Modelling approaches for rational solvent selection in drug development: enhancing the solubility prediction of small molecules

A thesis submitted to the University of Strathclyde for the degree of Doctor of Philosophy

by

Bruce James Wareham

2019

Strathclyde Institute of Pharmacy and Biomedical Sciences

University of Strathclyde

Glasgow

Declaration of authenticity and author's rights

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Signed:

Date:

Acknowledgements

I would like to express my thanks to my primary supervisor Dr Blair Johnston who has always given me his time when I have asked for it. He gave me the enormous freedom and trust to pursue my project independently and for that I am extremely thankful. I would also like to thank my second supervisor Prof. Alastair Florence.

I would like to thank EPSRC (Grant Ref: EP/K503289/1) for funding this work.

I would also like to thank Dr. Murray Robertson who had the misfortune of putting up with all my incessant questions. I'd also like to thank Dr. Antony Vassileiou for all the times I have annoyed him as I thoroughly enjoyed annoying him.

For supporting my fragile ego I'd like to thank Carlota Mendez, Sara Ottoboni, Stephanie Yerdelen, Alice Turner, Sarahjane Wood, Arabella McLaughlin, Michael Chrubasik, Carla Ferreira, Bilal Ahmed, Václav Svoboda and so many others that I could have mentioned but I can't spell their names correctly.

I'd also like to thanks my best friends' alcohol, coffee and chocolate for all the good times and seeing me through the bad times.

Last but certainly not least I would like to thank my family without whose help and support I would have been unable to complete this work.

Abstract

In the pharmaceutical industry crystallisation is the preferred method used to control crystal size and particle growth, and to purify the final product. It can also be used for the isolation of any impurities. Using the minimum amount of material is desirable. Therefore, the ability to predict solubility and to construct accurate and robust models for complex molecules has been of ongoing interest in the pharmaceutical industry. There are several methods to predict solubility *in silico,* including: COSMO*therm,* NRTL-SAC, UNIFAC, and SAFT-γ Mie. This work will focus on the use and appraisal of *ab initio* method COSMO*therm* and the application of a "correction factor" using the machine learning algorithm random forest to improve accuracy of predictions.

Chapter Two compares experimental data with COSMO*therm* to assess the robustness and reliability of the method. The influence of adjustable parameters required for predictions: enthalpy of fusion, and melting temperature, were assessed. These studies detail the importance of accurate measurements of these parameters and how deviations from their true value can affect the accuracy of solubility predictions.

Chapter Three details the building of a linear regression model using a design of experiment approach for almost instantaneous predictions using no specialised software, for non-experts and modellers.

Chapter Four details the building of machine learning models using random forest to apply a correction factor to the error between COSMO*therm* and experimental data.

Chapter Five uses predictive methods and a workflow approach to select crystallisation and wash solvents. A case study using paracetamol and its impurities is considered.

Abbreviations

API	Active pharmaceutical ingredient
BCUT	Burden eigenvalues
CDK	Chemistry development kit
CMAC	Continuous Manufacturing and Advanced Crystallisation
COSMO-RS	Conductor like screening model for real solvents
DFT	Density functional theory
DSC	Differential scanning calorimetry
EPSRC	Engineering and Physical Sciences Research Council
GSK	GlaxoSmithKline
GTO	Gaussian type orbitals
GUI	Graphical user interface
HPLC	High-performance liquid chromatography
HPLC-MS	High-performance liquid chromatography mass spectrometry
ICH	International conference on harmonisation
LCMS	Liquid chromatography with mass spectrometer
LLE	Liquid-liquid equilibrium
MOE	Molecular operating environment
NA	Not available
NMR	Nuclear magnetic resonance
NRTL-SAC	Non-random two-liquid segment activity co-efficient
OFAT	One factor at a time
OOB	Out of bag
PCA	Principle component analysis
PCM	Paracetamol
PC-SAFT	Perturbed chain statistical associating fluid theory
PEOE	Partial equalisation of orbital electronegativity descriptors
PSD	Particle size distribution
QuaSAR-Descriptors	Quality structure-activity relationship descriptors
QZVP	Quadruple zeta valence polarisation
RF	Random forest
RMSE	Root mean squared error
SAFT-γ Mie	Statistical associating fluid theory-gamma Mie
SD	Standard deviation
SLE	Solid-liquid equilibrium
SLESOL	Solid-liquid equilibrium solubility
SMILES	Simplified molecular-input line-entry system
SSE	Sum of squared errors
STO	Slater-type orbitals
SVM	Support vector machines
SVP	Split valence polarisation
TZVP	Triple-zeta valence polarisation
TZVPD-fine	Triple-zeta valance polarisation with diffuse functions

UK	United Kingdom
UNIFAC	Universal quasichemical functional group activity coefficients
vdW	Van der Waals
VLE	Vapour-liquid equilibrium
XRPD	X-ray powder diffraction

Table of contents

Contents

Table of figures

Table of tables

1 Introduction

1.1 CMAC (Continuous Manufacturing and Advanced Crystallisation)

The EPSRC Future Manufacturing Research Hub in Continuous Manufacturing and Advanced Crystallisation (CMAC) is a hub for medicines manufacturing, research and training. The CMAC Hub is located at the University of Strathclyde and with a multidisciplinary and collaborative academic team at the UK Universities of Bath, Cambridge, Imperial, Leeds, Loughborough and Sheffield. Established in 2011, it comprises of more than 130 staff and researchers, including more than 45 PhD students. CMAC has a close collaboration with industry and has the support of tier one partners, which include GlaxoSmithKline (GSK), AstraZeneca, Bayer, Lilly, Novartis, Pfizer, Roche and Takeda.

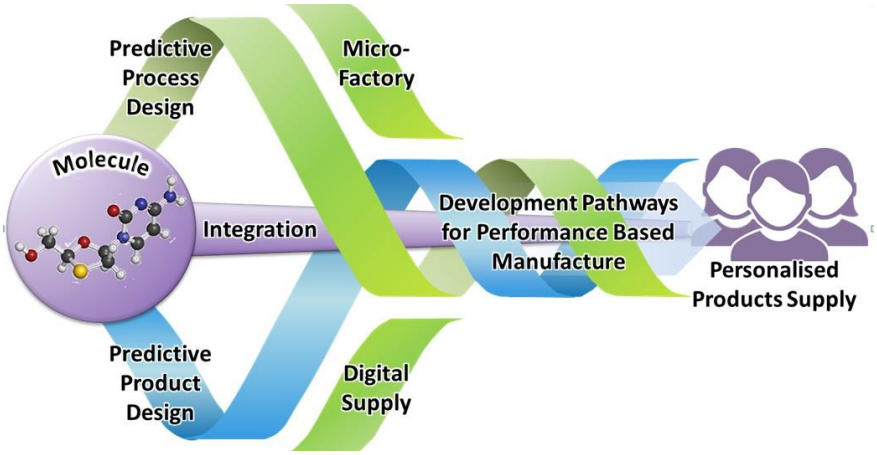


Figure 1-1 CMAC Future Manufacturing Research Hub areas of research and development

One of CMAC's research aims is to deliver new predictive tools and design approaches for drug products, processes and supply chains (Figure 1-1). This thesis will investigate predictive tool development and the use of machine learning (ML) to enhance *ab initio* and related modelling methodologies. The outputs will then be incorporated in digital workflows for speed and ease-of-use by non-experts.

1.2 Overview

Within the pharmaceutical industry, crystallisation is the preferred method used to control crystal size and particle growth, and to purify the final product. It can also be used for the isolation of any impurities encountered in the production process. Using the minimum amount of chemical material in the laboratory is desirable; predictive methods can help achieve this in many cases. As with any new product development, the availability of drug product for laboratory use in solubility studies is often restricted due to the high cost of manufacture and the competing demand for drug substance from formulation teams and for clinical trials. The thermodynamic driving force behind crystallisation is solubility. The solubility of any given substance in a solvent depends on the physical and chemical properties of the solute and solvent as well as on temperature, pH of the solution and atmospheric pressure. Pharmaceutical molecules can be complex with various functional groups exhibiting a range of intra- and inter-molecular interactions. The ability to predict solubility and to construct accurate and robust models for these complex molecules has been of particular ongoing interest throughout the pharmaceutical industry (Benazzouz *et al.*, 2014). To overcome the challenges of finding a reliable method for identifying suitable solvent systems for crystallisation, filtration and washing, several methods have been used; each with their own particular advantages and disadvantages. There are sophisticated quantum chemical computational methods such as non-random two-liquid segment activity co-efficient (NRTL-SAC) (Song, 2004) and universal quasichemical functional group activity coefficients (UNIFAC) which have been used for many years (Fredenslund,Jones and Prausnitz, 1975). Perturbed chain statistical

associating fluid theory (PC-SAFT) (Gross and Sadowski, 2001) has been available since the 1970s and conductor like screening model for real solvents (COSMO-RS) (F. Eckert and Klamt, 2002) has been used since the 1990s.

UNIFAC is a group-contribution method and uses parameters obtained from data reduction, enabling the activity co-efficient to be predicted with accuracy. Initially, the methodological objective was the prediction of vapour-liquid equilibria (VLE) which describes the distribution of chemical species between the vapour and liquid phase. This was expanded to include liquid-liquid equilibria (LLE), which is the distribution of a component in two liquid phases, and solid-liquid equilibria (SLE). SLE describe the distribution of a solid in a liquid. Group contribution methods assume that the mixture does not consist of molecules but a collection of functional groups such as aromatic CH and COOH (see Table 1-1).

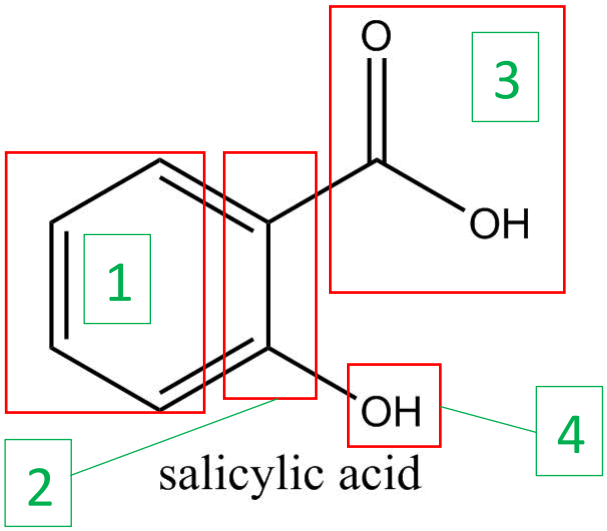


Figure 1-2 salicylic acid broken down into functional groups

Each molecule is divided into main groups and sub-groups (Figure 1-2 and Table 1-1). This has the advantage of there being a much smaller group of functional groups than there are compounds.

Table 1-1 UNIFAC functional groups for salicylic acid

Functional group	Main group name	Sub-group
1	Aromatic carbon	4 ACH
2	Aromatic carbon	2 AC
3	COOH	1 COOH
4	Aromatic carbon-alcohol	1 ACOH

In this method the activity co-efficient is calculated by a combinational part and a residual part with the combinational part taking into consideration parameters for area and volume and the residual part for binary interactions (Gmehling, 1998). This method is thought to be more accurate for smaller molecules but has trouble with larger molecules with more complex structures (C.C. Chen, 1993). One of the major drawbacks of this method is that if there are no data pertaining to a particular functional group, the method has insufficient information to be able to parameterise the molecule being studied, which then leads to unreliable results. This is of particular concern in the pharmaceutical industry as active pharmaceutical ingredients (API) can have functional groups that have no UNIFAC parameters and drugs can be, in some cases, large and complex molecules. There have been further modifications to UNIFAC designed by Gmehling (called modified UNIFAC(Do)) to tackle perceived weaknesses in the original method (Xue,Mu and Gmehling, 2012), such as the accuracy of the method across a particular concentration range. This method has resulted in a change to some of the equations for the combinational part with an additional temperature parameter for the group interaction term and the addition of two new main groups; cyclic amides and aromatic compounds containing sulphur (Gmehling, 1998). Temperature dependent group-interaction parameters were

introduced to modified UNIFAC(Do) to allow a better description of real behaviour over a wide temperature range; this has the drawback of requiring more parameters for each group. Generally, this method is not applicable to electrolytes.

NRTL-SAC is based on the polymer NRTL model (C.C. Chen, 1993) which itself is a derivation of the original NRTL method of Renon and Prausnitz (Renon and Prausnitz, 1968). NRTL-SAC is used to predict the physical behaviours of non-ideal systems. It is similar in many respects to UNIFAC as it uses experimental data to identify molecular parameters for solutes and these data are used to extrapolate to other solvent systems (C.C. Chen and Song, 2004).

Similar to UNIFAC, it uses a combinational and residual contribution to predict the activity co-efficient. Rather than using parameters for functional groups, this method uses a contribution from conceptual segments: those being hydrophobic (X); polar (Y+,Y-); and hydrophilic (Z). The hydrophilic segment represents the region of a molecule that has the characteristics of a hydrogen-bond donor or acceptor. The hydrophobic segment represents the region of a molecule that is unfavourable to hydrogen bonding. The polar segment is divided into two parts: polar-attractive is a segment, which shows attraction with a hydrophilic molecular surface; and the polar-repulsive segment, which exhibits repulsion with a hydrophilic molecular surface. NRTL-SAC differs from UNIFAC as it maps molecules into a few predefined conceptual segments where UNIFAC has a large set of predefined functional groups based on the chemical structure.

PC-SAFT (Gross and Sadowski, 2001) and statistical associating fluid theory-gamma Mie (SAFT-γ Mie) (Papaioannou *et al.*, 2014) are molecular based equation of state

models, which take into consideration the effect of hard chain (repulsive forces), shape dispersion (van der Waal interactions (vdW)) and association forces (hydrogen bonds). The theory is based on introducing a defined model fluid as a reference system then modelling a real fluid by adding perturbations to the reference system (Gross and Sadowski, 2001, Beret and Prausnitz, 1975). SAFT-γ Mie is studied in greater detail in section 1.3.7. The major drawback of NRTL-SAC, UNIFAC and PC-SAFT is that all three need a great deal of empirical data to establish parameters for their predictions.

1.3 Methods

Several computational methods have been used in this work to predict solubility and these are described in detail in the following sections. These include COSMO-RS theory, which was used for most of the solubility predictions in this thesis, and UNIFAC and SAFT-γ Mie, which were used for comparison of methods in section 2.6.2. Other modelling approaches discussed include random forest (RF), a ML method that was used in this thesis to enhance the predictive accuracy of existing solubility prediction approaches using descriptors from modelling software, Molecular Operating Environment (MOE). Additionally, the Joback and Reid method for the prediction of enthalpy of fusion and melting temperature (Joback and Reid, 1987), the Jain and Yalkowsky method for predicting enthalpy of fusion (Jain,Yang and Yalkowsky, 2004) and the COSMO*quick* linear regression method are also considered (Loschen and Klamt, 2012).

1.3.1 Equipment

For this project, a Dell Precision T7810 was used with two Intel® Xeon® CPU E5-2637 v3 @ 3.50 GHz and 32 GB RAM.

1.3.2 COSMO-RS

COSMO-RS is a quantum chemical method that predicts chemical potential in liquids. It brings together statistical thermodynamics methodology with an electrostatic theory of locally interacting molecular surface descriptors.

COSMO-RS is an *ab initio* method initially developed for the prediction of the thermophysical properties of liquids by Andreas Klamt and Frank Eckert in the 1990s (Klamt and Eckert, 2000). As it needs little experimental data to perform predictions, unlike NRTL-SAC and UNIFAC. It has a greater applicability than other models, although some accuracy may be lost as the other methods use fitted data. COSMO-RS takes a screening charge density (σ) from a molecule, which provides a discrete surface around a molecule in a virtual conductor. COSMO-RS has a greater emphasis on intermolecular interactions than the group contribution methods with hydrogen bonding and vdW interactions accounted for. The method also takes into consideration intramolecular hydrogen bonding which UNIFAC does not. COSMO*therm* is the computer application developed by Klamt and Eckert, which uses COSMO-RS theory. NRTL-SAC and UNIFAC do not explicitly consider the effects of multiple molecular conformers whereas COSMO*therm* does.

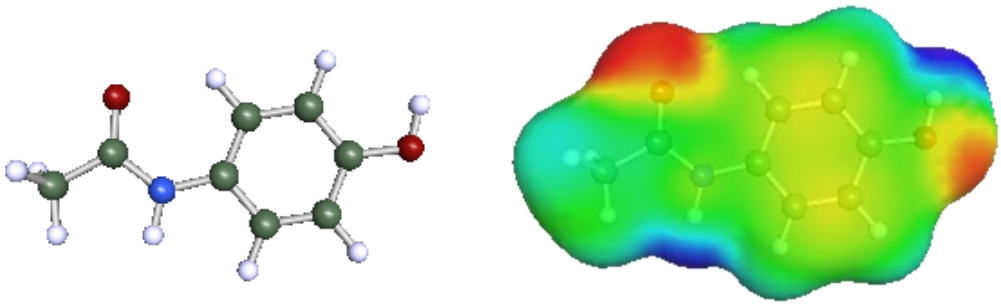


Figure 1-3 Visualisation of paracetamol (left) and the Sigma surface of paracetamol (right), colour coded by the polarization charge density, σ. Red areas denote strongly negative parts of the molecular surface and hence strongly positive values of σ. Blue areas denote strongly positive surface regions (strongly negative σ) and green denotes nonpolar surface.

COSMO-RS is a variant of dielectric continuum solvation models (CSMs) (Frank Eckert, 2015). A dielectric is a material which is an electrical insulator that can be polarized by an applied electric field. CSMs are models in which the solute is placed in a dielectric medium or cavity and which defines the interface between a solute molecule and the surrounding solvent molecules as a continuum. The calculations in COSMO*therm* are implemented in a virtual conductor environment. In such an environment, the solute molecule induces a polarization charge density, σ, on the interface between the molecule and the conductor, *i.e.* on the molecular surface. Figure 1-3 shows the sigma surface of paracetamol: the blue regions show electrostatically positive regions such as hydrogen; green are neutral areas and red are negative areas, which are located around the location of the oxygen lone pairs. The solute is treated as if inserted into a dielectric medium or cavity that is constructed around the molecule. The total energy of each screened molecule is calculated. In COSMO-RS theory, the solute and solvent molecules are considered to be a liquid of closely-packed, ideally screened molecules. To achieve this close packing, the system has to be compressed and the cavities of the molecules are

deformed slightly; although the volumes of the cavities do not change significantly. Every segment of the molecular surface is in close contact with another segment.

It is assumed that there is a conducting surface between molecules and that each molecule has net surface charge densities σ and σ' representing hydrogen bond acceptors and donors respectively. Where σ is positive surface charge density and σ' is negative surface charge density. It should be noted that, in reality, there is not a conducting surface between the surface contact areas (see Figure 1-4). The figure shows carbon dioxide and water with a hydrogen bond acceptor σ and donor σ' respectively.

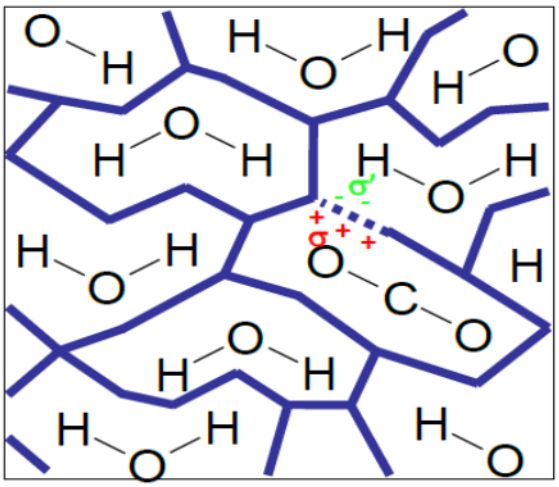


Figure 1-4 Electrostatic interactions arising from misfit screening charge densities σ acceptor and σ' donor for carbon dioxide and water

An electrostatic interaction arises from the contact of the two different surface charge densities (Figure 1-4). This specific interaction energy per unit area or "misfit" energy of surface charge densities is given by Equation 1:

Equation 1

$$E_{misfit}(\sigma,\sigma') = a_{eff}\frac{\alpha'}{2}(\sigma+\sigma')^2$$

(Equation 1) Where a_{eff} is the effective contact area between the two surface segments and α' is an adjustable parameter. If σ and σ' equal each other, they cancel out. Hydrogen bonding can also be described by the adjacent screening charge densities. Hydrogen bond donors have a strong negative screening charge density and hydrogen bond acceptors have a strong positive screening charge density. A hydrogen bond interaction can be expected if two pieces of segment are in contact and surfaces are sufficiently polar and of opposite polarity. This behaviour can be described by the following equation (Equation 2):

Equation 2

$$E_{HB} = a_{eff} c_{HB} min(0; min(0; \sigma_{donor} + \sigma_{HB}) max(0; \sigma_{acceptor} - \sigma_{HB}))$$

Where C_{HB} and σ_{HB} are adjustable parameters. The threshold for hydrogen bonding is $\sigma_{HB.}$ COSMO-RS, in addition to misfit and hydrogen bonding interactions, also considers vdW interactions between surface segments (Equation 3):

Equation 3

$$E_{vdW} = a_{eff}(\tau_{vdW} + \tau'_{vdW})$$

Where τ_{vdW} and τ'_{vdW} are element specific adjustable parameters. VdW energy is only dependent on the type of atoms of the elements involved in the surface contact.

Statistical thermodynamics provides the link between the surface of microscopic surface interaction energies and the macroscopic thermodynamic properties of a liquid. All molecular interactions within COSMO-RS consist of local pair wise interactions of the surface segments of molecules, therefore statistical averaging can

be applied to the interactions of surfaces. This is relatively computationally efficient. To describe the composition of the surface segment ensemble with respect to the interactions (which depend on σ only), only the probability distribution of σ has to be known for all compounds.

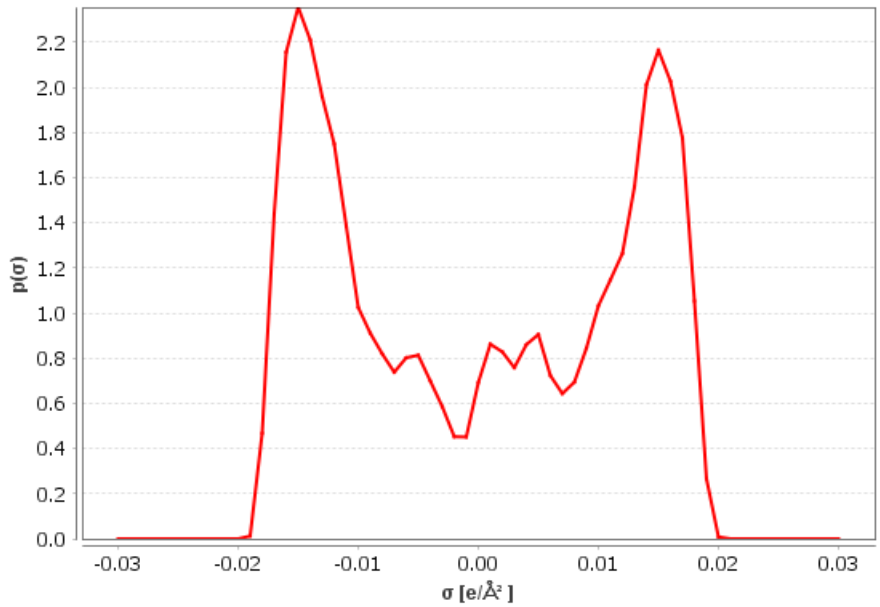


Figure 1-5 σ-profile showing probability distributions $p_i(\sigma)$ and surface charge density σ for water

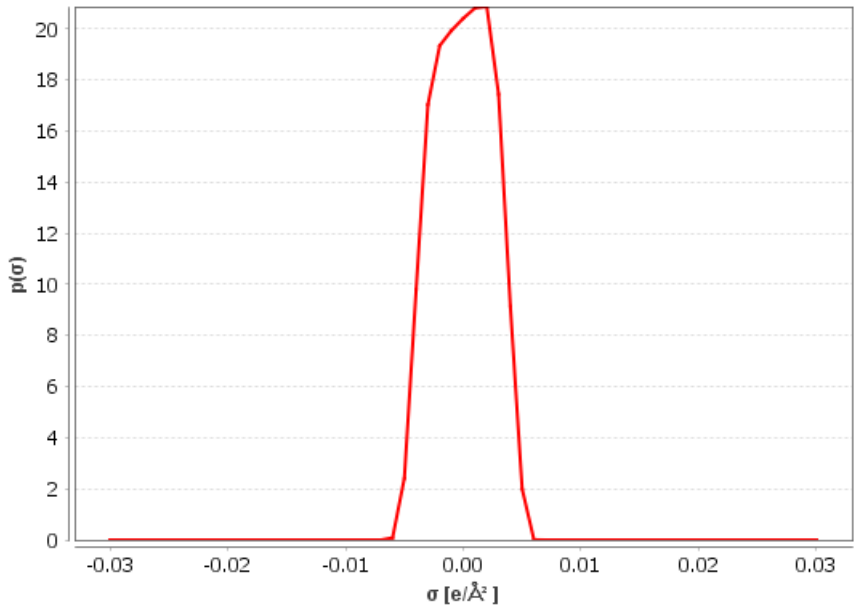


Figure 1-6 σ-profile showing probability distributions $p_i(\sigma)$ and surface charge density σ for hexane

These probability distributions $p_i(\sigma)$ are called σ-profiles (Figure 1-5 and Figure 1-6). The σ-profile of water shows a broad peak around –0.015 e/Å^2 arises from the two very polar hydrogen atoms whereas the peak around +0.015 e/Å^2 results from the lone pairs of the oxygen. The σ-profile for hexane reflects a non-polar compound with charge densities around zero. The peak in Figure 1-6 can be attributed to the carbon atoms for the positive σ and the hydrogen atoms for the negative σ. This is because negative charges of atoms cause a positive screening charge and vice versa. The σ-profile of the whole system $p_s(\sigma)$ is the sum of all the σ-profiles of all the components weighted with their mole fraction in the mixture x_i:

Equation 4

$$p_s(\sigma) = \sum_{i \epsilon S} x_i p_i(\sigma)$$

Using the chemical potential of a surface segment the screening charge density can be calculated by the normalised distribution function $p_s(\sigma)$ (Equation 4).

Equation 5

$$\mu_s(\sigma) = -\frac{RT}{a_{eff}} ln\left[\int p_s(\sigma') exp\left(\frac{a_{eff}}{RT}(\mu_s(\sigma') - e(\sigma,\sigma))\right) d\sigma'\right]$$

Where $\mu_s(\sigma)$ is the affinity of system *S* to a surface of polarity σ. This function is also called the σ-potential (Equation 5).

COSMO-RS represents molecular interactions in the form of a σ-profile and σ-potential of compounds and mixtures.

The chemical potential of compound *i* in system *S* can be calculated by integrating the σ-potential over the surface of *i* (Equation 6):

Equation 6

$$\mu_i^S = \mu_i^{C,S} + \int p_i(\sigma)\mu_s(\sigma)d\sigma$$

COSMO*therm* is, in the first instance, operated by a graphical user interface (GUI) called COSMO*therm*X. For this project COSMO*therm* has been automated using Python scripts to enable the calculations that are required for this project to be completed faster and on a larger scale than can be manually achieved by using COSMO*therm*X on a job-by-job basis. Only single conformers have been considered so far and conformers will be discussed more fully in the COSMO*conf* section 1.3.5.

1.3.3 Density Functional Theory

Density functional theory (DFT) (Kohn,Meir and Makarov, 1998) is an approach for modelling the electronic structure of atoms based upon a theory which states that all the ground-state properties of a system are a function of charge density. This theory uses electron density as a fundamental description of a molecular system. In this theory, the properties of a many-electron system can be studied using functionals *i.e.* a function of another function.

1.3.4 Basis sets

The basis set is a mathematical representation of the molecular orbitals within a molecule and are used for DFT calculations. The use of the set of *m* basis functions ($\Phi_{1...}\Phi_m$) can be interpreted as restricting each electron to a particular region of space. Basis functions are composed of a radial and an angular part. The radial part gives the

variation of probability amplitude (*X(r))* as the distance, *r*, from the nucleus is varied. The angular part gives the factor by which the radial part is scaled as the direction changes.

Types of basis functions include Slater-type orbitals (STO) and Gaussian type orbitals (GTO). GTOs are commonly used *ab initio*. The shapes of STOs and GTOs are different; STOs give a better representation of an atomic orbital but are more demanding computationally (Stewart,Hylton and Ravi, 2013). One of the advantages of Gaussian functions is that the product of two functions can be expressed as a single Gaussian (Magalhães, 2014). One of the disadvantages of Gaussian functions however is that they do not exhibit a cusp at the origin and they decay faster towards zero, whereas STOs have a cusp (Figure 1-7). This disadvantage is overcome by using a linear combination of Gaussian functions to represent each STO. It is common to use three times as many GTOs as STOs to achieve the same level of accuracy. Despite the need for more Gaussian functions, they are still more efficient computationally than STOs.

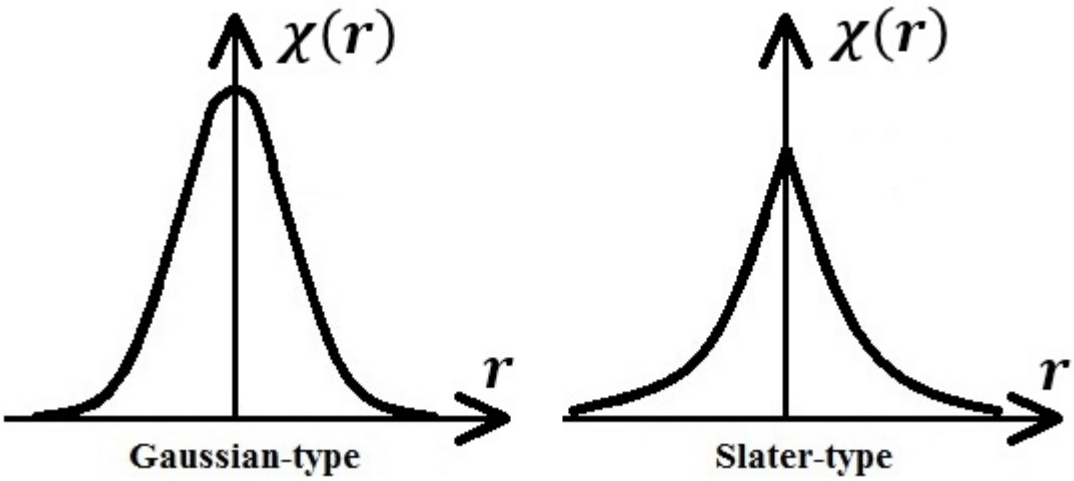


Figure 1-7 Plots of STO and GTO basis functions (Moradabadi, 2017)

Notation schemes have been devised to denote the basis set used in *ab initio* calculations (Leach, 1996). These notations, depend on the number and type of function used and are shown in Table 1-2. Other nomenclature is also used, notably Karlsruhe basis sets for COSMO*therm* (Zheng,Xu and Truhlar, 2011).

There are several major deficiencies in using a minimal basis set. Firstly, within a period, atoms on the right of the periodic table are described by the same number of functions as those from the left despite having more electrons. Secondly, the functions are unable to expand or contract in size in relation to the molecular environment. Thirdly, the minimal basis set cannot describe non-spherical characteristics of electronic distribution, occurring during polarisation. The solution to these three problems is to add more functions to the minimal basis set.

Table 1-2 Examples of some common basis sets, their notation and associated description

Class	Notation	Effect
Split basis set	3-21G 4-31G 6-31G SVP	More functions are added to describe valance electrons. Triple zeta basis set: triples functions from minimal basis set. Increases size of orbitals but does not allow modification of shapes under bond polarisation
Polarised	6-31G* or 6-31G(d) 6-31G** or G-31G(d,p) TZVP	Adds p orbitals on H-atoms and d orbitals on C. Takes into account changes in orbital shapes due to bond polarisation. 6-31G(d) only adds d functions to heavy atoms. 6-31G(d,p) adds d functions to heavy atoms and p functions to light atoms. This is important for polarised bonds.
Diffusion	6-31+G 6-31++G TZVPD-fine	Allows orbitals to occupy a larger region of space, which is important when orbital electrons are far away from the nucleus *e.g.* lone pairs or anions. 6-31+G only adds functions to heavy atoms. 631++G adds functions to heavy atoms and lighter ones.

The number of functions can be increased to better reflect the system being studied. As more functions are added, the more accurate the model calculation will become, however, this will increase the required computational time.

The three basis sets that COSMO*therm* uses are split valence polarisation (SVP), triple-zeta valence polarisation (TZVP) and triple-zeta valance polarisation with diffuse functions (TZVPD-fine).

1.3.5 COSMO*conf*

All the COSMO*therm* calculations used in this project were calculated using COSMO files obtained from the COSMO*logic* database, which were supplied with COSMO*therm*, or were parameterised using the software package COSMO*conf.* All

the molecules were parameterised at a TZVPD-fine basis set level. This basis set, when used for COSMO*therm* solubility predictions, was found to be the most accurate when compared with experimental solubility data points, as the study in section 2.8 details. COSMO*conf* is a software application for conformer generation. It uses pre-defined procedures that are optimised for the generation of the most relevant conformers for COSMO-RS. COSMO*conf* reduces the conformational space to a small set of relevant conformations by removing identical conformers, higher energy conformers and molecules with alternate stereochemistry. The conformers will be weighted internally by COSMO*therm* using their COSMO-energy and chemical potential. Lower energy conformations are assigned more weight in line with a Boltzmann form.

A conformer is a stereoisomer that can be converted by rotations around single bonds. Each conformer has a different energy, polarity and hydrogen bonding capacity. These properties are essential for the predictions that COSMO-RS performs. The correct conformational mixture should be used to achieve the most accurate solubility predictions as using only a single conformer can produce significant errors.

The key features of COSMO*conf* include an automatic conformer selection by relevance to the chemical potential (μ-clustering) in diverse solvents. Increased accuracy and robustness are achieved through the use of well-established density functional theory calculations. Finally, the ability to handle large molecules containing more than 100 atoms is feasible. A command line version, used for scripting and automation, makes it more amenable to integration within digital workflows as discussed in section 2.6 (Klamt, 2015).

1.3.6 UNIFAC

UNIFAC is a group contribution model that requires empirical data to parameterise the functional groups used to model the activity co-efficient in non-ideal mixtures. By utilising interactions of each functional group on both the solute and solvent molecules and also some binary interaction co-efficients, the activity co-efficient can be predicted. UNIFAC was developed for the prediction of VLE but has been expanded to predict other physical properties including SLE. UNIFAC separates the activity co-efficient (γ) into two parts: firstly, a combinational contribution (γ^C) to the activity co-efficient, which is due to the volume and surface area characteristics of a molecule; and secondly, a residual contribution (γ^R) reflecting energetic interactions (Equations 7 and 8).

Equation 7

$$ln\gamma_i = \ln {\gamma_i}^C + ln{\gamma_i}^R$$

Equation 8

$$ln{\gamma_i}^C = ln\frac{\varphi_i}{x_i} + \frac{z}{2}q_i ln\frac{\theta_i}{\varphi_i} + l_i - \frac{\varphi_i}{x_i}\sum_j x_j\, l_j$$

In Equations 8 to 13: x is the mole fraction; z is the co-ordination number and is usually 10; Θ_i is the area segment fraction; ϕ_i is the segment fraction for volume; r_i is the volume contribution for each molecule; k is the subgroup identification number; R_k is the vdW volume of group k (Equation 12); q_i is the area contribution for each molecule; Q_k is the vdW surface area of group k; $v_k^{(i)}$ is the number of groups of k in molecule i (Equation 13); and R_k and Q_k are taken from a list of fitted parameters.

Equation 9

$$l_i = \frac{z}{2}(r_i - q_i) - (r_i - 1)$$

Equation 10

$$\theta_i = \frac{q_i x_i}{\sum_j q_j x_j}$$

Equation 11

$$\varphi_i = \frac{r_i x_i}{\sum_j r_j x_j}$$

Equation 12

$$r_i = \sum_k \nu_k{}^{(i)} R_k$$

Equation 13

$$q_i = \sum_k \nu_k{}^{(i)} Q_k$$

The residual activity co-efficient is calculated by Equation 14. Γ_k is the group residual activity coefficient (Equation 15) and $\Gamma_k{}^{(i)}$ is the group residual activity coefficient in a reference solution of pure *i.* ψ is the group interaction parameter.

Equation 14

$$ln\gamma_i{}^R = \sum_k \nu_k{}^{(i)} \left[ln\Gamma_k - ln\Gamma_k{}^{(i)} \right]$$

Equation 15

$$ln\Gamma_k = Q_k \left[1 - ln \sum_m \theta_m \psi_{mk} - \sum_m (\theta_m \psi_{km} / \sum_n \theta_n \psi_{nm}) \right]$$

Θ_m is the area group fraction of group *m*. X_m is the mole fraction of group *m* in the solution and *n* is the molecular group (Equation 16).

Equation 16

$$\theta_m = \frac{Q_m X_m}{\sum_n Q_n X_n}$$

The group interaction parameter (ψ) is calculated by Equation 17 where a_{mn} and a_{nm} are taken from experimental data.

Equation 17

$$\psi_{mn} = e^{-(\frac{a_{mn}}{T})}$$

There are a significant number of parameterised values for functional groups available from the literature (Figure 1-8). The matrix below shows the binary interactions between functional groups which have been parameterised. However, as can be seen there are large gaps in the matrix.

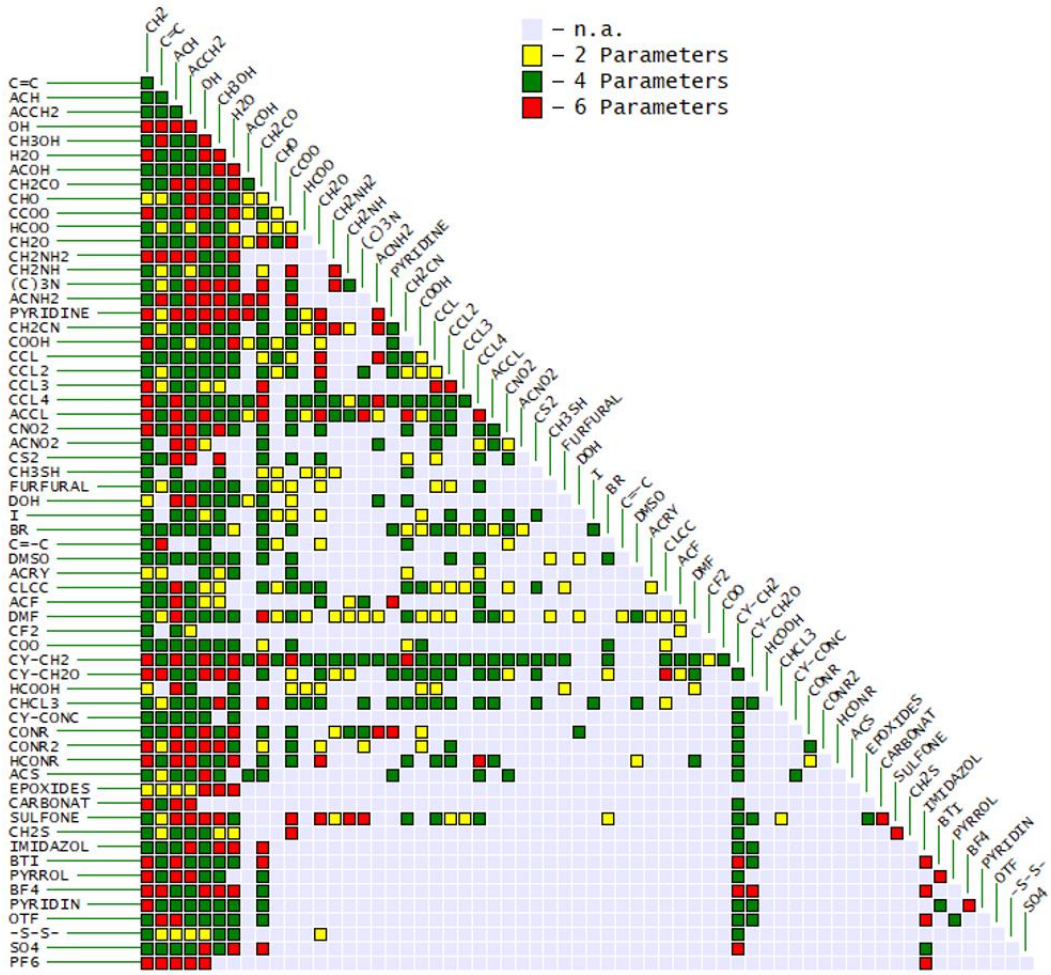


Figure 1-8 Available functional groups available for Modified UNIFAC (Do) (Gmehling, 2018)

1.3.7 SAFT-γ Mie

Similar to other group contribution methods such as UNIFAC, SAFT-γ Mie (Papaioannou *et al.*, 2014) determines molecular properties by sub-dividing molecules into functional groups. A value attributed to each group represents its contribution.

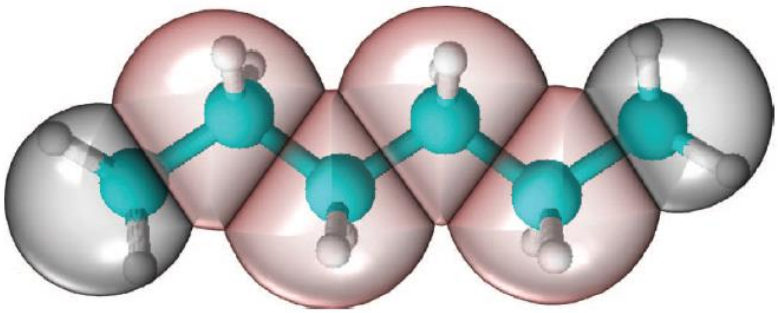


Figure 1-9 Representation of a fused heteronuclear molecular model employed within the SAFT- γ Mie. The example depicted is for n-hexane, comprising two instances of the methyl CH3 group (highlighted in grey), and four instances of the methylene CH2 group (highlighted in red) (Papaioannou et al., 2014)

A fused heteronuclear model (Figure 1-9) is employed where the molecules are constructed from distinct segments and potentials are calculated using various attractive and repulsive forces. One advantage of SAFT-γ Mie over other methods is that it can account for pressure effects that UNIFAC and COSMO-RS do not.

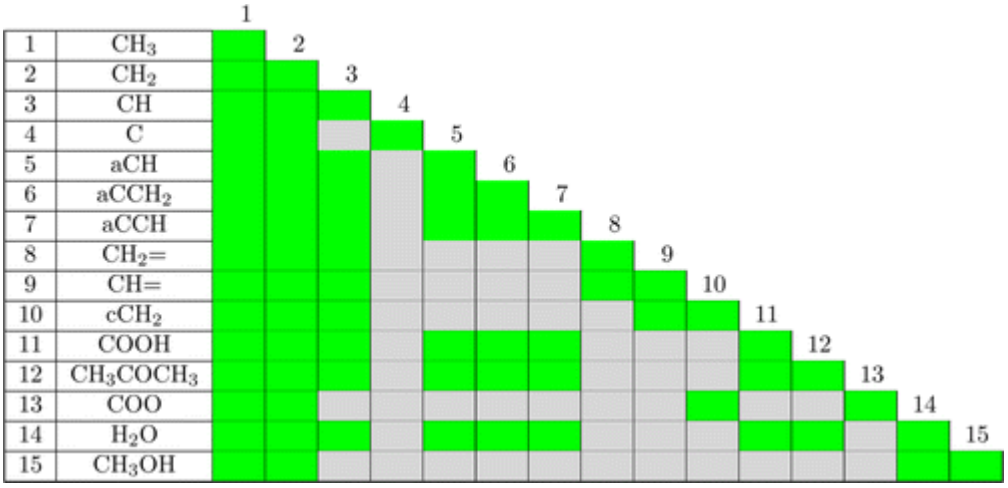


Figure 1-10 functional groups for SAFT-γ Mie with parameterised groups in green (Dufal et al., 2014)

The above figure (Figure 1-10) shows the functional groups for SAFT-γ Mie that are already parameterised. In comparison with UNIFAC groups there are some gaps which limit the predictive ability of this method.

1.3.8 The solubility equation

Solubility is the solute concentration in a solution that is in thermodynamic equilibrium with the solute in its solid state, denoted by Equation 18 (S. Gracin,Brinck and Rasmuson, 2002).

Equation 18

$$\mu^s = \mu^{sat}$$

The chemical potential is μ. The superscript *s* is the solid phase and is assumed to be pure and the superscript *sat* is the saturated solution. The chemical potential can be written thus:

Equation 19

$$\mu^{sat} = \mu^o + RTln(\gamma^{sat} x^{sat})$$

Where μ^o is the reference state (Equation 19). If an equal standard state is used for the solid and for the dissolved state then Equation 20 can be derived.

Equation 20

$$\ln a = lnx^{sat} + ln\gamma^{sat} = \frac{\Delta^{fus}H(T_m)}{R}\left(\frac{1}{T_m} - \frac{1}{T}\right) + \frac{1}{R}\int_{T_m}^{T}\frac{\Delta C_p}{T}dT - \frac{1}{RT}\int_{T_m}^{T}\Delta C_p\, dT$$

Where *a* is the activity of the pure solid, γ^{sat} is the activity co-efficient of the solute in the solution at point of saturation, x^{sat} is the mole fraction of the solute at saturation point. *T* is the temperature, T_m is the melting point temperature, $\Delta^{fus}H$ is the enthalpy of fusion and ΔC_p is the heat capacity and is the difference of heat capacity between the solid at temperature and the super-cooled melt. Since the heat capacity is difficult to measure and must be extrapolated from data, very little information regarding ΔC_p is available from the literature and it is often assumed ΔC_p=0. This results in Equation

21. This does not mean however that the effects of heat capacity are not negligible (see section 2.9).

Equation 21

$$\ln a = lnx^{sat} + ln\gamma^{sat} = \frac{\Delta^{fus}H(T_m)}{R}\left(\frac{1}{T_m} - \frac{1}{T}\right)$$

Equation 21 is most commonly used in solubility modelling and is the default equation used in COSMO*therm,* although there are two command line functions for heat capacity available within the software; one using a heat capacity estimate; and the other using a value from the literature.

1.3.9 Joback and Reid method

The Joback and Reid method, more commonly known as the Joback method, is a group contribution method that predicts eleven important and commonly used thermodynamic properties from a molecule's structure. In this project, only two of these properties have been used: enthalpy of fusion and melting temperature; as shown above in Equation 21. This method assumes that there is no interaction between functional groups and the contributions from each group are additive (Joback and Reid, 1987). The equation for enthalpy of fusion (Equation 22):

Equation 22

$$\Delta H_{fus}[kJ/mol] = -0.88 + \sum H_{fus,i}$$

Where ΔH_{fus} is the enthalpy of fusion and $H_{fus,i}$ is the contribution for functional group *i*.

The equation for the melting temperature (Equation 23):

Equation 23

$$T_m[K] = 122.5 + \sum T_{m,i}$$

Where T_m is the melting temperature and $T_{m,i}$ is the contribution from segment *i*.

This method was used when there was a lack of literature or experimental data pertaining to the enthalpy of fusion or melting temperature of a compound. A major drawback of this method is that it is cumulative and therefore larger molecules will be predicted to have a large enthalpy of fusion and melting temperature often resulting in unrealistically high values for each. In reality, most organic compounds do not have a melting temperature greater than 300°C. However, the Joback and Reid method can predict a melting temperature of 800°C and above. It also cannot distinguish between isomers or crystal polymorphs, which can have an effect on both values. A greater evaluation of the Joback method was carried out in Chapter Two (section 2.7).

1.3.10 Jain and Yalkowsky method

The Jain and Yalkowsky method (Jain and Yalkowsky, 2006) is another method for predicting enthalpy of fusion. It is similar to the Joback method in that it is a group contribution method. However, in this method proximity factors are taken into consideration for the enthalpy contribution values assigned to each functional group. In Equation 24, n_i is the number of times group *i* appears in a compound, n_j is the number of times proximity factor *j* appears in a compound, m_i is the contribution of group *i* to the enthalpy of fusion, and m_j is the contribution of proximity factor *j* to the enthalpy of fusion.

Equation 24

$$\Delta H_{fus}[kJ/mol] = \sum n_i m_i + \sum n_j m_j$$

A maximum of eight values can be attributed to each functional group, only one value being applied, depending on the proximity of another function group *e.g.* if a group is only attached to sp^3 atoms or it is an atom bridging two aromatic rings. A correction factor is also applied for intermolecular hydrogen bonding. Two hundred and nine enthalpic contribution values were obtained for the groups' environment and proximity values. This method, like the Joback and Reid method, does not take into consideration isomers or polymorphs. This method also predicts melting temperature but in this study, this prediction was not used due to the average error of 30°C reported in Jain and Yalkowsky's paper. An analysis of this method is included in Chapter Two (section 2.7).

1.3.11 COSMO*quick* linear regression model for enthalpy of fusion and melting temperature

COSMO*logic* provides a software package COSMO*quick* (Version 1.3 revision:996) in addition to COSMO*therm* and COSMO*conf*. This package has an inbuilt tool to predict both the enthalpy of fusion and the melting temperature using linear regression. A comparison with the Joback and Reid method and the Jain and Yalkowsky method is included in section 2.7.

1.3.12 Molecular Operating Environment

MOE is a drug discovery software and molecular simulation package developed by Chemical Computing Group. For this project, the software was used to generate

molecular descriptors for both solvents and solutes for use with the ML package RF, in R.

MOE uses its own “Quality” structure-activity relationship descriptors (QuaSAR-Descriptors). The purpose of QuaSAR-Descriptors is to calculate the properties of molecules that can then serve as their fingerprints or a digital representation. QuaSAR has a GUI that allows the user to select which descriptors to calculate. In principle, any molecular property can be used as a molecular descriptor and as such, there is no single calculation procedure for descriptors. Every descriptor is given a unique name, or code, which identifies it and is then used as database field names in the output.

Descriptors can be subdivided into two broad classes: 2D and 3D descriptors. 2D descriptors use only information from the atoms and how they are connected for the calculation (*e.g.* elements, charges and types of bonds). 3D descriptors use atomic coordinate information to perform the calculation (*e.g.* dipole moment) (MOE, 2018).

There are several other software packages available, both open source and proprietary, such as PaDEL-Descriptor, BlueDesc, ChemoPy, Rcpi, Cinfony, Modred and Dragon (Moriwaki *et al.*, 2018) and the Chemistry Development Kit (CDK) (Steinbeck *et al.*, 2003). Some of these packages have a large number of descriptors available and can be accessed using Python script or R.

1.3.13 Machine learning

ML is a field of computer science, which is concerned with making predictions or decisions based on supplied data. ML methods are defined by a particular algorithm

performed on a given dataset. Several ML algorithms have been developed such as artificial neural networks (Krogh, 2008), support vector machines (SVM) (Luts *et al.*, 2010) , fuzzy logic (Ross, 2010) and decision tree methods.

RF, a decision tree variant, was chosen as the ML algorithm for this project although, in practice, many other ML methods could have been applied to the model building process.

1.3.13.1 Random forest

RF is a ML method developed by Breiman and Cutler (Breiman, 2001). It is used for both classification and regression. Classification is used to predict which class or group a data point belongs to. Regression is used to predict continuous values and is the method used for this work. Within the CMAC research group, success has been achieved from the application of RF for predicting the outcomes of crystallisation experiments and for some image analyses (Johnston A, 2008, Bhardwaj *et al.*, 2015).

1.3.13.2 Example of random forest

For the purposes of demonstration, the simplified ML example below was constructed for demonstrating classification. The example regards classifying a person, based on their physical characteristics, with two outcomes being modelled: male or female. There are only a few descriptors for this example, Hair Length, Head Length, Weight, Build, Number of Arms, Number of Legs, Facial Hair, Height and Shoe Size. Using data for each person, a training set can be constructed with the response as Sex, allowing a RF model to be built. A separate test set can be used once the model has been built and this will establish the effectiveness of the model. The RF

algorithm takes a subset of these descriptors at random and tries to "guess" the response using only these descriptors. To do this, RF takes this subset and creates a number of decision trees. For example, if only two descriptors were used such as Height and Hair Length, a decision tree would be created with splits or nodes (Figure 1-11).

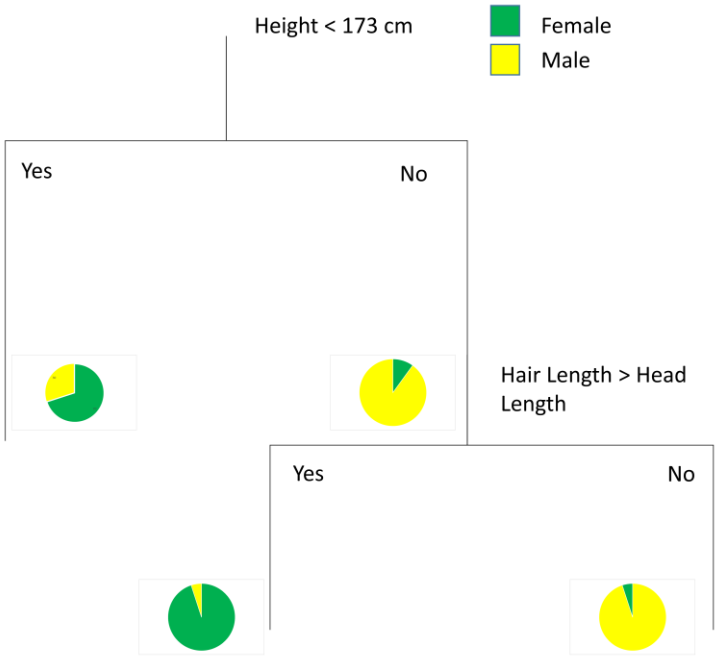


Figure 1-11 a simple RF decision tree

Descriptors such as Number of Legs and Number of Arms are very poor descriptors for this classification as they would be unlikely to split the data. At each of these nodes, a subset of the training set would be split according to some value in the descriptors. In this example, the model finds that the best split of the training set is below and above a threshold of 173 cm. The model does this by comparing the split with the response in the training set and moves the threshold up and down until the best split is achieved. A perfect descriptor would split the data exactly to match the

values of the response. In this example the split below 173 cm was 70 Females and 30 Males and above 90 Males and 10 Females. A further split with the descriptors Hair Length and Head Length then occurs at the next node. A real decision tree could have many more nodes and descriptors to achieve a split. More decision trees would be run with different subsets of descriptors and training set. Once all the trees were built, the model would vote on the classification for each point in the training set. Then the model could be applied to data that the model has not been exposed to, so long as the data contained the same descriptors used to build the model.

RF generates a collection, or ensemble, of independent decision trees (or a forest) each with an element of randomness. Through each tree, a query is run and the results are fed back as a prediction. In classification, the vote of each tree is taken and the majority result is given as the prediction. In regression, the mean of the prediction from each tree is taken as the result.

1.3.13.3 RF algorithm

One of the advantages of RF is its ability to perform internal validation. Since each tree is created using a sub-section of the dataset, data points that are excluded from the tree can be used for validation, this is called out-of-bag (OOB) data. The OOB error rate is calculated for each tree and aggregated. This provides an overall estimation of the model performance. Another benefit of RF is the ability to deal with less important descriptors as RF will select the best performing descriptor in a tree and use that one. With RF, it is also possible to retrospectively assess how well the model has performed as it is less of a “black-box” than other ML methods such as neural networks.

There are four main steps in the RF algorithm. A complete *n* x *p* data set is required where *n* is the number of individual cases and *p* is the number of descriptors. A response variable *y* is also required for each case.

1. A bootstrapped sample of the training set is drawn with replacement

Bootstrapping is the process of drawing a random subset of data from the training set. During construction of a bootstrap a single data point is sampled and returned to the original data pool. This process is repeated until the required sample size is reached. It is probable that the bootstrapped sample contains duplicated data points and some that have been left out. It is a random permutation of the original data.

2. The sample is split into two subsets by the best of m_{try} randomly selected descriptors, where $m_{try} << p$

For classification m_{try} has a default value of $\sqrt{p}$ and for regression it is a third of *p.* The data are then split by applying a threshold value. The quality of each split is assessed before repeating the split with every possible threshold. The best split for each descriptor is then determined with the overall best split for all descriptors retained. The quality of the split is determined by the split with the least sum of squared errors (SSE) calculated at each node for regression.

3. The splitting procedure is repeated with the subsets of cases at each node until full-length trees are formed.

4. New decision trees are generated each with a new bootstrapped sample until n_{tree}, the specified number of decision trees to be generated has been reached.

When these decision trees have been generated a test case can be passed through the model with each tree funnelling it through decisions to a terminal node with the outcome a numerical prediction in the case of regression.

1.3.13.4 Internal validation

An advantage of RF is the ability to perform internal validation. Since any one tree is constructed from the bootstrap of the training set, the cases excluded from the bootstrap, called out-of-bag (OOB) data, are used to conduct internal validation and storing the number of misclassified cases for classification. The OOB error rate is conducted for every tree and this provides a good estimate of the model's performance.

