Journal of Cheminformatics* **5,** 7. ISSN: 1758-2946 (2013).

77. Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures–A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation* **5,** 107–113. ISSN: 0021-9576 (1965/05/01 1965).

78. Dittmar, P. G., Stobaugh, R. E. & Watson, C. E. The chemical abstracts service chemical registry system. I. General design. *Journal of Chemical Information and Computer Sciences* **16,** 111–121. ISSN: 0095-2338 (1976).

79. Freeland, R. G., Funk, S. A., O'Korn, L. J. & Wilson, G. A. The chemical abstracts service chemical registry system. II. Augmented connectivity molecular formula. *Journal of Chemical Information and Computer Sciences* **19,** 94–98. ISSN: 0095-2338 (1979).

80. Murray-Rust, P. & Rzepa, H. S. Chemical markup, XML, and the Worldwide Web. 1. Basic principles. *Journal of Chemical Information and Computer Sciences* **39,** 928–942. ISSN: 0095-2338 (1999).

81. Holliday, G. L., Murray-Rust, P. & Rzepa, H. S. Chemical Markup, XML, and the World Wide Web. 6. CMLReact, an XML Vocabulary for Chemical Reactions. *Journal of Chemical Information and Modeling* **46,** 145–157. ISSN: 1549-9596 (2005).

82. Steinbeck, C. *et al.* The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *Journal of Chemical Information and Computer Sciences* **43,** 493–500. ISSN: 0095-2338 (2003).

83. Lawson, K. & Lawson, J. LICSS - a chemical spreadsheet in microsoft excel. *Journal of Cheminformatics* **4,** 3. ISSN: 1758-2946 (2012).

84. Guha, R. Chemical informatics functionality in R. *Journal of Statistical Software* **18,** 1–16 (2007).

85. Microsoft development team. *Microsoft Excel* Microsoft Corporation (2010).

86. *R: A Language and Environment for Statistical Computing* (ed R Core Team) ISBN: 3-900051-07-0. <`http://www.R-project.org`> (Vienna, Austria, 2012).

87. Wiener, H. Structural Determination of Paraffin Boiling Points. *Journal of the American Chemical Society* **69,** 17–20. ISSN: 0002-7863 (1947).

88. Randic, M. Characterization of molecular branching. *Journal of the American Chemical Society* **97,** 6609–6615. ISSN: 0002-7863 (1975).

89. Kier, L. B. & Hall, L. H. *Molecular connectivity in structure-activity analysis* (Research Studies, 1986).

90. Hall, L. H. & Kier, L. B. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure - Property Modeling. *Reviews in Computational Chemistry* **2,** 367–422. ISSN: 0470125799 (2007).

91. Hall, L. H., Mohney, B. & Kier, L. B. The electrotopological state: an atom index for QSAR. *Quantitative Structure - Activity Relationships* **10,** 43–51. ISSN: 1521-3838 (1991).

92. Burden, F. R. Molecular identification number for substructure searches. *Journal of Chemical Information and Computer Sciences* **29,** 225–227. ISSN: 0095-2338 (1989).

93. Mavridis, L., Hudson, B. D. & Ritchie, D. W. Toward High Throughput 3D Virtual Screening Using Spherical Harmonic Surface Representations. *Journal of Chemical Information and Modeling* **47,** 1787–1796. ISSN: 1549-9596 (2007).

94. Carhart, R. E., Smith, D. H. & Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences* **25,** 64–73. ISSN: 0095-2338 (1985).

95. Leo, A., Hansch, C. & Elkins, D. Partition coefficients and their uses. *Chemical Reviews* **71,** 525–616. ISSN: 0009-2665 (1971).

96. Rekker, R. F. *The hydrophobic fragmental constant* **977** (Elsevier, Amsterdam, 1977).

97. Rekker, R. F. & Mannhold, R. *Calculation of drug lipophilicity: the hydrophobic fragmental constant approach* (Wiley-VCH, 1992).

98. Leo, A. J. Calculating log P_{oct} from structures. *Chemical Reviews* **93,** 1281–1306. ISSN: 0009-2665 (1993).

99. Leo, A., Jow, P. Y. C., Silipo, C. & Hansch, C. Calculation of hydrophobic constant (log P) from π and f constants. *Journal of Medicinal Chemistry* **18,** 865–868. ISSN: 0022-2623 (1975).

100. Wildman, S. A. & Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *Journal of Chemical Information and Computer Sciences* **39,** 868–873. ISSN: 0095-2338 (1999).

101. Ghose, A. K. & Crippen, G. M. Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships I. Partition Coefficients as a Measure of Hydrophobicity. *Journal of Computational Chemistry* **7,** 565–577 (1986).

102. Ghose, A. K. & Crippen, G. M. Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. *Journal of chemical information and computer sciences* **27,** 21–35 (1987).