The OOB error can be used to monitor the growth of a RF model. As the number of decision trees is increased, the OOB error sharply decreases initially and levels off as sufficient trees have been added.

Validation by this method also mitigates the need for validation from a separate dataset. Normally, a model must be validated with a suitably sized set of external data, typically 20% the size of the training set. Having such a dataset requires a portion of the training set to be set aside and not used to build and train the model. Utilizing OOB error in RF, the model uses the whole training set.

Another advantage of RF is the ability to handle noise in the training data. Instead of handling all the predictors at the same time, each decision tree in RF looks at a small subset of predictors at each node and ignores the remainder. For each predictor, the one that produces the lowest SSE is taken forward. The process is then repeated. The predictors can be thought of as competing for selection at each node. Noisy predictors, less able to reduce the SSE, are selected-out as the tree is constructed. Including noisy predictors in the training set does not reduce the ability of the model to isolate the most useful ones, unless there are too many noisy predictors and useful predictors are very rarely selected.

This project uses two methods to validate the RF model; k-Fold cross-validation and solute-Fold cross-validation, which are discussed in detail in sections 4.2.3 and 4.2.4.

1.4 Literature review

1.4.1 Methods for solubility predictions

A 2018 study by Qiu and Albrecht investigated the correlation between the solubility of 905 distinct compounds (Qiu and Albrecht, 2018). Using 63240 pieces of data, analysis revealed correlation of solubilities between solvent pairs and allowed for clustering of most solvents. By using linear regression to calculate correlated solvents, it was possible to reduce the number of solvents required in solvent screening, resulting in a saving in both material and throughput.

Large-scale regression analysis of solvent pairs of interest was carried out to quantitate their correlations. It was shown that three solvents; water, dimethyl sulfoxide and acetonitrile do not correlate with any other solvents. Therefore, these

solvents should always be included in a solvent screen. Apart from alcohols, which require two disparate alcohols in a solvent screen, most other clustered solvents required only one solvent in the screening to be correlated. Twenty-four solvents were correlated and these could be excised from solvent screens.

Kan and Tomson's (Kan, 1996) study compared UNIFAC's predictions with the solubility of several compounds and solvents. The compounds had a range of solubilities spanning 11 orders of magnitude. The compounds were classed into several groups including short-chain alkanes, alkenes, alcohols, chlorinated alkanes and phenols. Good agreement was obtained between literature values of solubilities and UNIFAC predictions. The comparison of experimental data and predictions for the aqueous solubility of aliphatic and substituted aliphatic compounds had an absolute error of 0.43 of a log unit (log S). The solubility of 10 different compounds in 13 organic solvents gave an average error of log 0.18. The study, however, did use a limited dataset of 33 solubility data pointsand it is unknown whether this low error figure would have been maintained with a larger dataset.

A study comparing the predictions of the solubility of APIs by COSMO-RS with experimental data was completed by Ikeda et al. (Hirotaka Ikeda, 2005) 15 different APIs in four different solvents were compared at 25°C. Water, ethanol, acetone and chloroform were the solvents selected. The predictive method used for this study was COSMO*therm* with the TZVP basis set. The RMSEs were log 0.50, 0.61, 0.84 and 0.56 for water, ethanol, acetone and chloroform respectively with possibly a slightly larger error for acetone due to limited data points. The results that were obtained in this

study were satisfactory by the standards of the time. However, the study maybe could have improved the results by using an improved basis set such as TZVPD-fine (see study in section 2.8). The study emphasises the fact that COSMO*therm* does not rely on experimental data for the predictions, unlike UNIFAC.

Ruether *et al's* study (Ruether and Sadowski, 2009) compared experimental data and the prediction from two methods: PC-SAFT; and NRTL-SAC. The study compared five drug substances: paracetamol; ibuprofen; sulfaziazine, p-hydroxyphenylacetic acid and p-aminophenylacetic acid in pure solvents and in solvent mixtures. The PC-SAFT parameters for paracetamol were fitted with solubility data for paracetamol in water.

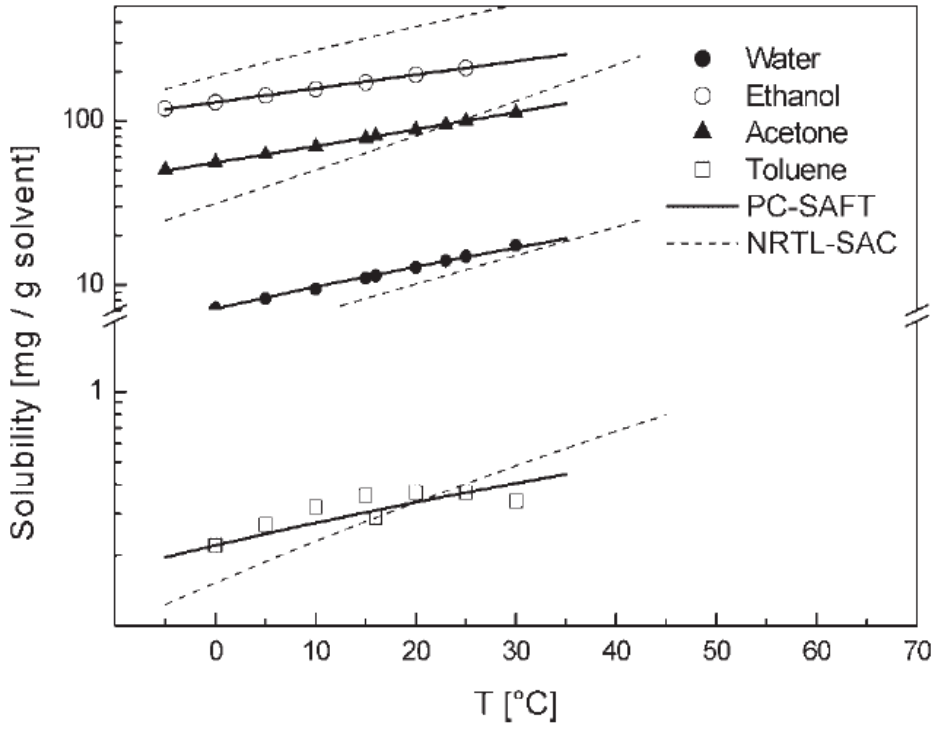


Figure 1-12 Comparison of PC-SAFT and NRTL-SAC solubility for the solubility of paracetamol in different solvent (Ruether and Sadowski, 2009)

The results in Figure 1-12 show the predicted solubilities of PC-SAFT and NRTL-SAC and give qualitative predictions that are similar. However, this is harder to quantify,

as exact figures were not quoted in the study. In the study, the enthalpy of fusion used for both methods was different with PC-SAFT using 27 kJ/mol and NRTL-SAC using 26 kJ/mol. Therefore, each model was not using the same input parameters and the comparison was not like-for-like. It is uncertain if NRTL-SAC's accuracy would be improved if both methods were using the same parameters.

A recent study by Bouilett et al (B. Bouillot *et al.*, 2017) compared the experimental solubility of seven pharmaceutical compounds in pure and mixed solvents with SciPharma (which implements a variation of PC-SAFT) and NRTL-SAC. In all, 386 pure solvent data points were used at one, or more, temperatures per solvent.

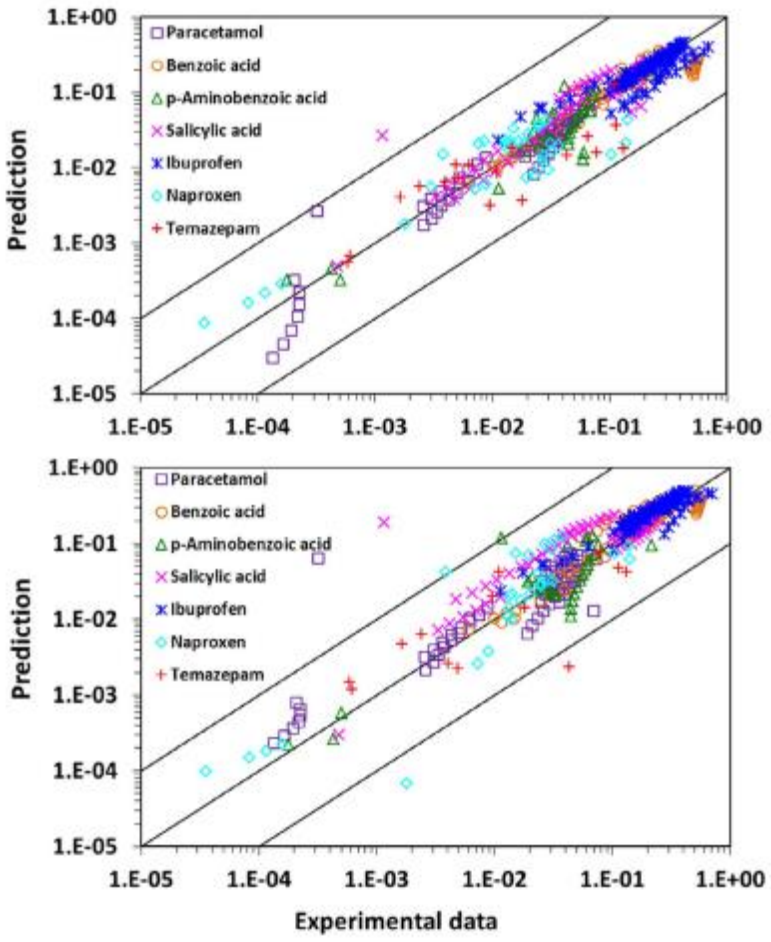


Figure 1-13 Predicted versus experimental (mole fraction) of the seven pharmaceutical compounds SciPharma (top) and NRTL-SAC (bottom) ref (B. Bouillot et al., 2017)

Figure 1-13 shows the results of the pure solvents: SciPharma had an RMSE (mole fraction) of 0.196 and NRTL-SAC had an RMSE of 0.227 with only benzoic acid and ibuprofen performing better using NRTL-SAC. In the mixed solvents, again SciPharma performed better than NRTL-SAC. In all SciPharma had an overall better accuracy in terms of RMSE.

There have been many studies conducted comparing the relative benefits and drawbacks of each method and there are several derivations of UNIFAC such as Original UNIFAC, Modified UNIFAC(Do) and Modified UNIFAC(Do) Consortium (Xue,Mu and Gmehling, 2012), and with NRTL-SAC such as original and temperature-dependant NRTL-SAC (Valavi,Svärd and Rasmuson, 2016) . There are also several derivations of COSMO-RS: COSMO*therm* which was developed by Klamt (F. Eckert and Klamt, 2002); COSMO-SAC which was developed by Lin (S. T. Lin and Sandler, 2002); and COSMO-RS(Ol) developed by Mu (Mu,Rarey and Gmehling, 2007).

In the study by Xue et al (Xue,Mu and Gmehling, 2012) the UNIFAC models were compared to two COSMO-RS models, COSMO-SAC and COSMO-RS(Ol), and experimental data. The paper concluded that the UNIFAC models were superior to the COSMO-RS models when comparing several thousand activity co-efficients at infinite dilution (see Table 1-3).

Table 1-3 RMSE's of COSMO-SAC, COSMO-RS(Ol), original UNIFAC, modified UNIFAC(Do) and modified UNIFAC(Do) Consortium for solubility

Method	**RMSE**
COSMO-SAC	0.835
COSMO-RS(Ol)	0.841
original UNIFAC	0.528
modified UNIFAC(Do)	0.340
modified UNIFAC(Do) Consortium	0.329

Klamt commented on this study with several points of criticism (Klamt, 2012). Firstly, that the study was biased in favour of the UNIFAC model as it did not use the most advanced version of COSMO-RS, namely COSMO*therm*, and since the most up-to-date version of UNIFAC was used that the study was unbalanced. Secondly, the comparisons were taken from the UNIFAC databank. As such the comparisons were made on the training set for UNIFAC models and that the database is biased towards mixtures of simple compounds such as alkanes/alcohols or alkane/ketones whereas more complicated compounds such as pharmaceuticals were greatly underrepresented. Klamt states that as UNIFAC is stronger on simple compounds this favours the results towards UNIFAC. Thirdly, that the graphs used in the study were not representative of the overall performance of COSMO*therm*. Klamt concluded that UNIFAC is a useful tool for simple structurally-related compounds and that COSMO*therm* provides more robust predictions for more complicated systems.

In Valavi *et al's* study (Valavi,Svärd and Rasmuson, 2016) UNIFAC and the two NTRL-SAC models were compared with experimental data. Original NRTL-SAC was extended to include temperature-dependant binary interaction parameters. The performance of each method depended on the solvent systems that were studied: 48 solute/solvent systems for the NRTL-SAC models; and 33 solute/solvent systems for the UNIFAC model. In evaluating these models, Valavi only compared systems where there were values for all three models with an observed RMSE error in mole fractions of 1.42, 1.06 and 0.87 for UNIFAC, NRTL-SAC original and the NRTL-SAC temperature-dependant model respectively. The study stated that there was an improvement in the model due to the temperature-dependant parameters and that

both NRTL-SAC models were better than UNIFAC. However, this was an extremely small dataset and it would be premature to draw that conclusion; a larger dataset would be required to make that assertion conclusively.

All these studies, except Xue's, used small datasets, therefore it is difficult to compare them. Xue's study, although having a larger dataset, was criticised for its lack of diversity.

1.4.2 Machine learning methods

There have been several attempts to use ML methods along with experimental data to predict the solubility of compounds in water.

Jorgensen and Duffy had some success predicting the aqueous solubility of drugs using a group contribution method with linear regression and neural networks (Jorgensen and Duffy, 2002). The training set consisted of 317 organic molecules, including some complex drugs, and the resultant model had an RMSE of log 0.63. This is better than some errors from predictive methods where between log 0.7 and log 1.0 RMSE for complex molecules is usually acceptable (Palmer and Mitchell, 2014).

Schroeter *et al* investigated four different ML approaches for predicting aqueous solubility in their 2007 study (Schroeter *et al.*, 2007) using approximately 4000 different compounds. The methods consisted of Gaussian Process, RF, SVM and Ridge Regression models. 1664 descriptors were generated using Dragon. Three different datasets were used giving different results for each ML method. Dataset 1 has 5,625 measurements from 3,307 compounds. Dataset 2 has 688 measurements from 632 compounds and Dataset 3 has 536 measurements.

Table 1-4 Results for Schroeter study using different machine learning approaches and datasets (Schroeter et al., 2007)

Schroeter's Dataset	Method	RMSE (logS)
1	Gaussian Process	0.747
	Ridge Regression	0.862
	SVM	0.803
	RF	0.840
2	Gaussian Process	0.846
	Ridge Regression	0.847
	SVM	0.848
	RF	0.855
3	Gaussian Process	0.579
	Ridge Regression	0.996
	SVM	0.600
	RF	0.660

As can be seen in (Table 1-4) the different datasets produce a variety of results with the Gaussian Process model being found more accurate in all three datasets. Without access to specific details and characteristics of each dataset, it would be difficult to make a detailed assessment of the advantages and disadvantages of each method.

Palmer *et al* (Palmer *et al.*, 2007) created four different ML models to predict aqueous solubility. Firstly, a RF model was created, next a partial least squares model, then a SVM model and finally a neural networks model. The RF model to predict aqueous solubility had a training set of 658 compounds and a test set of 330 molecules. MOE calculated over 200 2D and 3D descriptors. The RF was trained initially on all the 2D descriptors giving a RMSE of log 0.69 and then reduced to the 40 most important descriptors which gave a fit almost identical to the model using the full descriptors with an RMSE of log 0.695. The partial least squares (PLS) model was then constructed using twelve descriptors including SlogP and a_acc (number of hydrogen bond acceptors) and a_don (number of hydrogen bond donors). This model was not as successful as the RF model with an RMSE of log 0.787. Reducing the number of

descriptors can have the effect of destroying the inter-connectivity between descriptors. Eliminating descriptors that have a co-dependence on one another can reduce the ability for the model to find the complexities of the relationships between descriptor and response. The SVM model and the neural networks model performed a little better with an RMSE of log 0.726 and log 0.742 respectively. The last three models used different subsets of descriptors from the RF model but eight of the twelve descriptors were in the top 25 of ranked descriptors in the RF model.

Boobier et al (Boobier,Osbourn and Mitchell, 2017) compared 10 ML algorithms to predict aqueous solubility. From a dataset of 100 drug-like molecules the dataset was split using a training set of 75 molecules and a test set of 25 molecules. 123 descriptors were used from chemistry development kit (CDK), a software package developed by the CDK project.

Table 1-5 RMSE of 10 machine learning algorithms (Boobier,Osbourn and Mitchell, 2017)

Method	**RMSE (log S)**
Multi-layer perceptron	0.985
RF	1.165
Bagging	1.165
K nearest neighbours	1.204
ExtraTrees	1.227
AdaBoost	1.235
PLS	1.265
Stochastic gradient descent	1.280
SVM	1.429
Decision tree	1.813

Table 1-5 shows the results from the study; multi-layer perceptron performed the best with RF and related method bagging (m_{try} is equal to all the descriptors used) performing second best with the decision tree method performing worst.

There have been many attempts to predict solubility using the various methods stated previously. The methods that use some experimental data such as NTRL-SAC seem to be superior to the COSMO-RS derived methods. This might not be the case for all types of molecule however and more work needs to be done to compare and analyse the methods. A larger solubility database with a wide range of differing molecules with different properties, mass and functional groups would enable a more comprehensive analysis of the strengths and weakness of each method.

The attempts at using ML techniques to predict solubility have had some success but there is some difficulty in comparing studies with one another as each method is different due to the datasets that each use for training the algorithm. Without access to the dataset in each study, confirming the strengths and weaknesses of each is difficult.

1.5 Thesis structure

This thesis focusses on the automation of COSMO*therm* and assessing the predictive capabilities of the method. Tools were developed using COSMO*therm* predictions, exploiting design of experiment (DoE) approaches and linear regression to predict solubility at speed whilst maintaining the ease-of-use for the non-expert. ML algorithms were used to build a model to improve the accuracy of COSMO*therm* predictions by applying a "correction factor". Predictive tools for the selection of crystallisation and wash solvents were developed using COSMO*therm* predictions and with the ML algorithm "correction factor" applied.

Firstly, COSMO*therm*, UNIFAC and SAFT-γ Mie were compared to experimental data to assess the reliability and robustness of the methods. Secondly, the adjustable parameters required for COSMO*therm* predictions *i.e.* melting temperature and enthalpy of fusion, and the influence that inaccuracies in these parameters have on solubility predictions were assessed.

A model of COSMO*therm* was developed using DoE and linear regression for the almost-instant prediction of solubility for non-experts or modellers.

The selection of crystallisation and wash solvents for pharmaceutical manufacturing processes using predictive methods and a workflow approach was developed. This hybrid approach was implemented by comparison with experimental data from a study of the solubility of paracetamol and its impurities, with predictions from COSMO*therm*.

2 An assessment of errors and inconsistencies that arise when measuring and predicting solubility

2.1 Overview

This chapter considers the difficulties of comparing solubility predictions to experimental data with significant variance. Additionally, the problem of variance in literature values for enthalpy of fusion and melting temperature and the effect that can have on solubility predictions is studied. This chapter also describes building a database of molecules and associated data, the automation of COSMO*therm* workflows for a manufacturing environment and the validation of COSMO*therm* predictions by comparing those predictions with other predictive methods and with experimental data.

2.2 Solubility measurements

Before comparing solubility predictions from COSMO*therm* with experimentally determined data, there is a need to consider factors that give rise to significant variance on measured or computed values.

The empirical measurement of solubility can differ by a non-trivial quantity. An analysis of the aqueous solubility of 411 compounds by Katritzky et al reported an average standard deviation of log 0.58 (Katritzky *et al.*, 1998) attributing the variation to experimental error. The data collected by Kishi and Hashimoto (Kishi and Hashimoto, 1989) from 17 different laboratories for the compounds anthracene and fluoranthene, using the same experimental protocol (Jorgensen and Duffy, 2002) showed a wide range of aqueous solubility measurements. Solubility for each compound ranged over log 0.86 and with a standard deviation of log 0.19, which was

attributed to experimental uncertainty caused by crystal shape, polymorphism, pH and temperature control for the solution. Another study by Myrdal et al (Myrdal,Manka and Yalkowsky, 1995) used additional data for the two compounds and a range of approximately log 1.5 was observed from data compiled.

630 pure solvent solubility points were measured by GSK researchers using an automated protocol for a high performance liquid chromatography (HPLC) and taken from the GSK solubility database. Access to this database was obtained from a three-month industrial placement at GSK Stevenage in 2018. The dataset contained combinations of 13 compounds in 47 solvents. Most measurements were recorded in triplicate. Solubility was measured between 20-25°C with an accurate temperature being recorded. The average of each solute/solvent combination was taken and the standard deviation was calculated. Most combinations had deviations. Apart from an outlier the maximum deviation was approximately 2g/100g of solvent (Figure 2-1).

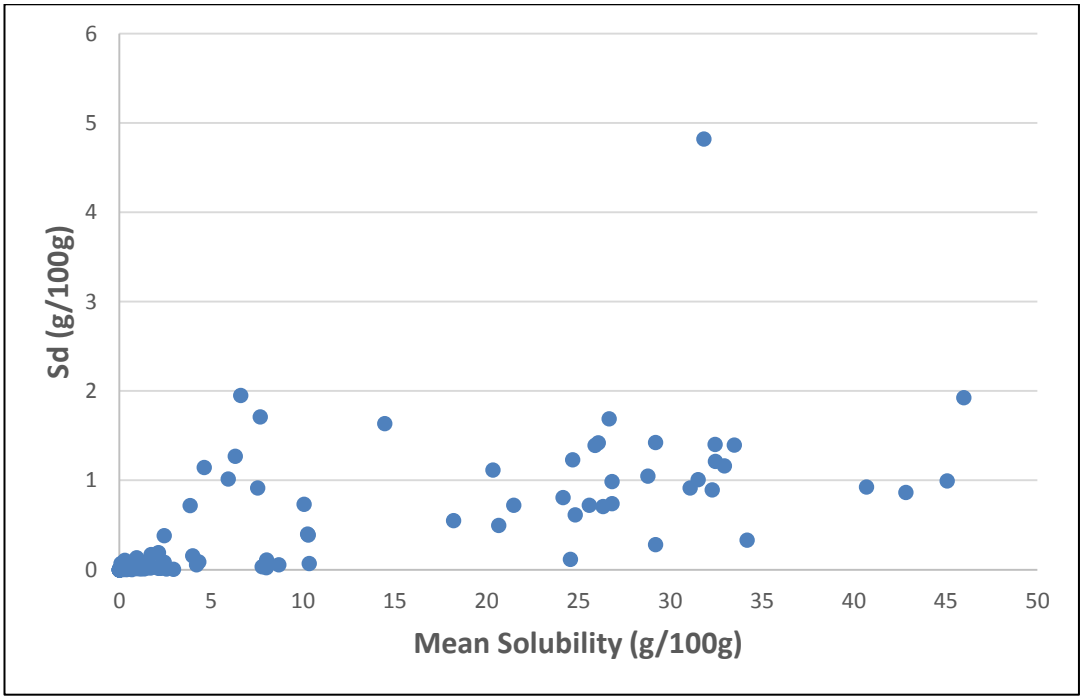


Figure 2-1 standard deviation in the mean solubility of HPLC solubility measurements

Figure 2-2 shows the percentage variance in the mean solubility for the GSK data points which were in triplicate. The greatest percentage variance is seen below 10g /100g which is generally considered to be low for solubility. This indicates that in terms of percentage variance it is much more difficult to obtain a consistent solubility measurement at low solubilities. This may explain some of, but not all, the error in predicting low solubility solute/solvent combinations as it is difficult to obtain a consistent experimental value.

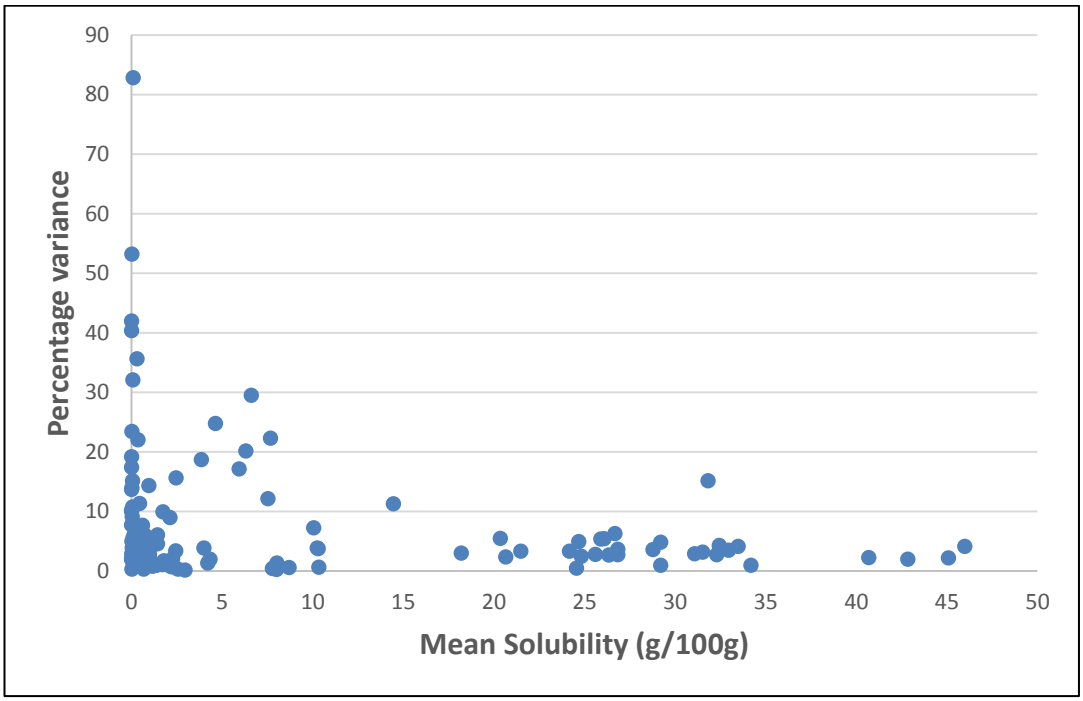


Figure 2-2 Percentage variance in experimental solubility points

2.3 Melting temperature and enthalpy of fusion variation

A study of the literature, was completed for this project, for melting temperatures and enthalpy of fusion of 73 compounds provides an inconsistency of values for each compound. The full results of this study can be found in Appendix One (section 7).

Figure 2-3 shows the difference in literature values of enthalpy of fusion for several API polymorphs. The variations in the results are most likely attributed to the reasons discussed below.

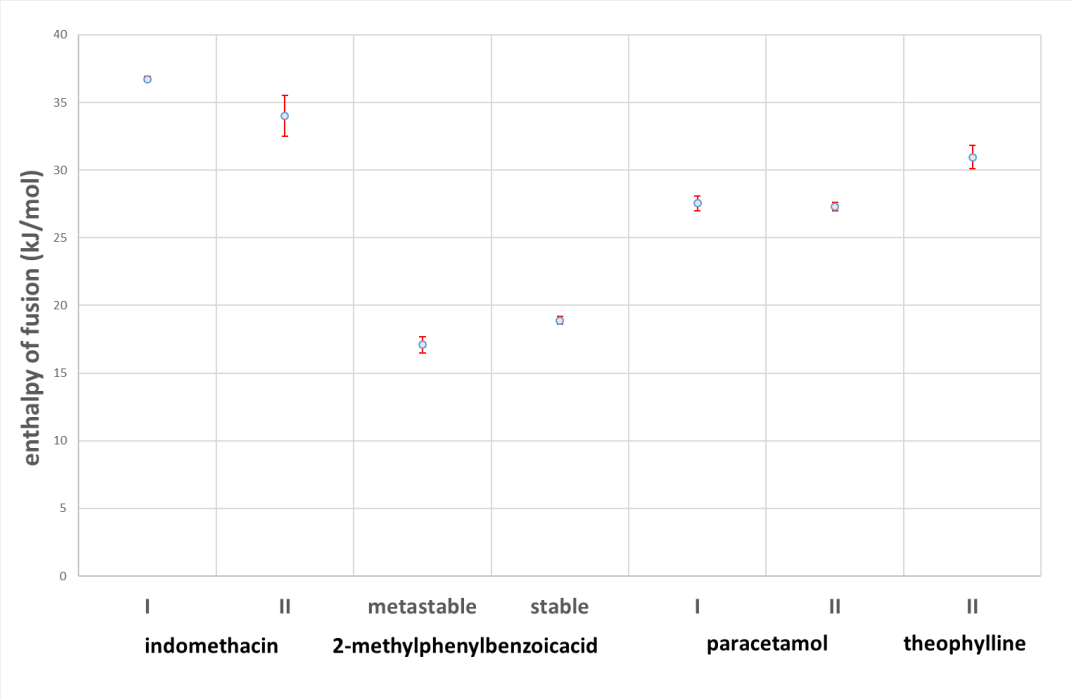


Figure 2-3 Mean enthalpy of fusion for polymorphs of indomethacin, 2-methylphenylbenzoic acid, paracetamol and theophylline

Figure 2-4 shows the mean enthalpy of fusion found in literature and Figure 2-5 shows the mean melting temperatures found in this study. The compounds in Figure 2-3 have also been included as there were some papers without polymorph details. As can be seen, there is a great deal of variation between literature values. This could be down to the simple fact that the compound does have different polymorphs and that the polymorph has not been specified in the paper. Although the number of polymorphs are not necessarily stated in the papers that present values for enthalpy of fusion, these polymorphs can be well known. Indomethacin, for example, has at least five polymorphs (Surwase *et al.*, 2013). D-mannitol has three known

polymorphs, with the beta form being the most stable (Cares-Pacheco *et al.*, 2014) and fenofibrate has at least three forms (Ying *et al.*, 2017).

A main reason for the variation in values is compounds having different polymorphs. However, in many literature papers, it is not stated which polymorph was studied; this creates a problem selecting the "correct" value for the COSMO*therm* model. Aside from polymorphism, sample purity, the equipment used for analysis, heating rate, and mass of sample can affect the value obtained for both the enthalpy of fusion and the melting temperature. Araujo et al (Adriano Antunes Souza *et al.*, 2010) compared different samples of zidovudine. The differing heating rates of the sample (1, 2, 5, 10 and 15°C per minute) gave a variation in enthalpy of fusion of 0.73 kJ/mol and a melting temperature variation of 0.51°C. Mass of sample also affected results with variations of 1.95 kJ/mol and 0.37°C. The degree of sample purity also affected results with purity ranging from 97.59-99.83% giving rise to a difference of 2.79 kJ/mol and 4.17°C. Man's study (Man and Tan, 2002) into the effects of differential scanning calorimetry (DSC) heating rate of 11 vegetable oils had a similar variation in results with one oil having a 7.8% variation in the enthalpy of fusion between heating rates.

The literature review was not exhaustive and it is possible that some of the values reported approximately match values for known polymorphs that are identified in other papers. It does show that caution must be taken accepting enthalpy of fusion and melting temperatures values from the literature as even a small deviation of a kJ/mol and a few degrees Celsius can have an effect on COSMO*therm* model

predictions, which are sensitive to these key input parameters. For model generation, it would be ideal if in-house DSC data were obtainable for the compound that solubility predictions were to be made. However, the raw material may not be available and as has been already highlighted, DSC results can vary even within the same batch of compound.

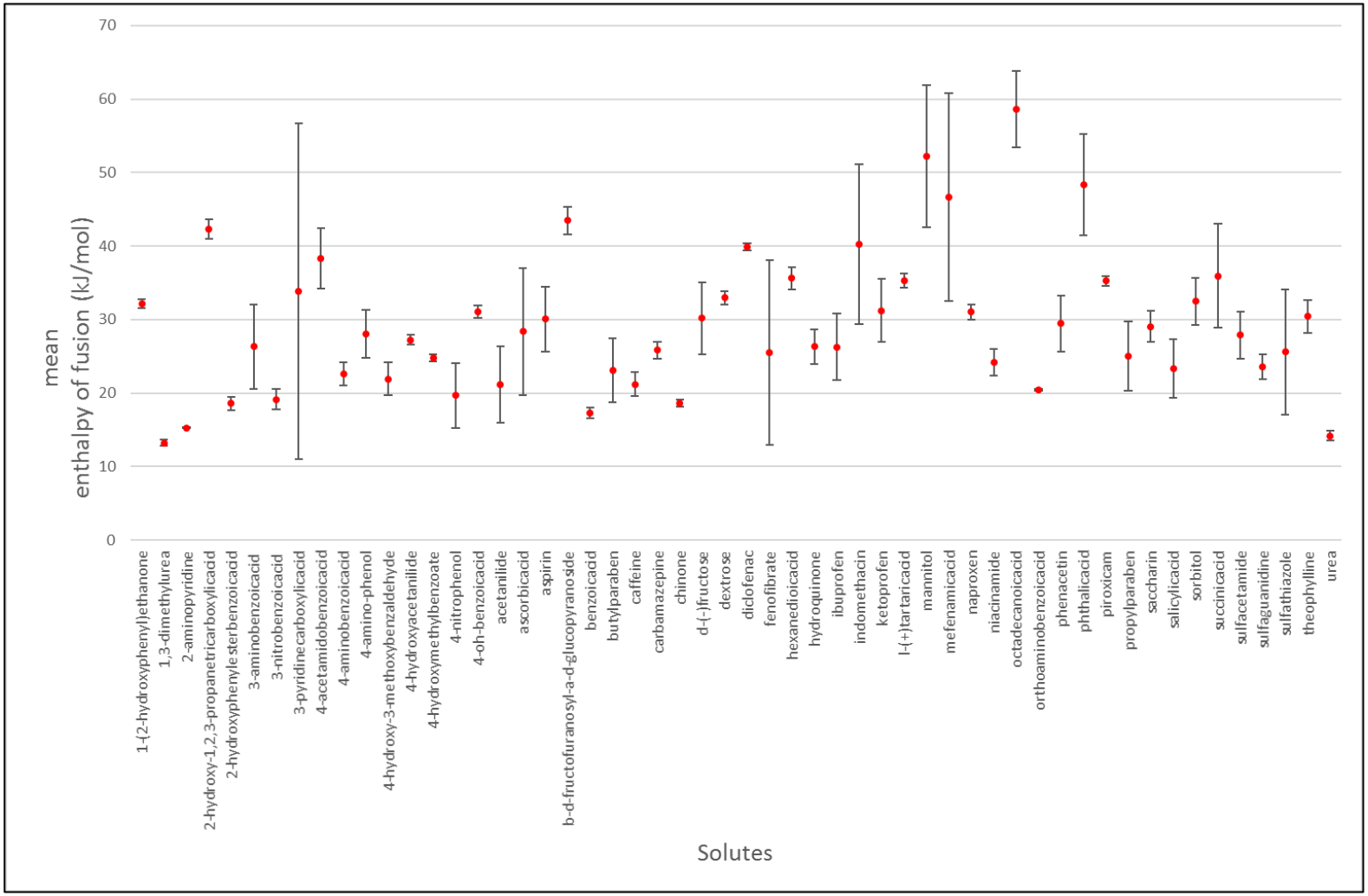


Figure 2-4 Mean enthalpy of fusion from literature with the bars showing the range of values

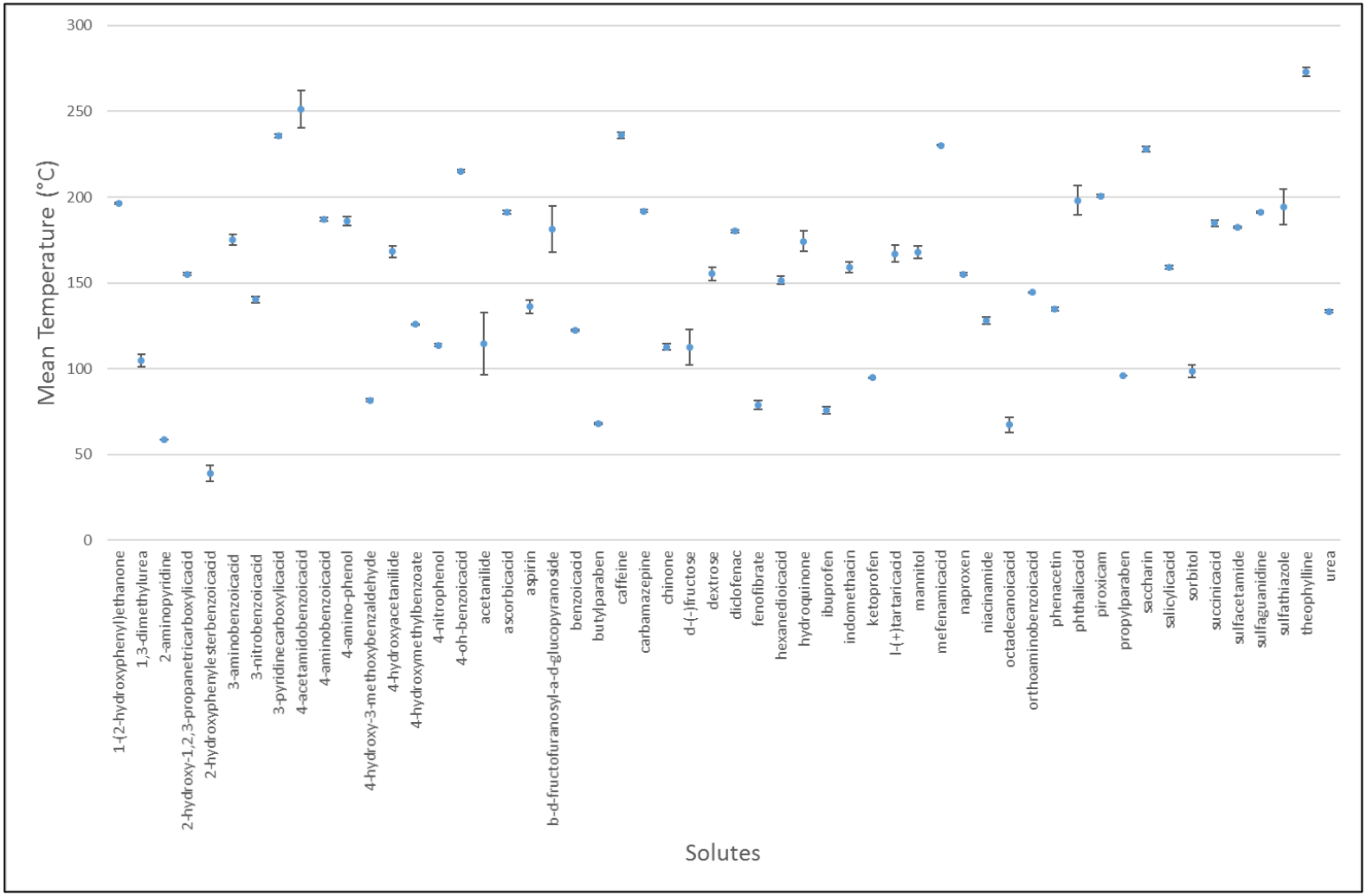


Figure 2-5 Mean melting temperature from literature with bars showing the range of values

2.4 Building a COSMO*therm* database

Before large-scale solubility predictions were carried out, a database of compounds commonly used by CMAC for crystallisation research was constructed. COSMO*therm* had a large database of compounds supplied with the software, however many of the compounds of interest to CMAC were not available and had to be parameterised using COSMO*conf.* The CMAC list had 41 solutes and 122 solvents that were either commonly used in industry or were thought to be of interest when this work began. This list grew in size as the project progressed to 115 solutes and 152 solvents due to an interest in solubility predictions for additional solutes, such as drug impurities, and the solubility of APIs in "green" solvents. Green solvents are environmentally friendly solvents derived from processing agricultural crops and are an alternative to petrochemical solvents.

The database in the first instance had COSMO files in three different basis sets for each compound. COSMO files in SVP, TZVP and TZVPD-fine basis sets were included. Later additions to the database only used the TZVPD-fine basis set as this was discovered to be the most accurate basis set for predictions when compared with experimental and literature data and any additional computational time was negligible. This analysis is reported later in this chapter (section 2.8). Each compound requires an additional file called a vapour pressure/property file or .vap file. This file contains compound-specific experimental information, some of which is essential input required for COSMO*therm* predictions. As solubility predictions require two specific pieces of experimental data: the enthalpy of fusion and the melting temperature, both of these must be added to the .vap file. Some of the .vap files

already have this information included. However, most did not and a literature search was required to obtain the missing information.

Some of the compounds in the database had neither literature data available or the compound was not available for experimental data to be obtained from DSC. Therefore, other methods were required to obtain this information such as the Joback and Reid method (Joback and Reid, 1987), the Jain and Yalkowsky method (Jain,Yang and Yalkowsky, 2004) or the COSMO*quick* regression model (Loschen and Klamt, 2012); the benefits and drawbacks of these methods are discussed in section 2.7. Predictive methods for obtaining enthalpy of fusion are far from ideal and any prediction made should be regarded with caution; however, in the absence of any experimental or literature data it is the only option to achieve a solubility prediction from COSMO*therm*.

2.5 COSMO*conf*

COSMO*conf* was used to parameterise any molecules not available in the database supplied by COSMO*logic*. Approximately one third of solutes of interest to CMAC were not available and had to be parameterised. COSMO*conf* can generate fully-parameterised COSMO files from a given compound structure. In all cases, simplified molecular-input line-entry system (SMILES) codes were used instead of inputting the structure manually; however, all structures were manually checked for accuracy following SMILES to structure conversion. All molecules were parameterised to use the TZVPD-fine basis set. The COSMO files were transferred to the database and a .vap file was created for each compound.

2.6 Automation of COSMO*therm*

COSMO*therm* is supplied with a graphical user interface (GUI), COSMO*thermX*, which was less practical for the volume of modelling predictions that were required for this project. Therefore, automation and connection of several steps in the COSMO*therm* workflow was required. These steps included: the *LLE calculation*, which predicts liquid-liquid equilibria (LLE) and the miscibility of solvents; *multiple solvents*, which predicts solid-liquid equilibria (SLE); and *solid-liquid*, which predicts the phases in solid-liquid extractions and can predict solubility in more complex systems with two, or more, solvents.

COSMO*therm* calculations can be fully automated via the command line. Python scripts needed to be written for each type of prediction. The scripts had to be simple to use and easy to connect into a computational workflow. The type of predictions described in the remainder of this section were all executed using automated scripts.

An input file must be created for each job instructing COSMO*therm* which type of prediction to run, with which solutes and solvents and at what temperatures and/or mole fraction in the case of using binary solvents. A Python script had to be written for each different type of prediction *e.g.* solubility in pure solvents, solubility of salts or solubility in binary solvents as a result of implementation differences in the input file. The input file can be broken into three sections; file directories, number of COSMO files required and job description. The purpose of the Python scripts was to produce these files and to start the predictions.

2.6.1 Miscibility

Miscibility information is needed for several processes in the pharmaceutical industry such as wash-solvent selection and anti-solvent crystallisations. Miscibility is the extent to which two liquids mix together. COSMO-RS theory was designed for liquid-liquid interactions instead of solid-liquid interactions therefore predicting miscibility is one of COSMO*therm*'s primary functions.

A study was completed by CMAC researcher Václav Svoboda using laboratory data from the DETHERM database (Detherm, 2016) for LLE to show the miscibility of solvents. 109 LLE phase separation data points were found. This study was compared with COSMO*therm* predictions. COSMO*therm* predicted all 109 combinations at 25°C. Of the data points taken from DETHERM, 80 of the 109 predictions had phase separations. When COSMO*therm* predictions were compared with the DETHERM data, 36 of the 109 had an error of less than 5% of a mole and 34 out of 109 had an error of more than 5% and less than 20% of a mole Table 2-1. These results from DETHERM did not always specify the starting mole fractions for each combination and some of the data points were from elevated temperatures therefore the comparisons were not always like for like.

Table 2-1 % error when phase separation predictions are compared with experimental data

No. of phase separations	**% error (mole fraction)**
36	<5
34	>=5 and <20
39	>=20

This study shows that COSMO*therm* is a useful tool for the prediction of miscibility and although COSMO*therm* does not eliminate the need for laboratory testing, it can give an indication of which solvents to test for miscibility.

2.6.2 Solubility of neutral compounds

COSMO*therm* was automated using Python script to predict the solubility of neutral compounds. To validate this the experimental solubility of lovastatin in three solvents, n-butyl acetate, 1-butanol, 1-pentanol, was compared with the predictions from COSMO*therm*, UNIFAC and SAFT-γ Mie. The experimental solubility was obtained by CMAC researcher Humera Siddique. UNIFAC predictions were obtained from CMAC researcher Václav Svoboda and the SAFT-γ Mie predictions were obtained from Alfonso Gonzalez Perez at Imperial College London.

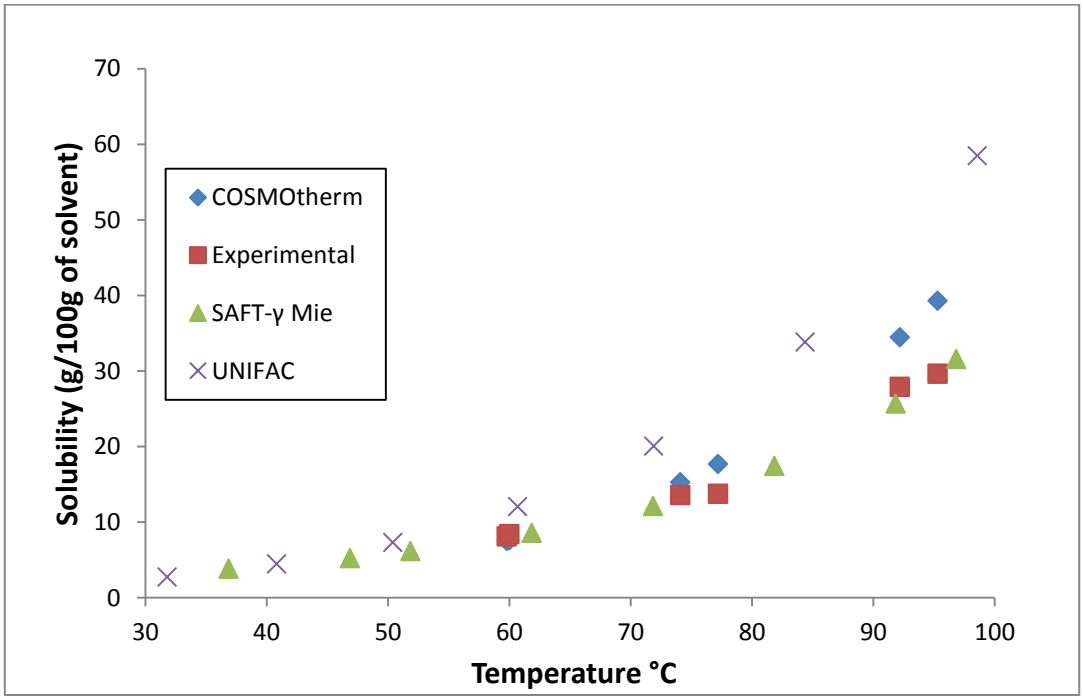


Figure 2-6 Solubility predictions for COSMOtherm, SAFT-γ Mie, UNIFAC and experimental solubility points for lovastatin and n-butyl acetate

In Figure 2-6 and Table 2-2, SAFT-γ Mie performs with the highest degree of accuracy with COSMO*therm* slightly over-predicting and UNIFAC over-predicting more than

the other methods. The extent of over-prediction increases with temperature, for this example, when compared with the experimental data.

Table 2-2 percentage error in predictive models when compared with experimental data for lovastatin in n-butylacetate

	% error lovastatin in n-butylacetate					
Temperature °C	59.8	60.0	74.1	77.2	92.2	95.3
COSMO*therm*	7.3	9.6	12.4	28.3	23.5	32.6
SAFT-γ Mie	7.5	4.5	4.5	14.5	6.1	1.7
UNIFAC	25.4	22.4	55.9	80.2	91.3	111.1

A value of 43.14 kJ/mol for enthalpy of fusion and a melting temperature of 172.35°C were used for these predictions (Nti-Gyabaah *et al.*, 2008).

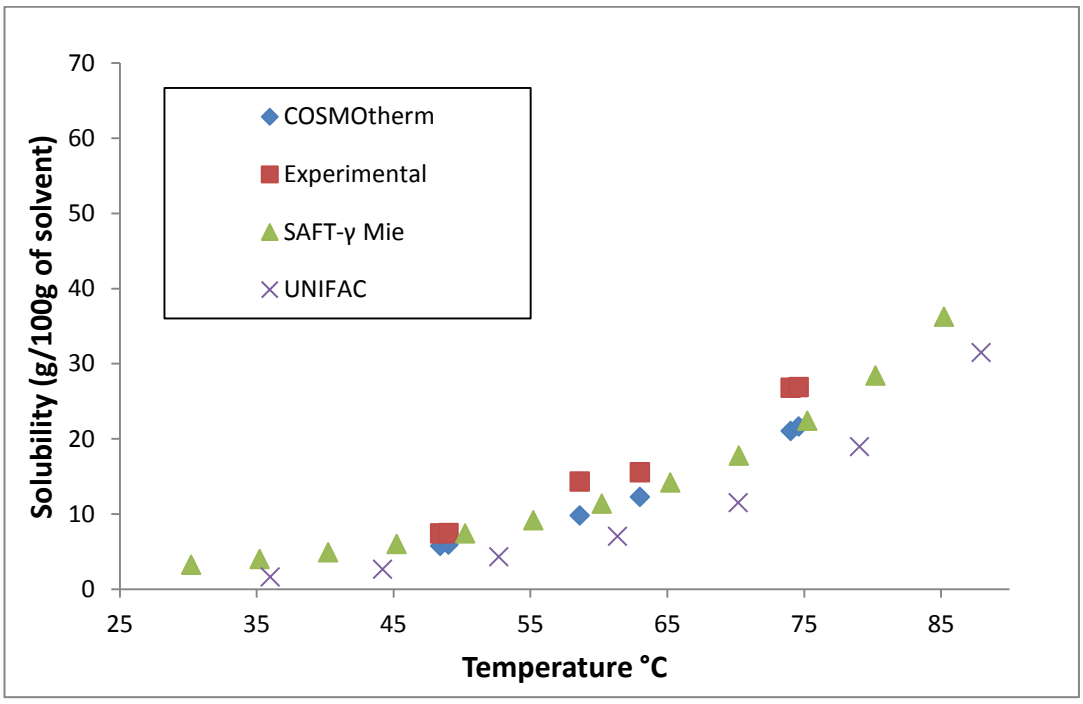


Figure 2-7 Solubility predictions for COSMOtherm, SAFT-γ Mie, UNIFAC and experimental solubility points for lovastatin and 1-butanol

Figure 2-7 shows the solubility predictions and experimental solubility values for lovastatin in 1-butanol. Table 2-3 shows the percentage error between experimental data and the predictive methods. In this example, all predictive methods have under-predicted when compared to the experimental values. SAFT-γ Mie has the least

percentage error for all temperatures in this example and UNIFAC has the largest percentage error.

Table 2-3 percentage error in predictive models when compared with experimental data for lovastatin in 1-butanol

	% error lovastatin in 1-butanol					
Temperature °C	48.4	49.0	58.6	63.0	74.0	74.6
COSMOtherm	22.6	20.7	31.3	21.0	21.5	19.4
SAFT-γ Mie	1.0	0.9	20.0	11.0	16.8	14.8
UNIFAC	57.5	56.2	59.7	51.9	46.6	44.8

Figure 2-8 shows lovastatin in 1-pentanol. In this example all three predictive methods have under-predicted and all have similar values.

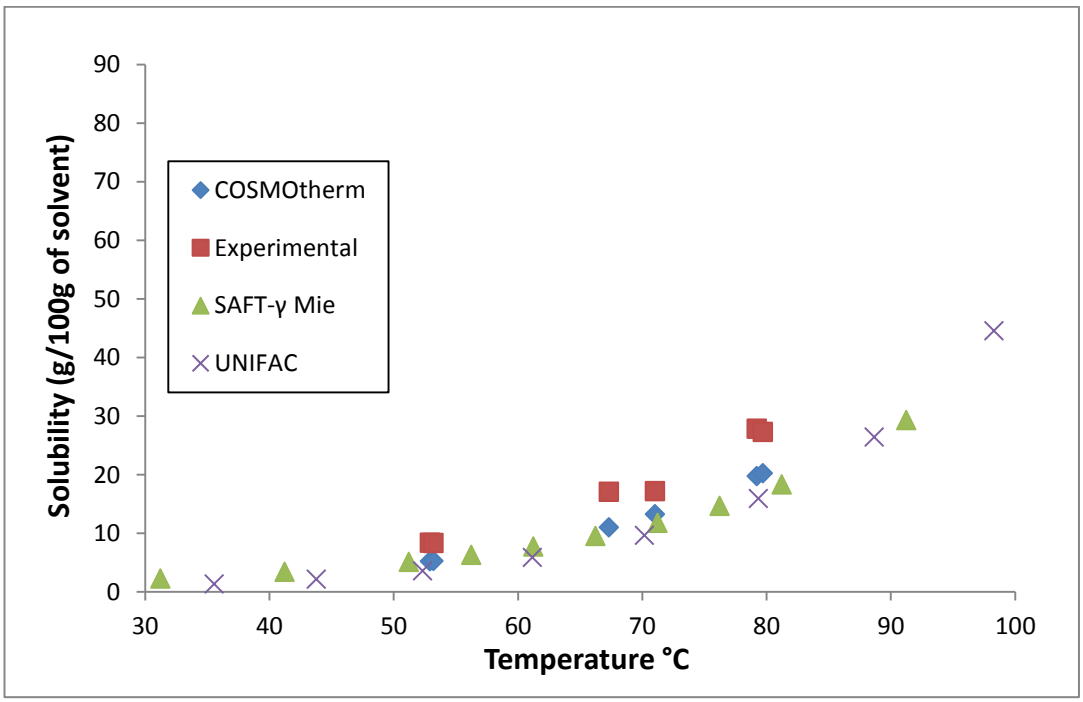


Figure 2-8 Solubility predictions for COSMOtherm, SAFT-γ Mie, UNIFAC and experimental solubility points for lovastatin and 1-pentanol

In the table below (Table 2-4) the percentage error between experimental and the predictive methods show that at lower temperatures SAFT-γ Mie has smaller errors than COSMO*therm* but as the temperature increases, SAFT-γ Mie's error increases. UNIFAC, again, gave the largest errors.

Table 2-4 percentage error in predictive models when compared with experimental data for lovastatin in 1-pentanol

	% error lovastatin in 1-pentanol					
Temperature °C	52.9	53.2	67.3	71.0	79.2	79.7
COSMO*therm*	37.7	36.5	35.4	23.0	29.0	25.8
SAFT-γ Mie	30.2	29.2	36.8	26.7	35.9	33.2
UNIFAC	58.7	57.9	53.0	42.1	42.3	39.4

In these three examples, all methods performed reasonably well when they were compared to experimental data. However, SAFT-γ Mie and UNIFAC are limited by the fact that they both rely on experimental data to parameterise functional groups or atoms; when these data are sparse, or not available, no prediction can be generated, limiting the utility of these methods. In contrast, with COSMO*therm*, an *ab initio* method, each molecule can be parameterised directly from the chemical structure; it is therefore simpler to apply COSMO*therm* "off the shelf" for any molecule.

The comparisons in this section show that COSMO*therm* can give accurate solubility predictions that are comparable with experimental and other predictive methods. Further analysis with a wide range of solutes and solvents would be needed to confirm this.

2.6.3 Solubility solvent/anti-solvent

Predicting the solubility of a compound in two solvents is essential for solvent/anti-solvent crystallisations. COSMO*therm* can predict the solubility of a compound in binary solvents and a Python script was written to automate this process. COSMO*therm* results were compared with literature results from Granberg and Rasmuson (Roger A. Granberg and Rasmuson, 2000); these results showed a good correlation between experimental and theoretical results. In the graphs, zero and one are the mole fractions of pure solvents.