103. Viswanadhan, V. N., Ghose, A. K., Revankar, G. R. & Robins, R. K. Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *Journal of Chemical Information and Computer Sciences* **29,** 163–172 (1989).

104. Wang, R., Fu, Y. & Lai, L. A new atom-additive method for calculating partition coefficients. *Journal of chemical information and computer sciences* **37,** 615–621 (1997).

105. Svetnik, V. *et al.* Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences* **43,** 1947–1958. ISSN: 0095-2338 (2003).

106. Berk,
R. *Classification and Regression Trees (CART)* Springer Science+Business Media. 2008.

107. Basak, D., Pal, S. & Patranabis, D. C. Support vector regression. *Neural Information Processing-Letters and Reviews* **11,** 203–224 (2007).

108. Parrella, F. Online support vector regression. *Master's Thesis, Department of Information Science, University of Genoa, Italy* (2007).

109. Wold, H. Soft modelling: the basic design and some extensions. *Systems Under Indirect Observation, PartII,* 36–37 (1982).

110. Wold, S., Martens, H. & Wold, H. in *Matrix Pencils* (eds Kagström, B. & Ruhe, A.) 286–293 (Springer Berlin Heidelberg, 1983). ISBN: 978-3-540-11983-8. doi:10.1007/BFb0062108.

111. Abdi, H. Partial Least Squares (PLS) Regression, 1–7 (2003).

112. Cramer, C. J. *Essentials of computational chemistry: theories and models* (John Wiley & Sons, 2013).

113. Jensen, F. *Introduction to computational chemistry* (John Wiley & Sons, 2007).

114. Azman, A. M. A Chemistry Spell-Check Dictionary for Word Processors. *Journal of Chemical Education* **89,** 412–413. ISSN: 0021-9584 (2012/02/14 2012).

115. Blinder, S. M. *Gaussian Approximations to 1s Slater-Type Orbitals from Wolfram Demonstrations Projects* <http://demonstrations.wolfram.com/GaussianApproximationsTo1sSlaterTypeOrbitals/> (2014).

116. Lewars, E. *Computational chemistry Introduction to the Theory and Applications of Molecular and Quantum Mechanics* Kluwer Academic Publishers. 2004.

117. Woon, D. E. & Dunning, T. H. Gaussian basis sets for use in correlated molecular calculations. V. Core-valence basis sets for boron through neon. *The Journal of Chemical Physics* **103,** 4572–4585 (1995).

118. Binkley, J. S., Pople, J. A. & Hehre, W. J. Self-consistent molecular orbital methods. 21. Small split-valence basis sets for first-row elements. *Journal of the American Chemical Society* **102,** 939–947 (1980).

119. Jónsson, H. *The Hartree-Fock Approximation* <http://theochem.org/CompChem11f/notesSysIdenB.pdf> (2011).

120. Grant, G. H. & Richards, W. G. *Computational chemistry* (Oxford University Press, 1995).

121. Sherrill, C. D. An introduction to Hartree-Fock molecular orbital theory. *School of Chemistry and Biochemistry Georgia Institute of Technology* (2000).

122. Møller, C. & Plesset, M. S. Note on an Approximation Treatment for Many-Electron Systems. *Physical Review* **46,** 618–622 (1934).

123. Mitchell, J. B. O. *Chemical application of electronic structure calculations* 2014.

124. Thomas, L. H. The calculation of atomic fields. *Mathematical Proceedings of the Cambridge Philosophical Society* **23,** 542–548 (1927).

125. Fermi, E. Un metodo statistico per la determinazione di alcune priorieta dell'atome. *Rend Accad Naz Lincei* **6,** 602–607 (1927).

126. Slater, J. C. A simplification of the Hartree-Fock method. *Physical Review* **81,** 385 (1951).

127. Dirac, P. A. *Note on exchange phenomena in the Thomas atom* in *Mathematical Proceedings of the Cambridge Philosophical Society* **26** (1930), 376–385.

128. Bloch, F. Bemerkung zur Elektronentheorie des Ferromagnetismus und der elektrischen Leitfähigkeit. German. *Zeitschrift für Physik* **57,** 545–555. ISSN: 0044-3328 (1929).

129. Hohenberg, P. & Kohn, W. Inhomogeneous Electron Gas. *Physical Review* **136,** B864–B871 (1964).

130. Kohn, W. & Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Physical Review* **140,** A1133–A1138 (1965).

131. Bühl, M. *Chemical application of electronic structure calculations* 2014.

132. Koch, W., Holthausen, M. C. & Holthausen, M. C. *A chemist's guide to density functional theory* (Wiley-Vch Weinheim, 2001).

133. Reimers, J. R. *Computational methods for large systems: electronic structure approaches for biotechnology and nanotechnology* (John Wiley & Sons, 2011).

134. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Physical Review Letters* **77,** 3865–3868 (1996).

135. Harrison, N. An introduction to density functional theory. *Nato science serise sub serise III computer and systems sciences* **187,** 45–70 (2003).

136. Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **38,** 3098–3100 (6 Sept. 1988).

137. Hamprecht, F. A., Cohen, A. J., Tozer, D. J. & Handy, N. C. Development and assessment of new exchange-correlation functionals. *The Journal of Chemical Physics* **109,** 6264–6271 (1998).

138. Lee, C., Yang, W. & Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B* **37,** 785–789 (2 Jan. 1988).

139. Perdew, J. P. & Yue, W. Accurate and simple density functional for the electronic exchange energy: Generalized gradient approximation. *Physical Review B* **33,** 8800–8802 (12 June 1986).

140. Perdew, J. P. *et al.* Atoms, molecules, solids, and surfaces: Applications of the generalized gradient approximation for exchange and correlation. *Phys. Rev. B* **46,** 6671–6687 (11 Sept. 1992).

141. Perdew, J. P. & Wang, Y. Accurate and simple analytic representation of the electron-gas correlation energy. *Physical Review B* **45,** 13244–13249 (23 June 1992).

142. Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics* **98,** 5648–5652 (1993).

143. Toulouse, J. *Extending Kohn-Sham density-functional theory : double-hybrid DFT and multiconfigurational hybrid DFT* <http://www.lct.jussieu.fr/pagesperso/toulouse/communications/presentation_strasbourg_12.pdf> (2012).

144. Grimme, S. Accurate description of van der Waals complexes by density functional theory including empirical corrections. *Journal of Computational Chemistry* **25,** 1463–1473. ISSN: 1096-987X (2004).

145. Brandenburg, J. G. & Grimme, S. Accurate Modeling of Organic Molecular Crystals by Dispersion-Corrected Density Functional Tight Binding (DFTB). *The Journal of Physical Chemistry Letters,* 1785–1789 (2014).

146. Tkatchenko, A. & Scheffler, M. Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data. *Physical Review Letters* **102,** 073005 (7 Feb. 2009).

147. Zhao, Y. & Truhlar, D. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)* **120,** 215–241. ISSN: 1432-881X (2008).

148. Cohen, A. J., Mori-Sánchez, P. & Yang, W. Challenges for Density Functional Theory. *Chemical Reviews* **112,** 289–320 (2012).

149. Wikipedia community. *Bloch wave* <http://en.wikipedia.org/wiki/Bloch_wave> (2014).

150. Wikipedia community. *Wigner-Seitz Cell* <http://en.wikipedia.org/wiki/Wigner%C3%A2%C2%80%C2%93Seitz_cell> (2014).

151. Stone, A. J. Distributed multipole analysis, or how to describe a molecular charge distribution. *Chemical Physics Letters* **83,** 233–239. ISSN: 0009-2614 (1981).

152. Williams, D. E. Nonbonded Potential Parameters Derived from Crystalline Aromatic Hydrocarbons. *The Journal of Chemical Physics* **45,** 3770–3778 (1966).

153. Williams, D. E. & Cox, S. R. Nonbonded potentials for azahydrocarbons: the importance of the Coulombic interaction. *Acta Crystallographica Section B* **40,** 404–417. ISSN: 0108-7681 (1984).

154. Williams, D. E. & Houpt, D. J. Fluorine nonbonded potential parameters derived from crystalline perfluorocarbons. *Acta Crystallographica Section B* **42,** 286–295. ISSN: 0108-7681 (1986).

155. Williams, D. E. & Hsu, L.-Y. Transferability of nonbonded Cl...Cl potential energy function to crystalline chlorine. *Acta Crystallographica Section A* **41,** 296–301. ISSN: 0108-7673 (1985).

156. Williams, D. E. Improved intermolecular force field for crystalline oxohydrocarbons including O-H...O hydrogen bonding. *Journal of Computational Chemistry* **22,** 1–20. ISSN: 1096-987X (2001).

157. Williams, D. E. Improved intermolecular force field for molecules containing H, C, N, and O atoms, with application to nucleoside and peptide crystals. *Journal of Computational Chemistry* **22,** 1154–1166. ISSN: 1096-987X (2001).

158. Williams, D. E. Improved intermolecular force field for crystalline hydrocarbons containing four- or three-coordinated carbon. *Journal of Molecular Structure* **485-486,** 321–347. ISSN: 0022-2860 (1999).

159. Ohare, B. *Phonons* Wikipedia. <`http://en.wikipedia.org/wiki/Phonon`> (2012).

160. Born, M. & Huang, K. *Dynamical Theory of Crystal Lattices* (Oxford Univ. Press, 2002).

161. Ochterski, J. W. *Thermochemistry in Gaussian* Gaussian, Inc. 2000.

162. Price, D. J. & Brooks, C. L. A modified TIP3P water potential for simulation with Ewald summation. *The Journal of Chemical Physics* **121,** 10096–10103 (2004).