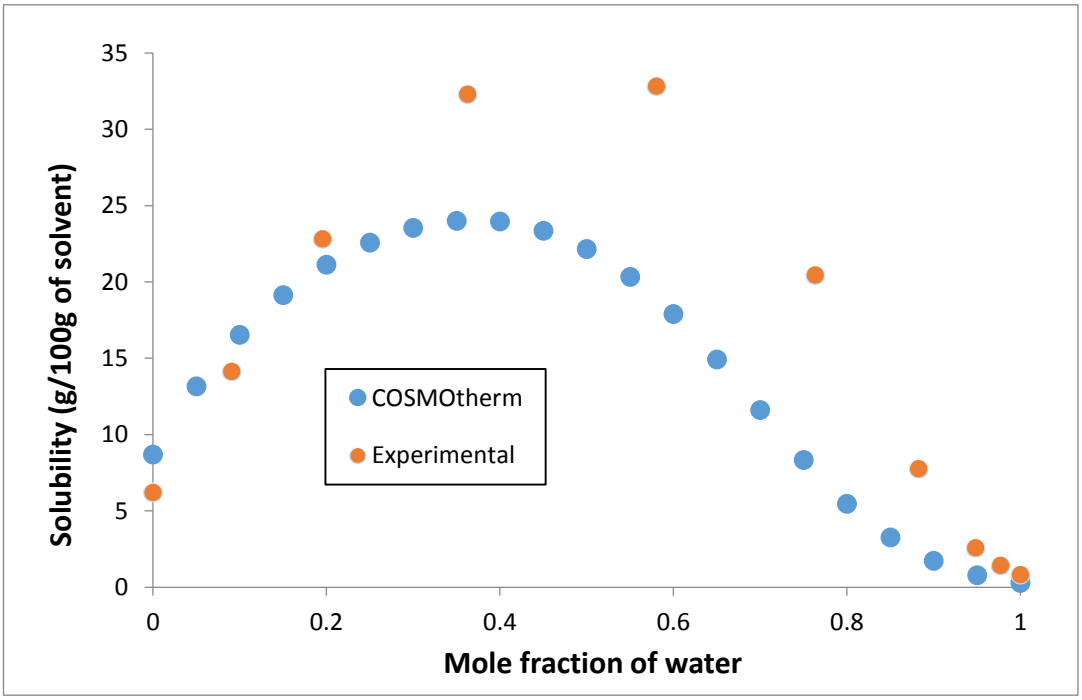


Figure 2-9 COSMOtherm solubility curve prediction and experimental solubility points for paracetamol in water and propanone at 5°C

The predictions in Figure 2-9 show the experimental maximum and the solubility in water. The experimental maximum is a third bigger at 32g/100g whereas the prediction gave 24g/100g. It would be expected that the maximum would occur at the full mole fraction of solvent with no anti-solvent. The occurrence of the maximum, with a mole fraction of both solvent and anti-solvent, has a complex

thermodynamic basis. It is a consequence of the influence of both enthalpy and entropy effects, and no definitive explanation has been achieved (Grant, 1990). However, COSMO*therm* has managed to predict the maximum's existence. A value of 28.12 kJ/mol for enthalpy of fusion and a melting temperature of 168.6°C was used for these predictions (Sacchetti, 2001).

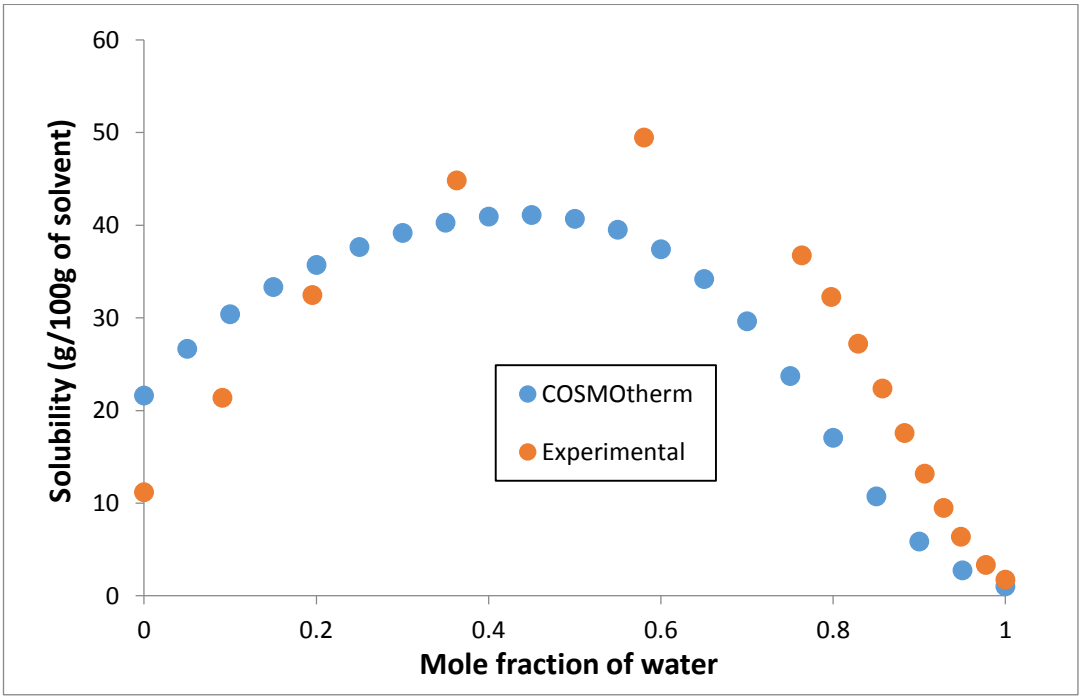


Figure 2-10 COSMOtherm solubility curve prediction and experimental solubility points for paracetamol in water and propanone at 30°C

Figure 2-10 shows the predicted and experimental solubility of paracetamol in water and propanone at 30°C. The position and the magnitude of the curve in the experimental is slightly to the right and higher. The experimental maximum at 0.58 mole fraction of water has an experimental value of 49g/100g compared with a predicted value of 37g/100g.

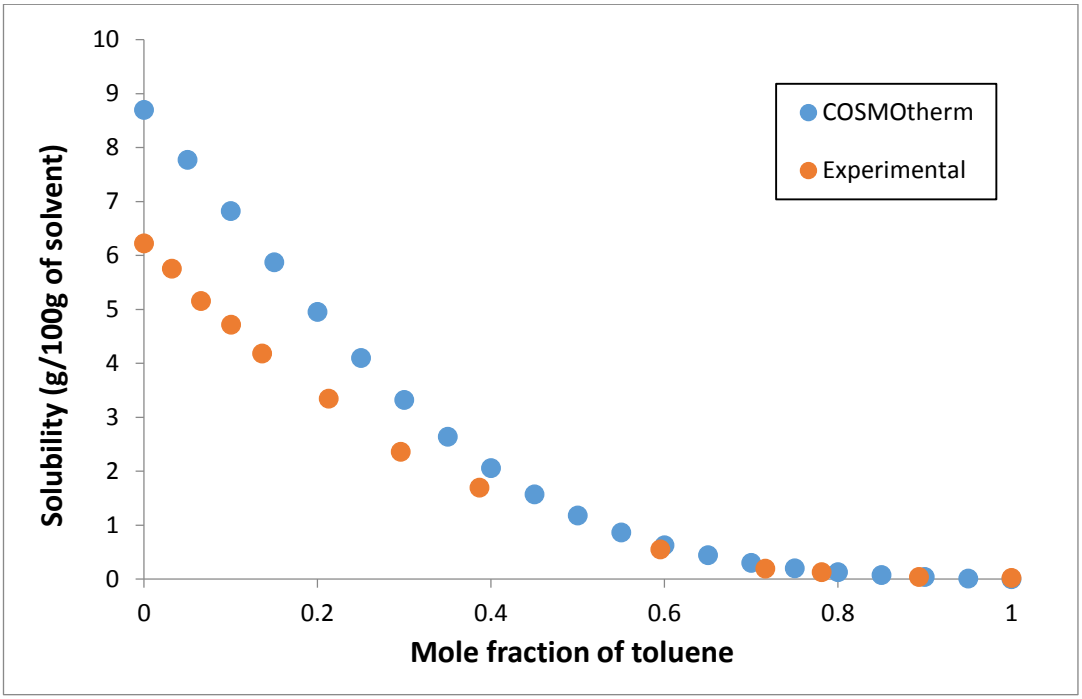


Figure 2-11 COSMOtherm solubility curve prediction and experimental solubility points for paracetamol in propanone and toluene at 5°C

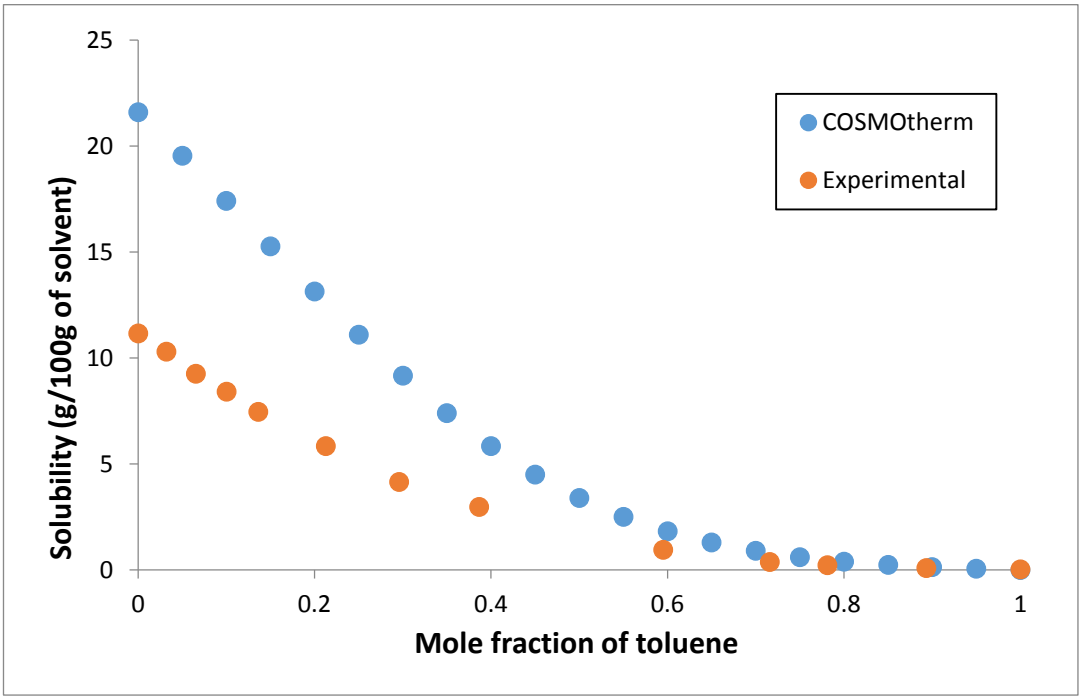


Figure 2-12 COSMOtherm solubility curve prediction and experimental solubility points for paracetamol in propanone and toluene at 30°C

Figure 2-11 and Figure 2-12 show a comparison for the predicted and experimental solubility of paracetamol in propanone and toluene (Roger A. Granberg and Rasmuson, 2000). The curves both have a similar shape as the literature value. The

initial gap between prediction and literature is 3g/100g at 5°C. However, as the temperature increases so does the error with 10g/100g error initially at 30°C although at higher mole fractions of water the data points start to converge. If the solubility predictions at zero and one mole fraction are not accurate then predicted solubilities in mixed solvents are likely to be inaccurate too.

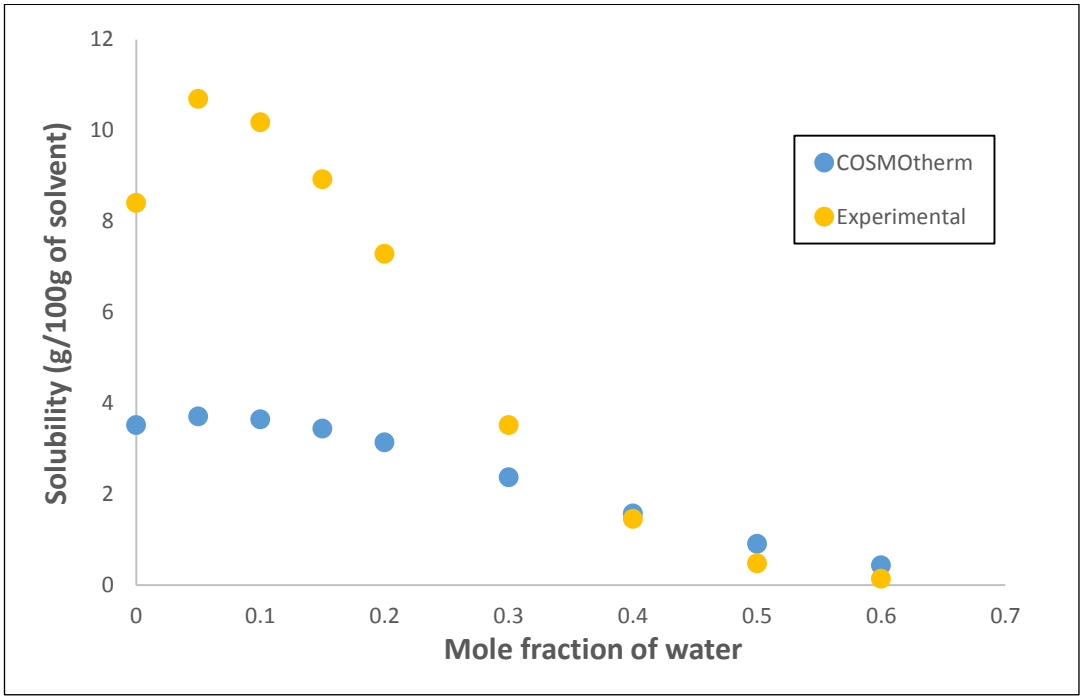


Figure 2-13 COSMOtherm solubility curve prediction and experimental solubility points for lovastatin in propanone and water at 25°C

Figure 2-13 shows the results of COSMO*therm* predictions and Dr John McGinty's experimental work done within CMAC with lovastatin. The largest error in this prediction is at low concentrations of water with an error of 6.5g/100g at 0.1 mole fraction of water. The predicted and experimental solubility points converge at approximately 0.6 mole fraction of water with an error of 0.3g/100g. The shape of the curve is, however, correct whilst the magnitude is not.

COSMO*therm* predictions for binary solvents can be a useful tool for anti-solvent crystallisations although they are not usually as accurate as a single solvent prediction due to the added dimension of a second solvent. These predictions do not remove the need for laboratory testing but can give a good starting point for solvent selection as the curves tend to be the correct "shape" and indicate the mole fraction for maximum solubility if not the solubility value itself.

2.6.4 Solvent screening

Finding a suitable solvent for cooling crystallisation follows a certain selection workflow based on the magnitude and the dependence on temperature of the solubility of the solute. For a cooling crystallisation, these criteria are critical in the pharmaceutical sector. To be selected, a compound requires, at low temperature (20°C), a solubility threshold of <5g/100ml and at high temperature (10°C below solvent boiling point) a threshold of >5g/100ml (Brown *et al.*, 2018). If the solubility at the higher temperature is in the range of 5-15g/100ml the system will be considered dilute for practical purposes. If the concentration is too high, the final slurry may be too dense or immobile (Muller,Fielding and Black, 2009).

COSMO*therm* has the potential to predict solubilities quickly for a compound in a large number of solvents. A comparison was made with the solubility classification scheme adopted in the above paper (Brown *et al.*, 2018). Using COSMO*therm* and a Python script to automate the process, predictions were obtained using the iterative job-type to decrease the time to about 15 minutes. The predictions for paracetamol were obtained for 43 solvents and the results were converted from mass to g/100 ml.

The graph below, Figure 2-14 is a subset of the 43 solvents that were screened for comparison with the solvents included in the paper. One solvent in the paper, o,m,p-xylene was not included as it is a solvent mixture and so does not have a discrete molecular structure to parameterise.

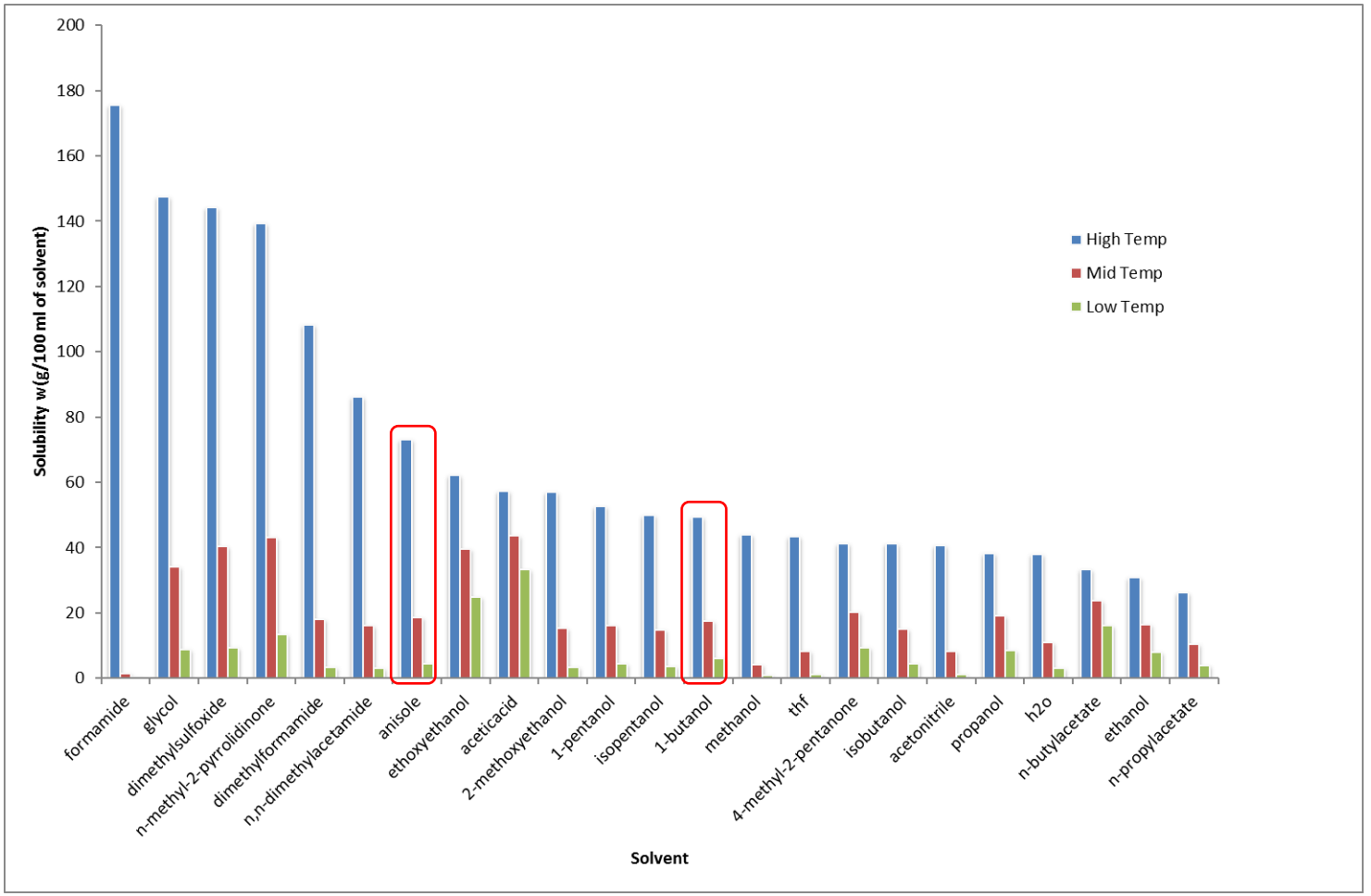


Figure 2-14 Sub-set of the solvent screen of paracetamol in multiple solvents with incorrect classifications with red boxes

The solvents in the paper were divided into three classes: low solubility at elevated and room temperatures; low solubility at room temperature and high solubility at elevated temperature; and finally high solubility at both temperatures and these have been assigned a class number for ease of discussion (Table 2-5).

Table 2-5 Classification of solvents at low and high solubility and low and high temperature

	Low solubility at room temperature	**High solubility at room temperature**
Low solubility at elevated temperature	Class 1	NA
High solubility at elevated temperature	Class 2	Class 3

Of the 43 solvents and predicted solubilities, 37 of the solvents agreed with the paper's classification and six disagreed. Of the six incorrectly classified solvents (anisole, 1-butanol, ethyl acetate, 2-propanol, methyl acetate, chlorobenzene) two solvents were misclassified as class two instead of class one (anisole and chlorobenzene). Anisole had an extremely high solubility prediction at elevated temperature. Two solvents were assigned class two instead of class three (1-butanol and 2-propanol) and one was put into class three instead of class two (methyl acetate). Ethyl acetate was wrongly classified as class three instead of class one as both predictions at elevated and room temperature were above the solubility criteria. This comparison with experimental data shows that COSMO*therm* has the ability to screen solvents with speed and accuracy. An accuracy of classification of 86% (37 out of 43) was obtained in this example.

2.6.5 Co-crystal solubility

Figure 2-15 shows the comparison of experimental work completed by Alex Cousen of CMAC using a Crystal 16, which detects turbidity, and with solubility predictions in COSMO*therm* using the co-crystal function.

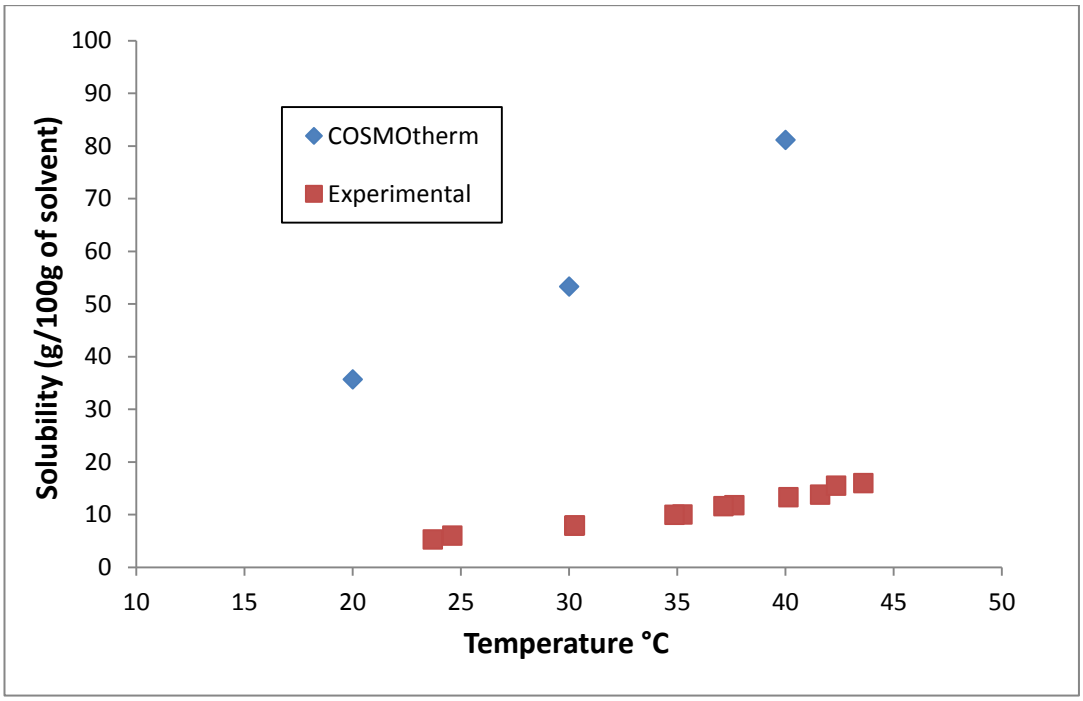


Figure 2-15 COSMOtherm solubility curve prediction and experimental data for naproxen and 2-aminopyridone 1:1 in 2-propanol

A value of 36.33 kJ/mol for enthalpy of fusion and a melting temperature of 102°C was used for these predictions (also provided by Alex Cousin). The results start with an over-prediction of around 20g/100g of solvent between experimental and predicted rising to 60g/100g of solvent. The general gradient of the curve is also wrong.

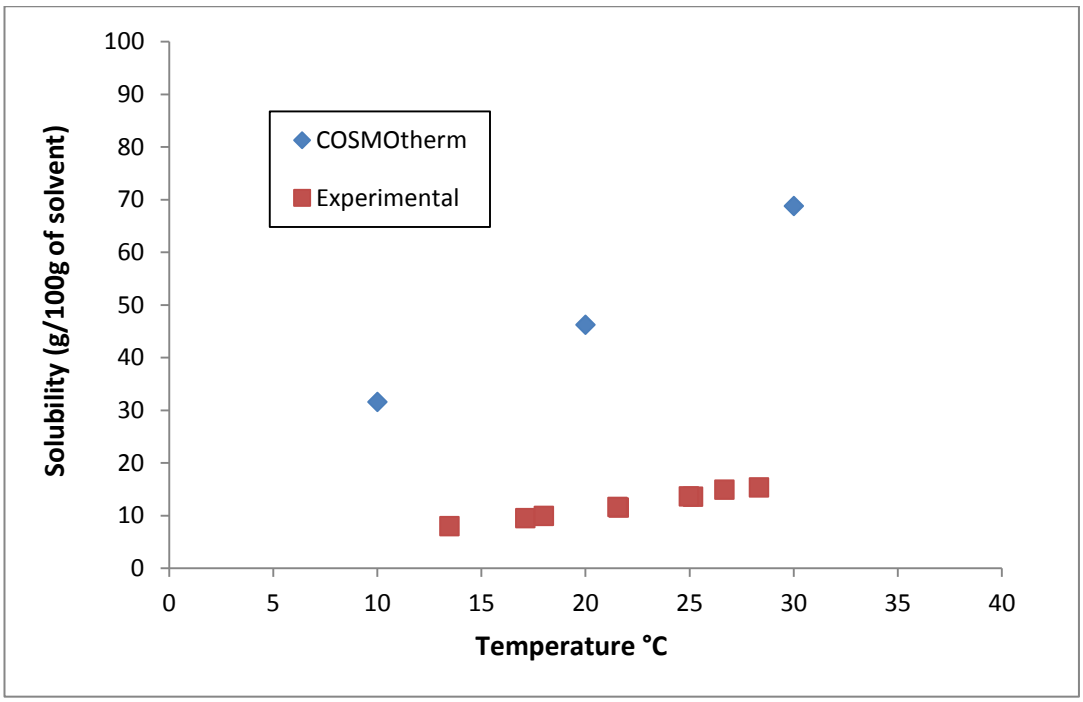


Figure 2-16 COSMOtherm solubility curve prediction and experimental data for naproxen and 2-aminopyridine 1:1 in ethanol

Figure 2-16 also shows the same problem of over-prediction and gradient mismatch. No further in-house examples of co-crystal laboratory solubility were available to make any further comparison with COSMO*therm* predictions. Co-crystal solubility predictions maybe more complicated than neutral compound solubility predictions as there are three components (two solute and one solvent) instead of two in neutral compound predictions.

Scripts for working with solid-liquid extractions, solvent screening for salts and computing ternary plots were also created but were more specialised for crystallisations and were not validated with laboratory data. These scripts are available for download (see section 9).

2.7 Enthalpy of fusion and melting temperature

For COSMO*therm* to work, the enthalpy of fusion and melting temperature must be provided for both the solute and the solvent. Ideally, the enthalpy of fusion and

melting temperature would be obtained directly from DSC data or another suitable method for each sample. If there is a lack of material, the data from literature is usually sufficient with some scrutiny. Lack of either data presents a significant challenge to predict solubility reliably. Even a small inaccuracy in the enthalpy of fusion will affect the accuracy of a prediction. There are several methods, which have already been discussed in Chapter One (section 1.3.9-1.3.11), that can be used for the prediction of both enthalpy of fusion and melting temperature. For this project a comparison between the three predictive methods was completed. The comparison between the methods and literature data showed the difference between each method. The predictions of enthalpy of fusion and the melting temperature for 60 compounds in the literature were compared using the Joback and Reid method and the COSMO*quick* method. The enthalpy of fusion predictions for the Jain and Yalkowsky method were compared for 56 compounds with literature data.

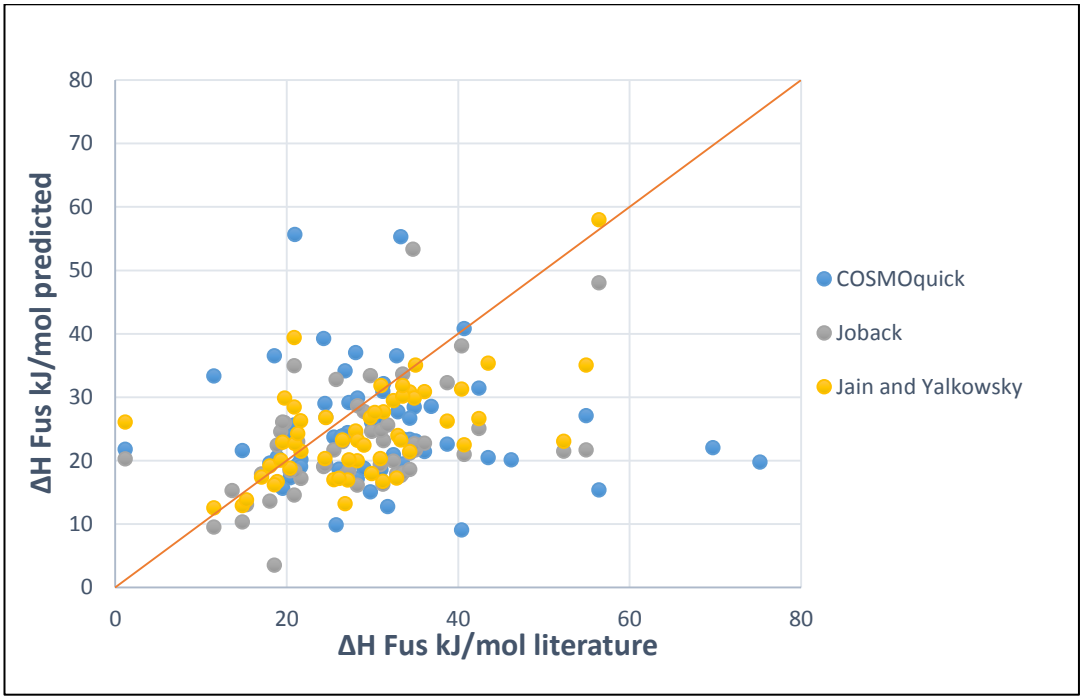


Figure 2-17 Prediction of Enthalpy of fusion with COSMOquick, Joback method and Jain and Yalkowsky method

Figure 2-17 shows the correlation between experimental results and the three modelling approaches. The nearer the data point is to the central line the more accurate the prediction. COSMO*quick* is the least accurate with Jain and Yalkowsky the most accurate. Summary statistics highlight this more clearly in Table 2-6. The Jain and Yalkowsky performed better but only had 56 molecules as a prediction could not be made for every compound due to not all functional groups having a value. Even a small change of 1 or 2 kJ/mol in the enthalpy of fusion can have a large impact on the accuracy of predictions (see co-efficients in section 3.2.2).

Table 2-6 Mean, SD deviation and Maximum deviation error of COSMOquick, Joback and Reid method and Jain and Yalkowsky for enthalpy of fusion

Method	**Error enthalpy of fusion kJ/mol**			**Range of predictions kJ/mol**
	Mean	**Standard deviation**	**Maximum deviation**	
COSMO*quick*	11.82	11.85	55.43	9.07 – 55.71
Joback and Reid	8.74	6.93	33.17	3.52 – 69.15
Jain and Yalkowsky	7.11	6.45	29.22	12.90 – 39.43

The COSMO*quick* regression model is clearly the inferior method for enthalpy of fusion prediction. The Jain and Yalkowsky method is more accurate than the Joback and Reid method. However, the latter method is quicker and easier to use. All that is required is the summation of the number of functional groups in the molecule, whereas the first method requires more parameters as the adjacent functional group or groups are accounted for. The main weakness of the Joback method, which has been previously stated in Chapter One (section 1.3.9), is that as the method is cumulative; a very large molecule will have a large enthalpy of fusion and melting temperature prediction and this is not always reflected in practice.

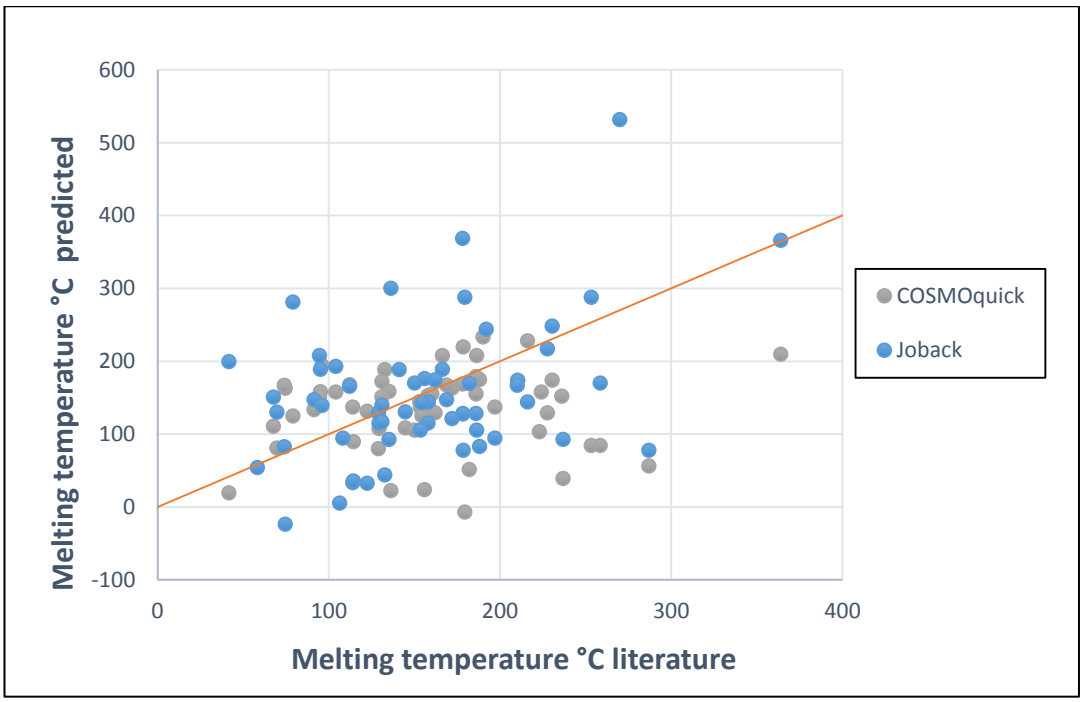


Figure 2-18 Correlation of literature and predictive methods for melting temperature

Figure 2-18 shows the correlation between experimental melting temperatures and predictions using COSMO*quick* and the Joback and Reid method. Here, the Joback method over-predicts melting temperature with the COSMO*quick* method under-predicting.

Table 2-7 Mean, SD deviation and Maximum deviation error of COSMOquick, Joback and Reid method for melting temperature

Method	**+/- Error in melting temperature (°C)**		
	Mean	**Standard deviation**	**Maximum deviation**
COSMO*quick*	59.08	53.95	230.54
Joback and Reid	73.41	67.42	271.57

When predicting melting temperatures, both methods (Table 2-7) have large standard deviations of 53.95°C and 67.42°C for COSMO*quick* and the Joback and Reid method respectively.

The COSMO*quick* method is more accurate than the Joback and Reid method. This is due to the Joback and Reid method being a contribution method: the larger the

molecule, the higher the melting temperature prediction leading to huge deviations in larger molecules.

The modelling approaches do not give a reliable figure for either enthalpy of fusion or melting temperature. Joback and Reid's study (Joback and Reid, 1987) had an average error of 8.4 kJ/mol for enthalpy of fusion with a standard deviation of 18 kJ/mol when predicting for 378 compounds and melting temperature error of 4.8°C with a standard deviation of 6.9°C when predicting for 409 compounds. Jain and Yalkowsky's study had an error of 2.91 kJ/mol when predicting enthalpy of fusion for 2230 compounds (Jain and Yalkowsky, 2006). Unfortunately, if for some compounds an enthalpy of fusion prediction is all that is available this can possibly make some solubility predictions inaccurate. However, most compounds have a melting temperature available in literature.

The experiment below (Figure 2-19) shows how sensitive the predictions from COSMO*therm* are to changes in the value of both enthalpy of fusion and melting temperature highlighting the need for accurate measurements of both values for solubility predictions.

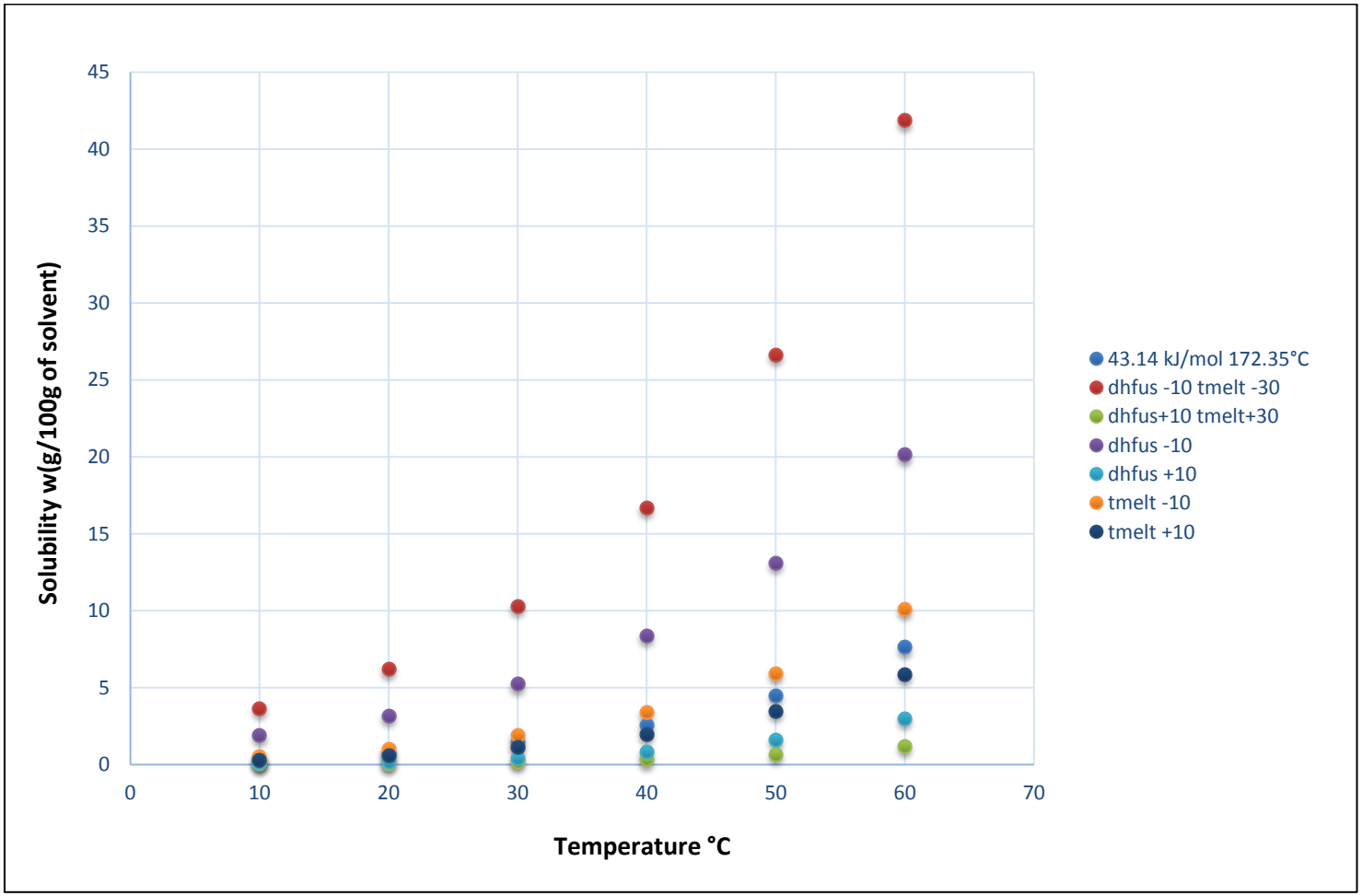


Figure 2-19 Plot showing the difference in solubility predictions when parameters are changed for lovastatin in 1-pentanol

The value of both enthalpy of fusion and melting temperature for lovastatin was varied by +/- 10 kJ/mol and +/- 10°C or +/- 30°C from literature values. A change of 10 kJ/mol in the enthalpy of fusion tripled the solubility prediction in this example. Therefore, even if the measured or predicted value of enthalpy of fusion, and to a lesser extent melting temperature, is slightly inaccurate this can have a large effect on the value of the prediction. If the melting temperature is also incorrect this will also make the predictions inaccurate although there is a lesser effect than for enthalpy of fusion. This experiment shows the need for accurate measurement of enthalpy of fusion and melting temperature.

2.8 Assessment of COSMO*therm*

A basis set in computational chemistry is a set of functions that is used to represent atomic (and subsequently molecular) wave functions and to turn them into algebraic equations suitable for use in computational calculations. The basis functions represent atomic orbitals, which combined to form molecular orbitals and are, by extension, how COSMO*therm* "views" the interacting surfaces of each compound in prediction calculations; these are described in Chapter One section 1.3.2.

COSMO*conf* can generate COSMO files with three different basis sets: SVP, TZVP and TZVPD-fine. There are "higher" basis sets available parameterised using other applications, such as quadruple-zeta valence polarisation (QZVP) but none were used in this project and COSMO*conf* does not parameterise or handle them readily.

COSMO*therm* has three different options for job-types for solubility calculations: solid-liquid equilibrium solubility (SLESOL); non-iterative; and iterative. SLESOL solves the SLE

and is the most expensive computationally. “Non-iterative” and “iterative” solubility predictions are approximations of the SLE. In non-iterative the chemical potential reference at infinite dilution is taken as an approximation of the chemical potential. For iterative the chemical potential at infinite dilution is taken as a starting point and the chemical potential is calculated by an iterative process; which is a repeated cycle of operations each new iteration using the answer from the previous iteration in the calculation. To establish the merits of both the different basis sets and the different job-types a comparative study was conducted using a dataset of 369 solubility points obtained from laboratory and literature data for seven solutes and 31 solvents at various temperatures.

Table 2-8 log RMSE error in solubility predictions for job-types and Basis Set

Job-type	**Basis set (RMSE log scale)**		
	SVP	**TZVP**	**TZVPD-fine**
Non iterative	0.91	1.01	0.71
Iterative	0.77	0.78	0.69
SLESOL	0.77	0.78	0.69

The jobs were run for each job-type and each basis set and were compared with the solubility data from literature (Table 2-8). TZVPD-fine was the best for all job-types with a large error decrease of 0.3 log units compared with TZVP for the non-iterative job-type. The other two job-types had an improvement for TZVPD-fine with an improvement in RMSE error of log 0.02. TZVP was the least accurate with the greatest log error for each job-type and the second most accurate was SVP. This was surprising as TZVP is supposed to perform better than SVP (Frank Eckert, 2015) although iterative and SLESOL jobs have identical results. This may be a result of the dataset used being too small. This

improvement in performance does have a computational cost with all TZVPD-fine calculations taking considerably longer than the same job-type using SVP and TZVP basis sets (Table 2-9).

Table 2-9 Total Time taken for Job-types and Basis sets for 269 solubility predictions

Job-type	**Basis set (time taken in hours, minutes and seconds)**		
	SVP	**TZVP**	**TZVPD-fine**
Non iterative	00:03:21	00:03:42	00:06:58
Iterative	00:06:31	00:08:45	00:21:32
SLESOL	01:25:38	01:08:12	05:43:27

For the rest of this project all predictions used TZVPD-fine basis set and, unless otherwise stated, used SLESOL job-type.

2.9 Heat capacity

The solubility equation has a heat capacity term ΔC_p (Equation 25) which has been discussed in section 1.3.8.

Equation 25

$$\ln a = lnx^{sat} + ln\gamma^{sat} = \frac{\Delta^{fus}H(T_m)}{R}\left(\frac{1}{T_m} - \frac{1}{T}\right) + \frac{1}{R}\int_{T_m}^{T}\frac{\Delta C_p}{T}dT - \frac{1}{RT}\int_{T_m}^{T}\Delta C_p\, dT$$

COSMO*therm* has a function to input the heat-capacity in the .vap file with the intention to increase the accuracy of predictions. The availability of heat capacity in the literature is very sporadic and for many compounds the figure was unavailable.

There are several options for handling heat capacity: firstly, a temperature dependant heat capacity for each temperature; secondly, a fixed heat capacity over a range of temperatures; another option is for a heat capacity estimate to be input into COSMO*therm* (Equation 26); and lastly to not include heat capacity in the predictions.

The estimate of heat capacity is thus:

Equation 26

$$\Delta Cp_{fus} = \frac{\Delta H_{fus}}{TmeltK}$$

Several fixed heat capacities for seven solutes were obtained from Schroder (Schroder *et al.*, 2010) and the values input into the appropriate .vap file for each solute and COSMO*therm* jobs were run for 364 solubility points (27 solvents at various temperatures); the jobs were rerun without the heat capacity in the .vap file for comparison (Table 2-10). These predictions were compared with laboratory and literature data.

Table 2-10 RMSE for COSMOtherm predictions with and without literature heat capacity

	Without heat capacity literature value	**With heat capacity literature value**
No. of solubility points	264	264
RMSE (log)	0.97	0.77
Standard Deviation	0.86	0.77
Max. Deviation	3.50	3.10

The log RMSE for the predictions decreased when the heat capacity was applied and the prediction was improved in around 60% of cases. This shows that the heat capacity should not be ignored, however due to the value for heat capacity not being readily available for most compounds heat capacity was not used for further predictions.

To establish whether the estimate for heat capacity improved the COSMO*therm* predictions, experimental data were compared with predictions for 1717 solubility points with and without the estimate (57 solutes and 48 solvents at various temperatures).

Table 2-11 RMSE for COSMOtherm predictions with and without heat capacity estimate

	Without heat capacity estimate	**With heat capacity estimate**
No. of solubility points	1717	1717
RMSE (log)	0.90	0.98
Standard Deviation	0.89	0.89
Max. Deviation	3.17	3.26

As can be seen from the table above (Table 2-11) using the estimate increased the RMSE of the COSMO*therm* solubility predictions from log 0.90 to 0.98 and increased the maximum deviation from log 3.17 to 3.26, therefore using the estimate has failed to improve the predictions and the estimate was not used for further predictions.

2.10 Conclusion

The aim of this chapter was to assess the suitability of COSMO*therm* for use in medicine manufacture and to compare this method with other methods available (UNIFAC and SAFT-γ Mie). To do this it was essential to automate COSMO*therm* and to assess uncertainty in experimental measurements of solubility and parameter sensitivity. When comparing COSMO*therm* with experimental solubility predictions there are some errors introduced into the system and it is important that these errors are minimised. The error associated when measuring solubility in the laboratory is likely a contributing factor to the error between experimental and predicted error. For each compound that requires solubility predictions, enthalpy of fusion and melting temperature data are needed. Ideally, for each batch of solute DSC data would be obtained. High quality DSC data for each compound is preferable to literature data, which may not be for the polymorph considered, or the method of measurement has not been specified. Error in measuring enthalpy of fusion and melting temperature could also introduce error

directly into the predictions, as they are two of the main adjustable parameters for COSMO*therm*. Even slight changes in the value of enthalpy of fusion and melting temperature could affect the prediction (see co-efficients in section 3.2.2) The error that accumulates and propagates throughout can, to some extent, be corrected by using a ML approach similar to the models discussed in Chapter Four.

From the comparisons of different modelling approaches all three predictive methods tested (UNIFAC, SAFT-γ Mie and COSMO*therm*) performed well in the examples in this chapter. However, performance would be dependent on the solute/solvent system and as this study only included lovastatin, the comparative performance of each method in other systems is unknown. COSMO*therm* does not rely on functional groups or atoms being parameterised using empirical data and therefore can be used for most molecules. COSMO*therm* has been fully automated for predicting the solubility of neutral compounds, binary solvent systems, solvent screening and salts/co-crystal systems. It is also an efficient and versatile tool that compares favourably with both UNIFAC and SAFT-γ Mie. COSMO*therm* can currently perform predictions on more molecules than the other two methods, as it does not rely on functional groups being parameterised, which for large pharmaceutical molecules with varied functional groups makes it a favourable choice for use in medicine development. This however might change in time as more functional groups are parametrised for the other methods.

3 Using a design of experiment approach to develop a model of COSMO*therm* using linear regression

3.1 Aims

The aim in this chapter was to use DoE and linear regression modelling to apply to COSMO*therm* predictions by changing the temperature, melting temperature and enthalpy of fusion. Using the simple equations obtained from linear regression, the solubility of any solute/solvent system in the set could be predicted for any value of temperature, melting temperature and enthalpy of fusion within range.

This method was intended for use by non-expert and modellers alike to reproduce predictions of COSMO*therm* without the use of the specialist. The adjustable parameters allowed for predictions highly specific to the individual use case: for example, when a new polymorph of a solute has been obtained with a different melting temperature and/or enthalpy of fusion.

3.2 Design of experiment

DoE is an approach for conducting experiments, which is used in numerous industries in the development of new products (Montgomery, 2008). This approach, when applied correctly, can decrease development and production costs and shorten time to market. DoE is used when a number of independent variables or factors need to be manipulated by changing the levels or settings of each factor to optimise one or more dependant variables. DoE addresses the factors to be tested, the levels of these factors and the structure and layout of experimental runs. DoE differs from the One Factor at a Time (OFAT) method in that it includes, in this case, all eight two-level factor combinations. With OFAT it is assumed that all factors are independent of one another. In reality,

factors do not always act independently and this is reflected in DoE. For example, an OFAT experiment with three factors would only have four experiments instead of at least eight (nine if the centre experiment is counted). The DoE approach would assess any binary interactions between factors.

DoE was used to change the experimental parameters in COSMO*therm* to build the linear regression models for a solubility interface and database. Three experimental factors were varied to model each system's solubility: temperature, enthalpy of fusion and melting temperature. A two-level factorial design approach was used: each factor was assigned low level (-1), high level (+1) and centre level (0) values and the solubility predicted *via* COSMO*therm* for specific combinations (Figure 3-1 and Table 3-1). The centre level was set to the value for enthalpy of fusion and melting temperature held in the .vap file (*i.e.* the real values found by experiment). Generally, the centre level is provided with an additional two replicates in order to factor in experimental variance, but as this experiment was a COSMO*therm* calculation with no variance, this was not applicable.

Table 3-1 Design matrix for COSMOtherm experiments

Experiment no.	Temperature	Enthalpy of fusion	Melting temperature
1	-1	-1	-1
2	+1	-1	-1
3	-1	+1	-1
4	+1	+1	-1
5	-1	-1	+1
6	+1	-1	+1
7	-1	+1	+1
8	+1	+1	+1
9	0	0	0

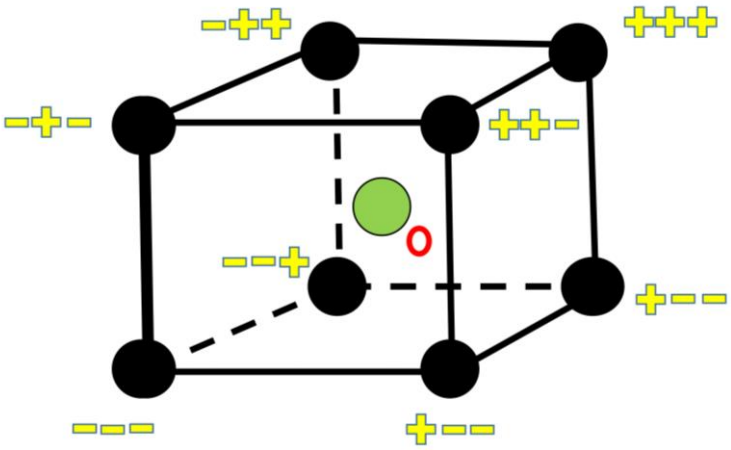


Figure 3-1 Design space for design of experiments

For some experiments COSMO*therm* returned a value of "NA", which means COSMO*therm* has been unable to give a solubility prediction, therefore one or more experimental results would be useless and regression models could not be fit. The experiments were then rerun with a reduction in the extremes for the DoEs with a maximum of three reruns for each combination.

Table 3-2 Change of extremes for design of experiment

Run number	**Melting temperature (°C) difference from centre point**	**Change in enthalpy of fusion (kJ/mol) from centre point**
1	30	20
2	30	15
3	20	10
4	10	10

These extremes changes (Table 3-2), where COSMO*therm* returned an "NA" value, were applied in most cases. There were exceptions if the enthalpy of fusion was below a certain number *i.e.* 20 kJ/mol. In those cases, the initial run was completed using the values for run two to four depending on the values of enthalpy of fusion and melting temperature. Temperature had "-1" at the laboratory temperature of 20°C and a "+1"

at 90°C if the boiling temperature of the solvent was above 100°C or 10°C below boiling point if it was below 100°C. The centre point for temperature would be the midpoint between low and high temperature in most cases 55°C.

The COSMO*therm* database discussed in Chapter Two (section 2.4), which was created initially using solutes and solvents commonly used by CMAC, and added to periodically with compounds of interest, contained 109 solutes and 136 solvents when the linear regression model was built, amounting to 14,824 different solute/solvent combinations. With nine experiments per combination there were therefore 133,416 COSMO*therm* solubility points required if solute/solvents regression models were to be completed for all permutations. While every point was attempted, due to COSMO*therm* returning an "NA" value for some solubility points even after reducing the extremes of the factors in the design of experiment, the final number of models generated was lower (Table 3-3).

Table 3-3 No of linear regression models completed

Solubility points completed	115,991
Solubility points "NA"	17,425
Models with all "NA" results	580
Models with some "NA" results	3,155
Successfully completed models	11,089
Total models	14,824

Table 3-4 shows the percentage of linear regression model with their R^2 categories. In total 98% of the models were above R^2 0.95 and performed similarly to the COSMO*therm* predictions. The 2% below R^2 0.95 had a significant number of outliers in the COSMO*therm* predictions that consequently lowered the quality of the models.

Table 3-4 percentages for categories of R^2 values for linear regression models

R^2 value	% of models
1	18
Between 1 and 0.95	80
Between 0.90 and 0.95	1
Between 0 and 0.90	1

Three of the generated models are presented in the following sections and analysed in detail. The first two show good agreement with the predictions generated by COSMO*therm*, while the third demonstrates the significant impact of a "bad" input value on the resulting model.

3.2.1 Indomethacin in 1,3-dioxolan-2-one regression model

Figure 3-2 shows the results file for indomethacin and 1,3-dioxalan-2-one with the 9 experiments each with the changes in value for all factors with the results. These results are used to calculate the linear regression model equations.

indomethacin
1,3-dioxolan-2-one
COSMOtherm 16 TZVP-D basis set solub=w(g/100g of solvent) Job Type: SLESOL

Exp_no	Temp	Dhfus	t_melt	log_solub	solub
1	20.0	16.852	130.1	1.15292838867	14.22094277
2	90.0	16.852	130.1	2.68844788864	488.03153842
3	20.0	56.852	130.1	-0.938296629996	0.11526657
4	90.0	56.852	130.1	1.47816391575	30.072111
5	20.0	16.852	190.1	0.792007669989	6.19452015
6	90.0	16.852	190.1	2.00570214383	101.3216244
7	20.0	56.852	190.1	-1.89330266456	0.0127849
8	90.0	56.852	190.1	0.355480042395	2.26714889
9	55.0	36.852	160.1	0.597947822467	3.96230427

Figure 3-2 indomethacin with 1,3-dioxolan-2-one model results file (Exp_no is experiment number, Temp is temperature, Dhfus is enthalpy of fusion, t_melt is melting temperature, log_solub is log solubility g/100g of solvent and solub is solubility in g/100g solvent)

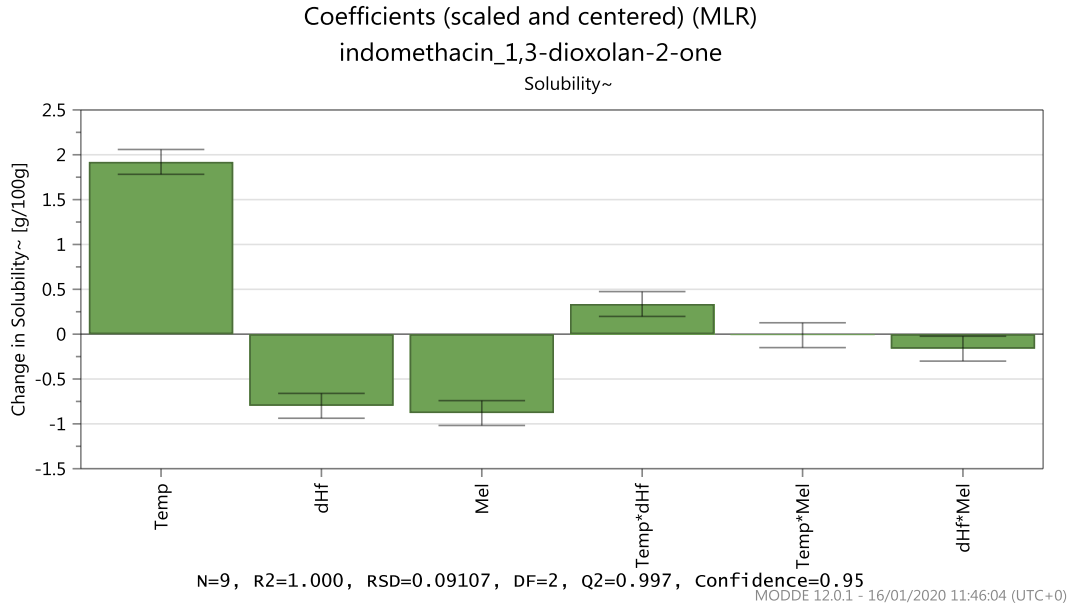


Figure 3-3 Coefficients for indomethacin and 1,3-dioxolan-2-one (Temp is Temperature, dHf is enthalpy of fusion, Mel is melting temperature)

Figure 3-3 shows the coefficient plot for indomethacin and 1,3-dioxolan-2-one. Each bar shows the influence of each factor coefficient on the solubility prediction. Temperature has the biggest influence on the solubility with an increase of temperature increasing the solubility. The size of the coefficient represents the change in the response when a factor varies from zero to one while the other factors are kept at their averages *e.g.* in the above example an increase in temperature of 1°C will increase the solubility by 1.8 g/100g. Melting temperature and enthalpy of fusion have an inverse influence with an increase of both factors leading to a decrease of solubility. The interaction terms for this solute/solvent combination have a varied influence with temperature*enthalpy of fusion increasing solubility with increased value. As enthalpy of fusion*melting temperature decreases solubility increases and for this combination temperature*melting temperature has little or no effect.