163. Berendsen, H. J. C., Grigera, J. R. & Straatsma, T. P. The missing term in effective pair potentials. *The Journal of Physical Chemistry* **91,** 6269–6271 (1987).

164. Mennucci, B., Cammi, R. & Interscience, W. *Continuum solvation models in chemical physics: from theory to applications* (Wiley Online Library, 2007).

165. Tomasi, J., Mennucci, B. & Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chemical Reviews* **105,** 2999–3094. ISSN: 0009-2665 (2005).

166. Cancès, E., Mennucci, B. & Tomasi, J. A new integral equation formalism for the polarizable continuum model: Theoretical background and applications to isotropic and anisotropic dielectrics. *The Journal of Chemical Physics* **107,** 3032–3041 (1997).

167. Pierotti, R. A. A scaled particle theory of aqueous and nonaqueous solutions. *Chemical Reviews* **76,** 717–726. ISSN: 0009-2665 (1976).

168. Zipse, H. *The Polarizable Continuum Model (PCM)* LMU.

169. Kamerlin, S. C. L., Haranczyk, M. & Warshel, A. Are Mixed Explicit/Implicit Solvation Models Reliable for Studying Phosphate Hydrolysis? A Comparative Study of Continuum, Explicit and Mixed Solvation Models. *ChemPhysChem* **10,** 1125–1134. ISSN: 1439-7641 (2009).

170. Bandyopadhyay, P. & Gordon, M. S. A combined discrete/continuum solvation model: Application to glycine. *The Journal of Chemical Physics* **113,** 1104–1109 (2000).

171. Pratt, L. R. & Chandler, D. Theory of the hydrophobic effect. *The Journal of Chemical Physics* **67,** 3683–3704 (1977).

172. Kovalenko, A. & Hirata, F. Self-consistent description of a metal-water interface by the Kohn-Sham density functional theory and the three-dimensional reference interaction site model. *The Journal of Chemical Physics* **110,** 10095–10112 (1999).

173. Gavezzotti, A. & Filippini, G. *Theoretical Aspects and Computer Modeling* (ed Gavezzotti, A.) 61–97 (Wiley and Sons, Chichester, 1997).

174. Day, G. M., Price, S. L. & Leslie, M. Atomistic Calculations of Phonon Frequencies and Thermodynamic Quantities for Crystals of Rigid Organic Molecules. *The Journal of Physical Chemistry B* **107,** 10919–10933. ISSN: 1520-6106 (2003).

175. Jensen, J. H., Li, H., Robertson, A. D. & Molina, P. A. Prediction and Rationalization of Protein pKa Values Using QM and QM/MM Methods. *The Journal of Physical Chemistry A* **109,** 6634–6643 (2005).

176. Gilson, M. K. & Zhou, H.-X. Calculation of Protein-Ligand Binding Affinities*. *Annual Review of Biophysics and Biomolecular Structure* **36.** PMID: 17201676, 21–42 (2007).

177. Geballe, M., Skillman, A., Nicholls, A., Guthrie, J. & Taylor, P. The SAMPL2 blind prediction challenge: introduction and overview. English. *Journal of Computer-Aided Molecular Design* **24,** 259–279. ISSN: 0920-654X (2010).

178. Nicholls, A. *et al.* Predicting Small-Molecule Solvation Free Energies: An Informal Blind Test for Computational Chemistry. *Journal of Medicinal Chemistry* **51,** 769–779 (2008).

179. USEPA. *Estimation Programs Interface Suite for Microsoft Windows v 4.10* United States Environmental Protection Agency (USEPA). <`http://www.epa.gov/opptintr/exposure/pubs/episuitedl.htm`> (2011).

180. Bergström, C. A. S., Wassvik, C. M., Norinder, U., Luthman, K. & Artursson, P. Global and Local Computational Models for Aqueous Solubility Prediction of Drug-Like Molecules. *Journal of Chemical Information and Computer Sciences* **44,** 1477–1488. ISSN: 0095-2338 (2004).

181. Palmer, D. S., Frolov, A. I., Ratkova, E. L. & Fedorov, M. V. Toward a Universal Model To Calculate the Solvation Thermodynamics of Druglike Molecules: The Importance of New Experimental Databases. *Molecular Pharmaceutics* **8,** 1423–1429. ISSN: 1543-8384 (2011).

182. Frolov, A. I., Ratkova, E. L., Palmer, D. S. & Fedorov, M. V. Hydration Thermodynamics Using the Reference Interaction Site Model: Speed or Accuracy? *The Journal of Physical Chemistry B* **115,** 6011–6022. ISSN: 1520-6106 (2011).

183. Frisch, M. J. *et al. Gaussian 09* Gaussian, Inc (Wallingford CT, 2009).

184. Viswanadhan, V. N., Ghose, A. K., Singh, U. C. & Wendoloski, J. J. Prediction of solvation free energies of small organic molecules: Additive-constitutive models based on molecular fingerprints and atomic constants. *Journal of chemical information and computer sciences* **39,** 405–412 (1999).

185. Guthrie, J. P. & Povar, I. A test of various computational solvation models on a set of "difficult" organic compounds. *Canadian Journal of Chemistry-Revue Canadienne De Chimie* **87,** 1154–1162. ISSN: 0008-4042 (Aug. 2009).

186. Ho, J., Klamt, A. & Coote, M. L. Comment on the Correct Use of Continuum Solvent Models. *The Journal of Physical Chemistry A* **114,** 13442–13444. ISSN: 1089-5639 (2010).

187. Chandler, D., Singh, Y. & Richardson, D. M. Excess electrons in simple fluids. I. General equilibrium theory for classical hard sphere solvents. *The Journal of Chemical Physics* **81,** 1975–1982 (1984).

188. Harano, Y., Imai, T., Kovalenko, A., Kinoshita, M. & Hirata, F. Theoretical study for partial molar volume of amino acids and polypeptides by the three-dimensional reference interaction site model. *The Journal of Chemical Physics* **114,** 9506–9511 (2001).

189. Ratkova, E. L., Chuev, G. N., Sergiievskyi, V. P. & Fedorov, M. V. An Accurate Prediction of Hydration Free Energies by Combination of Molecular Integral Equations Theory with Structural Descriptors. *The Journal of Physical Chemistry B* **114,** 12068–12079. ISSN: 1520-6106 (2010).

190. Marenich, A. V., Cramer, C. J. & Truhlar, D. G. Performance of SM6, SM8, and SMD on the SAMPL1 Test Set for the Prediction of Small-Molecule Solvation Free Energies. *The Journal of Physical Chemistry B* **113,** 4538–4543. ISSN: 1520-6106 (2009).

191. Ribeiro, R., Marenich, A., Cramer, C. & Truhlar, D. Prediction of SAMPL2 aqueous solvation free energies and tautomeric ratios using the SM8, SM8AD, and SMD solvation models. *Journal of Computer-Aided Molecular Design* **24,** 317–333. ISSN: 0920-654X (2010).

192. Guthrie, J. P. A Blind Challenge for Computational Solvation Free Energies: Introduction and Overview. *The Journal of Physical Chemistry B* **113,** 4501–4507. ISSN: 1520-6106 (2009).

193. Arrowsmith, J. A decade of change. *Nature Reviews Drug Discovery* **11,** 17–18 (2012).

194. Hopfinger, A. J., Esposito, E. X., LlinÃ s, A., Glen, R. C. & Goodman, J. M. Findings of the Challenge To Predict Aqueous Solubility. *Journal of Chemical Information and Modeling* **49,** 1–5 (2009).

195. Shareef, A., Angove, M. J., Wells, J. D. & Johnson, B. B. Aqueous Solubilities of Estrone, 17β-Estradiol, 17α-Ethynylestradiol, and Bisphenol A. *Journal of Chemical & Engineering Data* **51,** 879–881. ISSN: 0021-9568 (2006).

196. Hughes, L. D., Palmer, D. S., Nigsch, F. & Mitchell, J. B. O. Why Are Some Properties More Difficult To Predict than Others? A Study of QSPR Models of Solubility, Melting Point, and Log P. *Journal of Chemical Information and Modeling* **48,** 220–232. ISSN: 1549-9596 (2008).

197. Tetko, I. V. Computing chemistry on the web. *Drug Discovery Today* **10,** 1497–1500. ISSN: 1359-6446 (2005).

198. Tetko, I. V. *et al.* Virtual computational chemistry laboratory - design and description. *Journal of Computer-Aided Molecular Design* **19,** 453–63 (2005).

199. VCCLAB, V. C. C. L. *Virtual Computational Chemistry Laboratory* 2005. <http://www.vcclab.org>.

200. Warner, D. J., Griffen, E. J. & St-Gallay, S. A. WizePairZ: A Novel Algorithm to Identify, Encode, and Exploit Matched Molecular Pairs with Unspecified Cores in Medicinal Chemistry. *Journal of Chemical Information and Modeling* **50,** 1350–1357 (2010).

201. Stuart, M. & Box, K. Chasing equilibrium: measuring the intrinsic solubility of weak acids and bases. *Analytical chemistry* **77,** 983–990 (2005).

202. Narasimham, L. & Barhate, V. D. Kinetic and intrinsic solubility determination of some β-blockers and antidiabetics by potentiometry. *Journal of Pharmacy Research* **4,** 532–536 (2011).