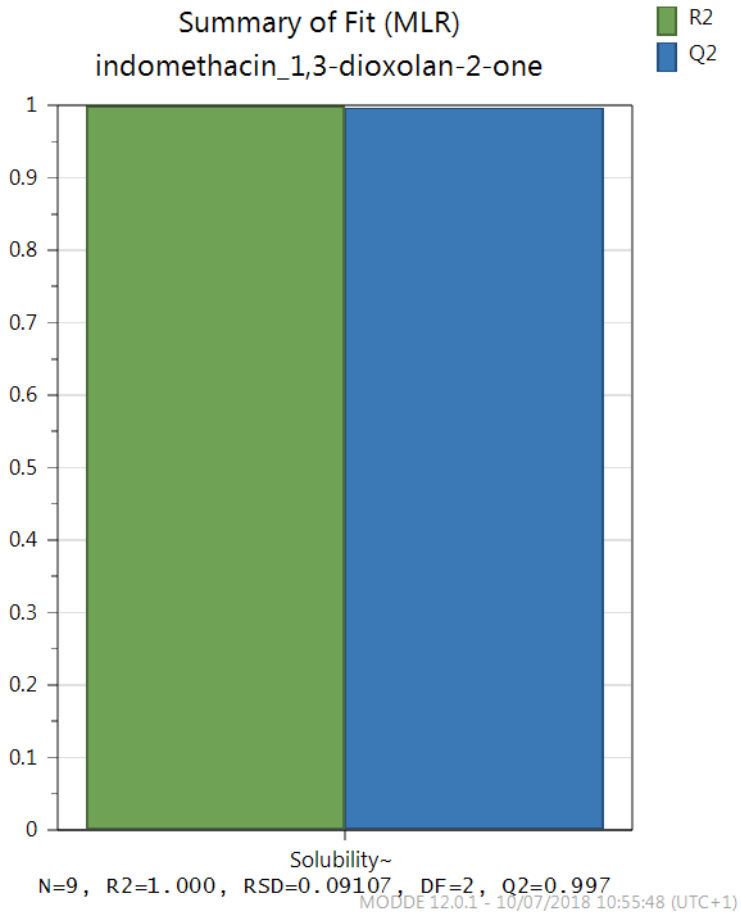


Figure 3-4 Summary of fit graph for indomethacin in 1,3-dioxolan-2-one

The summary of fit graph Figure 3-4 shows the R^2 and Q^2 values for indomethacin and 1,3-dioxolan-2-one. R^2 shows how well the model fits real data points and Q^2 estimates how well the model predicts new solubilty values: it is 1 minus the variation of the response predicated by the model according to cross-validation. A useful model will therefore have a large Q^2. Here, a R^2 value of 1 shows an excellent fit and a Q^2 value of 0.997 shows near-perfect prediction in cross-validation.

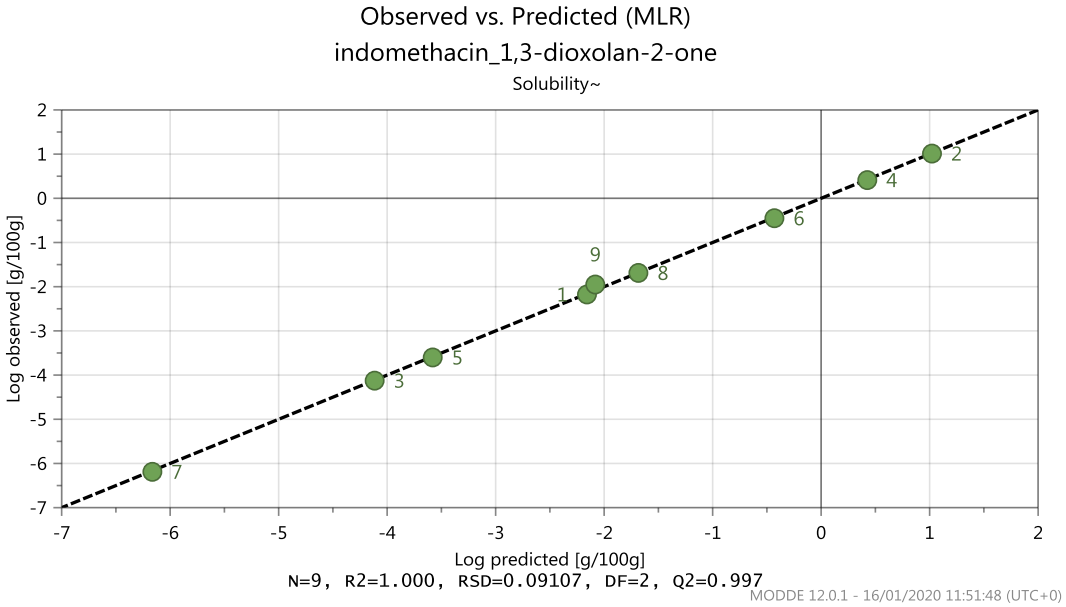


Figure 3-5 Plot of the observed v's predicted log solubilities for indomethacin in 1,3-dioxolan-2-one

The above figure (Figure 3-5) shows how well the solubility values obtained from COSMO*therm* fit with the regression model generated by MODDE. All of the experiments were on the regression line except for number nine which has a minor deviation.

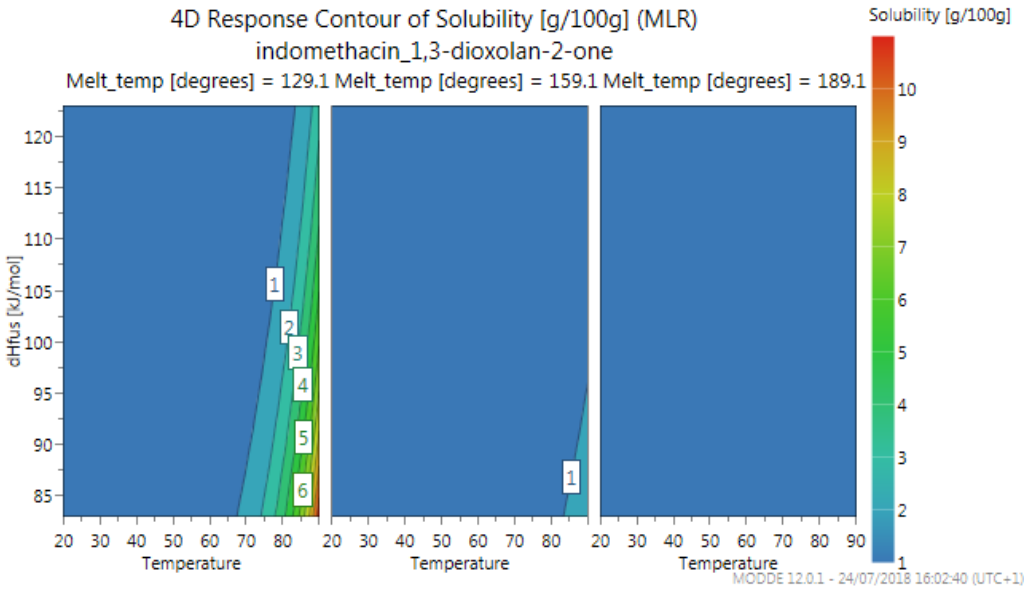


Figure 3-6 4D contour surface plot showing the impact of varying temperature, enthalpy of fusion and melting temperature on the solubility of indomethacin and 1,3-dioxolan-2-one. All temperatures are in °C

The 4D surface plot (Figure 3-6) helps visualise the effect of varying the input parameters on the solubility of indomethacin and 1,3-dioxolan-2-one. The first plot at 129.1°C clearly shows that as temperature increases so too does solubility but with the other two plots the solubility is very low due to the high melting temperatures. In the second plot which is at a higher temperature (159.1°C), a decrease in the value of enthalpy of fusion increases solubility.

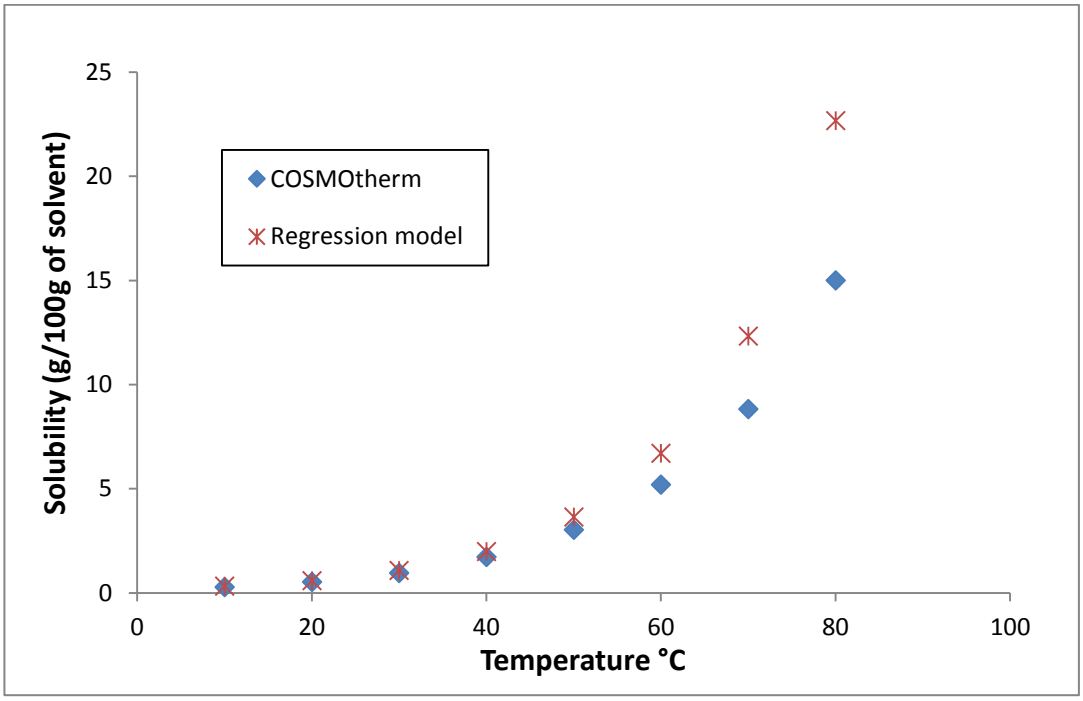


Figure 3-7 Comparison of COSMOtherm solubility curve with the regression model for indomethacin in 1,3-dioxolan-2-one

Figure 3-7 shows the results of a comparison of COSMO*therm* with the regression model for indomethacin in 1,3-dioxolan-2-one using the same input parameters of temperature, enthalpy of fusion and melting temperature. The model shows good comparison with COSMO*therm* at low temperature but starts to diverge around 50°C.

3.2.2 Saccharin in anisole regression model

Figure 3-8 shows the coefficients for the saccharin in anisole model with nine experiments.

saccharin
anisole
COSMOtherm 16 TZVP-D basis set solub=w(g/100g of solvent)

Exp_no	Temp	Dhfus	t_melt	log_solub	solub
1	20.0	19.89	207.59	0.18638138741	1.53596524
2	90.0	19.89	207.59	1.12146403016	13.22708152
3	20.0	39.89	207.59	-1.25824821281	0.0551762
4	90.0	39.89	207.59	0.263193617899	1.83313149
5	20.0	19.89	247.59	0.00570567401279	1.01322448
6	90.0	19.89	247.59	0.893903787517	7.83256103
7	20.0	39.89	247.59	-1.59340719648	0.02550309
8	90.0	39.89	247.59	-0.0938976967783	0.80556818
9	55.0	29.89	227.59	-0.041262873094	0.90936268

Figure 3-8 saccharin in anisole model results file (Exp_no is experiment number, Temp is temperature, Dhfus is enthalpy of fusion, t_melt is melting temperature, log_solub is log solubility g/100g of solvent and solub is solubility in g/100g solvent)

Figure 3-9 shows the coefficients for saccharin in toluene. The biggest influence for this solute/solvent combination is the enthalpy of fusion with solubility increasing with a reduction in this value.

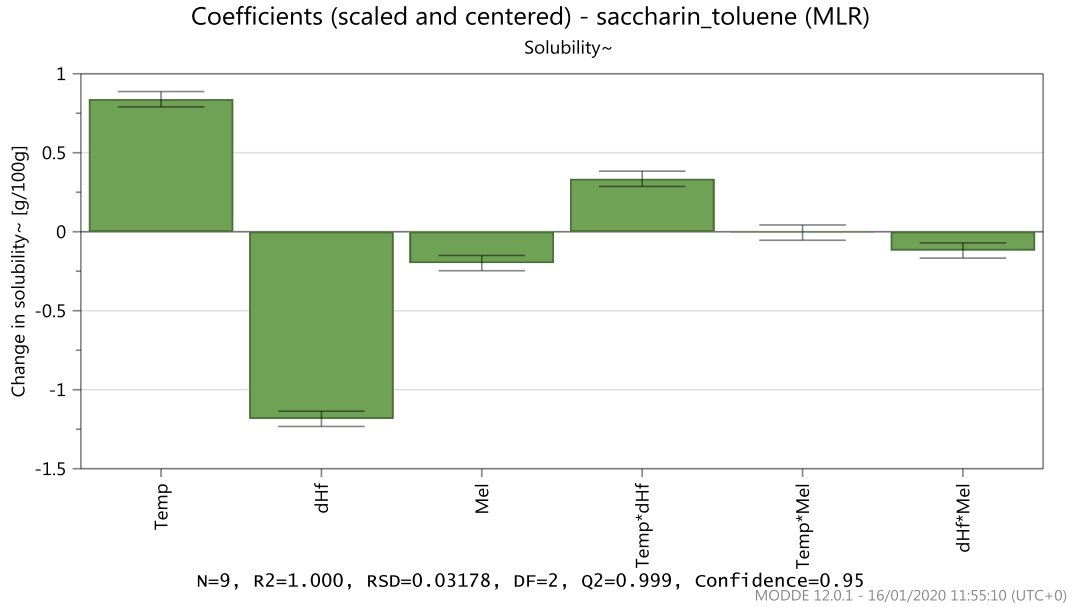


Figure 3-9 Coefficients for saccharin in toluene (Temp is Temperature, dHf is enthalpy of fusion, Mel is melting temperature)

Secondly, temperature is next largest coefficient with solubility increasing with an increase of temperature. The next biggest influence was temperature*enthalpy of fusion with a positive correlation and the melting temperature with a negative correlation. Enthalpy of fusion*melting temperature has a small negative correlation while temperature*melting temperature has little effect on solubility.

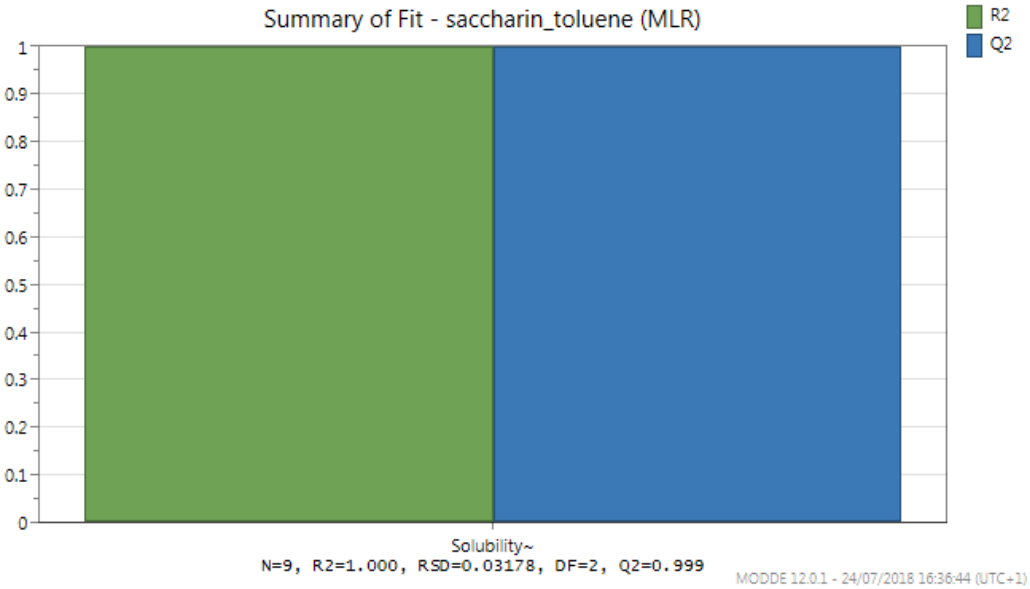


Figure 3-10 Summary of fit graph for saccharin in toluene

The summary of fit graph (Figure 3-10) shows an excellent R^2 value with a value of one and a good value for Q^2 with a value of 0.999 so this is a good model when compared with the results from COSMO*therm*.

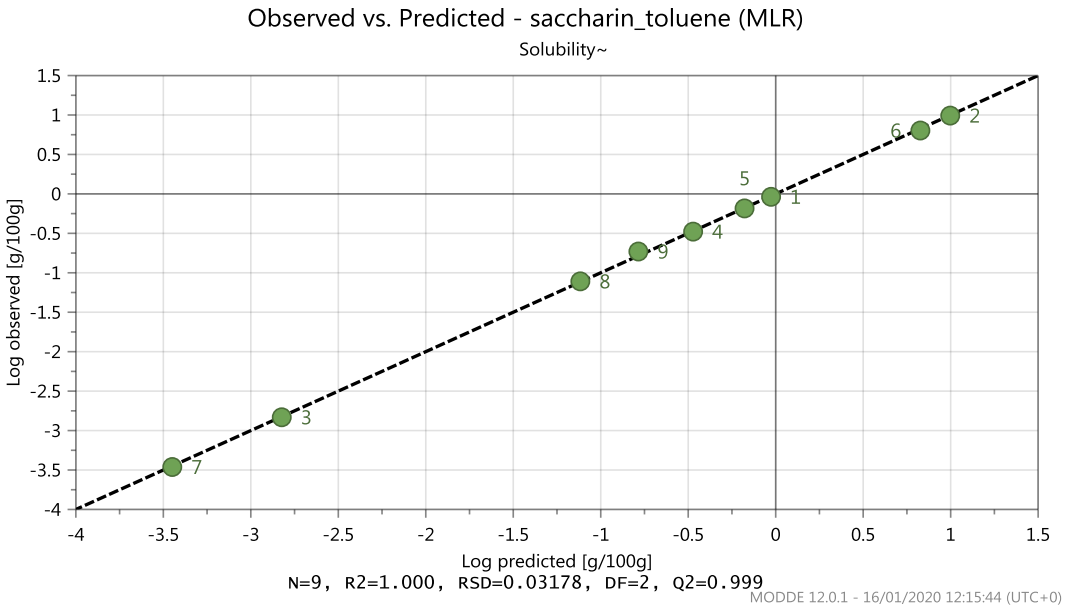


Figure 3-11 Plot of the observed v's predicted log values for saccharin in anisole

Figure 3-11 shows the observed versus predicted for saccharin in toluene. The model fits COSMO*therm* predictions well (R^2=1.000).

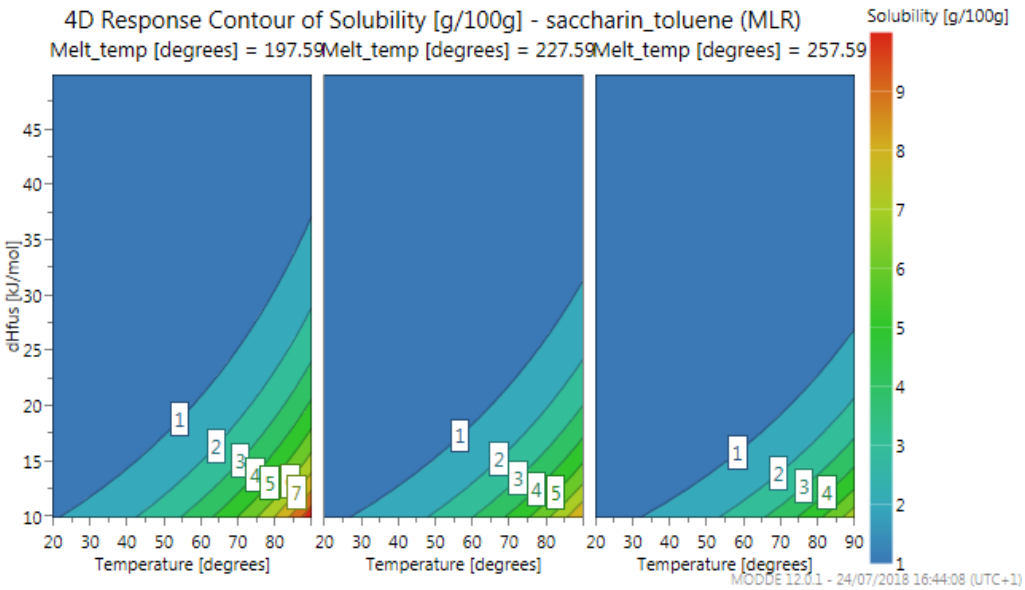


Figure 3-12 4D contour plot showing the impact of varying temperature, enthalpy of fusion and melting temperature on the solubility of saccharin in toluene. All temperatures are in °C

Figure 3-12 shows the visualisation of the effects of the input parameters on the solubility of saccharin in toluene. As the melting temperature and enthalpy of fusion increases the solubility decreases. With an increase of temperature, the solubility increases. This plot is comparable with the results in Figure 3-9.

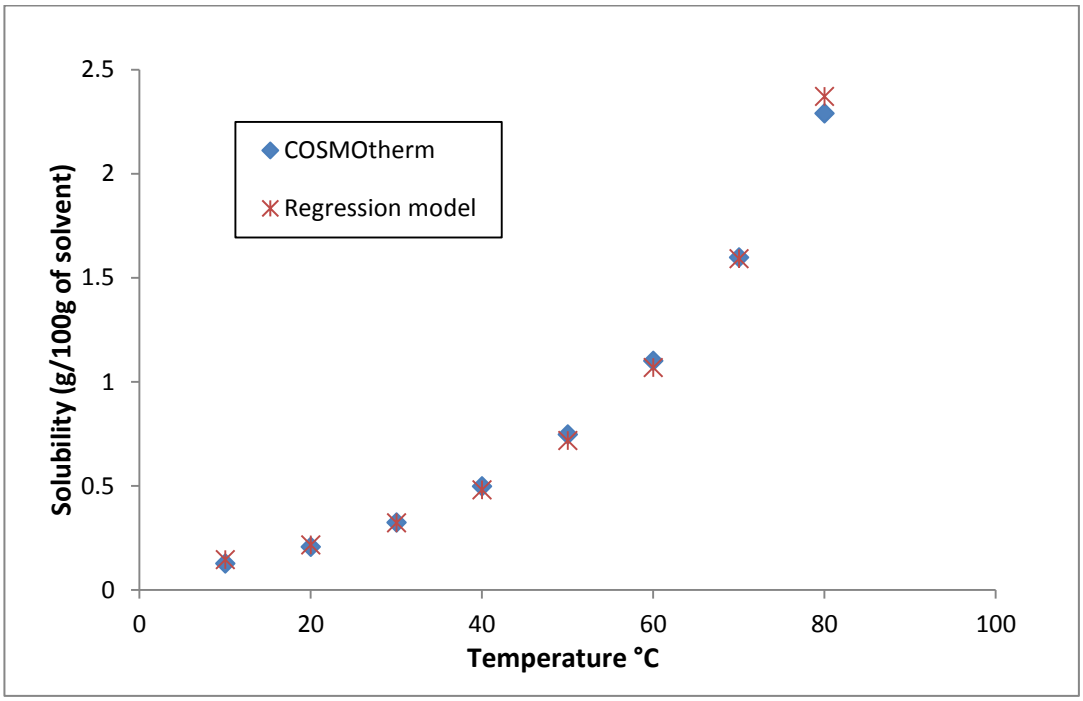


Figure 3-13 Comparison of COSMOtherm solubility curve with the regression model for saccharin in anisole

The above graph (Figure 3-13) shows the solubility curve of COSMO*therm* for saccharin in anisole and compares it to the solubility curve of the regression model. The model shows an excellent correlation with COSMO*therm* at all temperatures. The same input parameters of temperature, enthalpy of fusion and melting temperature were used.

3.2.3 4-pyridinecarbonitrile in water regression model

Figure 3-14 shows the results file for 4-pyridinecarbonitrile in water. Experiment two with a low enthalpy of fusion shows an outlier of over 43483g/100g of solvent.

4-pyridinecarbonitrile
h2o
COSMOtherm 16 TZVP-D basis set solub=w(g/100g of solvent) Job Type: SLESOL

Exp_no	Temp	Dhfus	t_melt	log_solub	solub
1	20.0	1.16	49.0	1.06421985832	11.59364127
2	40.0	1.16	49.0	4.63831824396	43482.8742916
3	20.0	31.16	49.0	0.401267321199	2.51922711
4	40.0	31.16	49.0	0.908314459395	8.09681953
5	20.0	1.16	109.0	1.00885917629	10.20608489
6	40.0	1.16	109.0	1.09874454207	12.55291367
7	20.0	31.16	109.0	-0.440040915837	0.36304385
8	40.0	31.16	109.0	-0.0285900371516	0.93628909
9	30.0	16.16	79.0	0.555330785627	3.59195416

Figure 3-14 4-pyridinecarbonitrile with water linear regression model results (Exp_no is experiment number, Temp is temperature, Dhfus is enthalpy of fusion, t_melt is melting temperature, log_solub is log solubility g/100g of solvent and solub is solubility in g/100g solvent)

It is likely that COSMO*therm* was unable to give a realistic value for solubility due to the low values of melting temperature and enthalpy of fusion and a slightly elevated temperature.

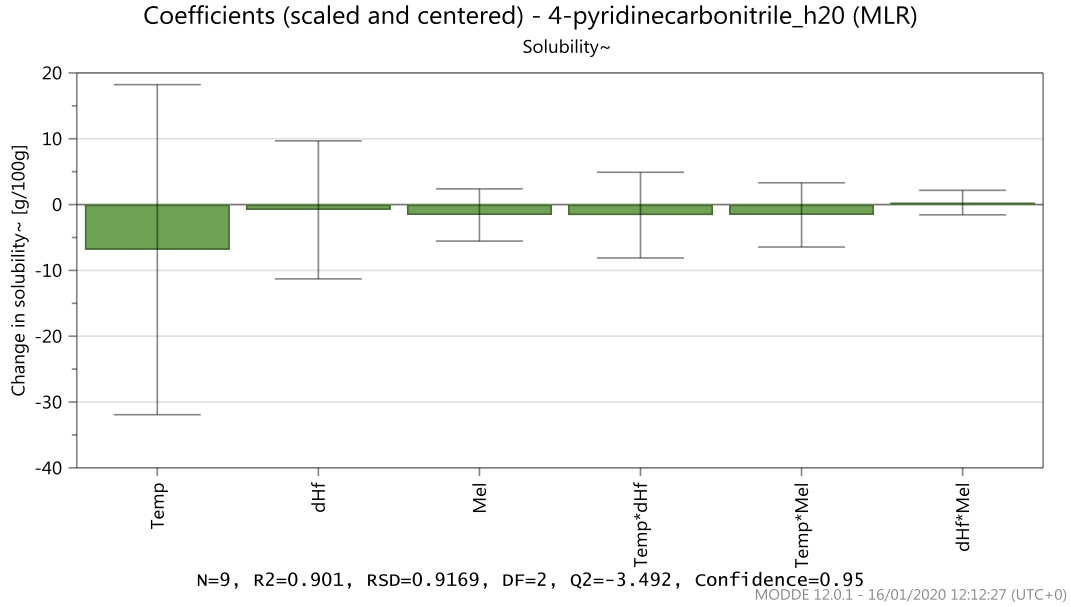


Figure 3-15 Coefficients for 4-pyridinecarbonnitrile and h2o (Temp is Temperature, dHf is enthalpy of fusion, Mel is melting temperature)

The plot showing the coefficients for this model (Figure 3-15) show negative correlation for temperature with large error bars. Usually solubility increases with temperature. All the other coefficients apart from enthalpy of fusion*melting temperature show a negative correlation for solubility with large error bars. These results are due to the outlier of 43483g/100g of solvent as shown in Figure 3-14 4-pyridinecarbonitrile with water linear regression model results (Figure 3-14).

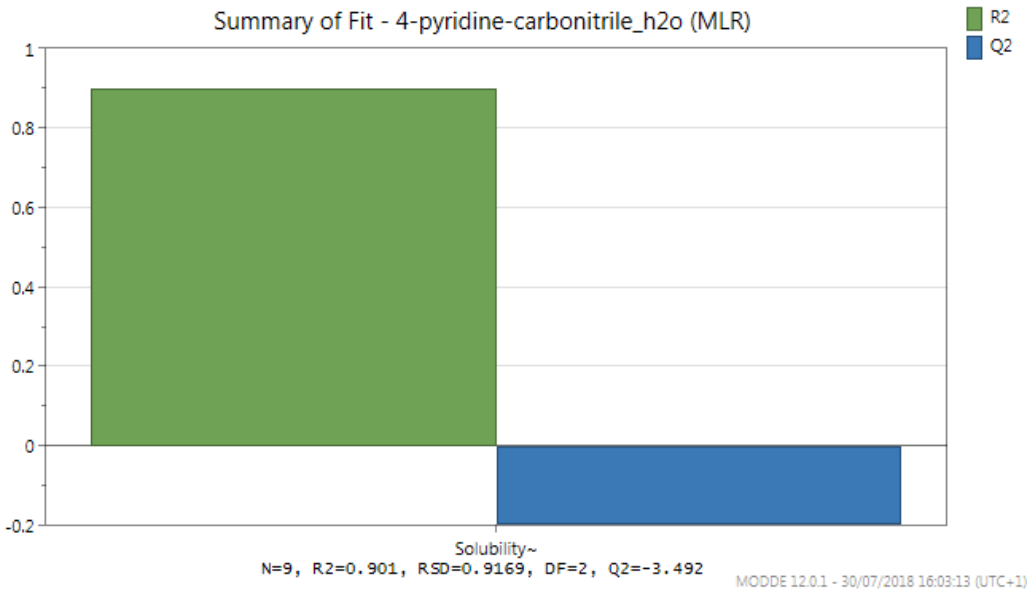


Figure 3-16 Summary of fit graph for 4-pyridine-carbonitrile in water

The summary of fit graph (Figure 3-16) shows a R^2 value of 0.9 and an extremely poor Q^2 of -3.5. Again this is due to the outlier.

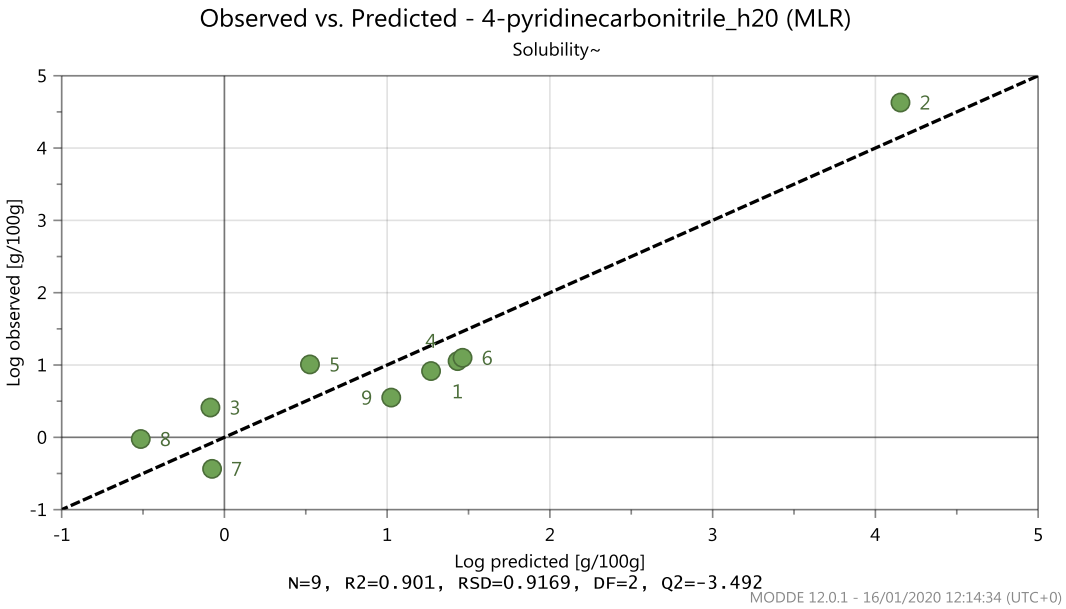


Figure 3-17 Plot of the observed v's predicted log solubilities for 4-pyridinecarbonnitrile in water

Figure 3-17 shows how well the model performs. As is shown this model has none of the experiments on the line.

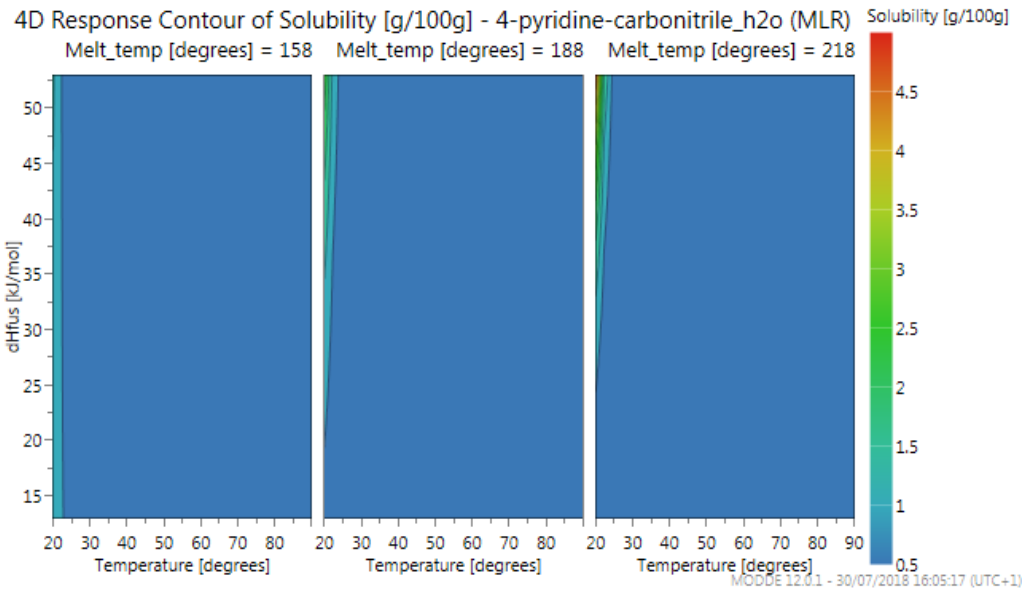


Figure 3-18 4D contour surface plot showing the impact of varying temperature, enthalpy of fusion and melting temperature on the solubility of 4-pyridinecarbonitrile in water. All the temperatures are in °C

The surface plot (Figure 3-18) shows that when temperature is low the solubility is higher. Usually when temperature increases the solubility increases. In this case the model was clearly invalid and was removed from the database.

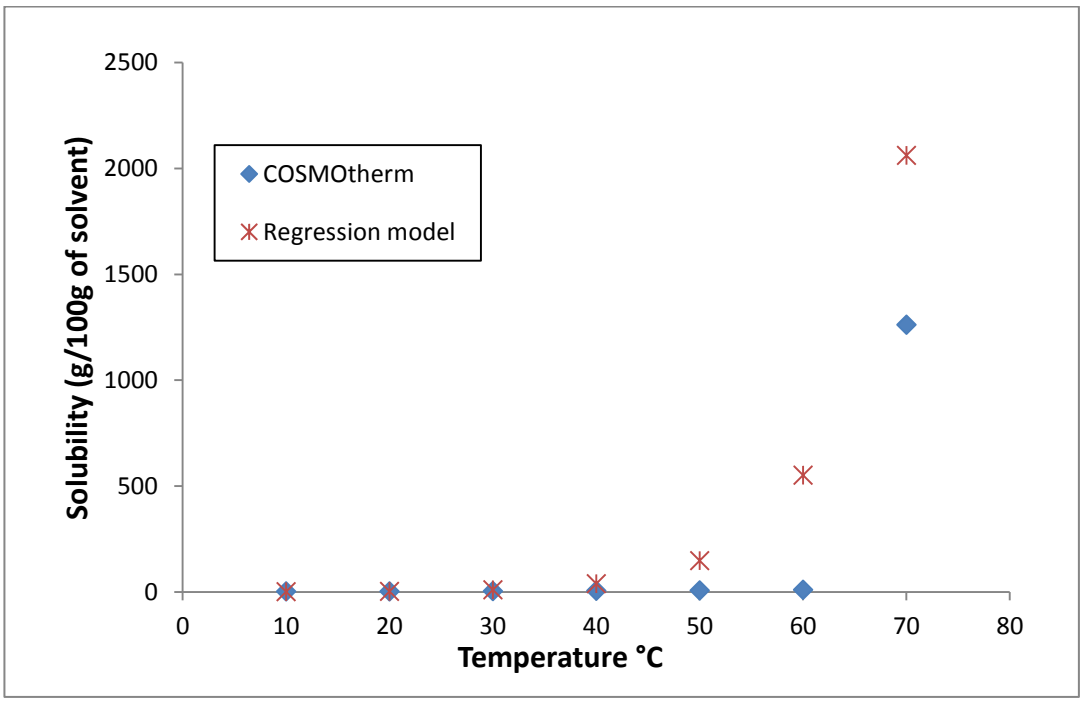


Figure 3-19 Comparison of COSMOtherm solubility curve with the regression model for 4-pyridinecarbonitrile in water

Figure 3-19 shows the solubility curve of the predictions from COSMO*therm* and compares them with the regression model for 4-pyridinecarbonitrile in water. As already stated the model had poor predictive ability because of the outlier and as can be seen at higher temperatures it diverts greatly from the predictions from COSMO*therm.* The same inputs of temperature, enthalpy of fusion and melting temperature were used for both methods.

3.2.4 Regression equations

Equation 27 shows the general form of the linear regression equation, *c1* through *c6* are factor coefficients. *Temp* is temperature, ΔH_{fus} is enthalpy of fusion, *Tmelt* is melting temperature and *C* is a constant.

Equation 27

$$Log\ Solubility = (c1)Temp + (c2)\Delta H_{fus} + (c3)Tmelt + (c4)Temp * Tmelt + (c5)Temp * \Delta H_{fus} + (c6)\Delta H_{fus} * Tmelt + C$$

The equation not only comprises experimental input values such as melting temperature, temperature and enthalpy of fusion but also three interaction terms *i.e. Temp*Tmelt, Temp*ΔH_{fus}* and ΔH_{fus} **Tmelt*.

Table 3-5 Coefficients of log solubility for (a) indomethacin in 1,3-dioxolan-2-one (b)saccharin in anisole (c) 4-pyridine carbonitrile in water

	Temp	**ΔHfus**	**Tmelt**	**Temp * Tmelt**	**Temp * ΔHfus**	**ΔHfus * Tmelt**	**C**
(a)	0.0232	-0.0321	-0.0019	-0.0000582	0.000342	-0.00022	1.988
(b)	0.00731	-0.0442	-0.0009	-0.000012	0.00043	-0.00018	1.726
(c)	0.2121	-0.0293	0.0142	-0.0015	-0.00229	-0.0015	0.0005

The three examples shown above (Table 3-5) show the diversity of the regression models obtained from the design of experiment. 98% of all the linear regression models had an R^2 value of greater than 0.95.

3.3 Conclusion

This project developed a database of linear regression equations by applying DoE approach to COSMO*therm* predictions. These simple equations require no sophisticated software and can be used by non-experts and modellers for almost instant solubility predictions. The input parameters of temperature, melting temperature and enthalpy of fusion can be changed if required.

Of the 10272 models that were produced in this project 98% had an R^2 value of 0.95 or above. For 3467 solute/solvent combinations a model could not be obtained due to the limitations of COSMO*therm.* This indicates that, for the majority of solute/solvent systems, a standardised set of initial COSMO*therm* predictions can be used to build a simple, linear regression model that can accurately reproduce any subsequent predictions desired. This is an important finding, particularly for large organisations,

because the initial COSMO*therm* predictions required for the DoE method could potentially be calculated automatically as soon as data is made available. When solubility predictions are required for a use case, a linear regression model could already be available to produce instant answers.

4 Applying machine learning to obtain a correction factor for solubility predictions

4.1 Introduction

Firstly, this chapter investigates the prediction of solubility by applying RF and using molecular descriptors. The study continues by studying the ability to predict the error between COSMO*therm* and experimental data using RF and COSMO*therm* data. The use of both 2D and 3D molecular descriptors combined and separated in both studies was compared to establish the best descriptors to use in the model. ML models had previously predicted solubility in log units. This chapter will also analyse the difference in predictions using logged and unlogged units.

4.1.1 Machine learning in the pharmaceutical industry

With the development of medicines costing several billions of pounds and around 12 years in getting the product to market (Van Norman, 2016), there is an increasing appreciation of the value of predictive modelling and its application. The application of any model is to reduce costs and time to market. ML relies on previous knowledge structures in a machine-readable format to be effective. There are currently a number of examples of ML being used in medicine development and manufacture such as disease identification/diagnosis (Keerrthega and Thenmozhi, 2016), drug discovery (Hongming Chen *et al.*, 2018) through to manufacturing (Tulsyan,Garvin and Ündey, 2018), analytical techniques, (Martinez *et al.*, 2018) clinical studies (de la Iglesia *et al.*, 2014) and drug safety (Ben Abacha *et al.*, 2015). With ML more data usually yields better results and the pharmaceutical industry (Domingos, 2012) has data that reaches back

decades. However much of this data is not structured in a standard format and thus not easily machine-readable. In addition, this data is rarely openly available.

ML is a method of data analysis that makes predictions or decisions based on the available data. ML is increasingly being used by many sectors of business and government, from financial services and the pharmaceutical industry to the oil and gas industry. ML applies an algorithm which is shaped by the dataset that is provided to build a model that can be compared with reality. There are many such algorithms that are available such as SVM (Schaathun, 2012), neural networks (Hammer, 2014) and RF (Breiman, 2001).

4.1.2 Descriptor calculation

For this work, molecular descriptors were calculated using MOE (MOE, 2018). The 3D structures were energy minimised using Pipeline Pilot (BIOVIA, 2017). MOE can calculate over 400 2D and 3D molecular descriptors such as LogP, pKa and pKb.

4.1.3 Aims

The aim of this chapter is to investigate the use of RF to predict both solubility and the error between COSMO*therm* predictions and experimental data. The ML response from the latter model can be used as a correction factor that is subtracted from the COSMO*therm* prediction to produce a new corrected solubility.

4.2 Methods

4.2.1 Dataset construction

The solubility database was compiled from a variety of sources, CMAC researchers, GSK solubility database (see section 2.4) and DETHERM database from the Royal Society of Chemistry (RSC) (http://detherm.cds.rsc.org/)(Detherm, 2016). The laboratory techniques used to obtain the solubility data were varied. However, for this work, it has been assumed that that these techniques were accurate, and these measurements are taken as "real" solubility values. The enthalpy of fusion and melting temperatures for many of the compounds were also obtained from the DETHERM website. In cases where duplicate data points were recorded, CMAC data points were retained over literature data points. If there were triplicates of a data point and two of the data points were similar, the dissimilar and one of the similar points was removed.

As this research progressed, three datasets were compiled and used. The datasets started at 25°C for use with the RF models to show that the model would work and to validate the hypothesis. The first dataset was used with 281 data points from CMAC researchers and literature sources via the DETHERM website. The second dataset added an additional 139 data points from GSK. A third dataset added another 150 data points from the DETHERM website compiled in collaboration with MSc student Mithushan Soundaranathan (Table 4-1). In producing this dataset, a tolerance of ±1°C was introduced, taking in solubility values between 24 and 26°C. This increased the dataset considerably as many temperatures from the website were converted from Kelvin, *e.g.* 298 K = 24.85°C. It was also assumed that a ±1°C variation in temperature was not sufficient to skew the data.

Table 4-1 Datasets used for machine learning

Dataset	Number of data points	Number of solutes	Number of solvents
Dataset 1 CMAC	281	52	34
Dataset 2 CMAC + GSK	420	60	65
Dataset 3 CMAC + GSK + DETHERM	529	97	65

4.2.2 Descriptor calculation

The COSMO*therm* predictions were calculated using COSMO*therm* Version C3.0 16.01 using the SLESOL job type and TZVPD-fine basis set.

The descriptors were calculated for both solutes and solvents using MOE version 2016.0802. Any descriptors produced by MOE that had zero variance were removed prior to running any of the RF algorithms. Solute and solvent descriptors were combined for each solute/solvent combination.

A maximum of 435 descriptors for each compound were calculated. The descriptors included both 2D and 3D descriptors. Some compounds returned a Null value for certain descriptors due to technical implementation. These descriptors were removed to ensure there were no gaps in the data. The descriptors were checked for cross-correlation and removed as required (see section 4.3.4). However, RF works in such a way that removal of correlated descriptors tends not to have any significant effect on the model as it is capable of disregarding descriptors with no useful information (Breiman, 2001).

4.2.3 k-Fold cross-validation

Cross-validation is a resampling technique used to evaluate ML models on a limited data sample. The dataset is shuffled randomly, then split into a number of groups, k. Each

group is taken as a test set while the remaining groups are entered into the training set. This cycle is repeated until all groups have been in the test set.

For the RF model, the default settings were accepted for this work except for n_{tree} which was set at n=1000. k-Fold cross-validation RF algorithms (Figure 4-1) were run as a comparison to the solute-Fold model which is described in section 4.2.4.

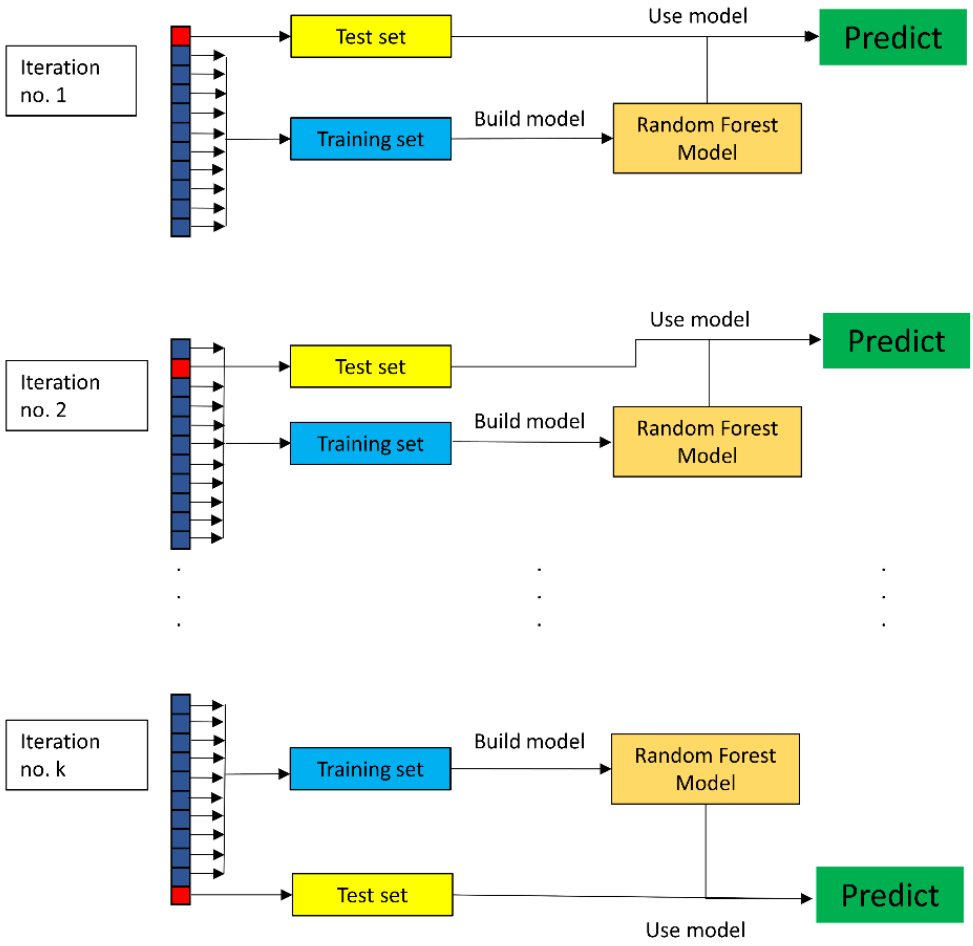


Figure 4-1 Graphical representation of the k-Fold RF model

The dataset in the k-Fold cross-validation model splits the data into k sections (k being 10 in this case) and uses one of the sections as a test set and k-1 sections as the training set, repeating this process until all sections have been used as a test set. This is to ensure

that there is consistency in the dataset regardless of which data are sampled for testing. While this is a standard, widely-accepted method for validating a model, an issue occurs for this dataset when assessing the effectiveness of a model for an unknown solute, which is a likely use case for any solubility prediction model. Since the sections are chosen entirely randomly, each split of a k-Fold cross-validation model will likely contain some solute and solvents included, in different combinations, in both training set and test set. This gives the model an advantage as it has "seen" the solute in the training set, whereas in reality for an unknown compound the model will not have "seen" the solute before.

4.2.4 Solute-Fold cross-validation

To fairly judge the performance of the RF models for unknown solutes it was decided to also build models excluding all data from the training set for the solute being tested. This was named the "solute-Fold" cross-validation model (Figure 4-2) and, unlike the k-Fold cross-validation model, selected all data points containing a given solute to be the test set. The procedure was repeated for each solute in turn. The reasons for developing this model is that if a new molecular entity is presented, the dataset that the RF model is built from would have no knowledge of the new compound. While the same procedure could also be performed for each solvent, it was determined that a new, unknown solvent represented a much less likely use case and was therefore not necessary.

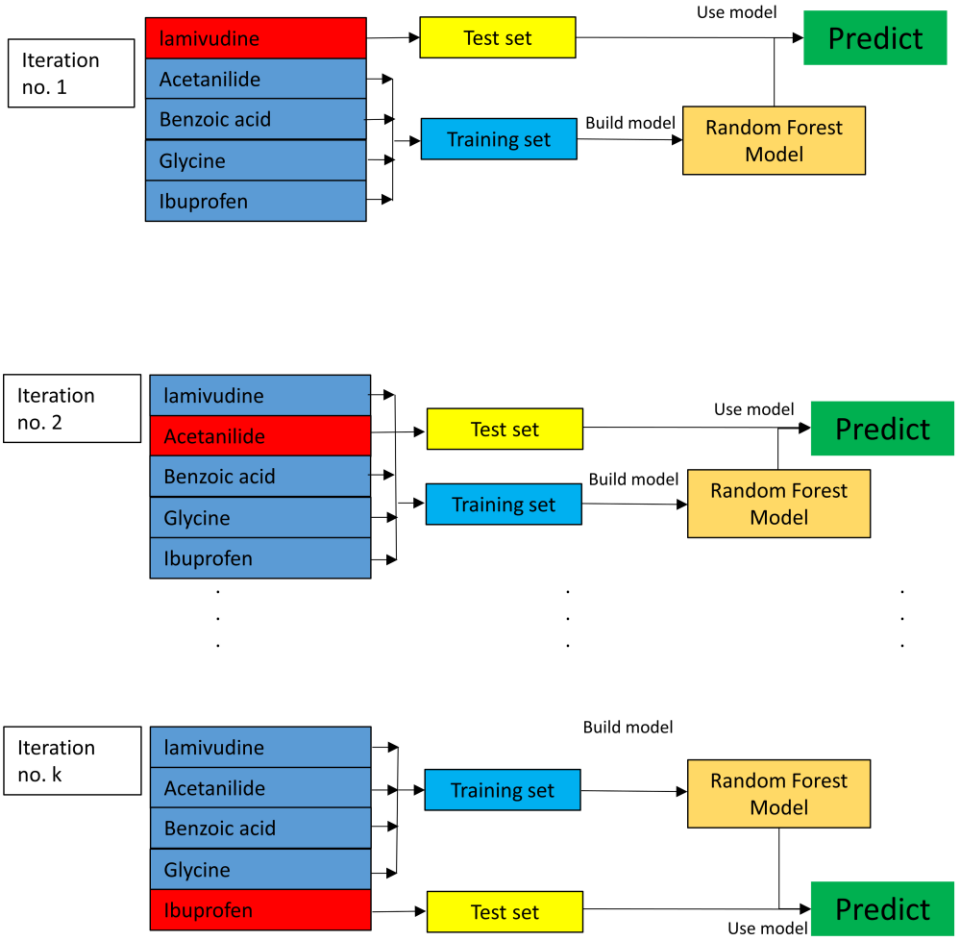


Figure 4-2 Graphical representation of the solute-Fold RF model

4.2.5 Drip-feed model

The solute-Fold cross-validation model used no data points of a solute in the training set that was in the test set. To investigate the significance of the model having limited prior knowledge of the solute (rather than none) a "drip-feed" model was developed. This would determine the average number of points required to have an optimum predictive model and to investigate potential points of diminishing return with respect to model performance. The drip-feed model (Figure 4-3) determines how many solubility data points of a solute are required to improve the model. Initially no data points for the

target solute were used in the training set of the model and then a data point was “drip-fed” into the training set.

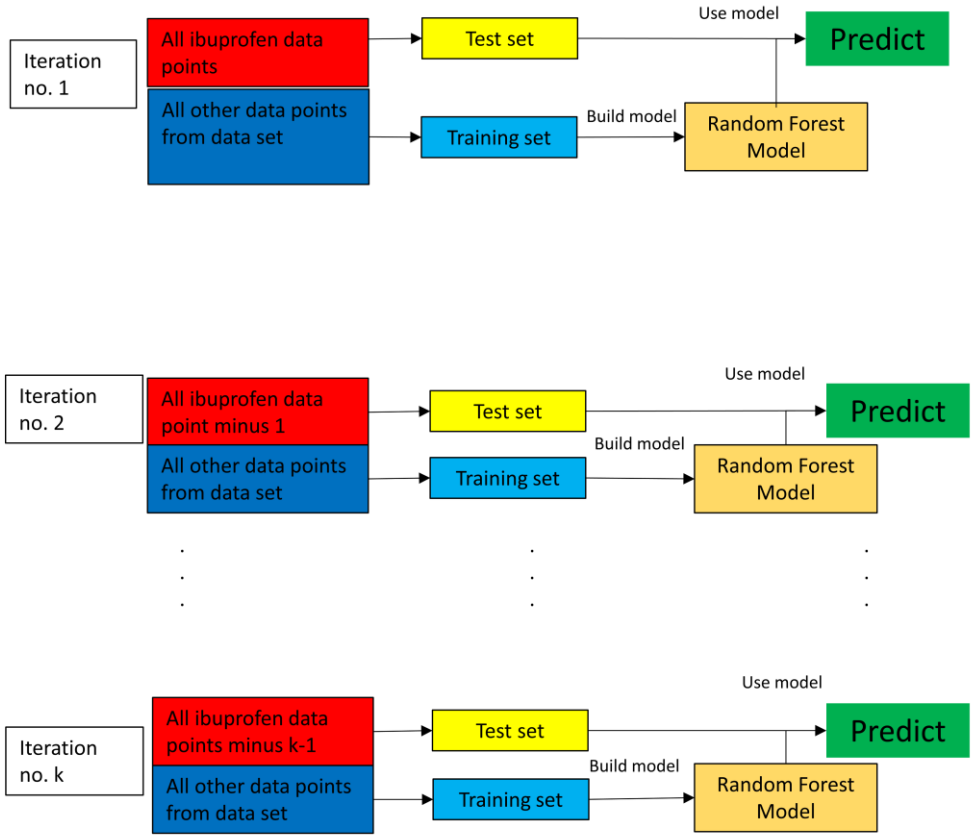


Figure 4-3 Graphical representation of the drip-feed RF model

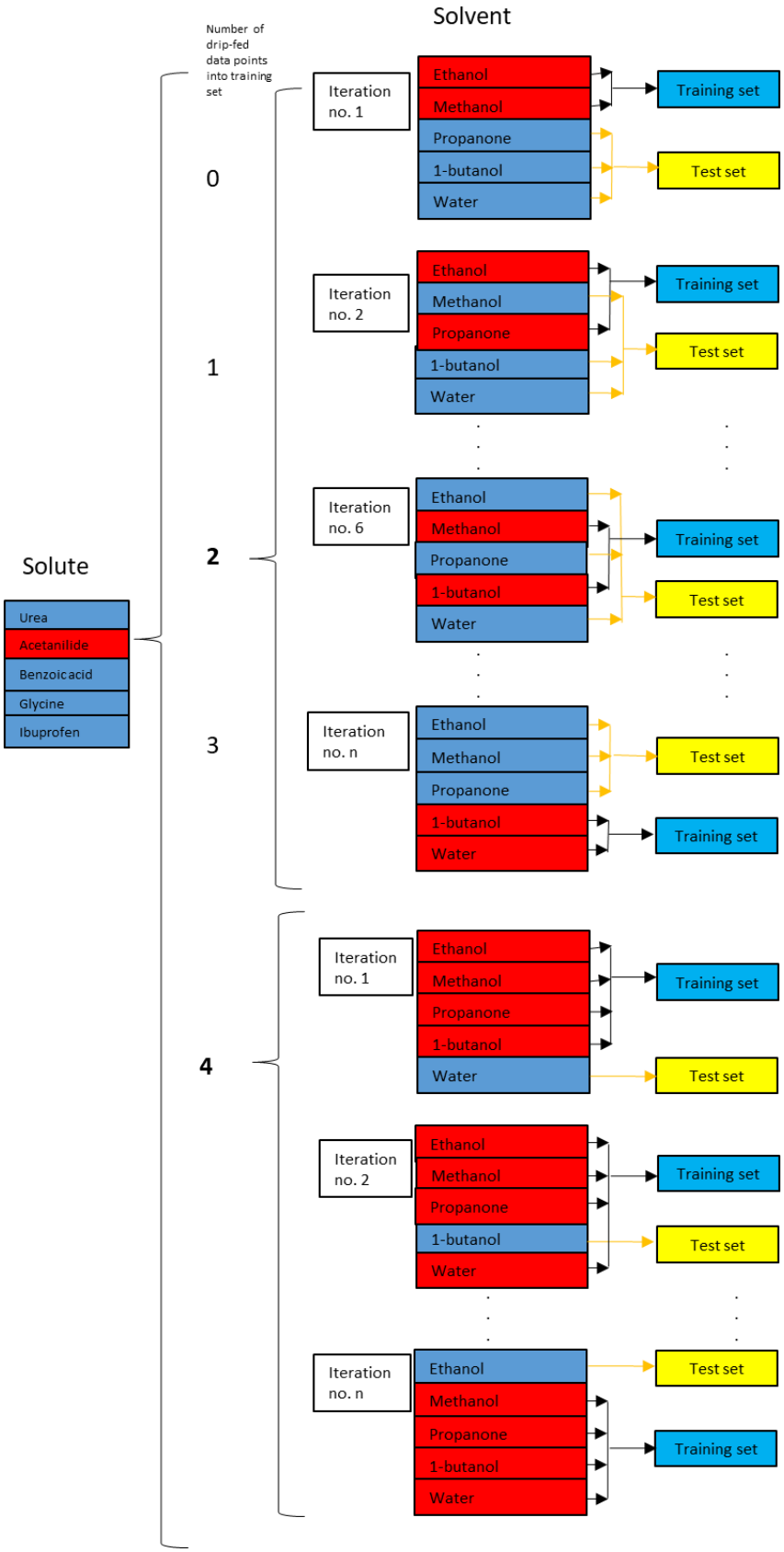


Figure 4-4 drip-feed model with combinations of different solvents for solute data points into the training set

Firstly, a RF was carried out with no examples of the solute being “drip-fed” into the training set, which is the same as the solute-Fold model. Then a RF model where only one data point for the target solute was drip-fed into the training set with all other points for that solute kept in the test set. This was repeated for each unique data point for that solute (Figure 4-4), before increasing the number of data points drip-fed by one. Next, each combination of two data points for the target solute were inserted into the training set. Then each combination of three data points per solute was inserted into the data set and the same with four data points. This increased the number of RF models generated exponentially.