203. Anderson, E., Veith, G. D. & Weininger, D. *SMILES, a line notation and computerized interpreter for chemical structures* (US Environmental Protection Agency, Environmental Research Laboratory, 1987).

204. O'Boyle, N. Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI. *Journal of Cheminformatics* **4,** 22. ISSN: 1758-2946 (2012).

205. RSC. *ChemSpider* <`http://www.chemspider.com/`>.

206. Brumfiel, G. Chemists spin a web of data. *Nature* **453,** 139 (2008).

207. Pence, H. E. & Williams, A. ChemSpider: an online chemical information resource. *Journal of Chemical Education* **87,** 1123–1124 (2010).

208. Rovner, S. L. *RSC Acquires Chemical Database* 2009.

209. Wang, R., Gao, Y. & Lai, L. Calculating partition coefficient by atom-additive method. *Perspectives in Drug Discovery and Design* **19,** 47–66 (2000).

210. Kier, L. *Molecular connectivity in chemistry and drug research* (Elsevier, 2012).

211. Moreau, G. & Broto, P. The auto-correlation of a topological-structure-a new Molecular Descriptor. *Nouveau Journal De Chimie-New Journal of Chemistry* **4,** 359–360 (1980).

212. Randic, M. On molecular identification numbers. *Journal of Chemical Information and Computer Sciences* **24,** 164–175 (1984).

213. Hewitt, M. *et al.* In Silico Prediction of Aqueous Solubility: The Solubility Challenge. *Journal of Chemical Information and Modeling* **49,** 2572–2587. ISSN: 1549-9596 (2009/11/23 2009).

214. Irmann, F. Eine einfache Korrelation zwischen Wasserlöslichkeit und Struktur von Kohlenwasserstoffen und Halogenkohlenwasserstoffen. *Chemie Ingenieur Technik* **37,** 789–798 (1965).

215. Ran, Y. & Yalkowsky, S. H. Prediction of Drug Solubility by the General Solubility Equation (GSE). *Journal of Chemical Information and Computer Sciences* **41,** 354–357. ISSN: 0095-2338 (2001).

216. Ran, Y., Jain, N. & Yalkowsky, S. H. Prediction of Aqueous Solubility of Organic Compounds by the General Solubility Equation (GSE). *Journal of Chemical Information and Computer Sciences* **41,** 1208–1217 (2001).

217. Bradley, J. C., Lang, A. & Williams, A. *Open Melting Point Data* <`http://lxsrv7.oru.edu/~alang/meltingpoints/download.php`>.

218. Ghose, A. K., Viswanadhan, V. N. & Wendoloski, J. J. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods:â€‰ An Analysis of ALOGP and CLOGP Methods. *The Journal of Physical Chemistry A* **102,** 3762–3772 (1998).

219. Karthikeyan, M., Glen, R. C. & Bender, A. General Melting Point Prediction Based on a Diverse Compound Data Set and Artificial Neural Networks. *Journal of Chemical Information and Modeling* **45,** 581–590 (2005).

220. Bergström, C. A. S., Norinder, U., Luthman, K. & Artursson, P. Molecular Descriptors Influencing Melting Point and Their Role in Classification of Solid Drugs. *Journal of Chemical Information and Computer Sciences* **43.** PMID: 12870909, 1177–1185 (2003).

221. Nigsch, F. *et al.* Melting Point Prediction Employing k-Nearest Neighbor Algorithms and Genetic Parameter Optimization. *Journal of Chemical Information and Modeling* **46,** 2412–2422 (2006).

222. Yang, G., Ran, Y. & Yalkowsky, S. H. Prediction of the aqueous solubility: Comparison of the general solubility equation and the method using an amended solvation energy relationship. *Journal of Pharmaceutical Sciences* **91,** 517–533. ISSN: 1520-6017 (2002).

223. Perlovich, G. L., Volkova, T. V. & Bauer-Brandl, A. Towards an understanding of the molecular mechanism of solvation of drug molecules: A thermodynamic approach by crystal lattice energy, sublimation, and solubility exemplified by paracetamol, acetanilide, and phenacetin. *Journal of Pharmaceutical Sciences* **95,** 2158–2169. ISSN: 1520-6017 (2006).

224. Perlovich, G. L., Kurkov, S. V., Hansen, L. K. & Bauer-Brandl, A. Thermodynamics of sublimation, crystal lattice energies, and crystal structures of racemates and enantiomers: (+)- and (±)-ibuprofen. *Journal of Pharmaceutical Sciences* **93,** 654–666. ISSN: 1520-6017 (2004).

225. Da Silva, M. A. R., Santos, L. M. & Lima, L. M. S. S. Standard molar enthalpies of formation and of sublimation of the terphenyl isomers. *The Journal of Chemical Thermodynamics* **40,** 375–385. ISSN: 0021-9614 (2008).