If four data points were drip-fed into the training set, then only solutes with at least five data points were used so that there would be at least one data point left for use in the test set. There were 36 solutes that satisfied that criterion.

The predicted result is then averaged for each data point when that data point is in the test set. With solutes with a large number of data points the possible combinations would increase exponentially and for some solutes the combinations numbered over 10,000. In total, 131,596 RF models were generated – a task that took several weeks over several CPUs (see Table 4-2). For each addition of a solvent the number of RF models increased exponentially.

Table 4-2 number of RF models required per number of solvents for each solute in training set

No. of solvents for each solute in training set	No. of RF models required
0	36
1	370
2	2,645
3	17,911
4	110,634

For this body of work, a script was written such that the job could be shared over all cores of a CPU and split over several computers to decrease the time taken.

4.3 Results and Discussion

4.3.1 Dataset analysis

4.3.1.1 Lipinski's rule of five

The chemical diversity and space of dataset 3 was compared, using Lipinski's rules as a guideline, with molecules from the DrugBank database (DrugBank, 2018). Lipinski's rule of five (Lipinski *et al.*, 2001) is a rule of thumb to determine whether a molecule has drug-like properties. In general, the rules state that a drug should have:

- No more than 5 hydrogen bond donors
- No more than 10 hydrogen bond acceptors
- A molecular weight of less than 500 daltons
- A cLogP not greater than 5

The chemical variation of the 97 solutes in dataset 3 was compared with a further 865 solutes taken as a subset from the approved drugs dataset from the DrugBank database.

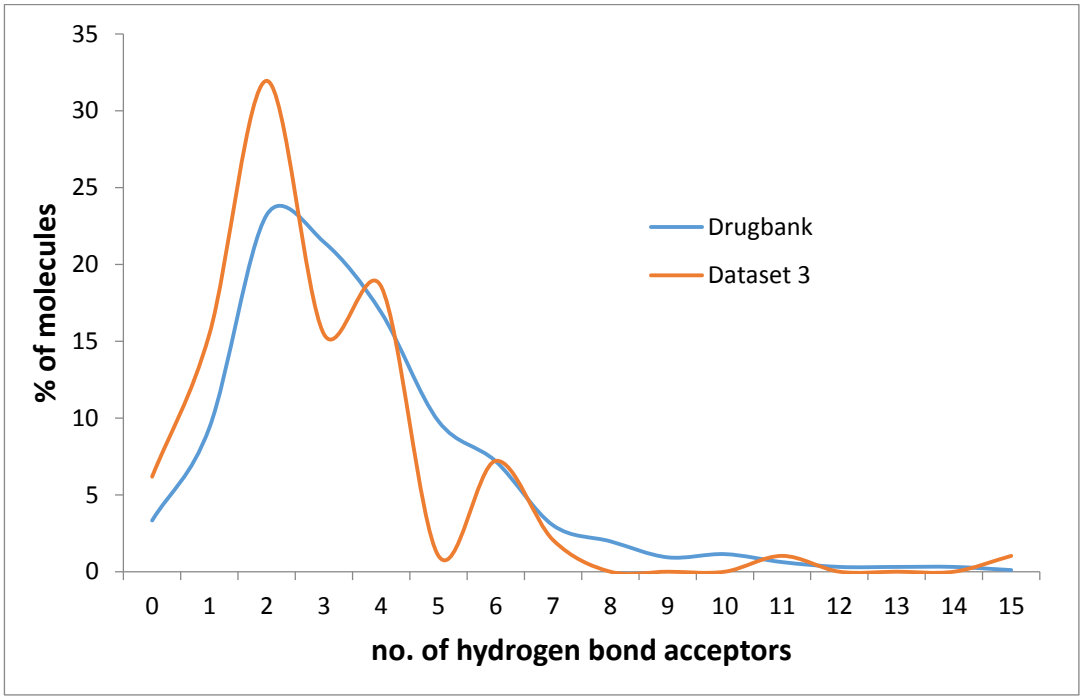


Figure 4-5 Percentage and number of hydrogen bond acceptors of dataset 3 and molecules from Drugbank

The above figure (Figure 4-5) shows the percentage of hydrogen bond acceptors for the molecules from both dataset 3 and the Drugbank database. The percentages are very similar with a slightly higher percentage in dataset 3 at two hydrogen bond acceptors. According to Lipinski's rule of five there should be, in general, no more than five hydrogen bond acceptors for a drug-like molecule and there is a small percentage in both of the databases with more than five. The 97 compounds have a very similar distribution of hydrogen bond acceptors as the much larger Drugbank dataset except for two acceptors which are overrepresented and at five acceptors which are underrepresented. This is unusual as 10% of Drugbank molecules have five acceptors.

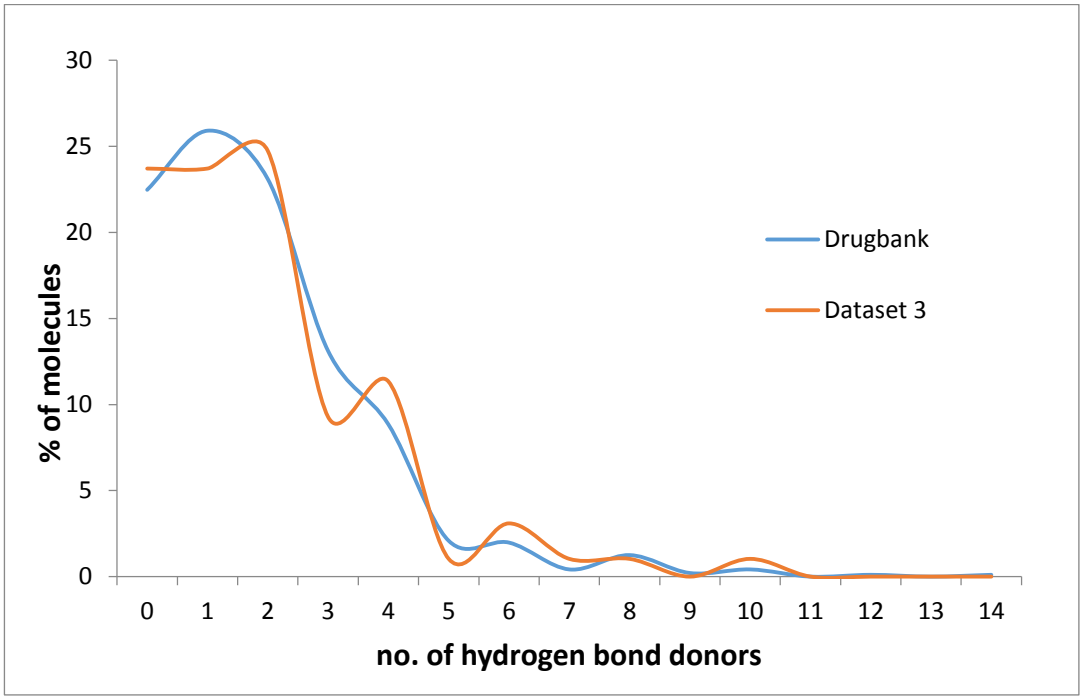


Figure 4-6 Percentage and number of hydrogen bond donors of dataset 3 and molecules from Drugbank

Figure 4-6 shows the percentage of hydrogen bond donors from both databases and shows a good overlapping representation for dataset 3. There are no molecules with more than 10 hydrogen bond donors in dataset 3, conforming to Lipinski's rule of no more than 10 donors.

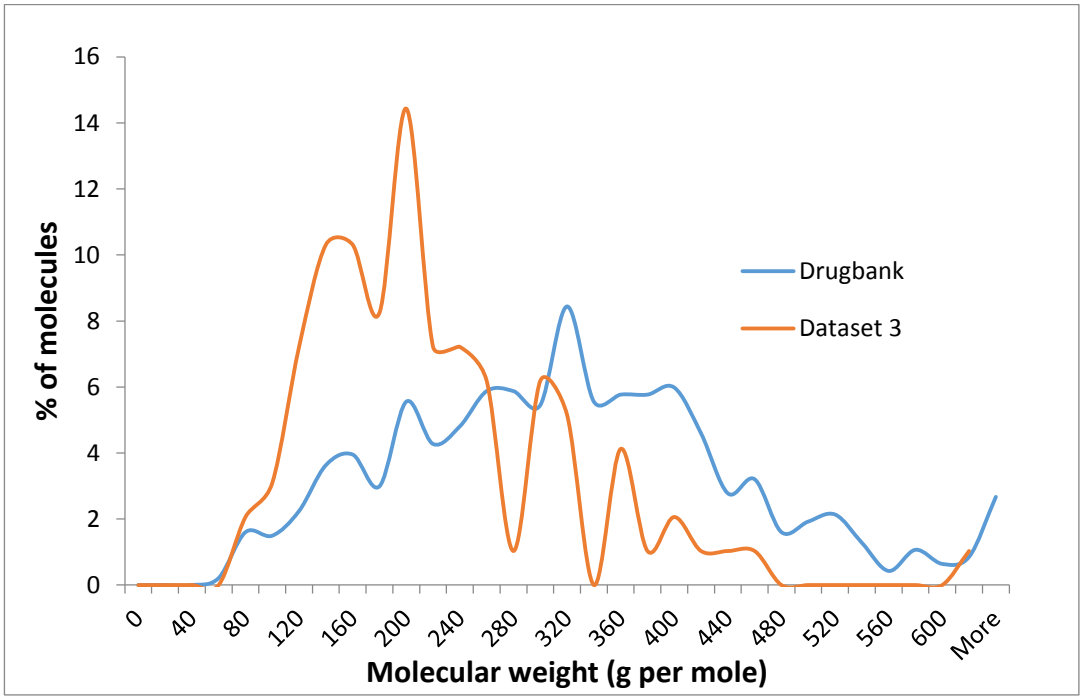


Figure 4-7 Percentage of molecules and molecular weight of molecules and from Drugbank

Figure 4-7 shows molecular weight of both datasets. Dataset 3 has an overrepresentation of molecular weights in the lower weights. All but one molecule has a molecular weight of under 500g mostly conforming to Lipinski's rules. This could be because there is more data available for the lower molecular weight range. Pharmaceutical companies, with the trend for heavier molecules (Bryant *et al.*, 2019), have a tendency not to have their newest data in the public domain. It can be assumed that available academic solubility studies use smaller, cheaper molecules.

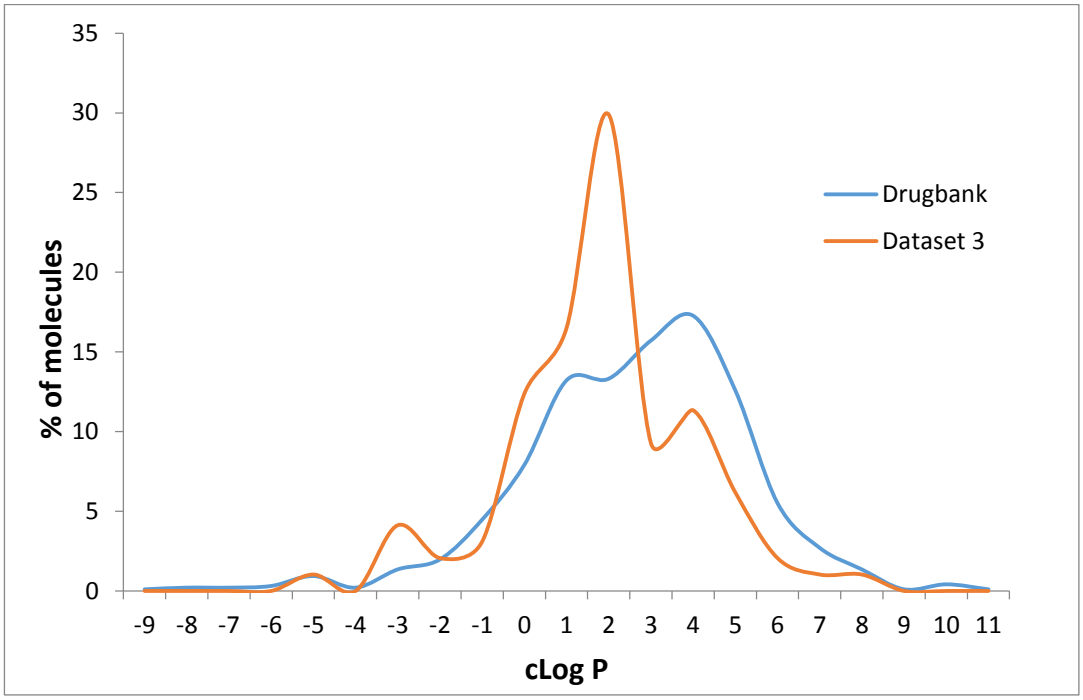


Figure 4-8 Percentage of molecules and cLogP of dataset 3 and molecules from Drugbank

Figure 4-8 shows dataset 3 having a broad range of clogP which is comparable with the Drugbank database but with dataset 3 having a larger percentage of molecules at clogP 2. Dataset 3 has a small percentage (~4%) of molecules above clogP 5 which is outwith Lipinski's rules. As a general rule small molecules are more water soluble than larger ones and dataset 3 has smaller molecules than the Drugbank dataset so has more of a concentration of molecules with a smaller clogP.

Table 4-3 Mean and standard deviation of Lipinski categories Number of Hydrogen Bond Acceptors, Donors, LogP and Molecular Weight for Dataset 3 and Drugbank dataset

		Number of Hydrogen Bond acceptors	Number of Hydrogen Bond donors	clogP	Molecular weight (g)
Dataset 3	**Mean**	2.9	2.0	1.4	212.5
	SD	2.2	1.9	2.1	93.6
Drugbank	**Mean**	3.6	1.9	2.1	313.7
	SD	2.3	1.8	2.6	133.0

The above table (Table 4-3) shows the mean and standard deviation in dataset 3 and the Drugbank dataset for the four Lipinski's rules. cLogP values were calculated using MOE. The results show that dataset 3 is within the limits of Lipinski's rules.

The analysis shows that dataset 3 is comparable to the Drugbank dataset and that dataset 3 is a reasonable representation of the distribution of properties that drugs have especially for hydrogen bond acceptors and donors. For the most part the dataset agrees with Lipinski's rule of five. This analysis justifies the use of this dataset in the machine learning model as it is representative of drug-like molecules.

4.3.1.2 COSMOtherm predictions

The COSMO*therm* prediction results were compared with the solubility data obtained from experiment and DETHERM (dataset 3) (Figure 4-9). The COSMO*therm* jobs were completed with SLESOL jobtype and TZVPD-fine basis set. The tramlines in the graph are set at zero. If the data point is on the zero line the prediction and experimental are the same value. If the data point is on one of the outer tramlines the prediction has either log -1 or log 1 error.

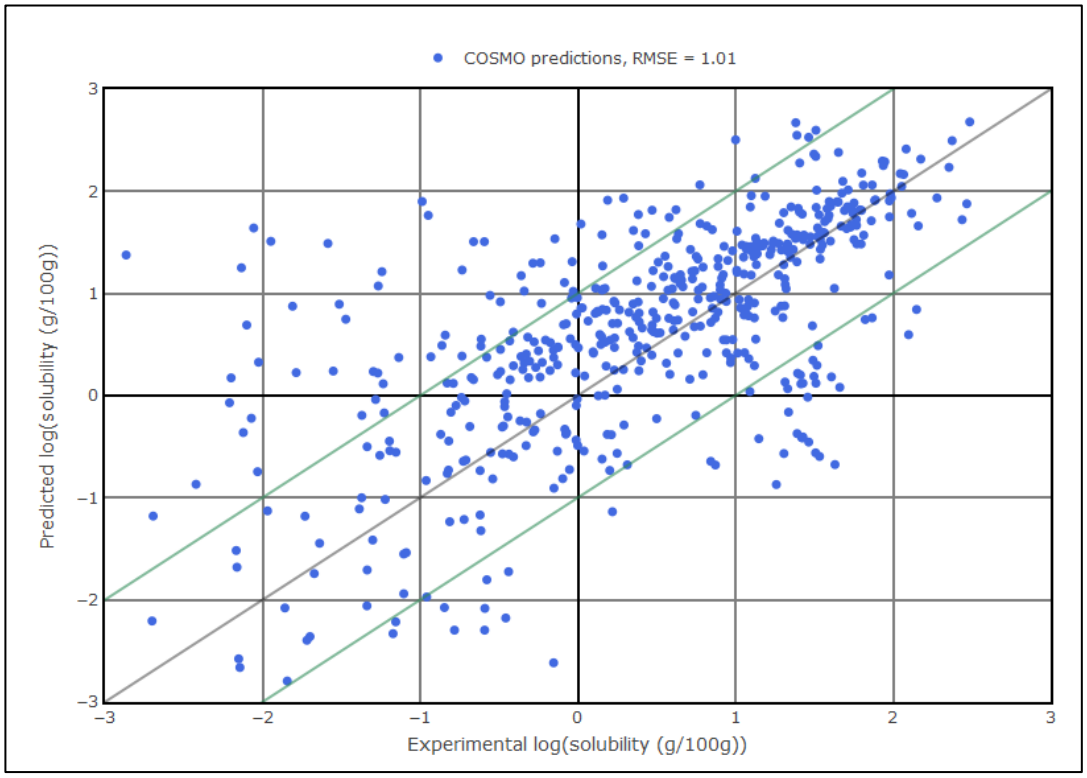


Figure 4-9 Correlation between COSMOtherm and experimental data

The above figure (Figure 4-9) shows the results of 529 solute and solvent combinations from dataset 3 with duplicates removed and at a temperature range of 24-26°C. If there was a duplicate data point the closest to 25°C was retained. The RMSE value for this data set was log 1.01.

High solubility predictions with an error of log 1 and low solubility predictions with the same logged error are not equivalent in real terms. For example, a high solubility prediction of 10g/100g with an error of log 1 could either be 1g/100g or 100g/100g whereas a low solubility prediction *e.g.* 0.01g/100g would either be 0.1g/100g or 0.001g/100g. The low solubility predictions can be deemed accurate as they still have low solubility with solubilities of below 1g/100g but the highly soluble predictions could not. In addition, log errors greater than one are more significant than less than one due

to the magnitude of the error in g/100g. For example an error of 0.1g/100g is less significant than 100g/100g.

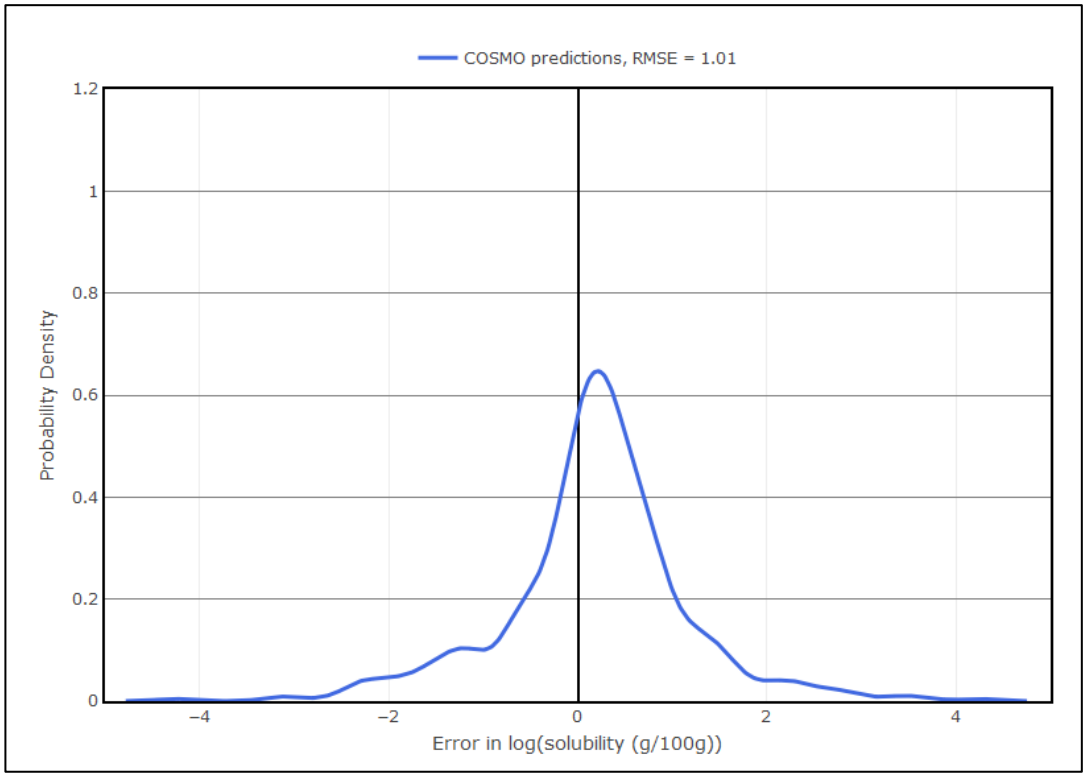


Figure 4-10 Density plot of COSMOtherm prediction error

The above figure (Figure 4-10) shows the distribution of the error from the COSMO*therm* with zero being accurate predictions. 261 (49%) of the predictions are within a log point of the experimental data (Table 4-4). Of the predictions 182 points (34%) are near experimental values between RMSE. The plot shows a tendency for COSMO*therm* to overpredict solubility. 73% of the data to the right of zero are over predictions with 27% to the left of zero under predicted. Reasons for the extent of error between experimental and predicted solubility have already been discussed in section 2.2.

Table 4-4 The number of residuals calculated for the range of log points

Residual range log(g/100g)	No. of points	% percentage of total points
-5 to -1	59	12
-1 to 0	79	15
0 to 1	182	34
1 to 3	209	39

529 data points from dataset 3 were analysed and the predictions were compared with experimental data. If the solubility was out by a margin of 5g per 100g of solvent this was deemed to be a "misclassified" prediction and within that limit was deemed to be a "correctly classified" prediction. The reason for this choice of limits was a result of what is considered useful information for cooling crystallisation (Brown *et al.*, 2018) . The graph below (Figure 4-11) shows the data points compared with increasing solubility from left to right.

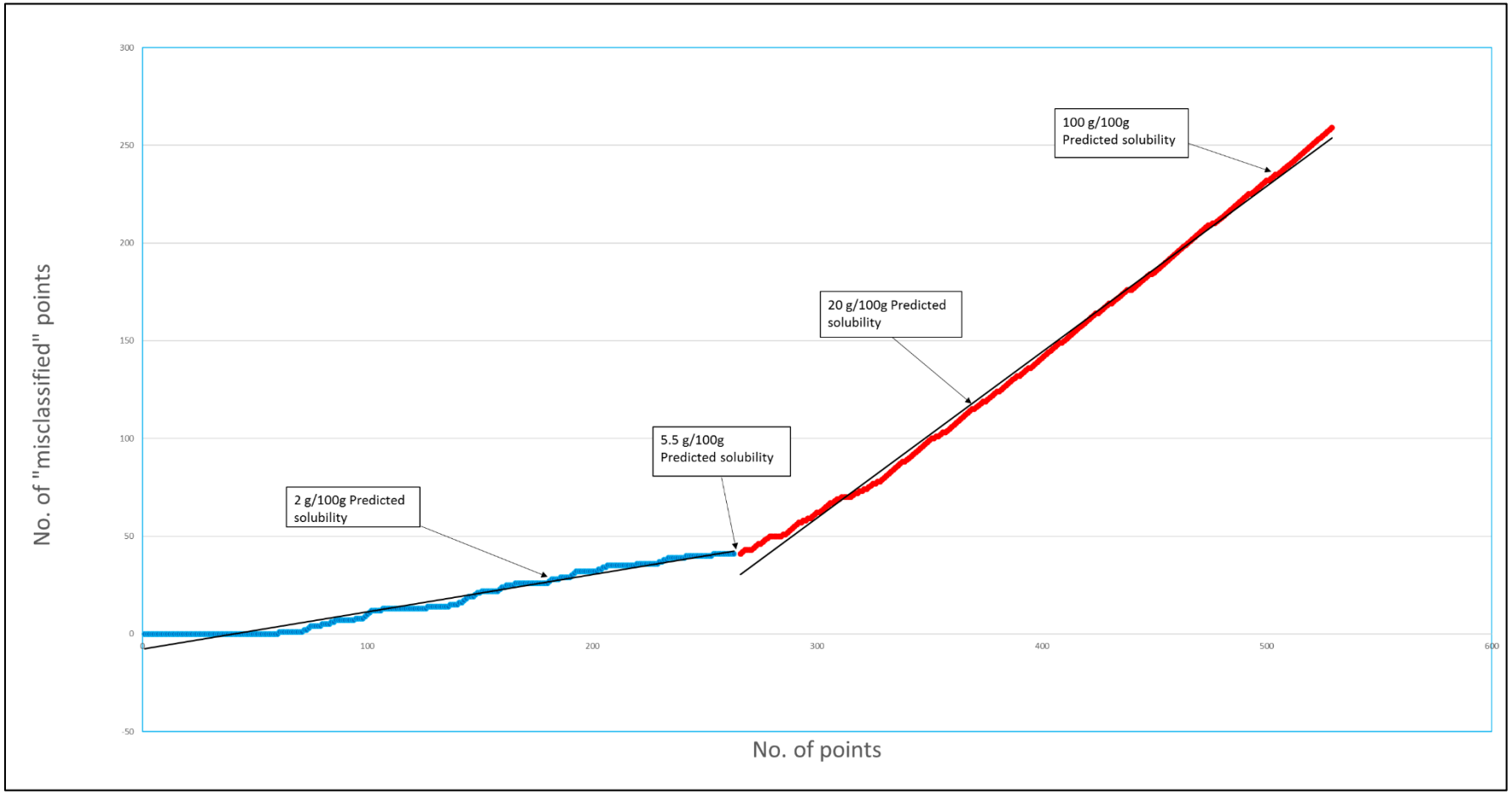


Figure 4-11 No. of COSMOtherm predictions and no. of "misclassified points"

The data has been split into two regions as the predictions of COSMO*therm* have a distinct change in accuracy at around 5.5g/100g as shown in the graph by the change of gradient. For cooling crystallisation purposes anything below around 5g/100g is "effectively insoluble" and above that can be termed "soluble" (Brown *et al.*, 2018). For the soluble predictions above 5.5g per 100g only 17% are classified accurately with 46 out of 265 points being classified correctly. This can be compared with the effectively insoluble (less than 5.5g per 100g solubility) where 83% of the compounds (264 data points) were classified correctly. The two regions on the graph are quite distinct with the rate of misclassified points visibly increasing after 5.5g per 100g. This analysis shows that COSMO*therm* can be relied on in over four fifths of examples to predict effective insolubility.

4.3.1.3 Tanimoto Coefficients and similar compounds

Similar chemical structures are expected in the majority of cases to have similar properties. The Tanimoto coefficient is one of the most popular similarity coefficients used to measure the similarity between molecules (Tanimoto, 1958). 2D molecular fingerprints are used to establish similarity. Fingerprints are fragmented substructures of molecules (Nikolova and Jaworska, 2003) and may have similar properties such as number of atoms or number of rotatable bonds. There are several methods of fingerprinting. The method used for this project was FCFP_4. The Tanimoto coefficient has a score between 0 and 1. If a molecule is very similar to the reference molecule then it will have a Tanimoto coefficient close to one. If the molecule is dissimilar the coefficient will be closer to zero.

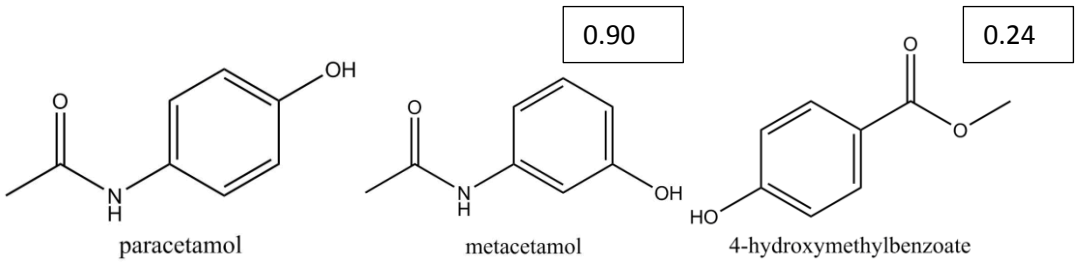


Figure 4-12 Tanimoto Coefficients for comparison with the structure of paracetamol

The purpose of this study was to establish if the RMSE between experimental and the RF model corrected solubility could be reduced if a compound was not in the dataset but had a compound of sufficient similarity in the training set. This would be useful to correct the solubility of impurities of compounds with similar structures to the compound where no solubility data was available.

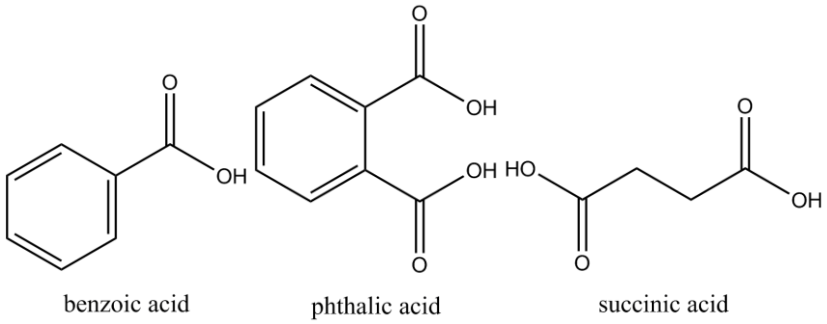


Figure 4-13 Structures of benzoic acid, phthalic acid and succinic acid

In this study, five compounds were used as a reference compound (paracetamol, metacetamol, benzoic acid, phthalic acid and succinic acid) (Figure 4-12 and Figure 4-13). These compounds were chosen as metacetamol and paracetamol are very similar and the others were chosen for being dissimilar to the aforementioned compounds. Each compound was compared with similar compounds that were in dataset 3 and the Tanimoto coefficient was calculated for the similar compounds. For each reference compound the similar compounds were removed from the training set and the RF model

was built. The reference compound was then removed from the training set and the model built again for comparison.

Paracetamol was the first compound and had 35 data points in the dataset. As can be seen from Figure 4-14 and Table 4-5 metacetamol was the most related compound with a Tanimoto score of 0.9. For this compound when paracetamol was in the training set it reduced the mean log RMSE from 0.93 to 0.52.

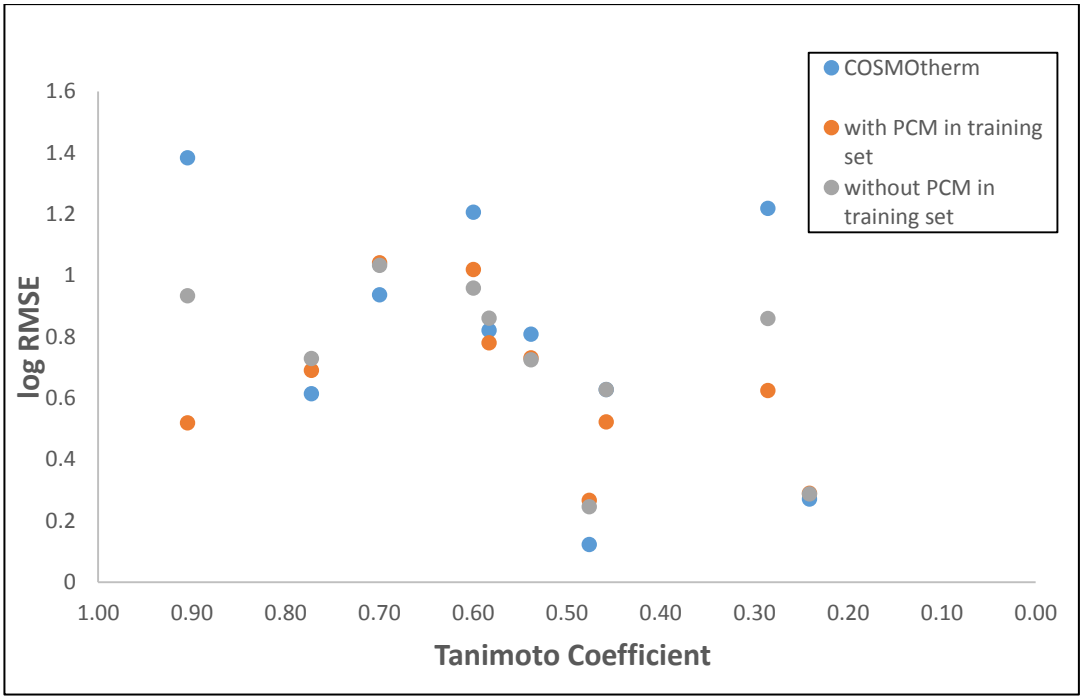


Figure 4-14 RMSE's for similar molecules with paracetamol both in and removed from the RF model training set

The Tanimoto coefficient in the x-axis is for a different compound and each compound has the corresponding coefficient in each table.

Table 4-5 paracetamol as the reference compound and related compounds with Tanimoto score

Compound name	**Tanimoto coefficient**	**Change in mean log RMSE with API in training set**
metacetamol	0.90	-0.41
o-hydroxyacetanilide	0.77	-0.04
acetanilide	0.70	0.01

4-acetamidobenzoic acid	0.60	0.06
p-chloroacetanilide	0.58	-0.08
acetaminophenacetate	0.54	0.01
4-amino-phenol	0.48	0.02
1-(4-hydroxyphenyl)ethanone	0.46	-0.11
4-nitrophenol	0.29	-0.23
4-hydroxymethylbenzoate	0.24	0.00

For most of the other compounds no significant decrease in RMSE was observed except for 4-amino-phenol and 4-nitrophenol. This initial finding aligns with the hypothesis that highly similar compounds can aid RF predictions.

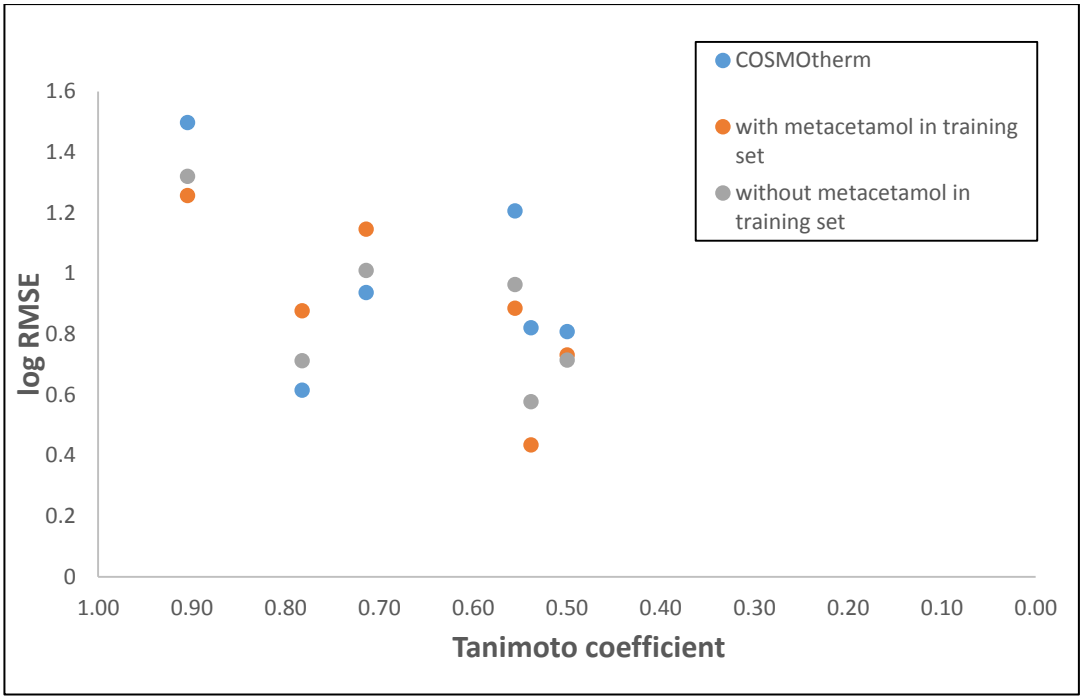


Figure 4-15 RMSE's for similar molecules with metacetamol both in and removed from the RF model training set

The above graph (Figure 4-15) and table (Table 4-6) above show the results for the RF model with metacetamol as the reference compound. Metacetamol had nine data points.

Table 4-6 metacetamol as the reference compound and related compounds with Tanimoto score

Compound name	**Tanimoto coefficient**	**Change in mean log RMSE with API in training set**
PCM	0.90	-0.06
o-hydroxyacetanilide	0.78	0.16
acetanilide	0.71	0.14
4-acetamidobenzoicacid	0.56	-0.08
p-chloroacetanilide	0.54	-0.14
acetaminophenacetate	0.50	0.02

As can be seen from the table paracetamol had a Tanimoto score of 0.90 and RMSE decreased slightly from log 1.32 to log 1.26 with metacetamol included in the training set. When metacetamol is in the training set compared to when paracetamol is in the training set the jump in RMSE significantly less. This may be due to the fact that paracetamol has 35 points and metacetamol only has nine points and the results above being mean RMSE. Any outlier in the metacetamol solubility data would have a larger effect than an outlier in the paracetamol solubility data in reducing the mean RMSE.

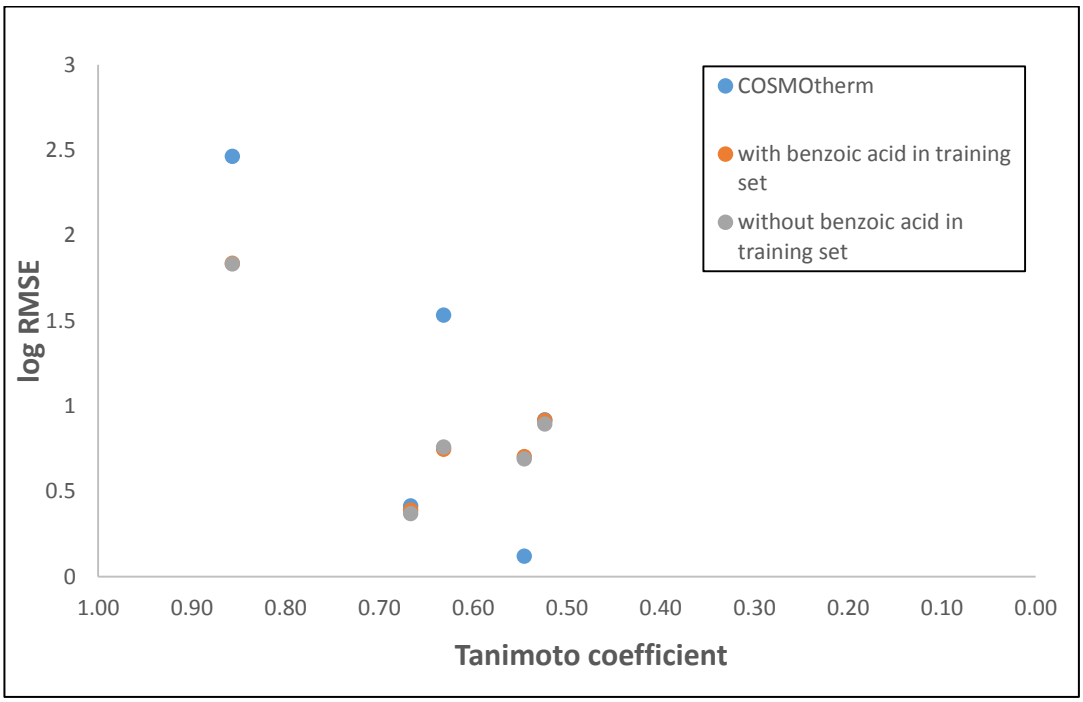


Figure 4-16 RMSE's for similar molecules with benzoic acid both in and removed from the RF model training set

Figure 4-16 and Table 4-7 show the results for benzoic acid as the reference compound with eight data points. Phthalic acid was the most similar compound with a Tanimoto score of 0.86.

Table 4-7 benzoic acid as the reference compound and related compounds with Tanimoto score

Compound name	Tanimoto coefficient	Change in mean log RMSE with API in training set
phthalic acid	0.86	0.00
4-aminobenzoic acid	0.67	0.02
orthoaminobenzoic acid	0.63	-0.01
aspirin	0.55	0.01
3-pyridinecarboxylic acid	0.52	0.02

The RMSE for phthalic acid did not decrease when benzoic acid was added into the training set and seemed to have little effect with other similar compounds. This may be due to the fact that the Tanimoto scores are lower for compounds with benzoic acid as a reference compound.

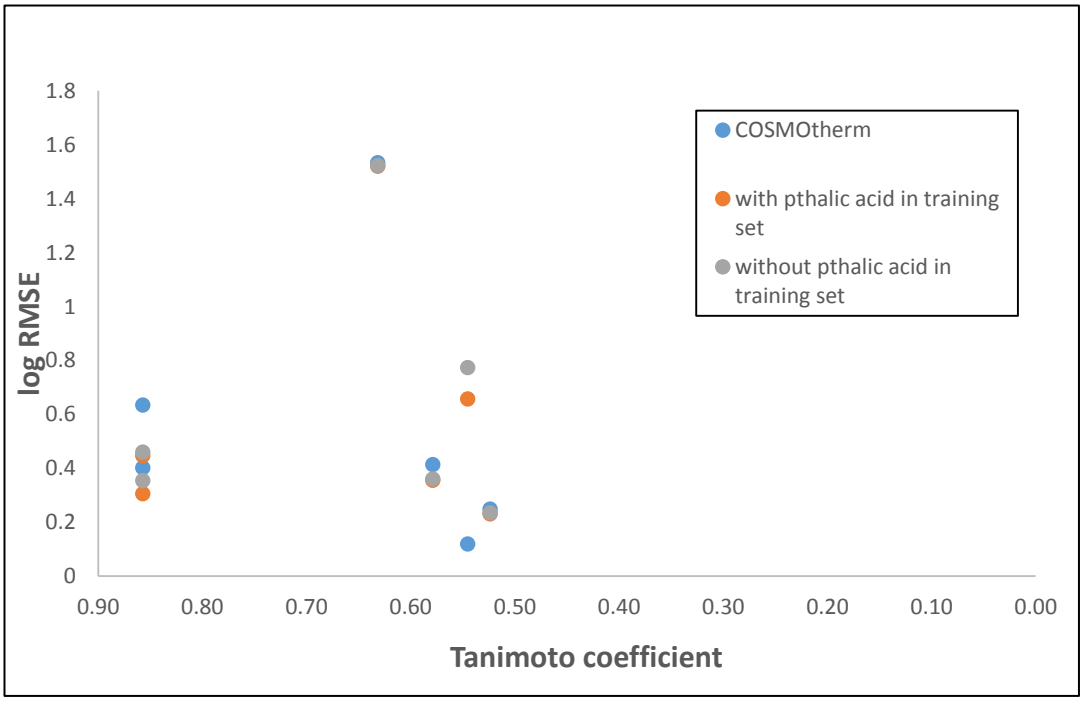


Figure 4-17 RMSE's for similar molecules with phthalic acid both in and removed from the RF model training set

Figure 4-17 and Table 4-8 show the results for phthalic acid as the reference compound and with one data point. This compound has two similar compounds 4-oh-benzoic acid and benzoic acid with a Tanimoto score of 0.86.

Table 4-8 phthalic acid as the reference compound wand related compounds with Tanimoto score

Compound name	Tanimoto coefficient	Change in mean log RMSE with API in training set
4-oh-benzoic acid	0.86	-0.05
benzoic acid	0.86	-0.01
orthoaminobenzoic acid	0.63	0.00
4-aminobenzoic acid	0.58	0.00
aspirin	0.55	-0.12
3-pyridinecarboxylic acid	0.52	0.00

Benzoic acid showed no significant reduction in RMSE however 4-oh-benzoic acid reduced the log RMSE from log 0.35 to log 0.30.

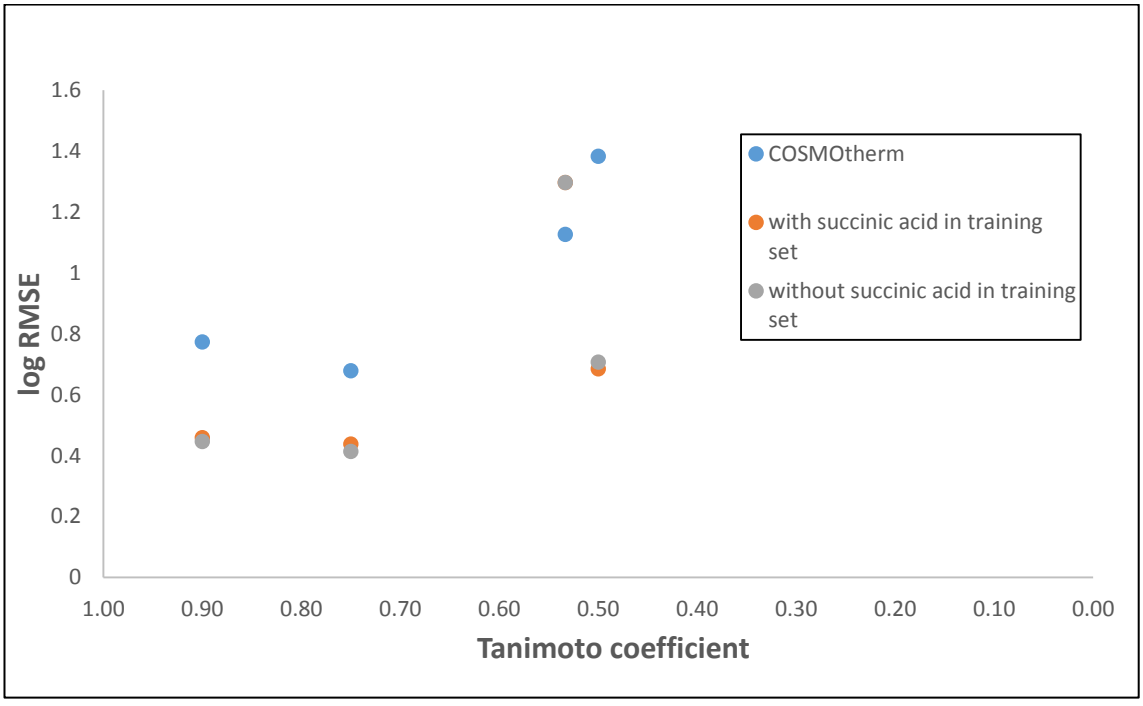


Figure 4-18 RMSE's for similar molecules with succinic acid both in and removed from the RF model training set

The above figure (Figure 4-18) and table (Table 4-9) show the results for succinic acid as the reference compound and with one data point.

Table 4-9 succinic acid as the reference compound and related compounds with Tanimoto score

Compound name	**Tanimoto coefficient**	**Change in log RMSE with API in training set**
hexanedioic acid	0.90	0.01
octadecanoic acid	0.75	0.02
2-hydroxy-1,2,3-propanetricarboxylic acid	0.53	0.00
fumaric acid	0.50	-0.02

Hexanedioic acid is the most similar compound with a Tanimoto score of 0.90. The RMSE for the RF models did not improve significantly for all points with the inclusion of succinic acid. This is probably due to their only being one data point for succinic acid.

It seems that for this dataset that when there is a high Tanimoto score > 0.86 and with at least some data points for the reference compound the RMSE for the RF model decreases although this is not validated due to the limited number of data points available in dataset 3.

4.3.2 COSMO*therm* predictions as a descriptor

Using COSMO*therm* predictions as a descriptor in the solute-Fold model shows a significant reduction in error as can be seen by the error density plot (Figure 4-19). The error is reduced from an RMSE of log 0.99 to log 0.91 this could possibly be due to some molecular information being available in the COSMO*therm* prediction that is not available in the molecular descriptors from MOE. The inclusion of the predictions as a descriptor gives the model an “anchor” from which to begin as there has been some minor improvements having a positive effect on the model.

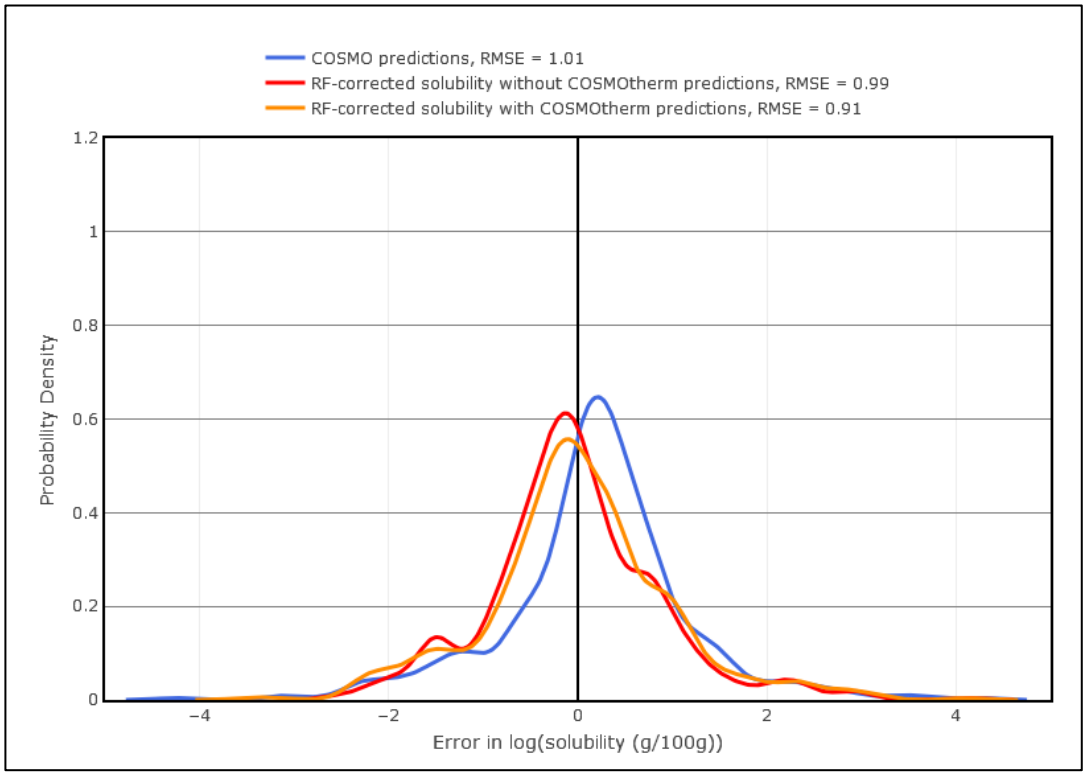


Figure 4-19 Density plot for errors from COSMOtherm predictions and RF corrected solubilities using 2D descriptors with and without COSMOtherm predictions as a descriptors

Both the COSMO*therm* predictions and the descriptors work in tandem to produce the correction factor for solubility.

4.3.3 Descriptor analysis

The results from the COSMO*therm* predictions were combined with descriptors from MOE and the descriptors were split into three groups: all descriptors (both 2D and 3D) and two subsets 2D and 3D descriptors and RF algorithms completed. The plots show the difference in error when experimental solubility data is compared with predicted solubility data.

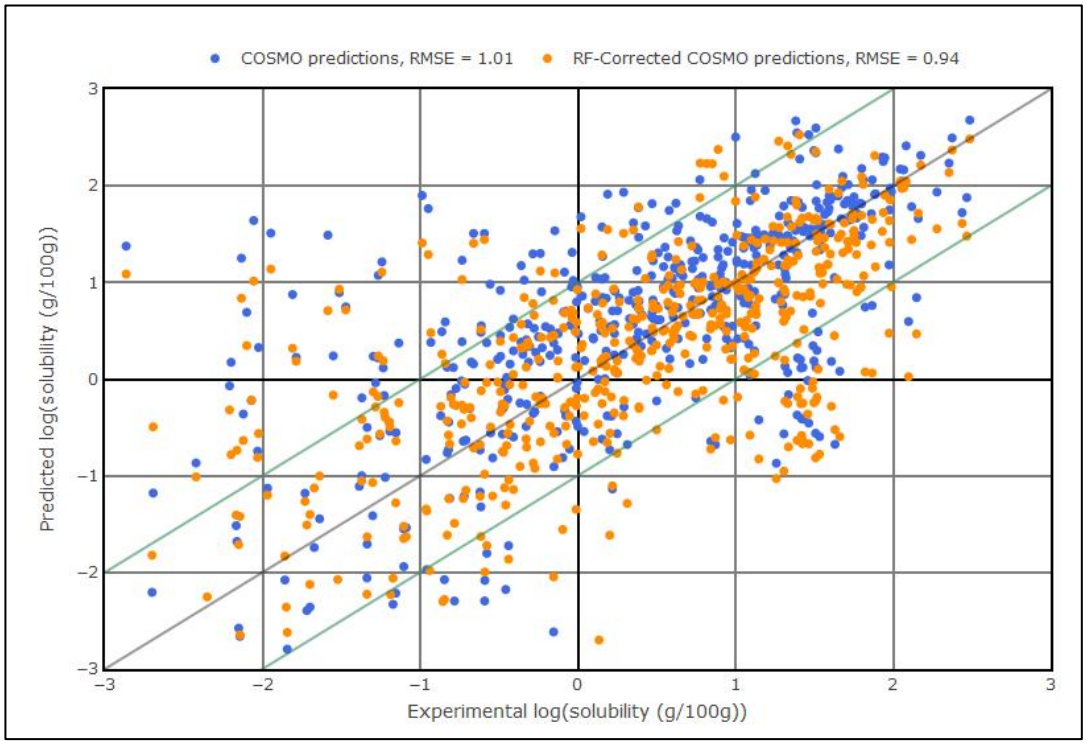


Figure 4-20 Correlation between COSMOtherm, experimental data and RF corrected solubility using both 2D and 3D descriptors

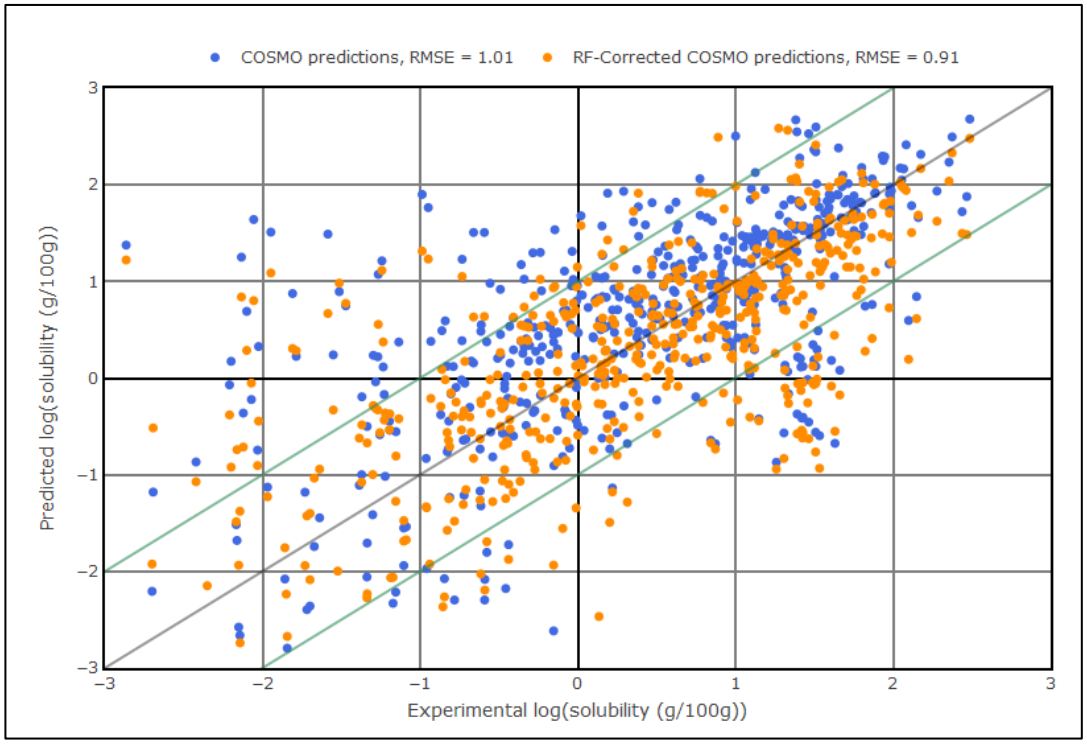


Figure 4-21 Correlation between COSMOtherm, experimental data and RF corrected solubility using 2D descriptors only

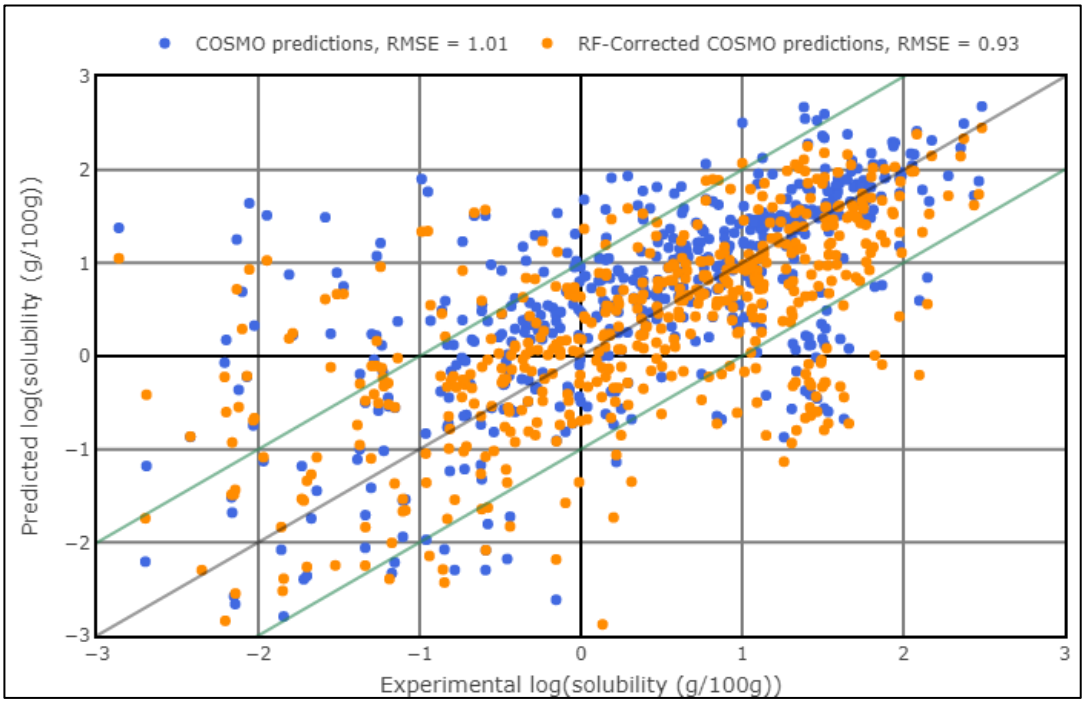


Figure 4-22 Correlation between COSMOtherm, experimental data and RF corrected solubility using 3D descriptors only

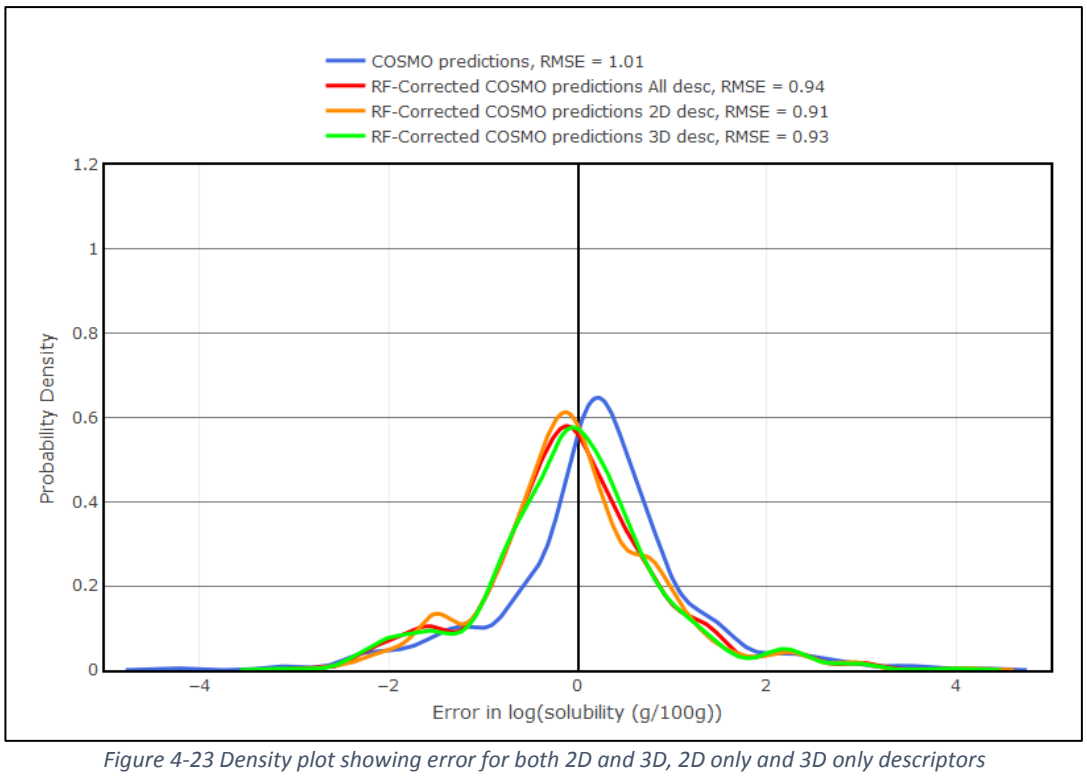


Figure 4-23 Density plot showing error for both 2D and 3D, 2D only and 3D only descriptors

Figure 4-23 and the other plots (Figure 4-20, Figure 4-21, Figure 4-22) shows the log RMSE for the three groups with not a significant difference between results. These

models have improved on the over-predictions from COSMO*therm*. The algorithm using all descriptors had an RMSE of log 0.94, 3D descriptors log 0.93 and 2D descriptors log 0.91.

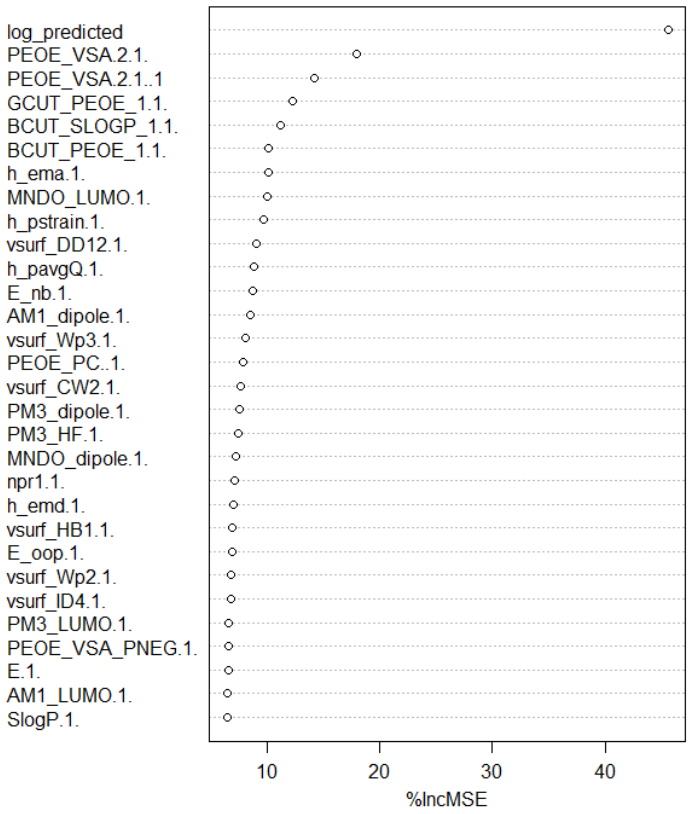


Figure 4-24 Variable importance plots from RF showing the most important descriptor used in the solute-Fold model for both 2D and 3D descriptors

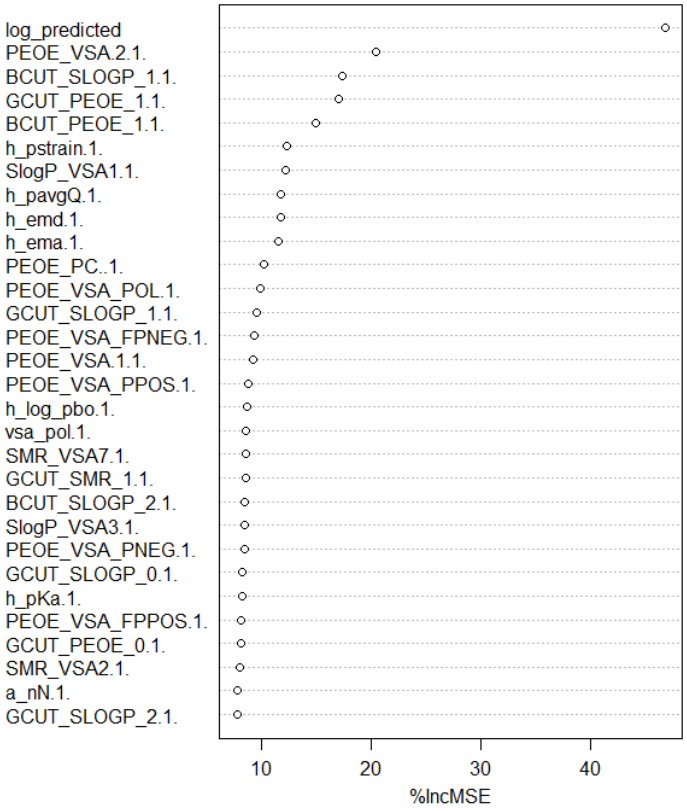


Figure 4-25 Variable importance plots from RF showing the most important descriptor used in the solute-Fold model for 2D descriptors

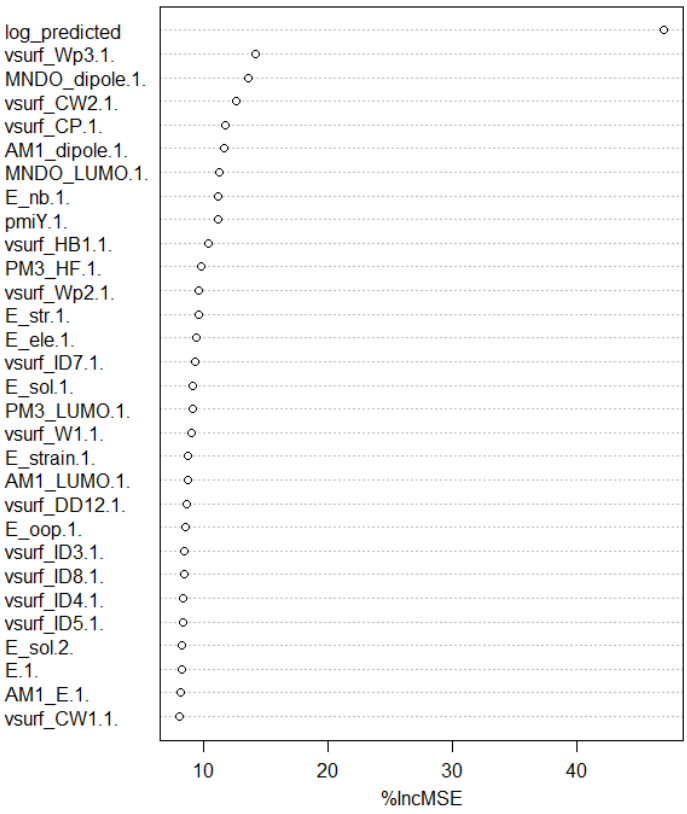


Figure 4-26 Variable importance plots from RF showing the most important descriptor used in the solute-Fold model for 3D descriptors

The above figures show the variable importance plots taken from each RF algorithm Figure 4-24 shows both 2D and 3D descriptors, Figure 4-25 shows 2D descriptors only and Figure 4-26 shows 3D descriptors only. Variable importance plots provide a list of the most significant variables in descending order. %IncMSE is the mean decrease in accuracy. All three algorithms show "log_predicted" as the most important descriptor; this descriptor being the log of the COSMO*therm* prediction included in the training set. Nearly all the descriptors in the above plots are descriptors for solutes; this could be because there are more solutes than solvents in the dataset and therefore more variance in the solutes. In Figure 4-24 the first five most important descriptors are 2D descriptors which match some of the top descriptors in Figure 4-25. The 3D descriptors in Figure 4-24 are further down the plot than the 2D descriptors. One of the most

important descriptors for the 2D model and the model that uses both 2D and 3D descriptors are partial equalisation of orbital electronegativity descriptors (PEOE). These descriptors are an abstract description of atomic charges through the partial equalisation of atom electronegativities of a molecule.

Partial dependence plots show the marginal effect that a descriptor can have on the response from the ML algorithm. It can show the complexity of the relationship between descriptor and outcome *i.e.* linear or a more complex outcome. Each descriptor although important is more important as a sum of parts when all descriptors are taken into consideration.

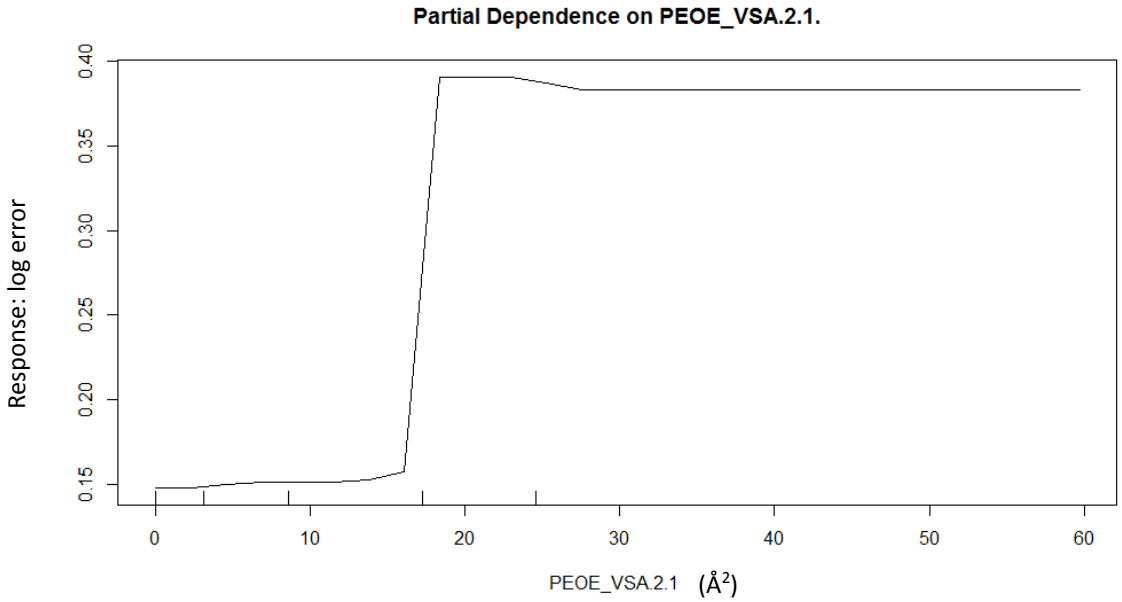


Figure 4-27 Partial dependence plot for 2D molecular descriptor PEOE_VSA 2 for solutes. Each tick mark represents 10% of the dataset

The above plot (Figure 4-27) shows the partial dependence plot for the descriptor PEOE_VSA 2 which is a 2D descriptor. It is one of the most important descriptors for both models. This descriptor describes the total vdWs surface area with units of square Angstroms (Å^2) of a molecule with a partial charge in the range between 0.10-0.15. The

atoms that have a partial charge of between 0.1-0.15 are carbon atoms which are hydrophobic. For dataset 3 this descriptor has a value of between 0-60 Å^2 and shows on the y-axis the range of the response for the RF algorithm of between just below an RMSE value of log 0.15 to around log 0.45. When the value of the descriptor is low the response is lower than at higher values.

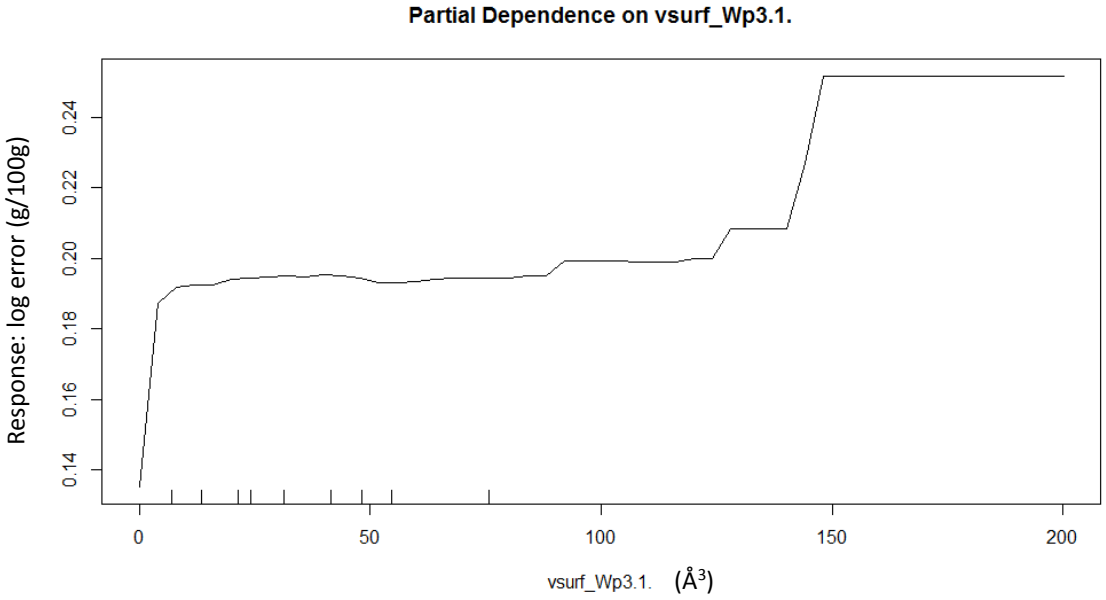


Figure 4-28 Partial dependence plot for 3D molecular descriptor vsurf_Wp3.1 for solutes. Each tick mark represents 10% of the dataset

Vsurf_Wp3 (Figure 4-28) is a 3D descriptor which describes the polar volume of a molecule with energies of -3 kcal/mol. It is one of the most influential descriptors in the 3D descriptor only model. It is also an influential descriptor in the 2D and 3D descriptor combined model. The y-axis variance correlates to an average response value of between zero to log 0.26. The low values up to approximately 125 Å^3 have an average response of log 0.20 which then increases to log 0.26 after 125 Å^3.

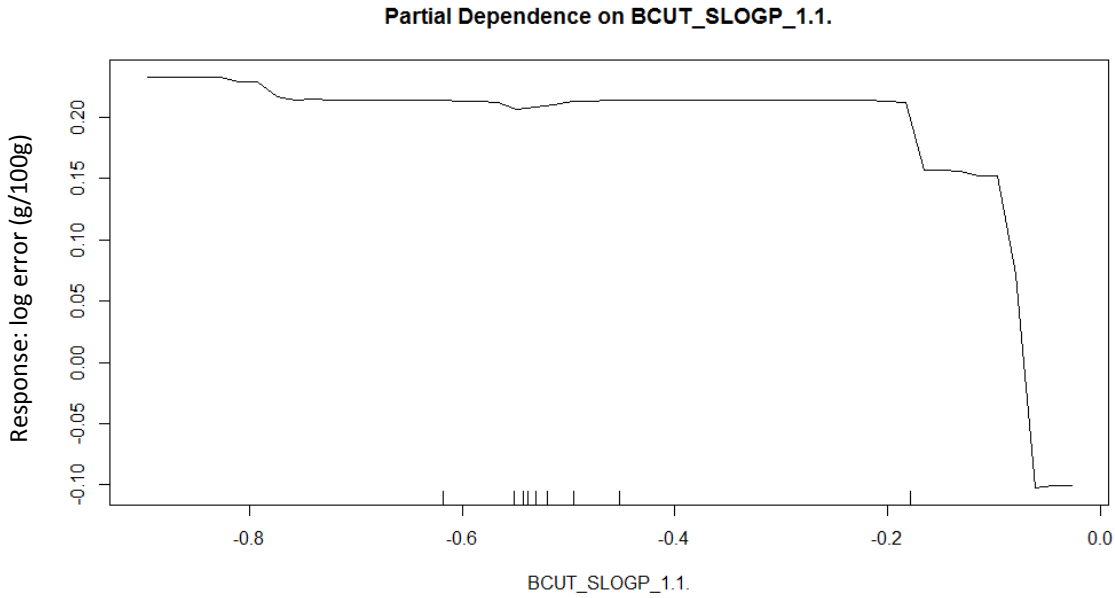


Figure 4-29 Partial dependence plot for 2D molecular descriptor BCUT_SLOGP_1 for solutes. Each tick mark represents 10% of the dataset

BCUT_SLOG_1 (Figure 4-29) is one of three BCUT descriptors using atomic contribution to logP instead of partial charge (Wildman and Crippen, 1999). An analysis of the difference in values of this descriptor identified three compounds caffeine, urea and a GSK compound were responsible for the change in value of the response. Although urea and caffeine share the N(C=O)N functional group the GSK compound does not and there are no distinct characteristics for the three compounds. These compounds were removed from the training set and a RF algorithm was rerun and the variable was re-examined.

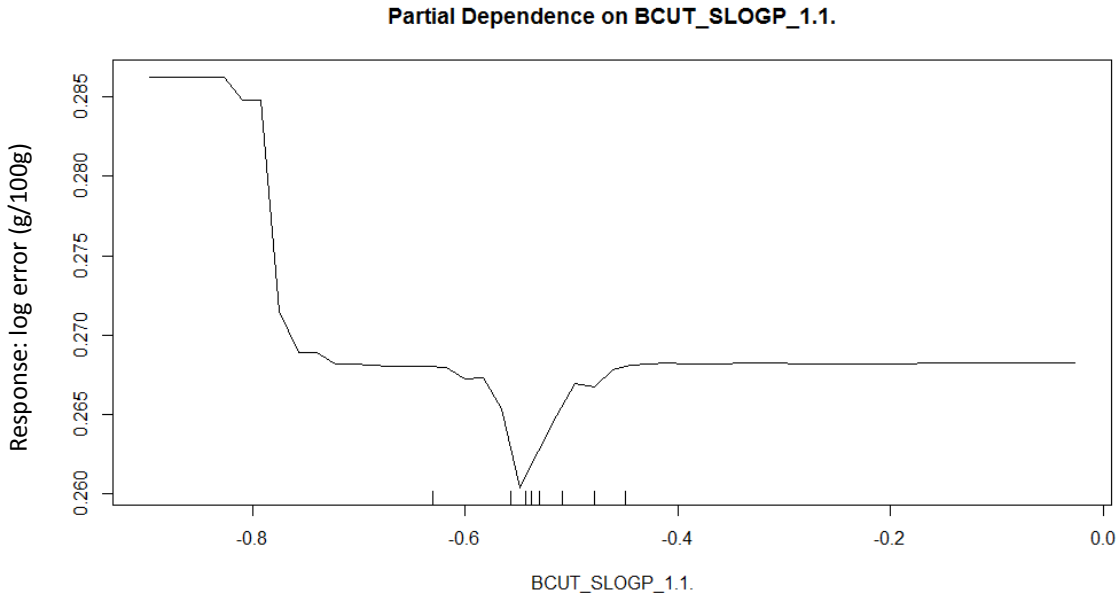


Figure 4-30 Partial dependence plot for 2D molecular descriptor BCUT_SLOGP_1 for solutes without caffeine, urea and a GSK compound in the training set. Each tick mark represents 10% of the dataset

Figure 4-30 shows a partial dependence plot for the RF model without the three compounds. The variance in response is insignificant when compared to Figure 4-29. This is because this descriptor has decreased in importance.

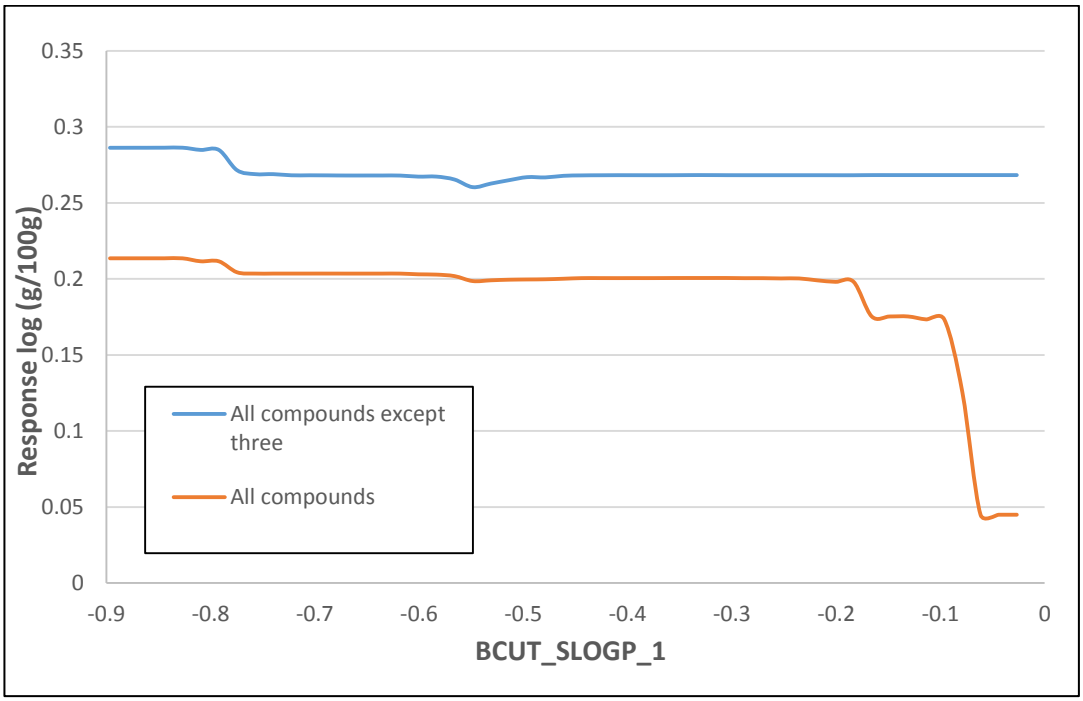


Figure 4-31 Partial dependence plot for 2D molecular descriptor BCUT_SLOGP_1 for solutes showing plot with and without caffeine, urea and a GSK compound in the training set

This can be seen more clearly on the above plot (Figure 4-31) as the plot without the three compounds displays little variance in value.

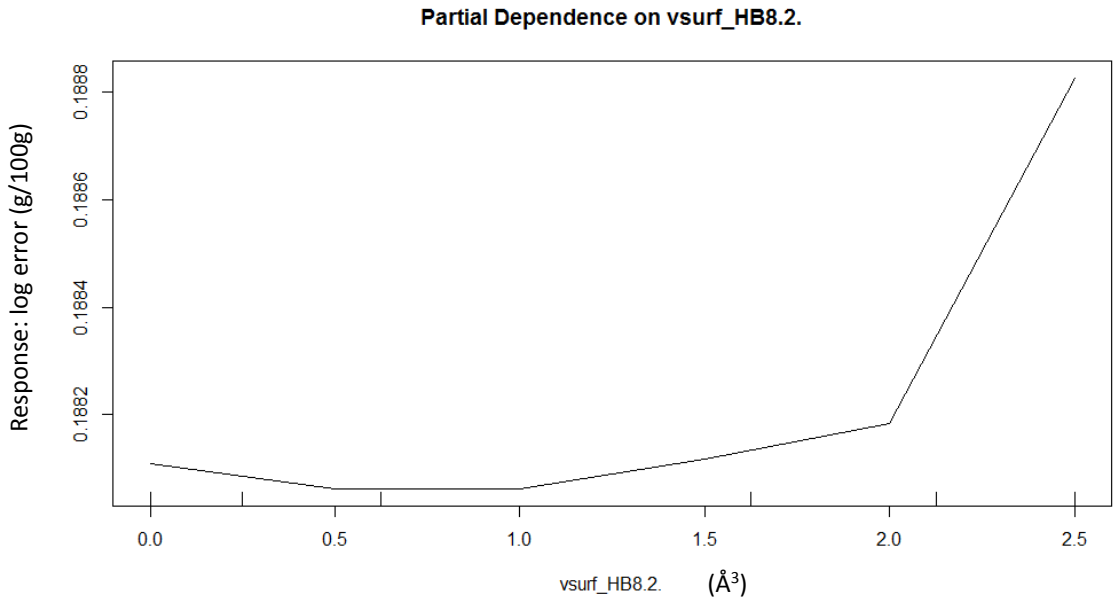


Figure 4-32 Partial dependence plot for 3D molecular descriptor vsurf_HB8 for solvents. Each check mark represents 10% of the dataset

Vsurf_HB8 (Figure 4-32) is a 3D descriptor and is the least influential descriptor for the 2D and 3D descriptors combined model and describes surface volume hydrogen bond donor capacity. The y-axis shows how little influence the values of this descriptor have on the response with a range of response from log 0.1882-0.1888. This is an insignificant descriptor for this model.

One of the major problems with 3D descriptors is that they are inconsistent as a molecule can have many conformers. Different values for 3D molecular descriptors can be obtained for each conformer whereas the 2D descriptors values are consistent for all conformers. Another advantage of 2D descriptors is that they take less computational time to generate than 3D descriptors (BIOVIA, 2017).

It is possible that you could remove the descriptors that have limited effect on the model and only include “more important” descriptors, however the removed descriptors could have a co-dependency with the more important ones *i.e.* the relationship between the descriptors called Shoe Size and Height in the example given in section 1.3.13.2. These descriptors could be co-dependant as Height increases so does, in many cases, Shoe Size. Although these descriptors might be correlated in most examples there will be examples where Height and Shoe Size are not correlated and therefore removing one of these descriptors would have a detrimental effect on the model. If too many descriptors were removed it could affect the accuracy of the model as collectively there could be some effect. In addition, if some descriptors were removed those descriptors might have to be re-introduced when new data points were added to the model as those descriptors might rise in importance.

4.3.4 Correlation of descriptors

A correlation script was used to remove descriptors if the descriptor correlated with another descriptor from the 2D descriptors *e.g.* if a descriptor had 90% (0.9) correlation or above that descriptor was removed. This was to establish if the model improved with less descriptors. As Table 4-10 shows the improvement of the solute-Fold model was not significant as descriptors were removed. As the model did not improve substantially with the removal of descriptors, except for descriptors with zero variance, descriptors were not removed for the models in this project. The reason for this is that RF deals with correlated descriptors well and will always choose the best descriptors to build the model (see section 1.3.13.3).

Table 4-10 2D Descriptor correlation level and RMSE values

Correlation level	No. of descriptors	solute-Fold model log RMSE
1	348	0.91
0.9	199	0.91
0.8	130	0.90
0.7	95	0.88
0.6	69	0.87
0.5	50	0.89

4.3.5 Unit analysis

Most of the research into solubility prediction uses the solubility value converted into log units. The analysis of units was obtained by comparing the results of COSMO*therm* error predictions by carrying out RFs using log units of g per 100g of solvent and in absolute units of g per 100g of solvent for the response. The RMSE of the error prediction in log units was converted back to g/100g to compare with the absolute value of g/100g RMSE. The model used 2D descriptors only.

Table 4-11 RMSEs from solute-Fold algorithms in g/100g and converted from log units of g/100g using 2D descriptors

	g/100g corrected solubility error	log g/100g converted to g/100g
RMSE	39.95	39.55

The above table (Table 4-11) shows the results for both models with the converted units model having an error of 39.55g/100g showing a very slight improvement, which can be considered insignificant, over the absolute units model at 39.95g/100g. This demonstrates that there is not much difference in the quality of the model when the units are changed. It was decided to use log units for the response in all models for ease of comparison.

4.3.6 Error prediction or solubility prediction

Most studies into solubility using ML algorithms have concentrated on the prediction of solubility (Kan, 1996, Ruether and Sadowski, 2009). The study in this section focuses on the prediction of the error between COSMO*therm* and experimental values. An analysis into the comparison of direct solubility predictions by RF only and COSMO*therm* predictions was carried out. For the solubility predictions by RF only, COSMO*therm* predictions were not required.

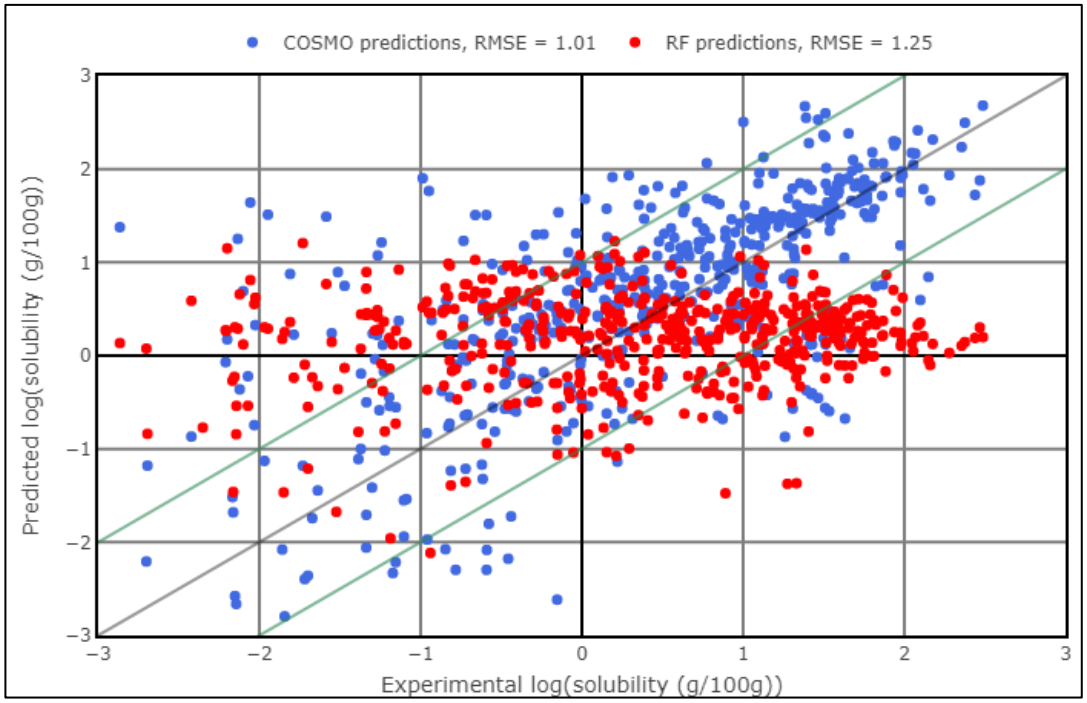


Figure 4-33 correlation between COSMOtherm predictions, RF solubility predictions and experimental data using 2D descriptors.

Figure 4-33 shows the distribution of COSMO*therm* predictions versus experimental solubility in blue and RF solubility predictions versus experimental solubility in red using only the RF model with 2D descriptors from dataset 3. COSMO*therm* is more accurate with an RMSE of log 1.01 when compared with an RMSE of log 1.25 for the solubility predictions. It is possible that the RMSE will improve with an increased dataset.

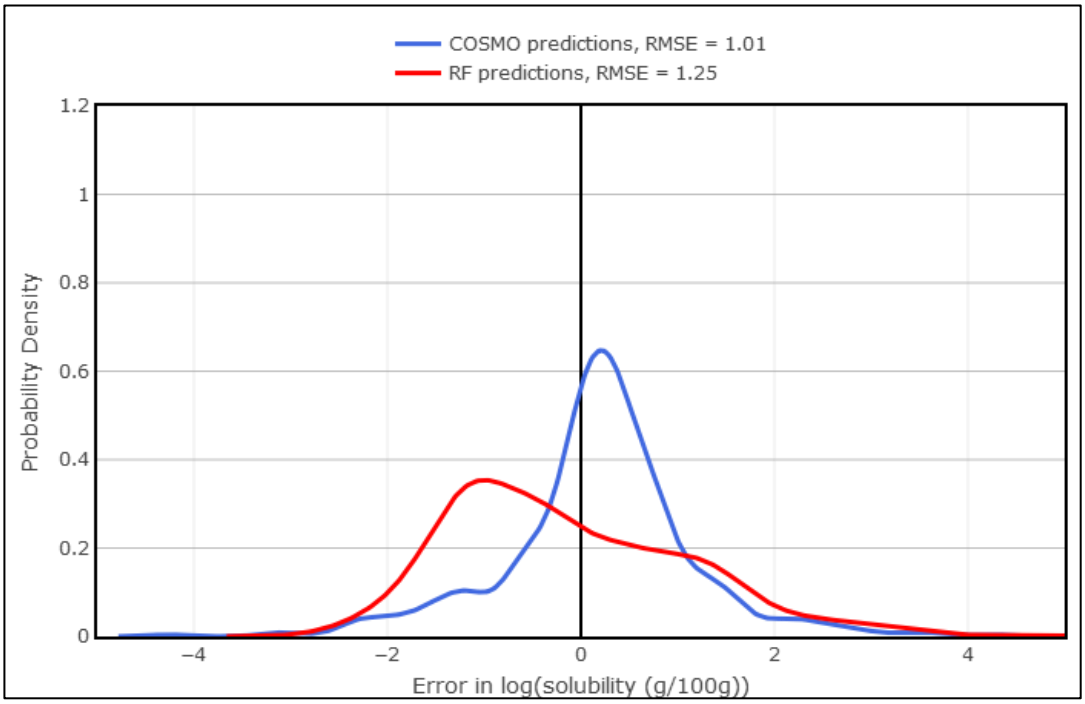


Figure 4-34 Density plot for errors from COSMOtherm predictions, RF solubility predictions and RF corrected solubility using 2D descriptors

Analysing the error distribution between the two methods Figure 4-34 shows the density plot for COSMO*therm* predictions and RF solubility predictions. It shows an over-prediction for COSMO*therm* and an under-prediction for the RF method. The distribution for the RF solubility prediction shows a broader spread of error than the other two methods with an emphasis on under prediction. The model is not consistent and in Figure 4-33 nearly all the points for the RF solubility model are between log 1 and -1 rather than having the more evenly distributed error for COSMO*therm* predictions. Clearly from the results the RF model is inferior to COSMO*therm* predictions. It is probable that the RF model will improve with an increase of data points as the trend is usually an improvement of the models when more data points are added.

4.3.7 Dataset comparison

Three datasets have been used in this project: the first two datasets are a subset of the third as experimental solubility data points were added over time. ML algorithms using the solute-Fold model to obtain a predicted error were calculated using each of the datasets with both 2D and 3D descriptors, 2D descriptors only and 3D descriptors only. Although it has been established in section 4.3.3 that 2D descriptors only produce the best model the results here, and in further sections, were obtained concurrently.

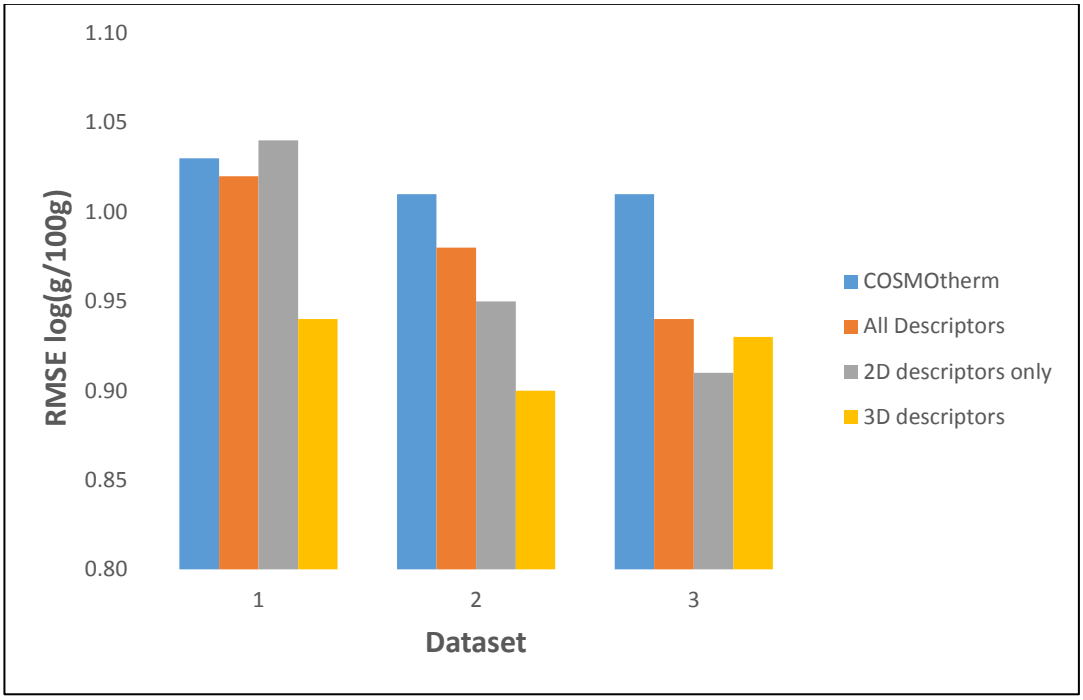


Figure 4-35 log RMSE for different datasets and their descriptor type

Table 4-12 and Figure 4-35 show the results from all data sets. As the number of data points increases there does seem to be, for most descriptors, a small reduction in log RMSE but it is difficult to say if this is a result of an increase in data points; a larger dataset maybe required to assert this definitively.

Table 4-12 Comparison of RMSE's for solute- Fold models from different datasets

	Data points	**COSMO*therm***	**All descriptors**	**2D descriptors only**	**3D descriptors only**
Dataset 1	281	1.03	1.02	1.04	0.94
Dataset 2	420	1.01	0.98	0.95	0.90
Dataset 3	529	1.01	0.94	0.91	0.93

The hypothesis for using only 2D descriptors is reinforced by these results. For all further models dataset 3 was used.

4.3.8 solute-Fold cross-validation

The solute-Fold model shows a minor improvement over the predictions of COSMO*therm* (Figure 4-36) with the 2D descriptors reducing the RMSE by 0.02-0.03 (Table 4-13) when compared with the models from 2D and 3D descriptors.

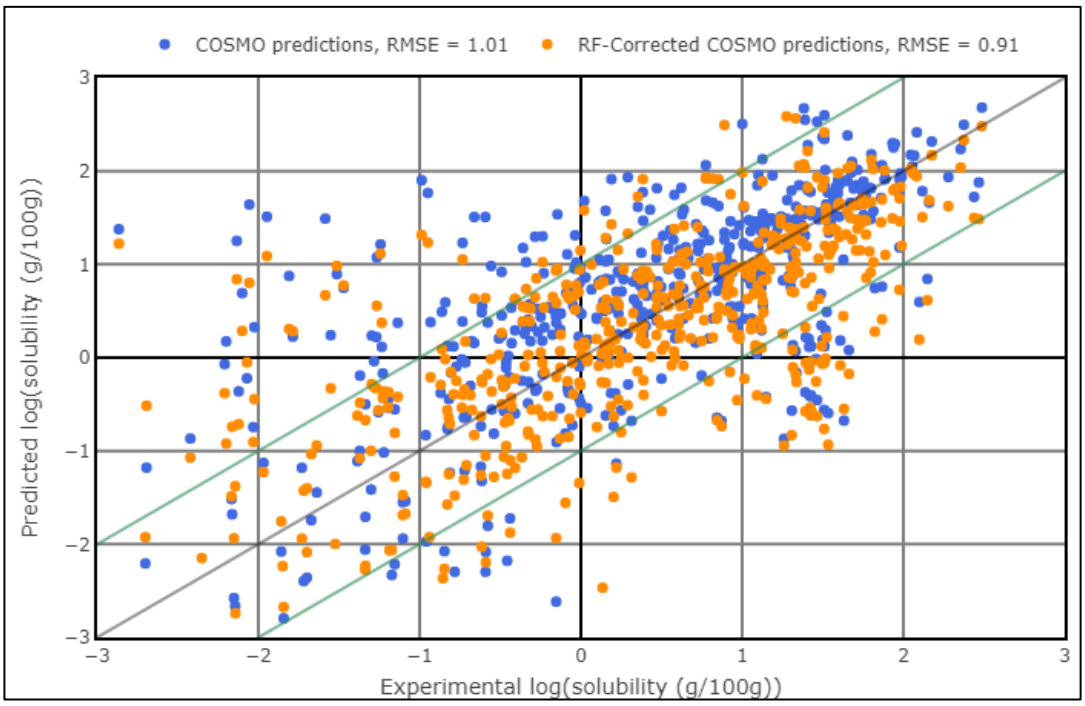


Figure 4-36 Correlation between experimental data, COSMOtherm predictions and solute-Fold RF corrected predictions

This is the type of model that would be used when obtaining a correction factor for solubility for a new compound that had no data points in the dataset. This is a more realistic model for a new compound as the model has no prior knowledge of the solute.

Table 4-13 RMSE for the solute-Fold model

Descriptor Type	**Log RMSE solute-Fold**
Both 2D and 3D	0.94
2D only	0.91
3D only	0.93

4.3.9 k-Fold cross-validation

Table 4-14 shows the results for the k-Fold cross-validation model for all descriptor types. All three k-Fold models show a significant improvement over the solute-Fold model with around log 0.3 of improvement for each descriptor type used.

Table 4-14 Comparison of the RMSE for k-Fold RF model

Descriptor type	Log RMSE k-Fold
Both 2D and 3D	0.65
2D only	0.64
3D only	0.64

The improvement over the solute-Fold model shows that for a solute that is in both the training set and test set there is an improvement in the model predictive ability.

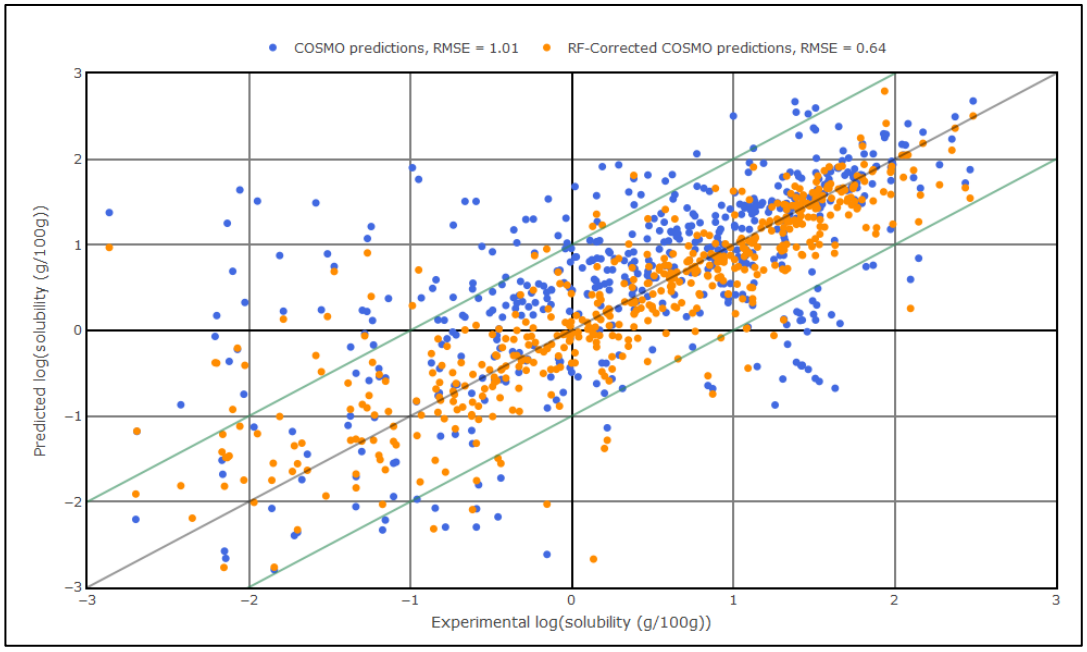


Figure 4-37 Correlation between experimental data, COSMOtherm predictions and K-Fold model RF corrected predictions for 2D descriptors

Figure 4-37 shows the RMSE error for the k-Fold model. The orange points show a clear improvement over the COSMO*therm* errors in blue.

4.3.10 Drip-feed model

Figure 4-38 shows the density plot for the drip-feed model. As shown when zero data points are included the RMSE is log 0.93 which matches the results in the solute-Fold model.

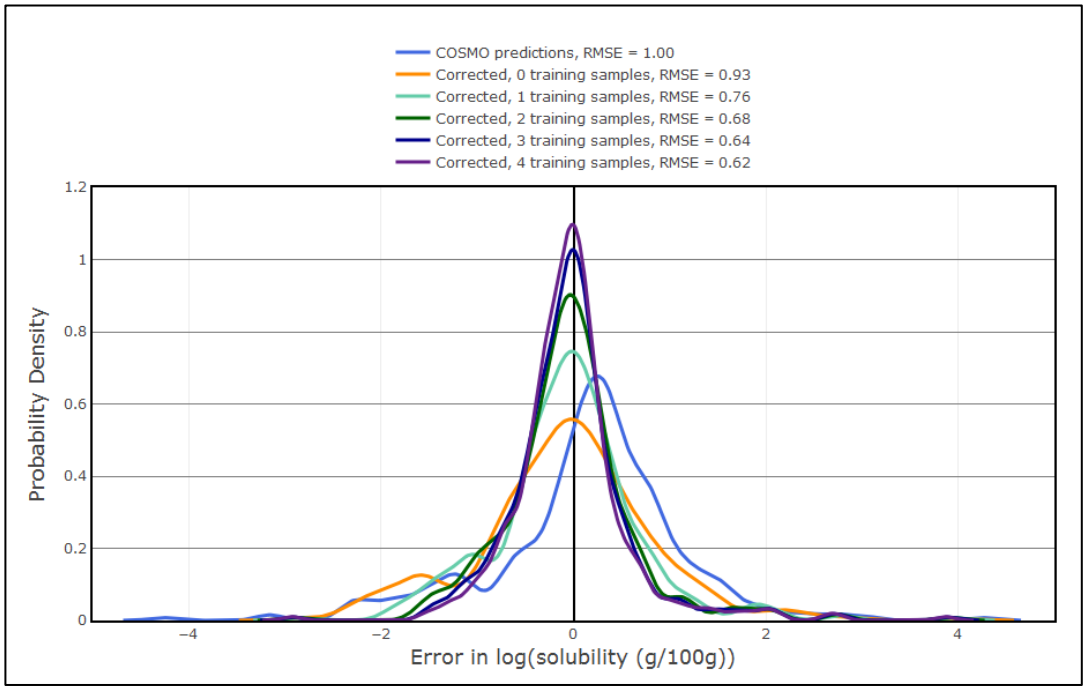


Figure 4-38 Density plot showing error for drip feed model

As another data point is inserted into the model there is a reduction in error. The biggest reduction of log 0.17 occurring with the addition of the one data point.