226. Monte, M. J. S. & Sousa, C. A. D. Vapor Pressures and Phase Changes Enthalpy and Gibbs Energy of Three Crystalline Monomethyl Benzenedicarboxylates. *Journal of Chemical & Engineering Data* **50,** 2101–2105. ISSN: 0021-9568 (2005).

227. Vecchio, S. & Brunetti, B. Vapor pressures and standard molar enthalpies, entropies, and Gibbs free energies of sublimation of 2,4- and 3,4-dinitrobenzoic acids. *The Journal of Chemical Thermodynamics* **41,** 880–887. ISSN: 0021-9614 (2009).

228. Perlovich, G. L., Kurkov, S. V., Kinchin, A. N. & Bauer-Brandl, A. Thermodynamics of solutions IV: Solvation of ketoprofen in comparison with other NSAIDs. *Journal of Pharmaceutical Sciences* **92,** 2502–2511. ISSN: 1520-6017 (2003).

229. Perlovich, G. L., Volkova, T. V. & Bauer-Brandl, A. Thermodynamic Study of Sublimation, Solubility, Solvation, and Distribution Processes of Atenolol and Pindolol. *Molecular Pharmaceutics* **4,** 929–935. ISSN: 1543-8384 (2007).

230. Manuel, A. V. *et al.* Experimental thermochemical study of 2,5- and 2,6-dichloro-4-nitroanilines. *The Journal of Chemical Thermodynamics* **41,** 1074–1080. ISSN: 0021-9614 (2009).

231. Vecchio, S. & Brunetti, B. Standard Sublimation Enthalpies of Some Dichlorophenoxy Acids and Their Methyl Esters. *Journal of Chemical & Engineering Data* **50,** 666–672. ISSN: 0021-9568 (2005/03/01 2005).

232. Lima, F. R. A. C. Carlos *et al.* Phenylnaphthalenes: Sublimation Equilibrium, Conjugation, and Aromatic Interactions. *The Journal of Physical Chemistry B* **116,** 3557–3570. ISSN: 1520-6106 (2012/03/22 2012).

233. Kurkov, S. V. & Perlovich, G. L. Thermodynamic studies of Fenbufen, Diflunisal, and Flurbiprofen: Sublimation, solution and solvation of biphenyl substituted drugs. *International Journal of Pharmaceutics* **357,** 100–107. ISSN: 0378-5173 (2008).

234. Perlovich, G. L., Surov, A. O. & Bauer-Brandl, A. Thermodynamic properties of flufenamic and niflumic acids–Specific and non-specific interactions in solution and in crystal lattices, mechanism of solvation, partitioning and distribution. *Journal of Pharmaceutical and Biomedical Analysis* **45,** 679–687. ISSN: 0731-7085 (2007).

235. Cundall, R. B., Frank Palmer, T. & Wood, C. E. C. Vapour pressure measurements on some organic high explosives. *Journal of the Chemical Society, Faraday Transactions 1: Physical Chemistry in Condensed Phases* **74,** 1339–1345. ISSN: 0300-9599 (1978).

236. Perlovich, G. L. *et al.* Thermodynamic and structural aspects of sulfonamide crystals and solutions. *Journal of Pharmaceutical Sciences* **98,** 4738–4755. ISSN: 1520-6017 (2009).

237. Perlovich, G., Volkova, T., Manin, A. & Bauer-Brandl, A. Influence of Position and Size of Substituents on the Mechanism of Partitioning: A Thermodynamic Study on Acetaminophens, Hydroxybenzoic Acids, and Parabens. *AAPS PharmSciTech* **9,** 205–216. ISSN: 1530-9932 (2008).

238. Monte, M., Santos, L., Fonseca, J. & Sousa, C. Vapour pressures, enthalpies and entropies of sublimation of para substituted benzoic acids. *Journal of Thermal Analysis and Calorimetry* **100,** 465–474. ISSN: 1388-6150 (2010).

239. Surov, A. O., Terekhova, I. V., Bauer-Brandl, A. & Perlovich, G. L. Thermodynamic and Structural Aspects of Some Fenamate Molecular Crystals. *Crystal Growth & Design* **9,** 3265–3272. ISSN: 1528-7483 (2009).

240. Ribeiro da Silva, M. A. V., Amaral, L. M. P. F., Santos, A. F. L. O. M. & Gomes, J. R. B. Thermochemistry of some alkylsubstituted anthracenes. *The Journal of Chemical Thermodynamics* **38,** 367–375. ISSN: 0021-9614 (2006).