Table 4-15 mean RMSE for drip-feed model

Drip-feed	Mean log RMSE
COSMO*therm*	1.01
0 points	0.93
1 point	0.76
2 points	0.68
3 points	0.64
4 points	0.62

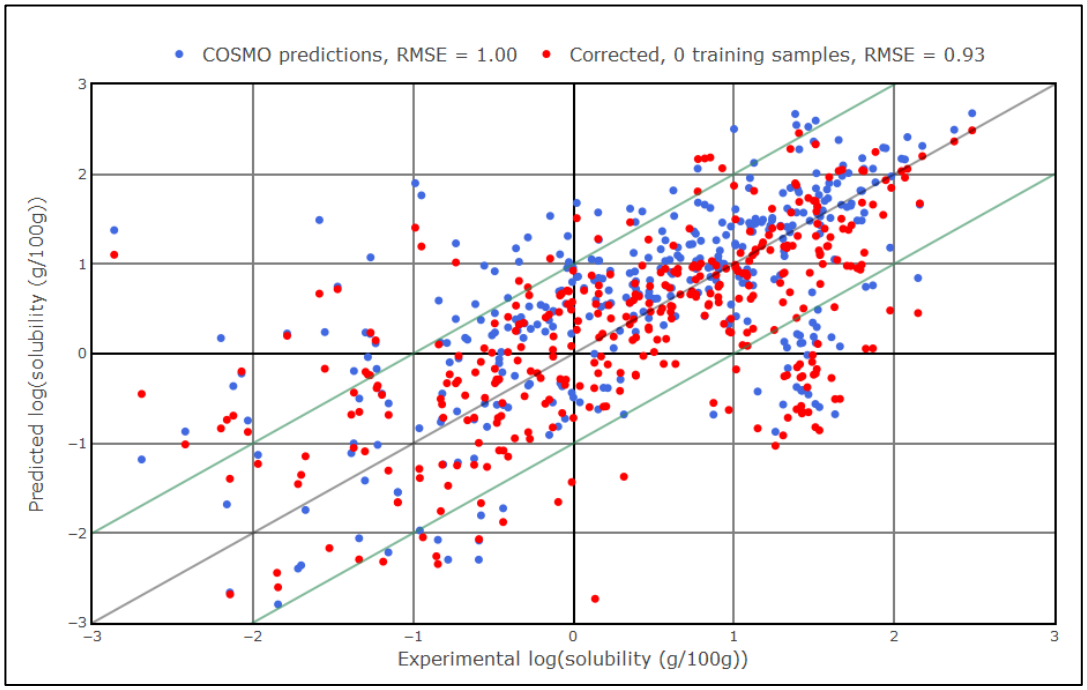


Figure 4-39 Comparison of errors COSMOtherm, drip-feed models and experimental data for 0 solvents in the training set

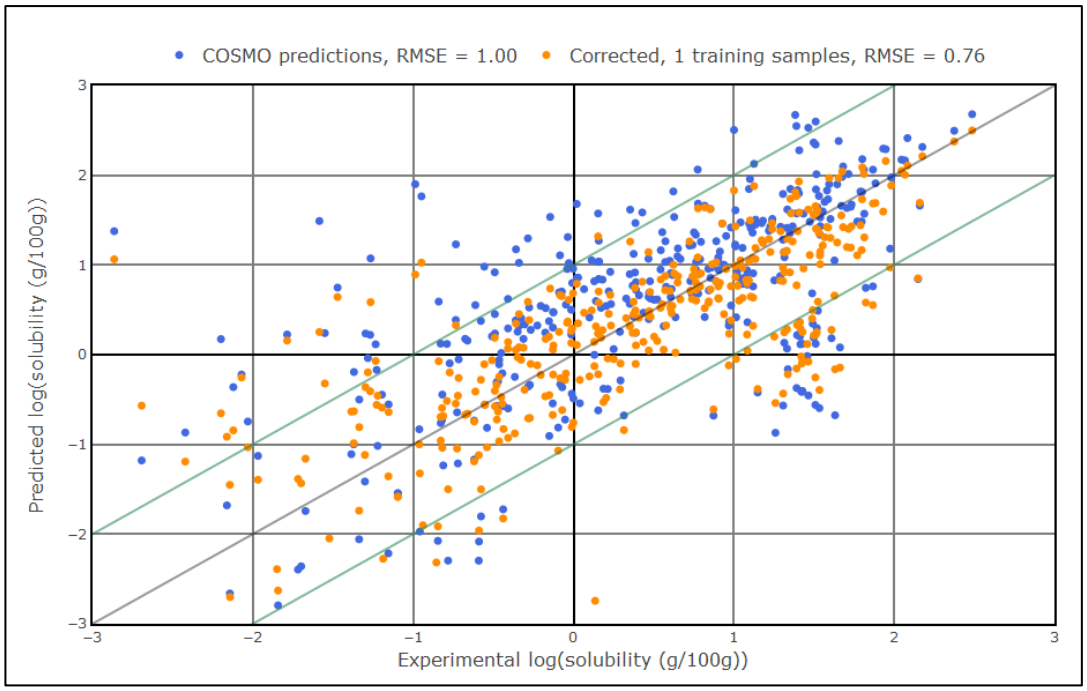


Figure 4-40 Comparison of errors COSMOtherm, drip-feed models and experimental data for 1 solvent in the training set

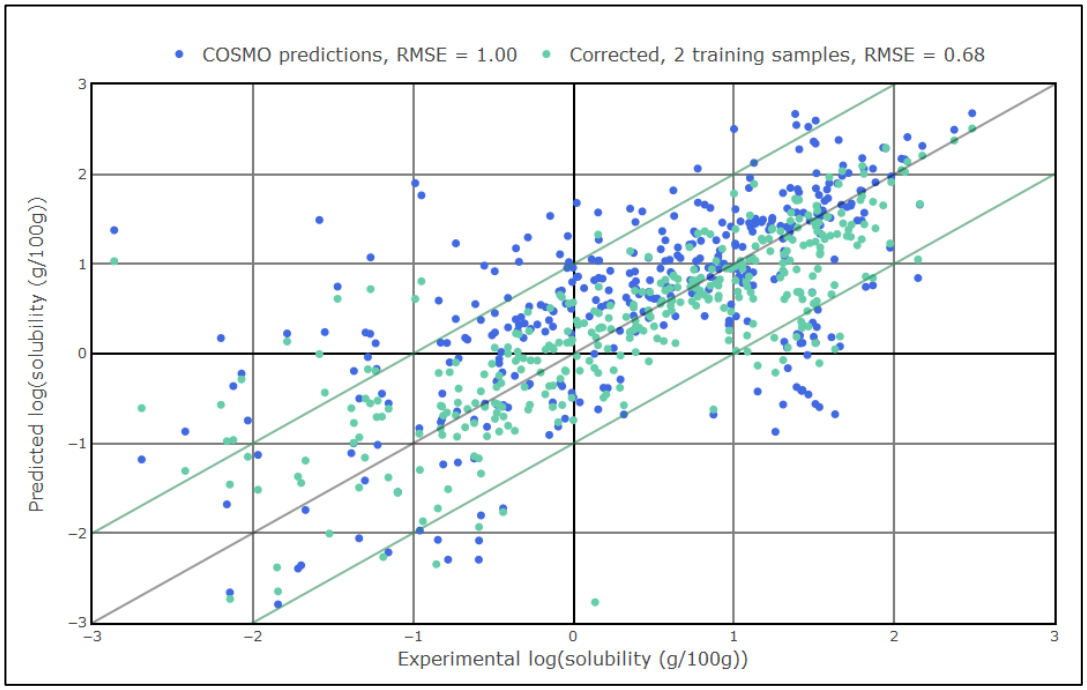


Figure 4-41 Comparison of errors COSMOtherm, drip-feed models and experimental data for 2 solvents in the training set

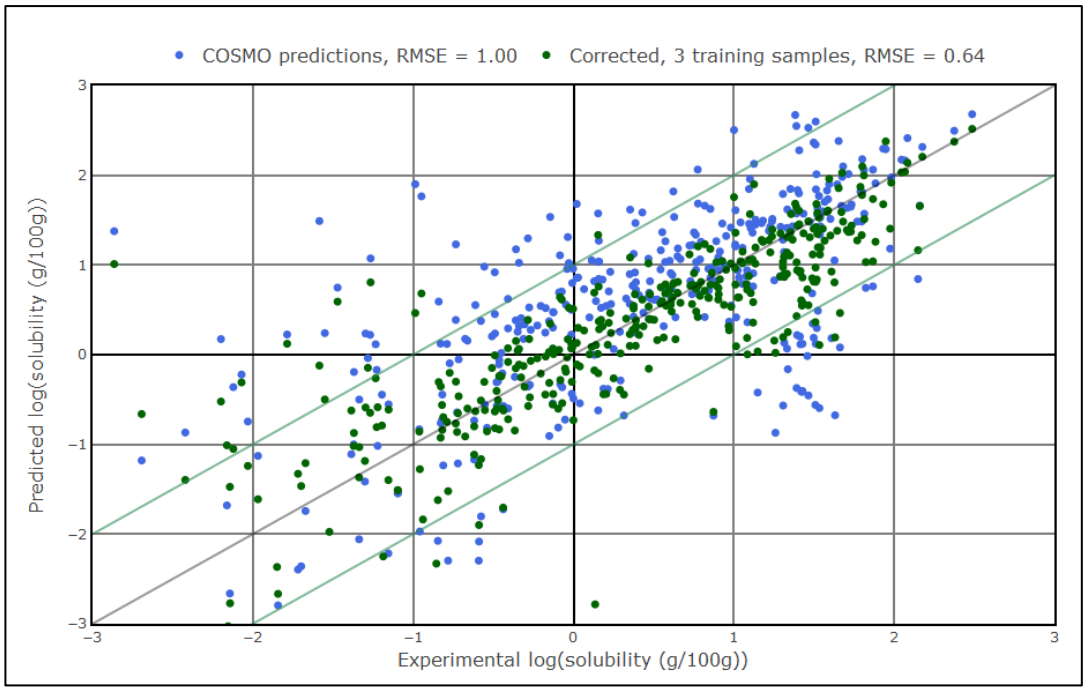


Figure 4-42 Comparison of errors COSMOtherm, drip-feed models and experimental data for 3 solvents in the training set

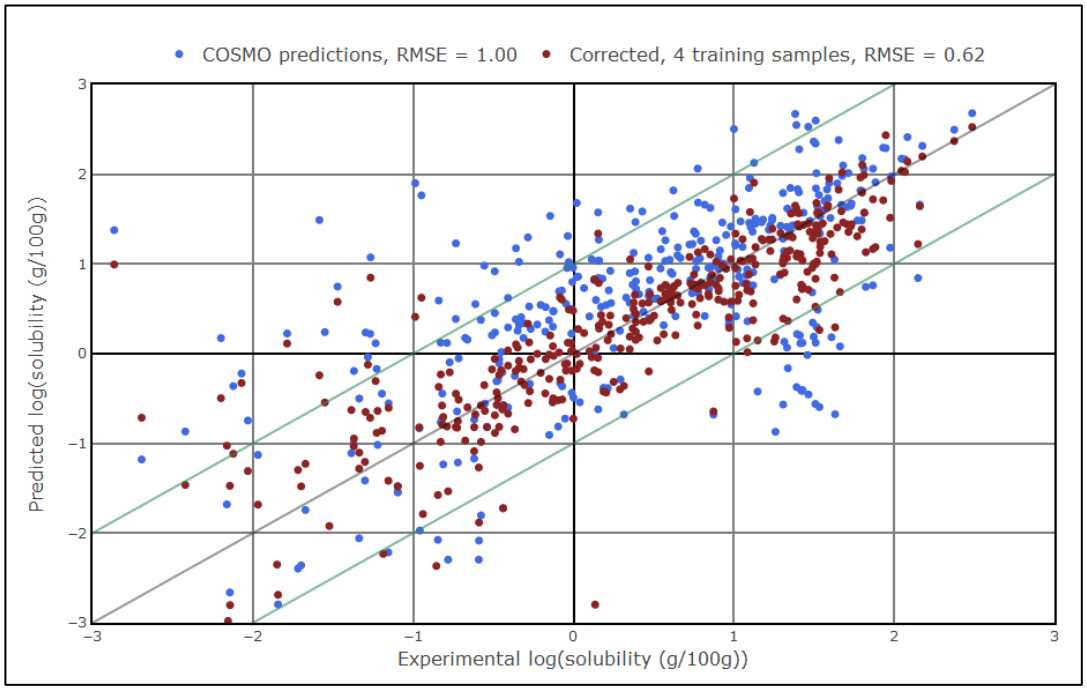


Figure 4-43 Comparison of errors COSMOtherm, drip-feed models and experimental data for 4 solvents in the training set

Figure 4-39 to Figure 4-43 and Table 4-15 show the average RMSE for the drip-feed model. It can be seen that with each addition of a data point for each solute that the RMSE error on average is reduced. This shows that for each additional point the model is improving but it seems to level off at around four points as the incremental improvements are getting smaller. Due to the limitations of the dataset only four data points for each solute were used so it is difficult to validate a greater improvement with additional data points. Four solubility points can be measured in the laboratory and fed back into the model for improved predictions. Also the computational time required for more than four points would be very expensive as the number of solvent combinations increases exponentially.

4.3.10.1 Drip-feed model results for solvent combinations

The RMSE has been calculated for each appearance in the training set of a solvent or a combination of solvents and the mean is given in the tables. Table 4-16 shows the results

with the lowest mean RMSE between the RF corrected solubility and experimental solubility for one solvent per target solute in the model training set. This has been described in detail in section 4.2.5. Methanol is the most accurate with a log RMSE of 0.764.

Table 4-16 mean RMSE for one solvent for the target solute in the drip-feed model training set

Solvents in training set	Mean log RMSE
methanol	0.764
ethanol	0.768
propanol	0.778
1-butanol	0.782
2-propanol	0.798
propanone	0.806

The top five solvents for accuracy are alcohols with propanone being sixth. Only solvents that appear over 100 times in the training set were included.

Table 4-17 mean RMSE for combinations of two solvents for the target solute in the drip-feed model training set

Solvents in training set	Mean log RMSE
ethanol propanone	0.649
2-propanol propanone	0.663
methanol propanone	0.673
1-butanol propanone	0.702
ethanol methanol	0.704
acetonitrile propanone	0.718

The above table (Table 4-17) show the results with the lowest mean RMSE for the drip-feed model with two solvents in the training set, ethanol and propanone are the combination with the lowest RMSE. Propanone is featured in all but one of the top six combinations.

Table 4-18 mean RMSE for combinations of three solvents for the target solute in the drip-feed model training set

Solvents in training set	**Mean log RMSE**
ethanol methanol propanone	0.635
2-propanol methanol propanone	0.640
2-propanol ethanol propanone	0.641
1-butanol ethanol propanone	0.650
1-butanol methanol propanone	0.652
acetonitrile ethanol propanone	0.655

Table 4-18 shows the results with the lowest RMSE for three solvents in the training set. Propanone is featured in all six combinations with either ethanol or methanol as at least one of the components in each combination.

Table 4-19 mean RMSE for combinations of four solvents for the target solute in the drip-feed model training set

Solvents in training set	**Mean log RMSE**
ethanol methanol n-heptane propanone	0.638
1-butanol ethanol methanol propanone	0.642
acetonitrile methanol n-heptane propanone	0.644
acetonitrile ethanol n-heptane propanone	0.644
2-propanol ethylacetate methanol propanone	0.649
2-propanol acetonitrile methanol propanone	0.649

The above table (Table 4-19) show the results with the lowest mean RMSE for four solvents in the training set of the drip-feed model. Propanone, ethanol and methanol feature with n-heptane in half the combinations. However, the RMSE's for these six have not changed significantly when compared to Table 4-18. This could mean that only three solvents are required instead of four in the RF model to achieve a significant improvement in RMSE.

4.3.10.2 Drip-feed model results for individual solvents in solvent combinations

Table 4-20 shows the mean RMSE for individual solvents with one solvent in the training set. When a particular solvent is drip-fed into the training set either singularly or part of a combination of solvents the RMSE is calculated. The mean is then calculated from the number of appearances in the training set of that solvent.

Table 4-20 mean RMSE for individual solvents with one data point for that solvent in the training set for drip-feed model

Solvent	**Mean log RMSE**
methanol	0.764
ethanol	0.768
propanol	0.778
1-butanol	0.782
2-propanol	0.798
propanone	0.806
acetonitrile	0.872
ethyl acetate	0.877
tetrahydrofuran	0.931
toluene	0.938
n-heptane	0.946
isopropyl acetate	0.954
methyl-t-butyl ether	0.955
butanone	0.966
Acetic acid	0.967
water	0.987
2-methyltetrahydrofuran	1.021
4-methyl-2-pentanone	1.025
cyclopentylmethylether	1.040
anisole	1.114

The five solvents with the lowest RMSE are all alcohols and do not have a significant difference in RMSE.

Table 4-21 mean RMSE for individual solvents with a two solvent combination in the training set for drip-feed model

Solvent	**Mean log RMSE**
propanone	0.762
methanol	0.784
ethanol	0.801
1-butanol	0.808
acetic acid	0.818
propanol	0.820
ethyl acetate	0.825
2-propanol	0.829
acetonitrile	0.854
butanone	0.888
tetrahydrofuran	0.889
methyl-t-butyl ether	0.891
isopropyl acetate	0.906
cyclopentylmethyl ether	0.909
2-methyltetrahydrofuran	0.915
n-heptane	0.916
toluene	0.918
water	0.933
4-methyl-2-pentanone	0.981
anisole	1.037

Table 4-21 shows the mean average RMSE for individual solvents with two solvents in the training set. Propanone has the lowest RMSE. This result is different from the previous table which has alcohols with the lowest RMSE. Propanone is better in combination with other solvents in the training set rather than alone.

Table 4-22 mean RMSE for individual solvents with a three solvent combination in the training set for drip-feed model

Solvent	**Log RMSE**
propanone	0.729
methanol	0.789
ethanol	0.799
ethyl acetate	0.803
1-butanol	0.804
propanol	0.822
2-propanol	0.822
acetonitrile	0.828
butanone	0.831
methyl-t-butyl ether	0.841
cyclopentylmethyl ether	0.844
n-heptane	0.848
tetrahydrofuran	0.849
isopropyl acetate	0.852
2-methyltetrahydrofuran	0.852
toluene	0.867
water	0.870
4-methyl-2-pentanone	0.939
anisole	0.964

The above table (Table 4-22) for the lowest mean RMSE for an individual solvent in a three solvent combination in the training set. As with the previous table propanone has the lowest RMSE with methanol and ethanol the second and third lowest. This shows agreement with the lowest RMSE in (Table 4-18) which is propanone, methanol and ethanol in combination.

Table 4-23 mean RMSE for individual solvents with a four solvent combination in the training set for drip-feed model

Solvent	**Log RMSE**
propanone	0.690
methanol	0.744
1-butanol	0.745
ethyl acetate	0.751
tetrahydrofuran	0.761
ethanol	0.773
2-methyltetrahydrofuran	0.786
acetonitrile	0.790
2-propanol	0.792
propanol	0.792
water	0.803
butanone	0.804
n-heptane	0.819
toluene	0.833
isopropyl acetate	0.845
methyl-t-butylether	0.847

Table 4-23 shows the RMSE for individual solvents with a four solvent combination in the training set. Again propanone has the lowest RMSE with methanol second.

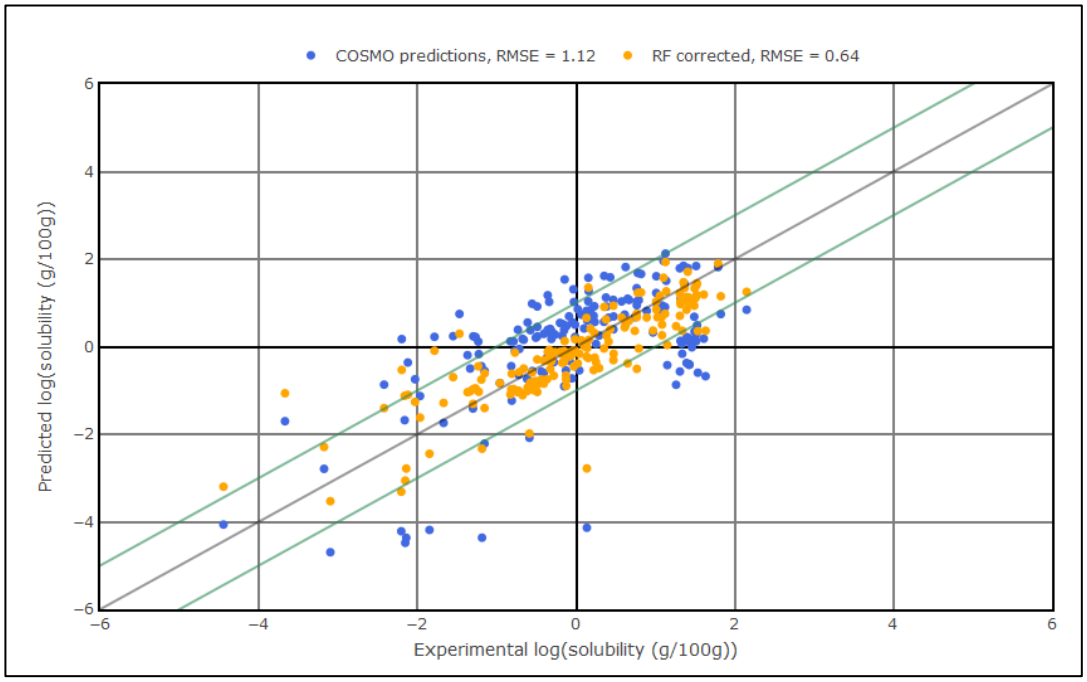


Figure 4-44 Correlation between experimental data, COSMOtherm predictions and drip-feed model RF corrected predictions for three solvents in the training set ethanol, methanol and propanone

To improve the predictions of the ML models at least three solvents are recommended to have the solubility measured experimentally for each solute. Propanone, methanol and ethanol as shown above (Figure 4-44) have the lowest RMSE for the model with all three solvents in the training set. The model has provided a solubility correction factor for both over predictions and under predictions and has improved the accuracy from an RMSE of log 1.12 to log 0.64. Using these three solvents to obtain solubility data for a new compound in the laboratory will inform the RF model efficiently, maximising the return on time and resources.

4.4 Workflow for a new molecular entity

When modelled solubility data for a new molecular entity are required a workflow (Figure 4-45) should be used. Firstly, the COSMO*therm* database would be checked to establish whether the cosmo file for the molecule was already available.

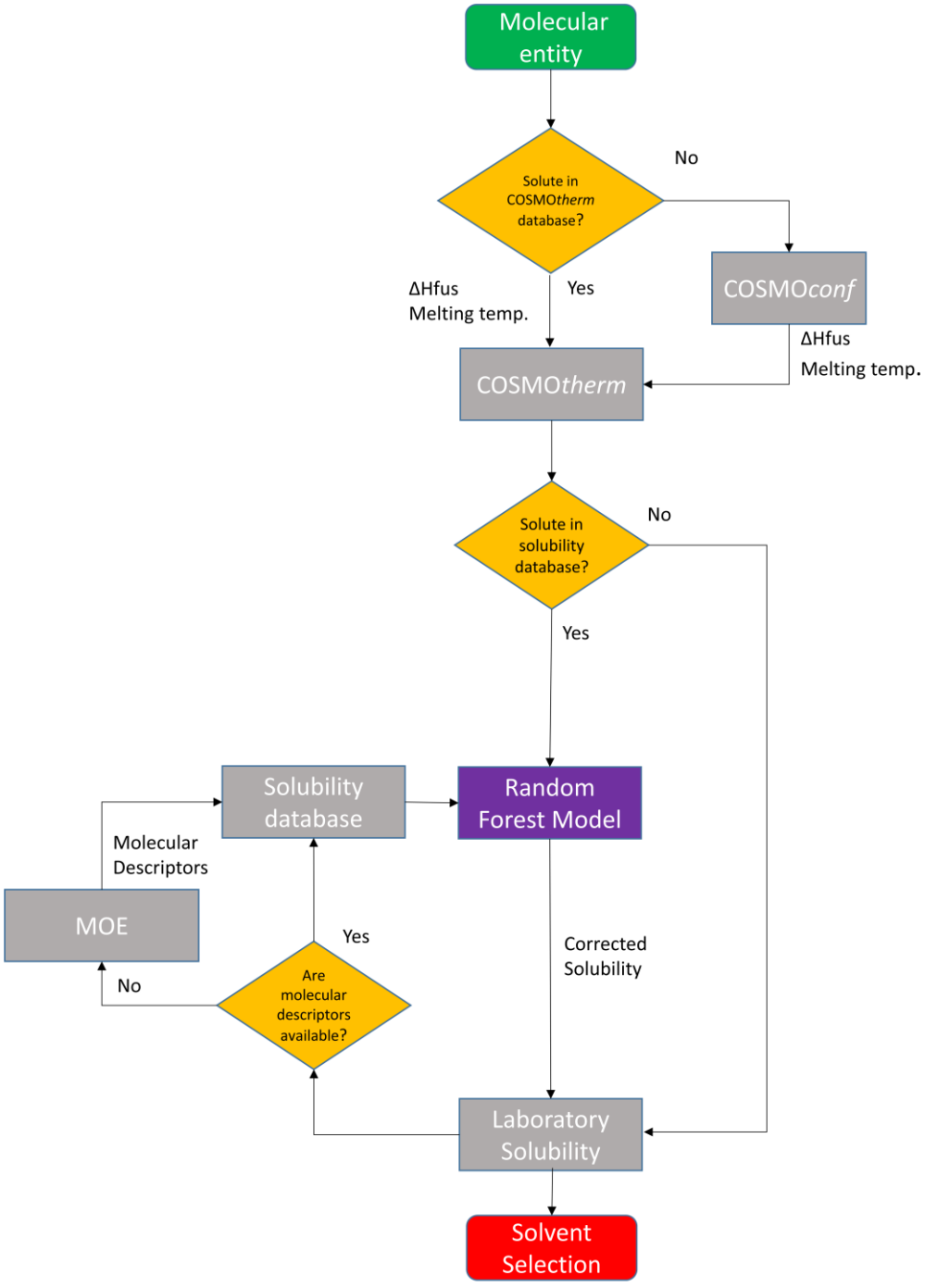


Figure 4-45 Workflow for correction factor for new molecular entity

If the molecule is not in the COSMO database, then COSMO*conf* must be used to obtain a COSMO file. The enthalpy of fusion and melting temperature of the compound must be obtained for use in COSMO*therm*. If the compound is in the solubility database a correction factor would be obtained and the solubility would be corrected. The selected solvents would be tested in the laboratory if the data wasn't already available in the database. This information would then be added to the database. If the compound is not in the solubility database, the COSMO*therm* predictions would be used to select which solvents would be used in the laboratory to get solubility data for the new compound. As shown in the drip-feed model these solvents should be propanone, methanol and ethanol. This data would be input into the database and MOE molecular descriptors generated if not already available. A correction factor would then be obtained for the COSMO*therm* predictions for all solvents. The required solvent would then be chosen. This would be an iterative process.

4.5 Conclusion

Using COSMO*therm* predictions and RF to obtain a correction factor has improved the RMSE of solubility predictions when compared with experimental data. For the dataset used within this work it was found that using an RF model with COSMO*therm* predictions included in the descriptors reduced the overall RMSE to a greater extent than not including predictions as a descriptor. Using 2D descriptors only with the COSMO*therm* is more accurate than using 3D only descriptors or both. This is due to 2D descriptors being consistent for all conformers whereas different values can be given for 3D descriptors for each conformer. For a new chemical entity solubility predictions are required. It is recommended that for any API that requires solubility predictions that

experimental data for at least three solvents with that API are preferred. The selection of these solvents has been shortlisted in an effort to maximise model performance. These measurements will be used train the RF model and improve the accuracy of the predictions. Ethanol, methanol and propanone are preferred but 2-propanol or 1-butanol could be used to replace either of the alcohols without much increase of RMSE. The application of ML algorithms is a powerful tool to improve the accuracy of solubility predictions. Ideally it is hoped that solvent specific models would be built, these however would require more solubility data than was available for this project.

5 Workflow procedure for crystallisation and wash solvent selection using predictive methods and machine learning

5.1 Background

A key step in pharmaceutical production is purification as it facilitates the isolation of impurities which are generated during synthesis and other manufacturing stages. According to the International Conference on Harmonisation (ICH) Guidelines (ICH, 2015) impurities are defined as any component present in the drug substance that are not defined as the chemical entity. Possible impurities can be unreacted material, degradants and by products. (Lakshmana Prabu and Timmakondu, 2010). After synthesis, purification is required to remove as far as possible all impurities formed in the synthesis process. It is important to remove impurities due to possible hazardous nature or due to the effect on solubility, crystal habit (Schmidt and Ulrich, 2012, Lakshmana Prabu and Timmakondu, 2010, Fiebig,Jones and Ulrich, 2007) and dissolution (Prasad *et al.*, 2001, Saleemi,Onyemelukwe and Nagy, 2013, Hendriksen *et al.*, 1998, Thompson *et al.*, 2004, Hendriksen and Grant, 1995, Hulse,Grimsey and De Matas, 2008, Lahav and Leiserowitz, 2001, Witschi and Doelker, 1997). A key aspect of removing impurities is to maintain the physical properties of the API along the manufacturing process (particle size, particle size distribution (PSD) and crystal habit). Selecting the best solvent is required for removal of impurities during purification. Purification, the post process of synthesis and workup, typically consists of two steps; crystallisation and isolation. Crystallisation, is the first purification step, which can define crystal properties such as shape and PSD, crystal habit and polymorph. The basic principle of purification by crystallisation is to crystallise your desired molecule only and

leave the impurities in the mother liquor. However, this is difficult as many impurities are structurally related to the target compound and have similar crystallisation properties. As such, it is often the case that some impurities are crystallised while trying to crystallise the API. Therefore, a second filtration step is required after crystallisation, to remove the mother liquor and to wash the crystals before drying. Drying is used to remove residual solvent molecules trapped between particles to get a dry product (Ottoboni, 2018).

Crystallisation solvent selection is usually done experimentally by testing solvents to get desirable crystal attributes, a solubility curve (metastable zone, solubility slope) for the crystallisation process selected and to get impurity rejection by leaving the impurity in the mother liquor. The wash solvent was usually selected experimentally by operator experience or previous knowledge of similar molecules. Firstly, solubility data was required for the API and of the impurities then secondly, an investigation was required to establish whether physical properties such as crystal habit or dissolution was affected.

Another factor that needed investigation was whether the impurity precipitated into the crystallisation or wash solvent. Miscibility between crystallisation and wash solvent is also important to enhance the diffusion and dilution washing (Ruslim *et al.*, 2007, Ruslim *et al.*, 2009). All of this experimental work takes time and consumes materials and solvent. To reduce the amount of time and materials several methods have been developed to generate a selection approach where solubility and other solute/solvent properties were simulated. Cheng applied a workflow approach but without using a method for predicting solubility (Cheng *et al.*, 2010). This workflow approach consisted

of using experimental data and the modelling of the washing process. A different approach was taken by Abramov (Abramov, 2018) who used COSMO*therm* predictions to select wash solvents to purge impurities. Abramov's publication took place at the same time as the development of this work. However, this method was not combined with the workflow approach taken in this chapter. The method developed in this chapter combines both a workflow approach and the predictive capabilities of COSMO*therm*.

5.2 Aims

To facilitate R & D and to save time, material and solvent consumption, a workflow was developed. The aim of the workflow was to obtain all the information required for cooling crystallisation and wash solvent selection. This workflow was then embedded within a tool that had the ability to rank solvents based on their suitability for crystallisation and the washing process. This tool will rank the solvents for cooling crystallisation with the objective to maximise yield and to minimise the amount of solvent used to crystallise the material. It will also to rank wash solvents in accordance with impurity rejection capability with minimisation of the amount of solvent used.

5.3 Materials

Paracetamol (4-actamidophenol, Bioxtra, ≥99% SLBR2060V), 4-nitrophenol (≥99% 1395915V), methyl-4- hydroxybenzoate (97% BCBL6776V), 4-acetamidobenzoic acid (≥98% 1395915V), 4'chloroacetanilide (97% MKBP5552V), acetanilide (99% STBB0193V), 4-hydroxy acetophenone (99% BCBH8862V), orthocetamol (97%), 4-aminophenol (98% BCBU5190V), metacetamol (≥99% MKCB2268V). Acetaminophen acetate (99% GLLHF-MQ) was supplied by Tokyo Chemical Industries, 4-

hydroxyacetaphenone oxime (Carbosynth FH675681550), 3-chloro-4-hydroxyacetanilide (>98% Acros Organics A0340054), 4-hydroxyphenyl-propanide (>95% Enamine R1989798).

Ethanol (purity ≥ 99.8% (GC), from Sigma Aldrich), 2-propanol (IPA) (purity ≥ 99.5 % (GC), from Sigma Aldrich), n-heptane (purity 99%, from Alfa Aesar), isopropyl acetate (purity 99+ %, from Alfa Aesar), toluene (purity 99%, from Alfa Aesar), anisole (purity 99%, from Alfa Aesar), n-dodecane (purity 99%, from Alfa Aesar), methyl-tert-butyl ether (TBME) (purity 98%, from Sigma Aldrich), cyclohexane (purity 99+ %, from Alfa Aesar), 4-methylpentan-2-one (purity ≥ 99.5% (GC), from Sigma Aldrich), 3-methyl-1-butanol (98% Sigma Aldrich) and acetonitrile (ACN) (purity 99.5%+, from Alfa Aesar).

5.4 Work flow aims and description

For this project a workflow approach was used with *ab initio* predictive models and ML model using predictive and experimental data. Predictive methods were validated using experimental solubility and filtration data was collated by Sara Ottoboni, a CMAC researcher (Ottoboni, 2018). The workflow shows the input parameters required from synthesis and workup. These parameters are the physical properties of API and impurities, such as impurity concentration, ideally the enthalpy of fusion and the melting temperature of all substances. As reported in the workflow stages one to four (sections 5.6.1 - 5.6.4), these properties can be obtained by the analytical characterisation of the material during synthesis and workup, either from literature or by predictive methods (Joback and Reid method (Joback and Reid, 1987)). In stage 5 (section 5.6.5), if solubility and miscibility data are not available, the data will be obtained by predictions using

COSMO*therm*. Stage 6 (section 5.6.6) verifies the data obtained by predicting the solubility of the compound and impurities can be verified experimentally. Once solubility is obtained from COSMO*therm* and from experimentation, this can be utilised by the RF model to obtain a correction factor to improve solubility predictions. Stage 7 (section 5.6.7) used the solubility predictions to rank the solvents for crystallisation and wash solvent selection. Stage 8 (section 5.6.8), screens both the crystallisation and wash solvent for suitability by using solubility predictions. Finally, in stage 9 (section 5.6.9), the workflow proposes the use of a laboratory scale batch filtration unit to verify the selection of solvents from predictions. To verify if the workflow was robust paracetamol was selected as the test compound as it is well researched and the impurities are commercially available.

Firstly, in the workflow process, information is required from synthesis and workup, in particular DSC for enthalpy of fusion and melting temperature of the compound and impurities. X-ray powder diffraction (XRPD) is required for the polymorph of the API. If DSC data are unavailable, the information could then be found in literature only if the polymorph of the compound is known (see Chapter Two section 2.3). Nuclear magnetic resonance (NMR) and high-performance liquid chromatography mass spectrometry (HPLC-MS) establish the identity of impurities and their concentration in the material. If the enthalpy of fusion and melting temperature are not available and if there is a solid sample available of the pure compound and/or impurities, the information can be obtained by DSC. If the sample of pure compound and/or impurities is not available a predictive method can be used such as Joback and Reid. A detailed description of this

method is reported in section 1.3.9. If sufficient quantity of sample is available, it can be used to establish the concentration of impurity in the material by liquid chromatography with mass spectrometer (LCMS). Once all this data has been obtained solubility predictions can commence, at a range of temperatures, by prediction using COSMO*therm* if solubility data of API and impurities are not available.

Once these predictions are available the solubility of the API is compared with the solubility of the impurities to choose a wash solvent. Ideally the solubility of the impurity in the wash solvent should be higher than the API. A crystallisation solvent is also selected from the COSMO*therm* predictions. Initially the solubility of the API at three temperatures; a target temperature of 10°C below the boiling point of the solvent, a low temperature of 25°C and the mid-point temperature between the two. 10°C below the solvent boiling point is used as a maximum operating range as any higher and the solvent begins to boil and this temperature will have higher solubility than lower temperatures. The midpoint temperature is required to obtain the shape of the solubility curve and the low temperature is used as no cooling or heating equipment is required at room temperature. Solubilities for all impurities at these temperatures are predicted. The target temperature will maximise the amount of material that will be soluble in the solvent.

The wash and crystallisation solvents also need to be miscible. Without miscibility you only have washing by displacement, miscibility maximises the washing efficiency. A miscibility screening procedure using COSMO*therm* is explained in section 2.6.1.

If no samples are available at this point the workflow cannot proceed as laboratory data are required to verify the prediction of COSMO*therm.* Solubility can be obtained from different methods *e.g.* iso-thermal cloud method (Meenan, 2001) or the polythermal method (Zimmerman, 1952, Brown *et al.*, 2018). In the workflow proposed here solubility was measured by iso-thermal cloud method (equilibration method). The correction factor for solubility using RF models has already been explained in Chapter Four. Once laboratory data is obtained solvents for cooling crystallisation and wash solvent selection can be classified and ranked according to the criterion in section 5.5.

In the case of miscible solvents, binary plots of the API in crystallisation and wash solvent from COSMO*therm* predictions are used to evaluate the amount of solvent used in washing and to evaluate the dissolution of API. Ideally the API would be less soluble in the wash solvent.

The results obtained by solvent selection inform the next stage in which the selection is verified using a laboratory scale isolation device. This verifies the extent of impurity removal and API dissolution with respect to the predictions.

5.5 Wash and crystallisation solvent classification

There are several criteria for choosing a wash solvent, a high yield, the process volume and the difference in solubility between the API and the impurities.

Three predictions for each solute/solvent combination are required; a low temperature, a target temperature and a midpoint temperature prediction (section 5.4).

The low temperature of 25°C is around room temperature is easy to achieve in the laboratory. The percentage of starting material theoretically recovered at the end of the cooling crystallisation can also be calculated (See Equation 28) by using the solubility predictions at the target temperature and low temperature.

Equation 28

$$Yield = 100 - (100 * \left(\frac{solubility\ at\ low\ temperature}{solubility\ at\ target\ temperature}\right))$$

The process solvent volume at the beginning and end of the crystallisation can be calculated if the density of the solvent is known. The process solvent volume is the amount of solvent used to obtain a gram of material (Equation 29).

Equation 29

$$process\ volume = \frac{100}{solubility * solvent\ density}$$

The process volume was split into nine categories as shown in the table below (Table 5-1). These categories are used to rank the crystallisation solvents. The best category is the one with the highest yield of greater than 90% and a process volume of less than 10g/ml of solvent as this will give the maximum yield of product using the least amount of solvent.

Table 5-1 Process volume and yield categories

Category	Process volume of solvent (g/ml of solvent)	Yield (%)
1	<10	>90
2	<20	>90
3	<30	>90
4	<10	>85
5	<20	>85
6	<30	>85
7	<10	>80
8	<20	>80
9	<30	>80

If the same solvent used for crystallisation is used also as a wash solvent, then two different scenarios are presented. The first scenario is that if the end temperature of the crystallisation is the same as room temperature then the yield will be the lowest of the two scenarios because the API will be highly soluble in the crystallisation solvent. However, if the solvent is cooled down further less impurity can be removed but the yield is improved. Another scenario is where the wash solvent is different from the crystallisation solvent, this is favourable to maximise yield and to remove impurity. The wash solvent that is selected would give a higher solubility to the impurity than the API. This would ensure that the impurity is rejected with minimal dissolution of API with consequential minimisation of particle size reduction.

Table 5-2 Categories assigned to differences in solubility between API and impurity, x is solubility

Category	Δ Solubility range
1	0g <= x < 1g/100g
2	1g/100g <= x < 10g/100g
3	10g/100g <= x < 20g/100g
4	20g/100g >= x

The above table (Table 5-2) shows the four categories assigned to the difference in solubility between API and each impurity in a solvent if the solubility of the impurity is

greater than the API. For crystallisation solvent selection category one would be preferable because the difference in solubility increases, dissolution of the impurity could become a problem. If the solubility of the impurity is less than the API in a particular solvent, then the concentration of the impurity is considered.

- Between 0% and 1% mole fraction ratio of the solubility of pure compound with respect to the solubility of the impurity.
- Between 1% and 2% mole fraction ratio of the solubility of pure compound with respect to the solubility of the impurity.
- Above 2% mole fraction ratio of the solubility of pure compound with respect to the solubility of the impurity.

If the concentration of impurities is in the first two categories, then it is possible that the impurity will be washed away in the wash solvent. If the impurity is in the third category than there is a risk of precipitation into the product.

If the binary system has a maximum (Figure 5-1) as in Jozwiakowski's experiment with lamivudine in ethanol and water (Jozwiakowski *et al.*, 1996). This would result in more solvent being used and more API being dissolved. As more water is added before the maxim this results in a a loss of yield as the API is more soluble. As such the starting position for the crystalisation would be at the maxima. A solubility curve without maxima is preferred as less material will be dissolved reducing costs.

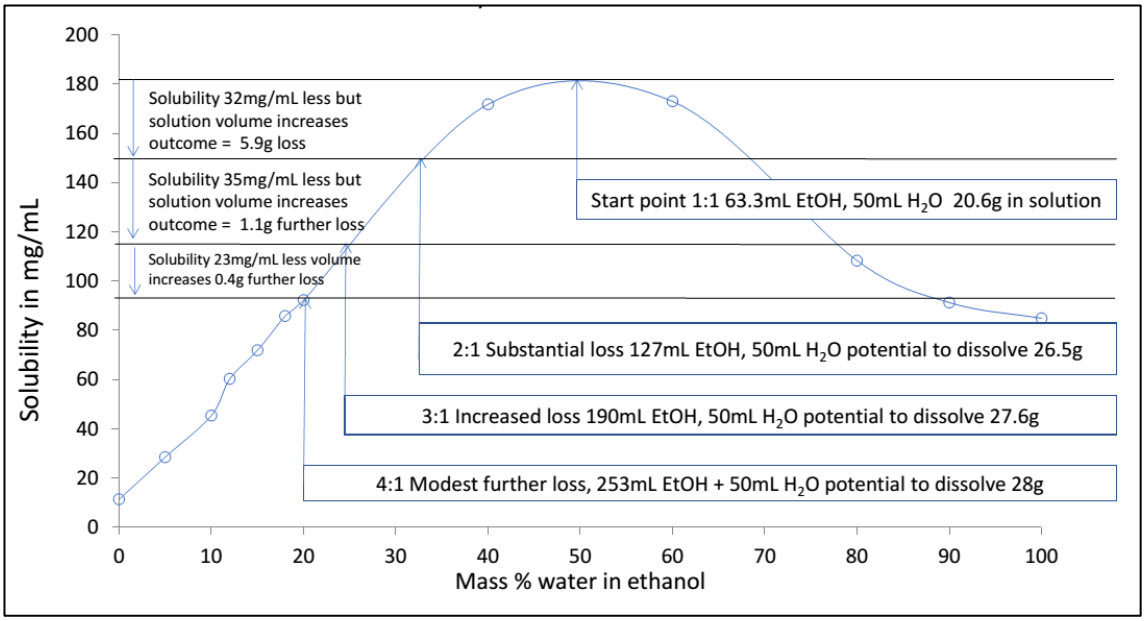


Figure 5-1 Impact of API solubility in solvent mixture, Lamivudine solubility in ethanol-water mixture at 25°C example (Jozwiakowski et al., 1996)

The aim of developing the workflow was to obtain all the information that is required for the selection of crystallisation solvent and wash solvent for a cooling crystallisation. This tool ranks solvents based on their suitability with the objective to maximise yield and minimise the amount of solvent used.

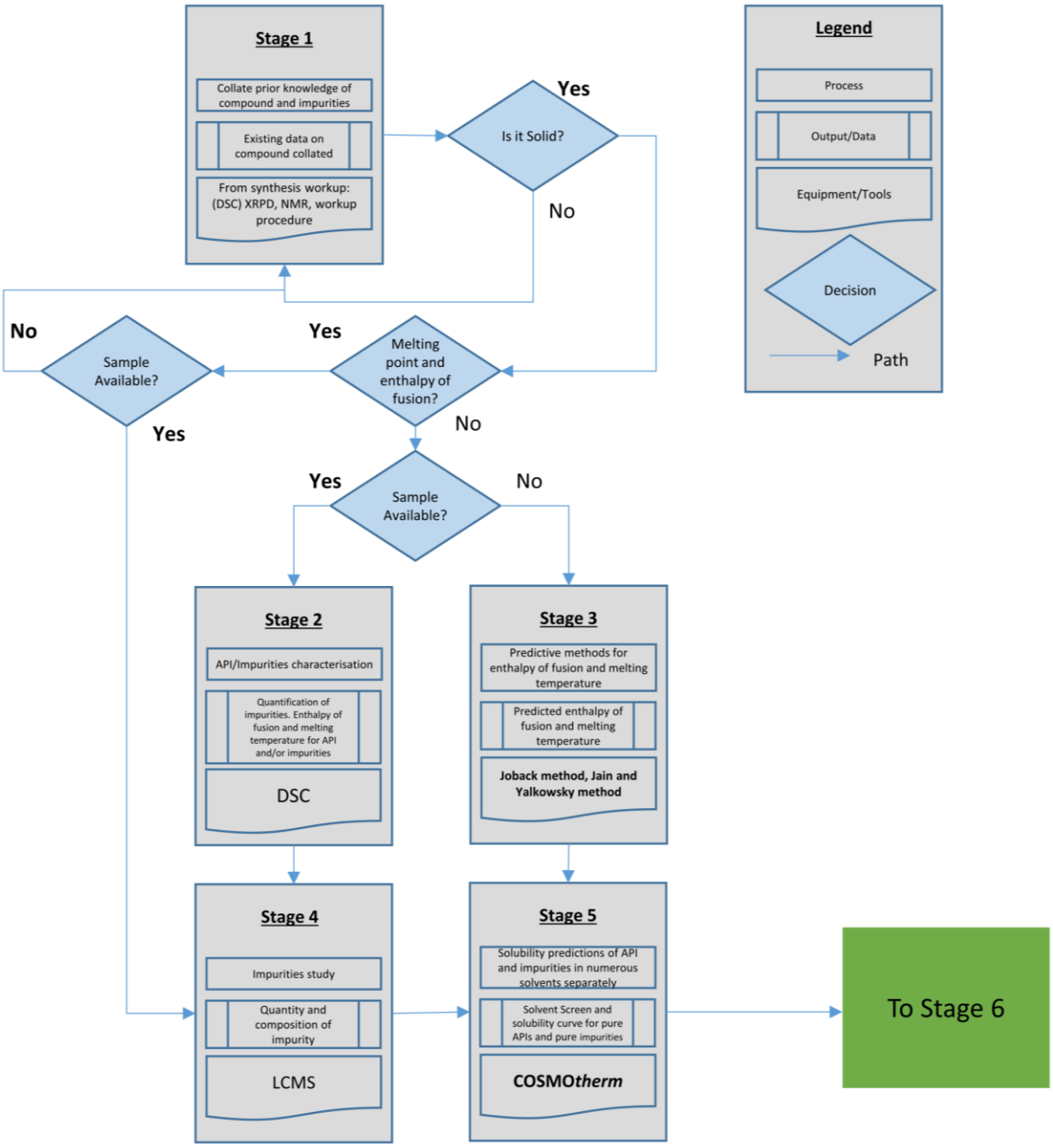
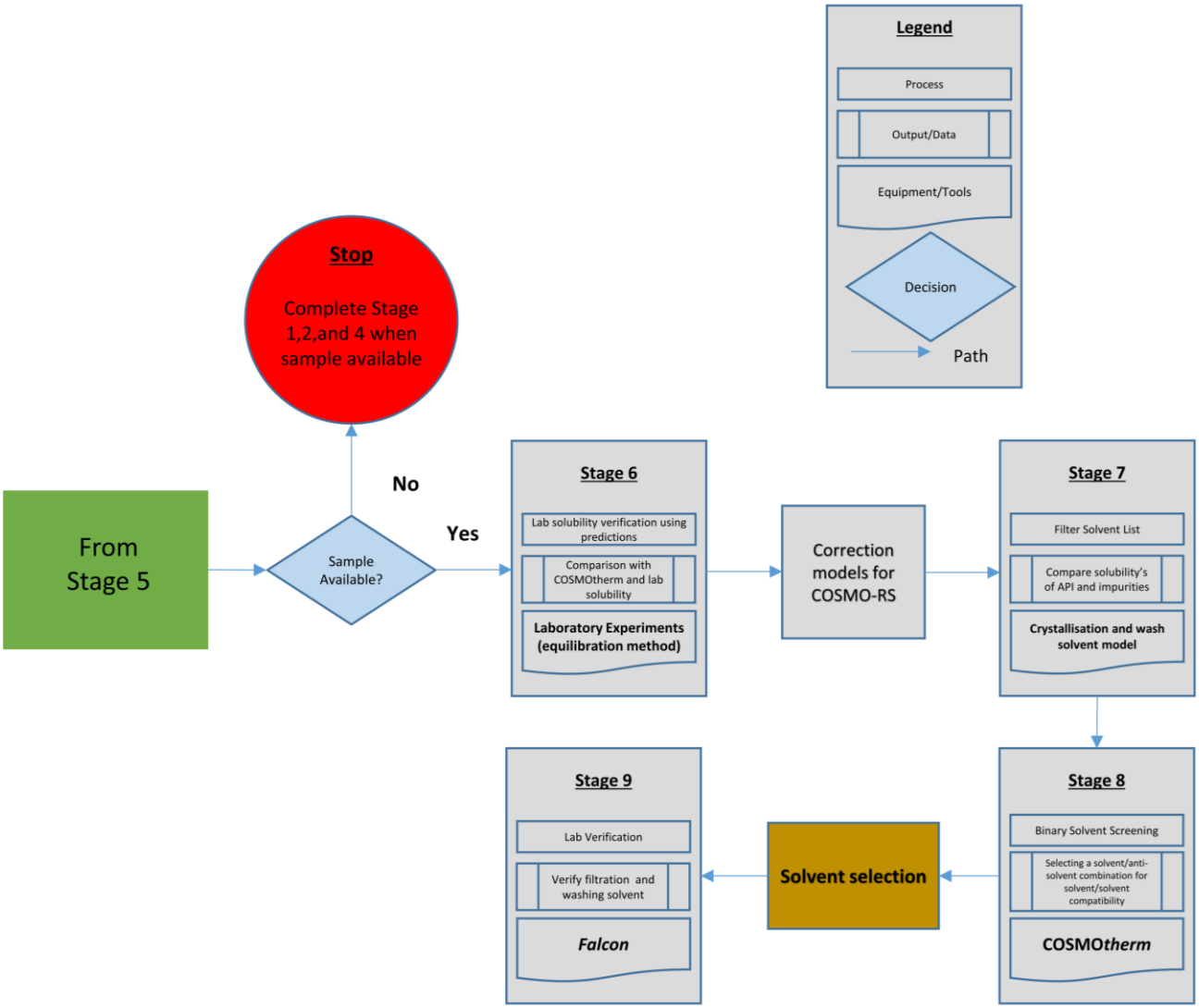


Figure 5-2 Workflow for cooling crystallisation and wash solvent selection



Continuation of Figure 5-2 Workflow for cooling crystallisation and wash solvent selection

Table 5-3 Stages for cooling crystallisation and wash solvent selection

Stage	Supplementary Information
1	Assessing extant information regarding API compound purity and physical properties. Addressing the purity of API information regarding synthesis route to estimate possible impurities.
2	Obtaining essential supplementary information of the pure compound *i.e.* enthalpy of fusion and melting temperature
3	Predicting thermal properties of the pure compound available *i.e.* enthalpy of fusion and melting temperature using group contribution methods.
4	Qualification and quantification of any impurity/degradation molecule present.
5	Using *ab initio*, *in silico* software package COSMO*therm* to predict the solubility of the API and the impurities in pure solvents to inform stage 7.
6	Using gravimetric methods to experimentally obtain the solubility of API/Impurities in pure solvents by verifying solubility predictions from COSMO*therm* data obtained from the correction model.
7	Assessing if a solvent is a good fit for a filtration and washing solvent using the equation to aid API recovery and yield using solubility differences in API/impurities in solvent, confirming impurity presence in solution and absence in cake (using Δsolubility of API/impurities).
8	Using COSMO*therm* with binary solvents to select wash solvent/crystallisation solvent.
9	Verification using an isolation device to confirm the high purity of the product and high yield using the selected binary combination.

5.6 Case Study: paracetamol and impurities

For this project paracetamol and impurities were chosen as a case study for validating the workflow (Figure 5-2 and Table 5-3). All the compounds were supplied from suppliers as detailed in section 5.3.

5.6.1 Stage 1 – Collate prior knowledge of compound and impurities

The idea of the workflow is to obtain all the synthesis and work up information for a new compound (DSC, XRPD, NMR and workup procedure) during this stage. The information that is required is API compound purity and physical properties; enthalpy of fusion, melting temperature and workup procedure. All the compounds for this chapter were commercially sourced and therefore no synthesis information or workup was available.

5.6.2 Stage 2 – API/Impurities characterisation

A literature review to find the enthalpy of fusion and melting temperatures for paracetamol and the impurities was carried out and the results are shown in Table 5-4. All the compounds had literature values except for orthocetamol. Some of the compounds had only one reference for literature values and many did not state the polymorphic form. However, if the polymorphic form was stated the value for the most stable form was used.

Table 5-4 Paracetamol and impurities enthalpy of fusion and melting temperatures

Compound	**Enthalpy of fusion (kJ/mol)**	**Melting temperature (°C)**	**Reference**
p-chloroacetanilide	41.00	178.40	(Gmehling, 2018)
orthocetamol	15.67	209.10	*DSC results
metacetamol	28.80	146.85	(Barrio *et al.*, 2017)
acetanilide	21.65	114.30	(Gmehling, 2018)
4-nitrophenol	18.88	112.00	(Gmehling, 2018)
4-hydroxymethylbenzoate	24.31	131.00	(Gmehling, 2018)
4-amino-phenol	31.20	186.30	(Gmehling, 2018)
4-acetamidobenzoicacid	20.93	186.00	(Monte *et al.*, 2010)
1-(4-hydroxyphenyl)ethanone	17.03	108.15	(Y. P. Chen,Tang and Kuo, 2005)
1-(2-hydroxyphenyl)ethanone	32.80	197.00	(Shiu Shiang Yang and Guillory, 1972)
paracetamol	26.49	168.74	(Xu *et al.*, 2006)

A sample of orthocetamol was obtained from Sigma Aldrich and a DSC experiment was carried out using a Netzsch DSC 214 POLYMA. The apparatus was used to obtain enthalpy of fusion and melting temperature for orthocetamol. An orthocetamol sample was heated from 20°C to 10°C above the melting point at a rate of 10°C per minute and then held for two minutes and then the temperature was dropped to 150°C. This was repeated cyclically three times. The graph (Figure 5-3) shows the results with a melting temperature of 209.1°C and an enthalpy of fusion of 15.67 kJ/mol. All the figures shown above were used in the COSMO*therm* predictions of stage 5 (section 5.6.5). Fortunately for this case study all the relevant data was gathered. However, for other cases when this data is not available predictive methods for enthalpy of fusion and melting temperature must be used.

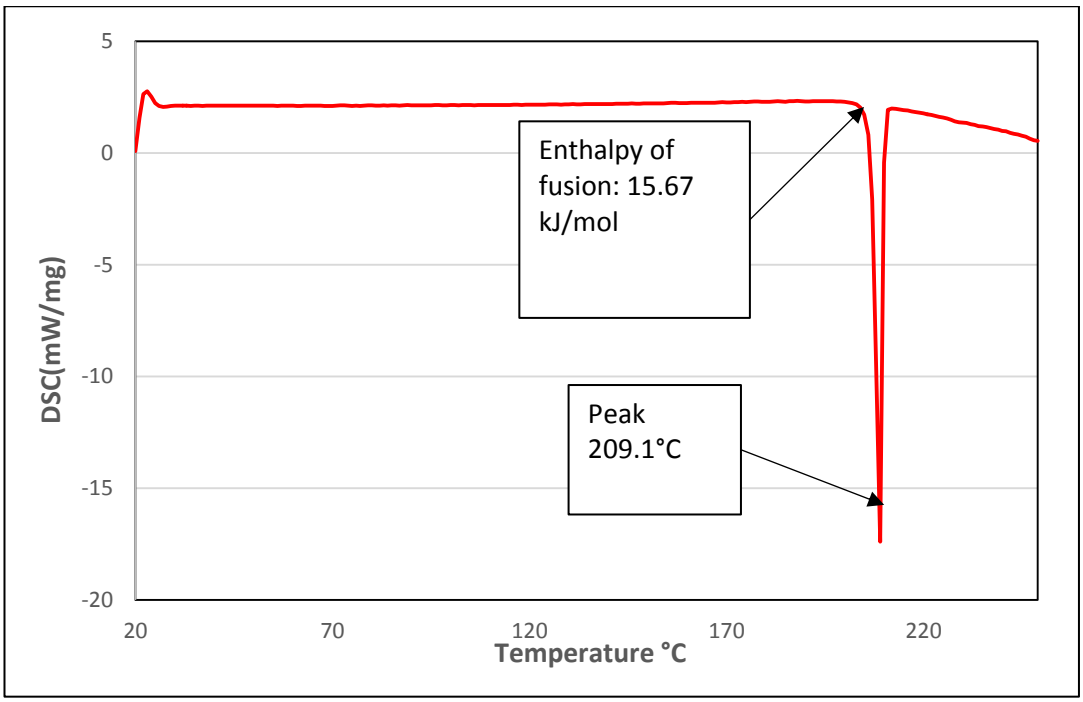


Figure 5-3 Plot of the enthalpy of fusion and melting temperature for orthocetamol using a NETZSCH DSC 214

5.6.3 Stage 3 – Predictive methods for enthalpy of fusion and melting temperature

None of the compounds used enthalpy of fusion or melting temperature predictions from the Joback and Reid method (see section 1.3.9) or the Jain and Yalkowsky (see section 1.3.10). These estimations come with a rather large degree of error. As such, this should be factored into solubility predictions using these methods. COSMO*therm* used enthalpy of fusion and melting temperature data from DSC measurements or literature for this case study.

5.6.4 Stage 4 – Impurities study

The aim of this stage is to use LCMS to establish the quantity and composition of impurities in the sample. No LCMS data was obtained for this chapter. The API and impurities used for this project were not from a real synthesis so no such data was available.

5.6.5 Stage 5 – Solubility predictions of API and impurities in numerous solvents

Using the information obtained from the previous stages the solubility of the API compound and impurities was predicted using COSMO*therm*. COSMO*therm* was used to predict the solubility of paracetamol and three structurally related impurities of synthesis *i.e.* p-chloroacetanilide, acetanilide and metacetamol (Figure 5-4) in 136 solvents to indicate wash solvent suitability for paracetamol. The 136 solvents used were from the database discussed in section 2.4.

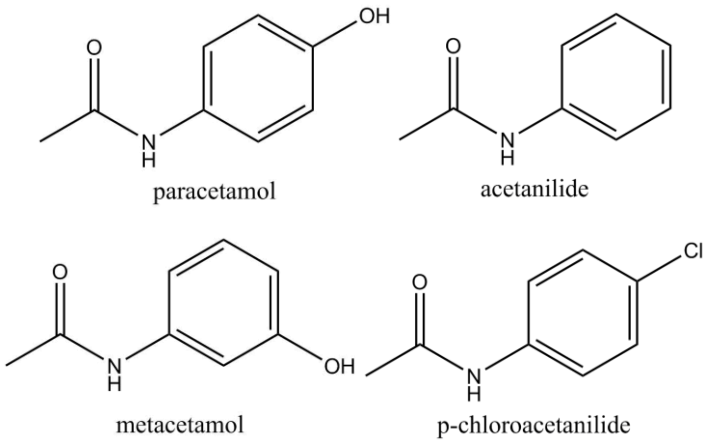


Figure 5-4 structures for paracetamol and related impurities

COSMO*therm* was used to predict the solubility of paracetamol and impurities at an initial target temperature (10°C below the boiling point of the solvent) and a final crystallisation low temperature (laboratory temperature is normally 25°C) over a range of solvents. Some of the solubility predictions returned a value of "NA" which means that the solute is very soluble or very insoluble. For this project if COSMO*therm* returned a value of "NA" it was assumed that the solute is soluble and was given the artificial value for solubility of 500g per 100g of solvent. This was just for visualisation and solvent ranking purposes only. The reason for COSMO*therm* returning a value of "NA" is because initially the software was designed for LLE with

SLE added on later. COSMO*therm* assumes that the solute/solvent combination are "miscible" and can't reconcile the equations and therefore returns "NA" as a prediction value.