241. Da Silva, R., Manuel, A. V., Monte, J. S. Manuel & Ribeiro, R. José. Thermodynamic study on the sublimation of succinic acid and of methyl- and dimethyl-substituted succinic and glutaric acids. *The Journal of Chemical Thermodynamics* **33,** 23–31. ISSN: 0021-9614 (2001).

242. Monte, J. S. Manuel, Almeida, R. R. P. Ana & Ribeiro da Silva, A. V. Manuel. Thermodynamic study of the sublimation of eight 4-n-alkylbenzoic acids. *The Journal of Chemical Thermodynamics* **36,** 385–392. ISSN: 0021-9614 (2004).

243. Ribeiro da Silva, M. A. V., Lima, S. Luís M. Spencer, Moreno, A. R. G., Ferreira, A. I. M. C. L. & Gomes, J. R. B. Combined experimental and computational thermochemistry of isomers of chloronitroanilines. *The Journal of Chemical Thermodynamics* **40,** 155–165. ISSN: 0021-9614 (2008).

244. Perlovich, G. L., Volkova, T. V. & Bauer-Brandl, A. Towards an understanding of the molecular mechanism of solvation of drug molecules: A thermodynamic approach by crystal lattice energy, sublimation, and solubility exemplified by hydroxybenzoic acids. *Journal of Pharmaceutical Sciences* **95,** 1448–1458. ISSN: 1520-6017 (2006).

245. Perlovich, G. L., Surov, A. O., Hansen, L. K. & Bauer-Brandl, A. Energetic aspects of diclofenac acid in crystal modifications and in solutions—mechanism of solvation, partitioning and distribution. *Journal of pharmaceutical sciences* **96,** 1031–1042 (2007).

246. Ribeiro da Silva, A. V. Manuel, Monte, J. S. Manuel & Santos, M. N. B. F. Luís. The design, construction, and testing of a new Knudsen effusion apparatus. *The Journal of Chemical Thermodynamics* **38,** 778–787. ISSN: 0021-9614 (2006).

247. Perlovich, G. L., Ryzhakov, A. M., Tkachev, V. V. & Hansen, L. K. Sulfonamide Molecular Crystals: Thermodynamic and Structural Aspects. *Crystal Growth & Design* **11,** 1067–1081. ISSN: 1528-7483 (2011).

248. Perlovich, G. L. *et al.* Thermodynamic and Structural Aspects of Hydrated and Unhydrated Phases of 4-Hydroxybenzamide. *Crystal Growth & Design* **7,** 2643–2648. ISSN: 1528-7483 (2007).

249. Clark, S. J. *et al.* First principles methods using CASTEP. *Zeitschrift für Kristallographie* **220,** 567–570. ISSN: 0044-2968 (2005).

250. Reilly, A. M. & Tkatchenko, A. Seamless and Accurate Modeling of Organic Molecular Materials. *The Journal of Physical Chemistry Letters* **4,** 1028–1033. ISSN: 1948-7185 (2013).

251. Truchon, J.-F., Pettitt, B. M. & Labute, P. A Cavity Corrected 3D-RISM Functional for Accurate Solvation Free Energies. *Journal of Chemical Theory and Computation.* ISSN: 1549-9618. doi:10.1021/ct4009359 (2014).

252. Sergiievskyi, V. P., Jeanmairet, G., Levesque, M. & Borgis, D. Fast Computation of Solvation Free Energies with Molecular Density Functional Theory: Thermodynamic-Ensemble Partial Molar Volume Corrections. *The Journal of Physical Chemistry Letters* **5,** 1935–1942 (2014).

253. MedChemica. *SALT knowledge Sharing* MedChemica. <http://www.medchemica.com/services.html> (2014).

254. Green, D. *Challenges and Perspectives on Computational Chemistry & Informatics in Drug Discovery* GSK. <https://www.ebi.ac.uk/training/sites/ebi.ac.uk.training/files/materials/2013/131209DrugDiscovery/2_-_darren_green_-_challenges_and_perspectives_on_computational_chemistry_informatics.pdf> (2014).

255. RSC. *CSID:1999: Asmol* RSC Chemspider. <http://www.chemspider.com/Chemical-Structure.1999.html> (2014).

256. IUPAC. *The IUPAC International Chemical Identifier (InChI)* <http://www.iupac.org/home/publications/e-resources/inchi.html>.

257. Chis, V. *Atomic units; Molecular Hamiltonian; Born-Oppenheimer approximation* Babes-Bolyai University. <http://www.phys.ubbcluj.ro/~vchis/cursuri/cspm/course2.pdf> (2014).

258. Physics-Forums. *Orthogonality and orthonormality ?* online website. 2011. <http://www.physicsforums.com/showthread.php?t=473005>.

259. Hirst, D. *Mathmatics for Chemists* (The Macmillan Press Ltd, 1978).

260. Holzner, S. *Quantum physics for dummies* (John Wiley & Sons, 2009).