Table 5-5 wash solvent selected by COSMOtherm for paracetamol and three impurities acetanilide, metacetamol p-chloroacetanilide

Wash solvent	ICH class	Viscosity (mPas)	Boiling point °C	Enthalpy of vaporisation (kJ/mol)
1,1,1-trichloroethane	1	0.8582 (20°C) (Prakash *et al.*, 1996)	74.00	33.342 (13.38°C) (Rubin,Levedahl and Yost, 1944)
1,1-diethoxypropane	NA	NA	102.00	NA
1-chlorobutane	NA	0.45 (20°C) (Mumford and Phillips, 1950)	78.50	33.51 (24.85°C) (Majer and Svoboda, 1985)
1-methylnaphthalene	NA	3.44 (20°C) (Luther and Waechter, 1949)	243.00	54.1 (25°C) (Hopfe, 1990)
2,2,4-trimethylpentane	NA	0.503 (20°C) (Smyth and Stoops, 1928)	99.00	35.15 (25°C) (Svoboda *et al.*, 1982)
3-fluorotoluene	NA	0.608 (20°C) (Swarts, 1931)	239.00	NA
4-fluorotoluene	NA	0.6215 (20°C) (Swarts, 1931)	241.90	NA
bromobenzene	NA	1.0597 (25°C) (Rodríguez *et al.*, 1997)	156.00	44.4 (25°C) (Hopfe, 1990)
chloroform	1	0.965m(25°C) (Titani, 1927)	76.72	32.4 (25°C) (Hopfe, 1990)
cyclohexane	2	0.89 (20°C) (Vorländer and Walter, 1925)	81.00	33.1 (25°C) (Pedley,Naylor and Kirby, 1986)
cyclopentane	2	0.423 (25.05°C)(Fischer and Weiss, 1986)	49.00	28.5 (25°C) (Hopfe, 1990)
diisopropylether	NA	0.329 (20°C)(Gartenmeister, 1890)	69.00	32.6 (25°C) (Hopfe, 1990)
di-n-butylether	NA	0.644 (25°C) (Lutskii,Obukhova and Sidorov, 1958)	140.80	43.9 (25°C) (Hopfe, 1990)
dipentene	NA	0.9 (20°C) (Bardyshev, 1948)	176.00	NA
dodecane	NA	1.3877 (25°C) (Awwad,Jabara	216.20	61.3 (25°C)(Hopfe, 1990)

		and Salman, 1988)		
hexane	2	0.294 (25°C) (Bardavid *et al.*, 1996)	68.00	31.4 (25°C)(Hopfe, 1990)
iodobenzene	NA	1.64 (20°C) (Toropov and Kim, 1961)	188.00	49.6kJ/mol (25°C) (Hopfe, 1990)
isopropylbenzene	2	0.786 (20°C) (Schmack,Rother and Bittrich, 1973)	152.40	45.1 (25°C) (Hopfe, 1990)
methylcyclohexane	2	0.72 (20°C) (Vorländer and Walter, 1925)	101.00	35.4 (25°C) (Hopfe, 1990)
n-heptane	3	0.409 (20°C) (Smyth and Stoops, 1928)	98.42	36.58 (25°C) (Hala *et al.*, 1979)
pentane	3	0.214 (25°C) (Acevedo,Pedrosa and Katz, 1993)	36.10	26.4 (25°C) (Hopfe, 1990)
tetralin	2	2.01 (20°C) (Vorländer and Walter, 1925)	406.00	54.3 (25°C) (Hopfe, 1990)
triethylamine	NA	0.369 (20°C) (Kokkonen and Nissema, 1979)	89.00	36.8 (25°C) (Hopfe, 1990)

The above table (Table 5-5) shows the wash solvents selected by COSMO*therm.* Crystallisation solvent selection is discussed in section 5.6.7. There are also some additional important criteria such as cost, viscosity, boiling point, toxicity and enthalpy of vaporisation that have to be considered. The boiling point and enthalpy of vaporisation of the wash solvent affect the ease of solvent removal during drying. Solvents with higher boiling points (>100°C) were discarded due to the longer drying time required. Higher viscosities may be associated with increased solvent retention in the filtration cake. During filtration, particles of solid are deposited on the filter

medium forming a filter cake. As time passes the cake thickness increases and the cake becomes more compacted increasing the resistance to fluid flow through it. Ideally a solvent with an ICH class of three would be selected however a class two solvent might be considered. When these criteria were applied cyclohexane, cyclopentane, n-heptane and pentane were shortlisted as possible wash solvents.

5.6.6 Stage 6 – Laboratory solubility verification using predictions

Once the predictions using COSMO*therm* were completed experimental data was then obtained by CMAC researchers for the solubility of paracetamol and the impurities in the shortlisted solvents. The results (Table 5-6) are a subset of over 100 solvents predicted for and show the difference in solubility between paracetamol and the impurities with the solvents ranked in order of suitability for wash solvent selection. Out of 21 laboratory results only five (24%) were wrongly classified, using the classification system in section 5.5 above, by COSMO*therm.* All misclassified were for acetanilide as an impurity in toluene, isopropyl acetate, 4-methyl-2-pentanone, 2-propanol and ethanol respectively which are shown in red. Each misclassified solvent was only one classification different from the correct classification. The results are of a qualitative nature but that is all that is required for wash solvent selection. If the results from experimental and COSMO*therm* are compared, n-heptane and dodecane are ranked as the best solvents in both sets of data as the solubility differences between paracetamol and each impurity are low which is one of the main criterion for wash solvent selection. The wash solvents selected were also checked for miscibility with the crystallisation solvent selected (see section 5.6.6). For the remaining solvents in the table there is no change in the ranking using

COSMO*therm.* Although these results are only a subset of a larger number of COSMO*therm* predictions it gives confidence that the predictions are mostly accurate and can be relied on. The model predicted 23 possible wash solvents for this API and impurities. Both n-heptane and dodecane were the top ranked wash solvents. The other candidates are out of the scope of this project and would require further laboratory results.

Table 5-6 Ranking of wash solvents with difference in solubility between paracetamol and acetanilde, metacetamol and p-chloroacetanilde from laboratory testing and COSMOtherm predictions (incorrect classification in red) classifications are taken from Table 5-2

Impurity	Solvent	Δ laboratory solubility (g/100g)	Δ predicted solubility (g/100g)	Measured class (predicted class)
acetanilide	n-heptane	0.050	0.066	1(1)
metacetamol	n-heptane	0.020	0.000	1(1)
acetanilide	dodecane	0.125	0.042	1(1)
metacetamol	dodecane	0.050	0.000	1(1)
acetanilide	toluene	0.775	3.692	1(2)
metacetamol	toluene	0.040	0.003	1(1)
acetanilide	anisole	2.950	4.798	2(2)
metacetamol	anisole	0.190	0.021	1(1)
acetanilide	methyl-t-butylether	2.748	3.113	2(2)
metacetamol	methyl-t-butylether	0.573	0.122	1(1)
acetanilide	isopentanol	11.069	16.285	3(3)
acetanilide	isopropylacetate	8.205	14.798	2(3)
metacetamol	isopropylacetate	1.705	1.832	2(2)
acetanilide	acetonitrile	18.470	16.161	3(3)
metacetamol	acetonitrile	5.085	2.828	2(2)
acetanilide	4-methyl-2-pentanone	11.440	2.828	3(2)
metacetamol	4-methyl-2-pentanone	5.280	3.302	2(2)
acetanilide	2-propanol	4.776	16.967	2(3)
metacetamol	2-propanol	7.050	1.120	2(2)
acetanilide	ethanol	13.576	24.820	3(4)
p-chloroacetanilide	ethanol	18.665	11.216	3(3)

5.6.6.1 Applying RF model to API and impurities solubility

The RF solute-Fold model with 2D descriptors to obtain a solubility correction factor, as described in section 4.2.4, was applied to the data in the above section.

Using dataset 3 and the ML algorithm with 2D descriptors a correction factor was calculated for paracetamol and impurities by using all paracetamol and impurities

solubility data as the test set and the remaining solubility data from dataset 3 as the training set.

Table 5-7 Ranking of wash solvents with difference in solubility between paracetamol and acetanilde, metacetamol and p-chloroacetanilde from laboratory testing and RF corrected predictions without paracetamol in training set (incorrect classification in red) classifications for ranking are taken from Table 5 2

Impurity	Solvent	Δ laboratory solubility (g/100g)	Δ predicted solubility (g/100g)	Δ corrected solubility (g/100g)	Measured class (corrected class)
acetanilide	n-heptane	0.050	0.066	0.297	1(1)
metacetamol	n-heptane	0.020	0.000	0.001	1(1)
acetanilide	dodecane	0.125	0.042	0.167	1(1)
metacetamol	dodecane	0.050	0.000	0.000	1(1)
acetanilide	toluene	0.775	3.692	3.915	1(2)
metacetamol	toluene	0.040	0.003	0.013	1(1)
acetanilide	anisole	2.950	4.798	4.579	2(2)
metacetamol	anisole	0.190	0.021	0.031	1(1)
acetanilide	methyl-t-butylether	2.748	3.113	2.448	2(2)
metacetamol	methyl-t-butylether	0.573	0.122	0.088	1(1)
acetanilide	isopentanol	11.069	16.285	14.027	3(3)
acetanilide	isopropylacetate	8.205	14.798	10.497	2(3)
metacetamol	isopropylacetate	1.705	1.832	1.330	2(2)
acetanilide	acetonitrile	18.470	16.161	12.251	3(3)
metacetamol	acetonitrile	5.085	2.828	2.376	2(2)
acetanilide	4-methyl-2-pentanone	11.440	2.828	10.536	3(2)
metacetamol	4-methyl-2-pentanone	5.280	3.302	2.423	2(2)
acetanilide	2-propanol	4.776	16.967	13.685	2(3)
metacetamol	2-propanol	7.050	1.120	0.483	2(2)
acetanilide	ethanol	13.576	24.820	21.409	3(4)
p-chloroacetanilide	ethanol	18.665	11.216	18.279	3(3)

The above table (Table 5-7) shows the difference in solubility between API and impurities predictions with a correction factor applied. Although in the example the

correction factor did not result in any solvents being ranked differently due to the application of the solubility correction factor, the absolute values of predicted solubility were improved upon when compared with COSMO*therm* predictions.

Table 5-8 Ranking of wash solvents with difference in solubility between paracetamol and acetanilde, metacetamol and p-chloroacetanilde from laboratory testing and RF corrected predictions with paracetamol in training set and using experimental solubility values for paracetamol. (incorrect classification in red) classifications are taken from Table 5-2

Impurity	**Solvent**	**Δ laboratory solubility (g/100g)**	**Δ predicted solubility (g/100g)**	**Δ experimental API and corrected Impurity solubility (g/100g)**	**Measured class (corrected class)**
acetanilide	n-heptane	0.050	0.066	0.298	1(1)
metacetamol	n-heptane	0.020	0.000	0.007	1(1)
acetanilide	dodecane	0.125	0.042	0.208	1(1)
metacetamol	dodecane	0.050	0.000	0.053	1(1)
acetanilide	toluene	0.775	3.692	3.967	1(2)
metacetamol	toluene	0.040	0.003	0.018	1(1)
acetanilide	anisole	2.950	4.798	4.279	2(2)
metacetamol	anisole	0.190	0.021	0.025	1(1)
acetanilide	methyl-t-butylether	2.748	3.113	1.809	2(2)
metacetamol	methyl-t-butylether	0.573	0.122	0.027	1(1)
acetanilide	isopentanol	11.069	16.285	11.899	3(3)
acetanilide	isopropylacetate	8.205	14.798	7.316	2(3)
metacetamol	isopropylacetate	1.705	1.832	0.734	2(2)
acetanilide	acetonitrile	18.470	16.161	9.330	3(3)
metacetamol	acetonitrile	5.085	2.828	0.289	2(1)
acetanilide	4-methyl-2-pentanone	11.440	2.828	7.552	3(2)
metacetamol	4-methyl-2-pentanone	5.280	3.302	0.971	2(2)
acetanilide	2-propanol	4.776	16.967	11.480	2(3)
metacetamol	2-propanol	7.050	1.120	3.277	2(2)
acetanilide	ethanol	13.576	24.820	21.349	3(4)
p-chloroacetanilide	ethanol	18.665	11.216	19.964	3(3)

The above table (Table 5-8) shows the difference in solubility with experimental solubility values of paracetamol and RF corrected solubility predictions for impurities. There is no change in the ranking of solvents when the corrected solubility is applied.

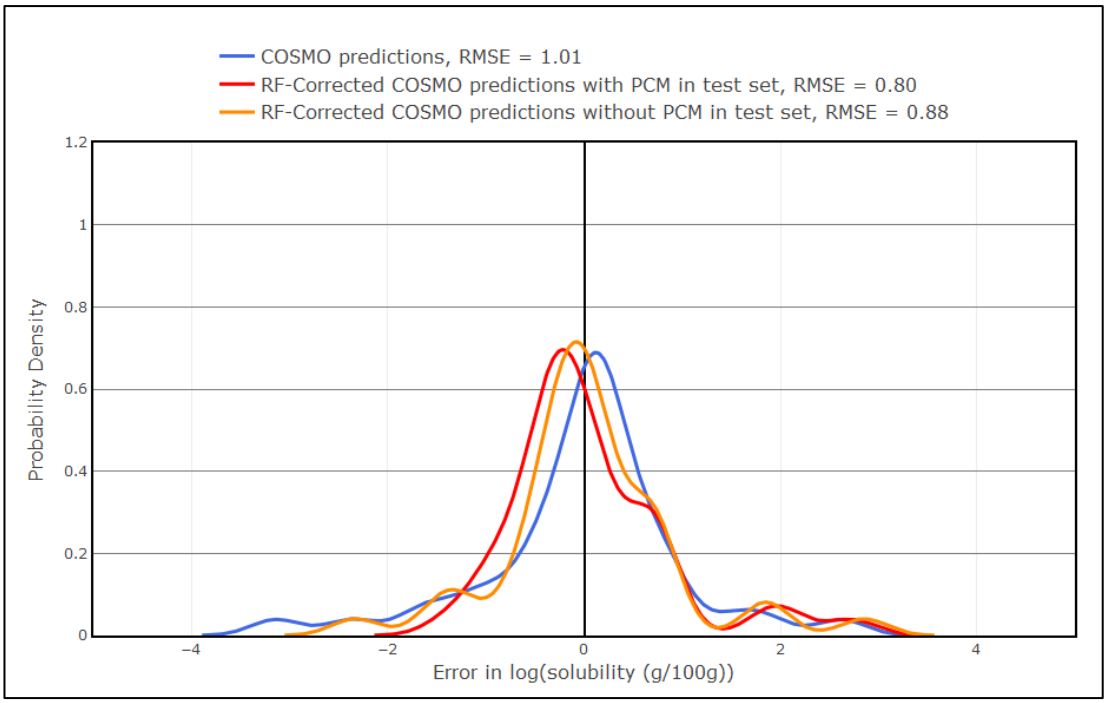


Figure 5-5 Density plot showing RMSE for COSMOtherm predictions (blue), RF-corrected solubility with paracetamol in the training set (red), RF-corrected without paracetamol in the training set (orange)

The error density plot (Figure 5-5) above shows the results of the RF algorithms. The RMSE for COSMO*therm* predictions was log 1.01. The RF corrected solubility without paracetamol in the training set improved the error to log 0.88, and the RF corrected solubility with paracetamol in the training set reduced the error to log 0.80. This is also consistent with the model in the previous chapter where a molecule of similar structure to the target molecule in the training set improved the RF model.

The results show that COSMO*therm* even without the RF corrected solubility is a powerful tool for crystallisation and wash solvent selection. The methods here chose n-heptane and dodecane as the best wash solvents.

5.6.7 Stage 7 – Filtering solvent list

This stage is required to assess if the solvent is a good fit for a crystallisation solvent. The crystallisation solvent can be selected either on process volume or product yield.

Table 5-9 Crystallisation solvents selected by yield

Crystallisation solvent	**Process volume at ml of solvent for g of solute and % recovery classification**	**ICH class**	**Viscosity (mPas)**	**Yield (%)**	**Reference**
2-amino-1-butanol	<10ml/g and >90%	NA	NA	93.4	NA
trifluoroacetic acid	<10ml/g and >90%	NA	1.8 (20°C)	93.0	('*www.solvay.us*', 2019)
glycerol	<20ml/g and >90%	NA	1490 (20°C)	90.9	(Lux and Stockhausen, 1993)
propanoic acid	<20ml/g and >90%	NA	1.101 (20°C)	90.3	(Lutskii,Obukhova and Sidorov, 1958)
4-methyl-2-pentanone	<30ml/g and >90%	3	0.575 (20°C)	90.3	(Riggio *et al.*, 2011)
acetonitrile	<30ml/g and >90%	2	0.3568 (25°C)	92.7	(Walden, 1906)
butyric acid	<30ml/g and >90%	NA	1.78 (20°C)	90.4	(Irany, 1943)

These solvents were selected by COSMO*therm* by using the two criteria; percentage recovery of solute (Table 5-9) and process volume (Table 5-10). Some of the solvents in these lists cannot be used due to the ICH class (RSC, 2019), these substances being toxic. If the viscosity of the crystallisation solvent and the wash solvent are significantly different then these solvents must also be ruled out as the crystallisation solvent cannot be removed (Ottoboni, 2018). Therefore, glycerol and propylene

glycol were rejected for high viscosity. Additionally, for the selection of a crystallisation solvent a crystallisation screening must be carried out to select a solvent that obtains the polymorph or crystal habit desired. This is out with the scope of this chapter.

Table 5-10 Crystallisation solvent selected by process volume

Crystallisation solvent	**Process volume at ml of solvent for g of solute and % recovery classification**	**ICH class**	**Viscosity (mPas)**	**Reference**
2-amino-1-butanol	<10ml/g and >90%	NA	NA	NA
trifluoroacetic acid	<10ml/g and >90%	NA	1.8 (20°C)	('*www.solvay.us*', 2019)
1,3-dioxane	<10ml/g and >85%	2	1.0567 (20°C)	(Parks,LeBaron and Molloy, 1941)
formamide	<10ml/g and >85%	2	6.92 (20°C)	(Walden, 1906)
formicacid	<10ml/g and >85%	3	1.78 (20°C)	(Tsakalotos, 1908)
propyleneglycol	<10ml/g and >85%	2	54.62 (20°C)	(Detherm, 2016)
2-hydroxypropanoicacidethylester	<10ml/g and >80%	NA	2.61 (20°C)	(Rehberg and Dixon, 1950)
2-methoxyethanol	<10ml/g and >80%	2	1.708 (20°C)	(Detherm, 2016)
dioxane	<10ml/g and >80%	2	1.439 (15°C)	(Timmermans and Hennaut-Roland, 1937)
furfural	<10ml/g and >80%	NA	0.38 (20°C)	(Detherm, 2016)
piperidine	<10ml/g and >80%	NA	1.06 (20°C)	(Detherm, 2016)
tetrahydrothiophene-1,1-dioxide	<10ml/g and >80%	2	10.3 (29.85°C)	(Ponomarenko *et al.*, 1995)

For both yield and process volume 2-amino-1-butanol and trifluoroacetic acid were the best solvents. However as class three solvents are preferred, 4-methyl-2-pentanone and formic acid were selected as potential crystallisation solvents.

5.6.8 Stage 8 – Binary solvent screening

Solubility predictions at 25°C for mole fractions of the selected crystallisation and wash solvents were calculated using COSMO*therm*. For the final selection of crystallisation and wash solvents the solubility curve should not have a maxima (Figure 5-6) which can lead to dissolution and a reduction of yield. However, if the maximum was small or close to zero, the system could be considered.

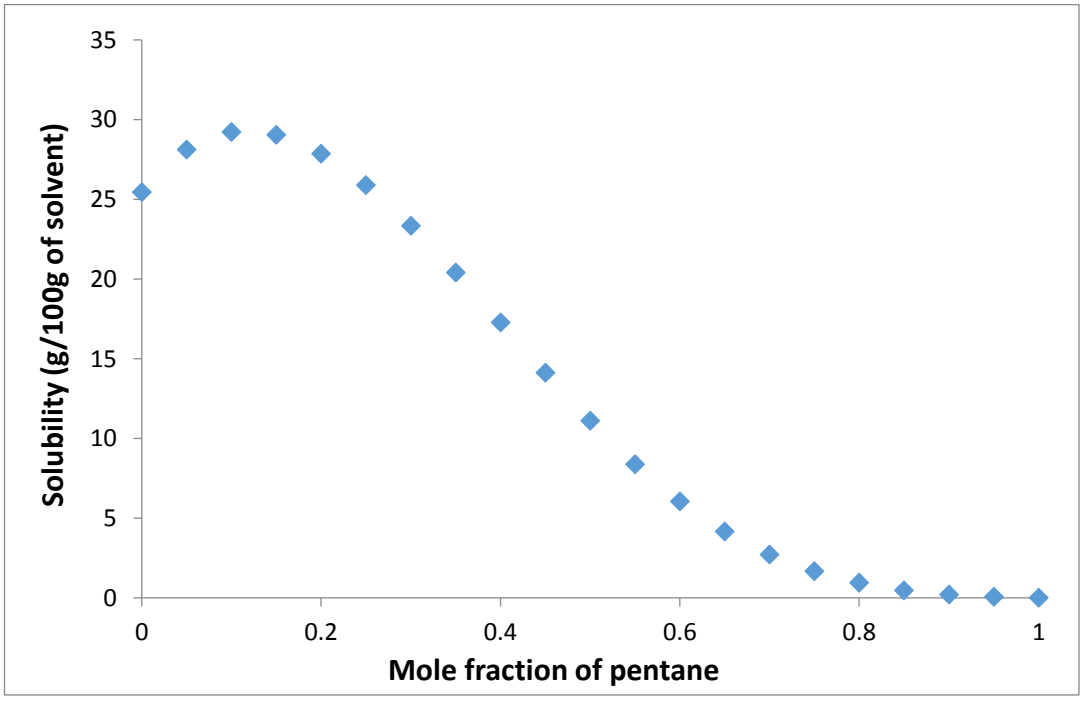


Figure 5-6 predicted solubility curve for paracetamol in formic acid and pentane at 25°C using COSMOtherm

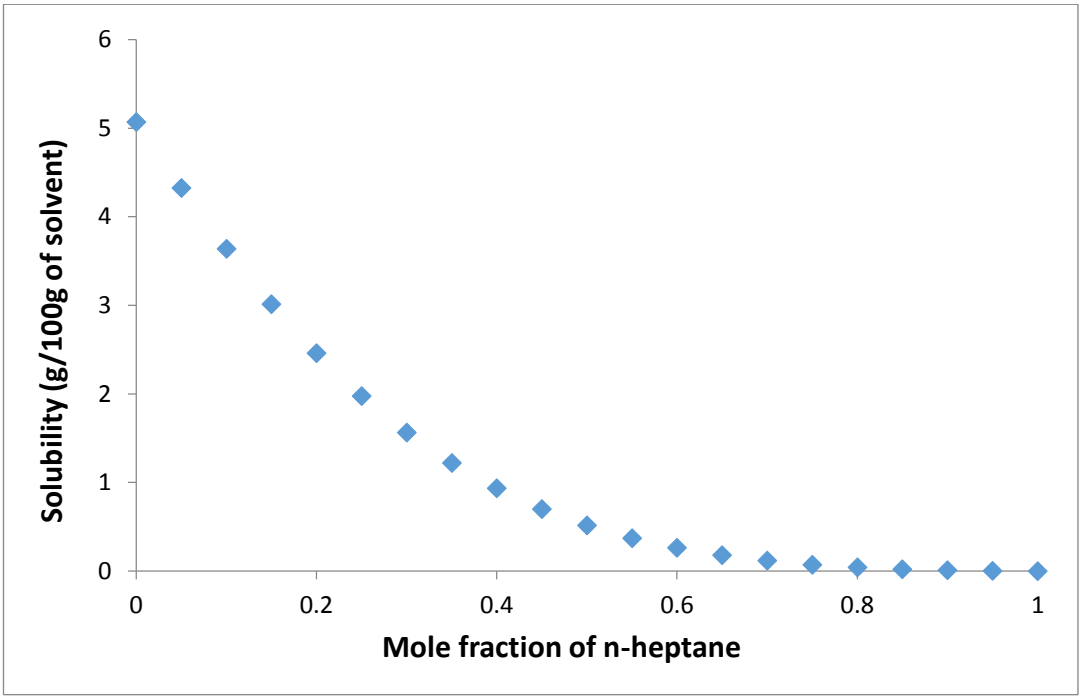


Figure 5-7 predicted solubility curve for paracetamol in 4-methyl-2-pentanone and n-heptane at 25°C using COSMOtherm

The above plot (Figure 5-7) shows the binary plot for paracetamol in 4-methyl-2-pentanone and n-heptane and this has a slope that decreases from 0 to 1 mole fraction of n-heptane which is what is required for wash solvent selection. The best solvents that the workflow has chosen are 4-methyl-2-pentanone for crystallisation and n-heptane for the wash solvent this was confirmed by solubility experimental data. However, this selection will not be used in the following section as no crystallisation screening was available for paracetamol and 4-methyl-2-pentanone.

5.6.9 Stage 9 – Laboratory verification

As the first choice for crystallisation solvent did not have the required crystallisation screening data the yield requirements were lowered and two crystallisation solvents with an ICH class of three were chosen; ethanol and 2-propanol (Table 5-11) (Thompson *et al.*, 2004).

Table 5-11 Crystallisation solvent selection data for ethanol and 2-propanol

Crystallisation solvent	Process Volume at ml of solvent for g of solute and % recovery classification	ICH class	Viscosity (mPas)	Ref.
Ethanol	<30ml/g and >70%	3	1.1	(Detherm, 2016)
2-propanol	<30ml/g and >80%	3	2	(Detherm, 2016)

The binary solvent solubility curve showed no maxima which is suitable for selection of wash and crystallisation solvent (Figure 5-8).

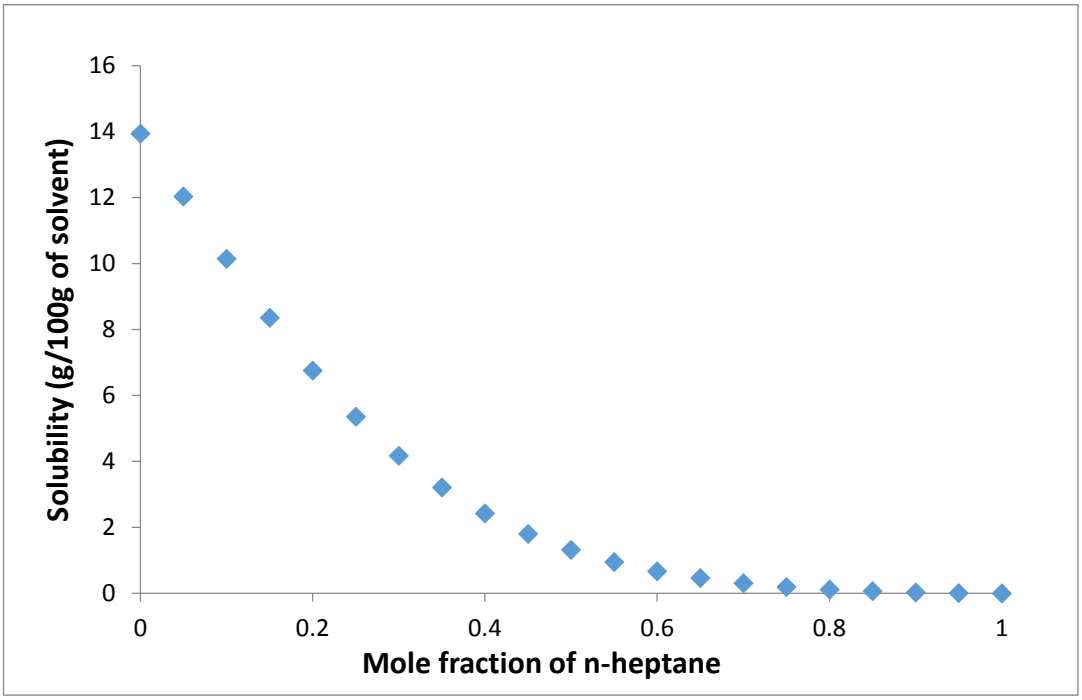


Figure 5-8 solubility curve for paracetamol in ethanol and n-heptane at 25°C

Laboratory verification was carried by Sara Ottoboni and all the results are shown in Table 5-12 (Ottoboni, 2018).

Table 5-12 residual wash and crystallisation (ethanol) solvent remaining in cake, drying time, extent of agglomeration

Wash solvent	Residual wash solvent (%)	Residual crystallization solvent (%)	Drying time (min)	Extent of agglomeration (%)
n-heptane	99.1	0.09	25	91.42
dodecane	99.8	0.20	3500	94.59

The important factors for selecting a suitable wash solvent are residual impurities present due to residual crystallisation solvent in the cake, drying time and agglomeration. The better solvent was n-heptane on all three criteria. This suggests the ability of the workflow to choose a suitable wash solvent and crystallisation solvent for paracetamol. More experimentation is needed on other wash solvents to confirm this.

5.7 Conclusion

A workflow approach using predictive methods for the selection of both crystallisation and wash solvents has been shown to be a powerful tool. Choosing a suitable wash solvent can improve impurity rejection, risk of API dissolution (maintaining or increasing yield) and reduce the cost of manufacturing. The combining of predictive methods of COSMO*therm*, an ML algorithm to apply a "correction factor" to predicted solubility and experimental methods into a workflow has been shown to be effective for selecting crystallisation and wash solvent for paracetamol and the impurities. Although more experimental work will be required to confirm this. Accurate concentration and composition data of both API and impurities are essential. Enthalpy of fusion and melting temperature data would ideally be obtained from DSC experiments but literature data could be used if that

was unavailable. Predictive methods to obtain enthalpy of fusion and melting temperature can be unreliable and their use should be avoided and experimental or literature values used if possible. The predictions from COSMO*therm* were favourably compared with the laboratory data that confirmed the crystallisation and wash solvent selection.

The workflow was designed to give useful information and guidance to the user for the selection of crystallisation and wash solvent.

In the example for paracetamol and impurities in this project it was shown that COSMO*therm* can help with the selection of wash solvents. Applying the correction factor did not improve the overall accuracy of the wash solvent selection but did improve the absolute value of the solubility predictions. The application of the correction factor could improve the selection of other wash solvents for other API and impurities.

6 Conclusions and Future Work

Obtaining predictions of molecular solubility accurately and quickly is a major focus of this thesis. Applying ML algorithms to improve on existing theoretical techniques has been shown to improve the accuracy of those predictions. Building a linear regression model for solute/solvent systems using design of experiments approaches can increase accessibility and the speed of prediction for non-experts. There are several predictive methods that could be used in conjunction with ML such as COSMO*therm*, UNIFAC, NRTL-SAC and SAFT-γ Mie. Each method has its strengths and weaknesses: COSMO*therm* does not always give an accurate prediction and sometimes is unable to give a prediction at all; UNIFAC and SAFT-γ Mie relies on experimental data being available to parametrise functional groups and atoms; and NRTL-SAC needs experimental data for the conceptual segments to inform the model. COSMO*therm* needs minimal experimental data parameterisation and usually needs only the molecular structure, enthalpy of fusion and melting temperature. These methods could complement each other and fill in gaps in the linear regression model described in Chapter Three as there were many solute/solvent combinations that COSMO*therm* was unable to give a prediction for. If one of the methods was better at predicting the solubility of compounds with particular physical properties this method would be used primarily.

Obtaining heat capacity by experimental methods for every solute has also the possibility of improving the predictions of COSMO*therm*. The study in section 2.9 showed its importance to improving predictions. However, values for most solutes were unavailable in literature.

Accurate measurement of enthalpy of fusion and melting temperature is essential to maximise the possibility of an accurate solubility prediction. Obtaining laboratory measurements for each compound would be ideal. As has been shown in Chapter Two, the predictive methods for enthalpy of fusion and melting temperature (Joback and Reid method, Jain and Yalkowsky method) can be inaccurate but if laboratory or literature data are not available for solubility predictions using a predictive method is the only option; although this is the weaker option for solubility predictions. Using a ML approach to improve the accuracy of the enthalpy of fusion and melting temperature predictions, similar to the approach taken to improve solubility predictions, is a possibility. The predictive methods would be used initially and then a correction factor would be applied to those predictions. The major drawback of this approach can be that for some solutes, there can be more than one answer for the response. This is because different polymorphs will have different values for enthalpy of fusion and melting temperature. The predictive methods do not take polymorphs into consideration and is a weakness of each method. However, any improvement in the accuracy of predictions of enthalpy of fusion and melting temperature cannot be dismissed and would improve solubility predictions where laboratory or literature data for enthalpy of fusion and melting temperature were unavailable.

One of the most pressing problems for any ML model is the amount of data required to build it. Industry has a large quantity of solubility data (Qiu and Albrecht, 2018), enthalpy of fusion data and even some heat capacities and it would be of great benefit to obtain access to it as more data would potentially inform better models and to enable the construction of solvent cluster or solvent specific models. Solvent

cluster models would include solvents with similar physical properties. Solvent specific models would focus exclusively on the API rather than solvent and would reduce the number of molecular descriptors by around a half. It is hoped these solvent specific models would improve on the type of models already shown in Chapter Four.

There is much work that can be done in improving the ML models described in Chapter Four. This work concentrated on obtaining a correction factor at 25°C. Developing a correction factor at higher temperatures would be of tremendous value in obtaining more accurate solubility curves. This would require more solubility data at higher temperatures to build the model. Building models using descriptors from other applications such as Dragon instead of MOE can be compared with the models already constructed in this work. Molecular fingerprints could also be used instead of descriptors. The ML algorithm used for building the model in this project was RF but other algorithms such as SVM and neural networks could be used instead and compared with RF. It is possible that all or some of these changes could result in improvements in the model.

The DoE method used in Chapter Three showed that COSMO*therm* predictions can be accurately reproduced using linear regression models in most cases. The initial predictions required for building these models is fairly standardised, and so could be prepared as soon as enthalpy of fusion and melting temperature data becomes available. This is particularly important when predictions are required quickly. The procedure could then be implemented with a user interface connected to a database,

providing users with effectively instant COSMO*therm* predictions where data are available.

Since the outputs of the linear regression models closely mirror COSMO*therm* predictions, the principle of applying a correction factor to COSMO*therm* predictions, as described in Chapter Four, applies equally to these.

As has already been stated, COSMO*therm* does not always give a prediction. Therefore, building ML models with UNIFAC, NRTL-SAC and SAFT-γ Mie to obtain a correction factor instead of or in conjunction with COSMO*therm* would have some benefit.

Solubility data are inconsistent in their availability across a range of temperatures. One of the major reasons for having the ML models at 25°C was that there were more data points at that temperature than others. It is possible that the correction factor at 25°C could be scaled and applied for higher temperatures. More data points would be required to test this hypothesis.

Further work would also include the construction of a correction factor for binary solvents. The amount of laboratory solubility data required for building this model would be considerably larger than the amount of data required for single solvent models. Each temperature has a number of data points as the binary solvents have a range of mole fraction ratios. The ML model will have another level of complexity when compared to the models already described as more descriptors will be needed for the extra solvent.

The case study for the selection of wash solvents using paracetamol as the API used COSMO*therm* and ML to obtain a correction factor. Each solvent was classified using solubility predictions and according to its suitability. In this study, because COSMO*therm* was able to classify the wash solvents with a high degree of accuracy, the use of ML did not improve the accuracy of the classifications but did improve the absolute value of solubility predictions. To further demonstrate the use of ML using another case study with a different API such as ibuprofen and its impurities could be completed. COSMO*therm* has not been as accurate at predicting the solubility for ibuprofen as with paracetamol. This could give the ML approach an opportunity to apply the correction factor to the predictions and reclassify any solvents misclassified by COSMO*therm*.

Manufacturing is increasingly using data analytics and data driven models to inform future manufacturing. The methodology developed in this thesis, including the workflow approach to solvent selection, can optimise manufacturing protocols. This can reduce costs, carbon footprint and time to market.

7 Appendix One

Table 7-1 Enthalpy of fusion and melting points taken from literature with references

COSMOtherm model compound name	melting point range °C	Δfus range kJ/mol	melting point °C	Δfus kJ/mol	Reference
1-(2-hydroxyphenyl)ethanone	196-197	31.59-32.80	196	31.589	(Shiu Shiang Yang and Guillory, 1972)
			197	32.802	(Shiu Shiang Yang and Guillory, 1972)
1,3-dimethylurea	97.85-106.72	12.64-13.62	106.3	13.62	(Gmehling, 2018)
			106.72	12.64	(Zordan *et al.*, 1972)
			97.85	12.76	(Kabo *et al.*, 1990)
			106.25	13.6	(Ferro *et al.*, 1987)
			106.35	13.62	(Della Gatta and Ferro, 1987)
2-aminopyridine	58.35-58.38	15.3	58.35	15.31	(Gmehling, 2018)
			58.38	15.3	(R. Sabbah and Gouali, 1996)
2-hydroxy-1,2,3-propanetricarboxylic acid	153.85-156	40.32-43.48	156	43.484	(Gmehling, 2018)
			153.85	43.455	(Booth *et al.*, 2010)

			155.4	41.84482	(Klímová and Leitner, 2012)
			154.65	40.32	(Meltzer and Pincu, 2012)
2-hydroxyphenylesterbenzoic acid	30-43	16.5-19.5	41.6	19.5	(Gmehling, 2018)
			43	19.28	(Dongwei Wei *et al.*, 2009)
			30	16.5	(Moura Ramos,Correia and Diogo, 2004)
			41	18.6	(Moura Ramos,Correia and Diogo, 2004)
			41.9	19.2	(Lazerges *et al.*, 2010)
			30.8	17.7	(Lazerges *et al.*, 2010)
			39.55	18.4	(Perisanu *et al.*, 2006)
			41.05	18.98	(Murthy,Paikaray and Arya, 1995)
			41.82	19.16	(Hanaya *et al.*, 2002)
3-aminobenzoic acid	171.95-179.75	21.81-36.04	174.4	21.836	(Gmehling, 2018)
				21.84048	(Andrews,Lynn and Johnston, 1926)

			172.55	33.7	(Rotich,Glass and Brown, 2001)
			179.75	21.84	(Acree, 1991)
			178.23	27.24	(Fredrik Nordström, 2008)
			171.95	36.04	(Fredrik Nordström, 2008)
				21.81	(Nielsen *et al.*, 2001)
3-nitrobenzoic acid	136.5-142.1	15.9-21.4	142.1	19.305	(Gmehling, 2018)
				19.33	(Andrews,Lynn and Johnston, 1926)
			141.4	19.246	(Lebedeva,Ryadnenko and Kuznetsova, 1971)
			139.85	21.4	(U.S. Rai and Mandal, 1990)
			141.15	19.33	(Domalski and Hearing, 1996)
			136.5	15.9	(R.K. Gupta and Singh, 2004)
				19.28	(Nielsen *et al.*, 2001)
				19.3	(Dean, 1992)
3-pyridinecarboxylic acid	233.85-236.85	13.01-30	236.01	13.01	(R Sabbah and Ider, 1999)
			235.95	27.57	(S. X. Wang *et al.*, 2004)

			233.85	27.7	(Goncalves and Da Piedade, 2012)
			236.85	30	(Allan *et al.*, 1989)
			236.85	26.7	(El Moussaoui,Chauvet and Masse, 1993)
			236.76	28.2	(Joseph,Bernardes and Minas Da Piedade, 2012)
			234.51	27.56	(Gonçalves,Rego and Minas da Piedade, 2011)
				26.7	(Nielsen *et al.*, 2001)
			236.1	97.097	(Liu-Cheng Wang and Wang, 2004)
4-acetamidobenzoic acid	240.55-262.15	34.2-42.4	262.15	34.2	(Manin,Voronin and Perlovich, 2014)
			240.55	42.4	(Monte *et al.*, 2010)
4-aminobenzoic acid	186-188.5	20.9-25	186	20.934	(Gmehling, 2018)
			188.25	20.92	(Andrews,Lynn and Johnston, 1926)
			187.3	24.016	(Ho-Meei Lin and Nash, 1993)
			188.5	24.03	(B. Bouillot, 2011)

			185.55	22.62	(Sandra Gracin and Rasmuson, 2002)
			185.55	25	(Blokhina *et al.*, 2015)
				24.03	(Nielsen *et al.*, 2001)
				20.9	(Dean, 1992)
			188.25	20.92	(Jia *et al.*, 2007)
4-amino-phenol	182.05-189.35	23.8-31.2	186.3	31.2	(Bret-Dibat and Lichanot, 1989)
			182.05	23.8	(Rotich,Glass and Brown, 2001)
			189.35	26	(R. Sabbah and Gouali, 1996)
			186.3	31.2	(Gmehling, 2018)
4-hydroxy-3-methoxybenzaldehyde	79.85-82.35	16.13-24.8	81.75	21.364	(Gmehling, 2018)
				20.9	(Lebedeva *et al.*, 1976)
				16.13	(Sharma,Sharma and Rambal, 1992)
			79.85	21.35	(Grady *et al.*, 1973)
			82.35	23.964	(Sumarokova and Nurmakova, 1960)

			81.12	22.62466	(Thakur *et al.*, 2012, Sandra Gracin and Rasmuson, 2002)
			81.45	22.35	(Preeti Gupta *et al.*, 2012)
			82.25	22.4	(Temprado,Roux and Chickos, 2008)
			81.4	23.522	(Draucker *et al.*)
			81.6	24.8	(Jayram Singh and Singh, 2015)
4-hydroxymethylbenzoate	125.35-126.05	24.31-25.3	126.05	25.3	(Giordano *et al.*, 1999)
			125.35	24.31	(Manzo and Ahumada, 1990)
4-nitrophenol	112.99-115.05	11-30.12		18.25	(Nielsen *et al.*, 2001)
				15.9	(Dean, 1992)
			112.99	18.882	(Gmehling, 2018)
			113.6	19.3	(Booss and Hauschildt, 1972)
			113.25	11	(Musuc,Razus and Oancea, 2002)
			115.05	18.25	(Domalski and Hearing, 1996)
			113.85	24.271	(Campbell and Campbell, 1941)
			113.85	18.254	(Poeti,Fanelli and Braghetti, 1982)

			112.99	18.87	(Donnelly *et al.*, 1990)
			112	30.118	(N. B. Singh and Kumar, 1986)
			113.9	24.309	(Nigam and Dhillon, 1970)
			114	19.22	(Kant,Rai and Rai, 2012)
			115	18.97	(Manjeet Singh *et al.*, 2013)
4-oh-benzoic acid	214-216.25	29.3-32.5	216	30.9	(Gmehling, 2018)
			214.9	30.9	(Armstrong,James and Wong, 1979)
			214	31.4	(Sandra Gracin and Rasmuson, 2002)
			214.85	32	(Heath,Singh and Ebisuzaki, 1992)
			215.55	29.3	(German L. Perlovich,Volkova and Bauer-Brandl, 2006)
			214.45	30.85	(Fredrik Nordström, 2008)
			215.8	30.85	(Fredrik L. Nordström and Rasmuson, 2006)
				30.86	(Nielsen *et al.*, 2001)
			216.25	32.5	(Monte *et al.*, 2010)

acetanilide	113.75-115.8	20.11-22.1	114.3	21.653	(Gmehling, 2018)
			113.75	21.44	(Y. P. Chen,Tang and Kuo, 2005)
			114.38	21.65	(Connett, 1979)
			115.8	20.301	(Bustamante *et al.*, 1998)
			114.17	20.112	(Mantheni *et al.*, 2012)
			113.75	22.1	(Umnahanant and Chickos, 2012)
			115.8	20.29	(Peña *et al.*, 2006)
				21.65	(Nielsen *et al.*, 2001)
ascorbic acid	190-192	19.72-37.00	192	19.723	(Gmehling, 2018)
			190	37.004	(Klímová and Leitner, 2012)
aspirin	130.85-143.12	19.10-34.34	131	34.343	(Gmehling, 2018)
			135.6	19.096	(Campanella *et al.*, 2010)
			132.8	34.32048	('Thermal Analysis T127: Application to Medical and Pharmaceutical Products (Melting Point and Fusion Heat)', 2018)
			143.12	31.874	(Kleineberg, 2009)

			136.05	29.17	(Xu *et al.*, 2004a)
			139.55	31.01	(German Perlovich and Bauer-Brandl, 2001)
			135.8	32.555	(Hahnenkamp, 2008)
			141.9	25.9	(Gorniak *et al.*, 2011)
				25.6	(Nielsen *et al.*, 2001)
			135.5	33.509	(Matsuda *et al.*, 2015)
			134.23	33.85	(Almeida *et al.*, 2015)
			130.85	29.8	(Kirklin, 2000)
b-d-fructofuranosyl-a-d-glucopyranoside	151.25-190.5	40.39-46.19	186	46.1867	(Gmehling, 2018)
			190	40.39141	(Roos, 1993)
			190.05	43.30096	(Diedrichs, 2005)
			186.05	42.88	(Paduszynski,Okuniewski and Domanska, 2013)
			151.25	43	(Magoń *et al.*, 2014)
			184.07	45.21	(Magoń *et al.*, 2014)
benzoic acid	121.8-122.55	15.53-18.06	122.4	17.452	(Gmehling, 2018)

				17.5	(Grigor'ev *et al.*, 1994)
				17.3	(Sharma,Sharma and Rambal, 1992)
				17.317	(U.S. Rai and Mandal, 2011)
			121.25	16.99	(Brittain, 2009)
			122.25	17.1	(Roy,Riga and Alexander, 2002)
			122.55	17.1	(Kennedy and Carr, 1973)
				17.321	(Andrews,Lynn and Johnston, 1926)
			122.5	18.7	(Murray,Cavell and Hill, 1980)
			122.4	17.44	(David, 1964)
			122.44	18.06	(R. Sabbah and Antipine, 1987)
			121.8	15.53	(Hrynakowski and Smoczkiewiczowa, 1937)
butylparaben	67.34-68.65	15.64-26.6	67.57	24.567	(Gmehling, 2018)
			68.65	26.6	(Giordano *et al.*, 1999)
			67.34	25.535	(Huaiyu Yang,Thati and Rasmuson, 2012)

				15.64	(Nielsen *et al.*, 2001)
caffeine	232.25-239	17.9-23.43	236	21.6	(Gmehling, 2018)
			238.85	23.4304	(Cesàro, 1980)
			238	21.12	(Shufen Li,Varadarajan and Hartland, 1991)
			234.85	20.98	(Grady *et al.*, 1973)
			234.55	24.8	(Guo *et al.*, 2010)
			237.05	21.9	(Pinto and Diogo, 2006)
			237.05	19.38	(Klous *et al.*, 2005)
			235.15	20.95	(Weinstein,Leffler and Currie, 1984)
			239	21.1041	(Adjei, 1980)
			235.85	19.86	(Dong *et al.*, 2007)
			236	21.6	(Bothe and Cammenga, 1980)
			235	21.963	('Thermal Analysis T127: Application to Medical and Pharmaceutical Products (Melting Point and Fusion Heat)', 2018)

			232.25	17.9	(Manic *et al.*, 2012)
			235.6	20.37	(Klímová and Leitner, 2012)
carbamazepine	190.71-192.59	24.55-27.41	190.71	26.33	(Liu *et al.*, 2013)
			192.15	25.06	(Good and Naír, 2009)
			192	24.551	(Subrahmanyam and Sarasija, 1997)
			192.59	27.41	(Kikic *et al.*, 2010)
chinone	111.92-114.85	18.35-19.6	114	18.542	(Gmehling, 2018)
				18.451	(Andrews,Lynn and Johnston, 1926)
			111.92	18.4	(Rojas-Aguilar *et al.*, 2004)
			112.5	18.35	(Rojas-Aguilar *et al.*, 2004)
			114.85	18.45	(Acree, 1991)
d-(-)fructose	104-127	24.11-36.03	104	36.03	(Gmehling, 2018)
			127	30.446	(Roos, 1993)
			106.75	24.11	(Paduszynski,Okuniewski and Domanska, 2013)

dextrose	150-158	32.25-34.19	150	32.422	(Gmehling, 2018)
			158	32.24828	(Roos, 1993)
			157.65	34.19	(Paduszynski,Okuniewski and Domanska, 2013)
diclofenac	179.85-181	39.39-40.4	181	39.388	(Pasquali,Bettini and Giordano, 2007)
			179.45	40.4	(Surov *et al.*, 2009)
fenofibrate	74-81.4	27.3-33.55	78.9	33.553	(Gmehling, 2018)
			81.4	27.3	(Gorniak *et al.*, 2011)
			74	0.9	(Gorniak *et al.*, 2011)
			80.55	32.4	(Zhou *et al.*, 2002)
			78.9	33.53	(Watterson *et al.*, 2014)
hexanedioic acid	145.85-153.4	33.7-39.7	153.4	34.851	(Gmehling, 2018)
				39.7	(Babinkov *et al.*, 1979, Donnelly *et al.*, 1990)
			151.49	36.74	(Donnelly *et al.*, 1990)
			152.35	34.88	(Cingolani and Berchiesi, 1974)

			145.85	33.7	(Roux,Temprado and Chickos, 2005)
			149.85	35.891	(Booth *et al.*, 2010)
			153.16	35.2	(Tzu-Chi Wang,Lai and Chen, 2010)
			153.15	35.2	(T. C. Wang,Li and Chen, 2012)
			152	35.15	(Y. Li *et al.*, 2014)
			152.35	34.852	(Mao *et al.*, 2009)
hydroquinone	171.8-191.35	21.09-30.8	172	27.108	(Gmehling, 2018)
			172.6	27.108	(D.W. Wei,Li and Zhang, 2004)
			172.83	21.09	(R. Sabbah and Buluku, 1991)
				27.112	(Andrews,Lynn and Johnston, 1926)
			171.8	26.5	(Bret-Dibat and Lichanot, 1989)
			173.55	30.8	(Naoki *et al.*, 1999)
			191.35	24.3	(Naoki *et al.*, 1999)
			170.55	27.2	(S. P. Verevkin, 1999)

			171.95	27.23	(Sergey P. Verevkin and Kozlova, 2008)
			172.3	24.618	(Fall and Luks, 1965)
ibuprofen	73.2-80.8	17.62-39.5	75.29	26.085	(Gmehling, 2018)
			80.8	23.681	(Lerdkanchanaporn,Dollimore and Evans, 2001)
			74	25.5	(S. Gracin,Brinck and Rasmuson, 2002)
			75.2	25.04	(Xu *et al.*, 2004b)
			77.25	39.5	(Cilurzo *et al.*, 2010)
			73.25	26.6	(Wassvik *et al.*, 2006)
			77.75	25.7	(Z. Jane Li *et al.*, 1999)
			74.45	27.94	(Hong *et al.*, 2010)
			75.21	17.623	(Graubner, 2008)
			76.65	25.119	(Kleineberg, 2009)
			74.55	27.7	(Domańska *et al.*, 2009)
			76.44	25.07	(Baoguo Wang *et al.*, 2012)

			74.35	25.692	(Baptiste Bouillot,Teychené and Biscans, 2013)
indomethacin	153-162.05	32.91-75.4	159.85	37.656	(Grady *et al.*, 1973)
			161.35	43.5	(Aceves-Hernandez *et al.*, 2009)
			162.05	36.5	(Murdande *et al.*, 2010)
			160.85	39.99	(Basavoju,Bostrom and Velaga, 2008, Paus *et al.*, 2015)
			160.45	75.4	(Paus *et al.*, 2015)
			159	36.852	(Hamdi *et al.*, 2004)
			153	32.934	(Hamdi *et al.*, 2004)
			156	36.137	(Hancock and Parks, 2000)
			162	36.494	(Hancock and Parks, 2000)
			159.8	37.9	(Wassvik *et al.*, 2006)
			160.1	36.852	(Legendre and Feutelais, 2004)
			153	32.91668	(Legendre and Feutelais, 2004)
ketoprofen	94.5-94.8	28.23-37.3	94.5	28.245	(Gmehling, 2018)

			94.5	28.226	(Espitalier,Biscans and Laguérie, 1995)
			94.8	37.3	(Wassvik *et al.*, 2006)
l-(+)tartaric acid	162-171.89	34.32-36.31	171.89	36.31	(Meltzer and Pincu, 2009)
			162	34.325	(J. Li,Zhou and Huang, 1991)
mannitol	164.1-176.9	33.52-61.57	169.05	54.7265	(Gmehling, 2018)
				59.3	(Siniti *et al.*, 1993)
			166	53.58	(Spaght,Thomas and Parks, 1932)
			176.9	33.51983	(Arias,Moyano and Ginés, 1998)
			165.95	56.1	(Barone *et al.*, 1990)
			169	59.3884	(Siniti,Jabrane and Létoffé, 1999)
			165	61.574	(Gombás *et al.*, 2003)
			164.1	54.69	(Bo Tong *et al.*, 2010)
			166.9	33.52	(Campanella *et al.*, 2010)
			165.95	56.1	(Barone *et al.*, 1990)
mefenamic acid	229.95-230.45	71.2-38.7	229.95	71.2	(Domańska *et al.*, 2010)
			230.45	38.25	(Romero *et al.*, 2004)

			230.35	38.7	(Surov *et al.*, 2009)
				38.7	(Nielsen *et al.*, 2001)
naproxen	155-156.32	29-32.2	155.85	31.751	(Gmehling, 2018)
			154.55	31.4	(Türk and Kraska, 2009)
			155.35	31.5	(Neau,Bhandarkar and Hellmuth, 1997)
			155	30.11	(Kikic *et al.*, 2010)
			153.55	29	(Saini and Murthy, 2014)
			155.6	32.2	(Wassvik *et al.*, 2006)
			156.32	31.5	(Paus *et al.*, 2015)
niacinamide	122.19-130.65	20.49-26.94	128.85	22.84464	(Grady *et al.*, 1973)
			130.65	23.8	(Good and Naír, 2009)
			128.45	25.5	(Nicoli *et al.*, 2008)
			128.85	26.94	(Negoro *et al.*, 1960)
			129.2	26.08	(Lvova,Garber and Kozlov, 1988)
			122.19	23.4304	(Wyrzykowska-Stankiewicz and Szafranski, 1975)

			128.45	20.49	(Wu,Dang and Wei, 2014)
			128.05	23.7	(Almeida,Oliveira and Monte, 2015)
			128.02	25.2	(Almeida,Oliveira and Monte, 2015)
octadecanoic acid	52.95-70.95	45.23-68.44	69.35	56.5854	(Gmehling, 2018)
				45.22904	(Lebedeva, 1964)
			69.34	61.209	(Schaake,van Miltenburg and De Kruif, 1982)
			67.11	57.67	(Donnelly *et al.*, 1990)
			69.6	56.4	(Danilin *et al.*, 2001)
			69.68	68.44	(Singleton,Ward and Dollear, 1950)
			65	58.6	(Vold, 1949)
			69.6	61.3	(Sato *et al.*, 1990)
			65.15	60.4	(Moore *et al.*, 2007)
			69.65	63.2	(Moreno *et al.*, 2007)
			70.95	57.8	(Teixeira,Gonçalves Da Silva and Fernandes, 2006)

			67.05	50.93	(Yu *et al.*, 2000)
			69.35	61.21	(Domalski and Hearing, 1996)
			52.95	64.643	(Eykman, 1889)
			64	56.7	(Bruner, 1894)
			68.85	57.78	(Berchiesi,Cingolani and Leonesi, 1974)
orthoaminobenzoic acid	144.1-144.65	20.38-20.66	144.6	20.659	(Gmehling, 2018)
			144.65	20.37608	(Andrews,Lynn and Johnston, 1926)
			144.1	20.38	(Jia *et al.*, 2007)
paracetamol	156.4-172.45	26.25-28.15	168.05	27.619	(Gmehling, 2018)
			170.45	27.1	(R. A. Granberg and Rasmuson, 1999)
			168.6	28.1	(Sacchetti, 2001)
			156.4	27.6	(Sacchetti, 2001)
			170.05	27.6	(Mota *et al.*, 2009)
			167.15	27	(S. Vecchio and Tomassetti, 2009)

			168.75	26.49	(Xu *et al.*, 2006)
			169.13	26.25	(Bustamante,Romero and Reillo, 1995)
			169.7	27.708	(Bustamante,Romero and Reillo, 1995)
			172.45	27	(Kleineberg, 2009)
			168.55	27	(Neau,Bhandarkar and Hellmuth, 1997)
			168.05	26.024	(Manzo and Ahumada, 1990)
			168.72	28.151	(Graubner, 2008)
			169.03	27.852	(Kleineberg, 2009)
phenacetin	133.5-137.05	21.4-32.45	135	31.275	(Gmehling, 2018)
			133.85	32.45	(E.E. Marti, 1972)
			133.5	24.51	(Hrynakowski and Smoczkiewiczowa, 1937)
			136.45	30	(S. Vecchio and Tomassetti, 2009)
			135.15	28.75	(Peña *et al.*, 2009)
			134.25	34.1	(Wassvik *et al.*, 2008)

			137.05	21.4	(Stefano Vecchio *et al.*, 2004)
			134.05	31.254	(Manzo and Ahumada, 1990)
			134.7	31.4529	(Thakur *et al.*, 2012)
phthalic acid	190.3-210	36.5-52.3	210	52.2994	(Gmehling, 2018)
			190.3	36.5	(R Sabbah and Perez, 1999)
				52.3	(Dean, 1992)
			193.85	52.3	('Thermophysical data', 1936)
piroxicam	199.8-202.1	34.13-36.3	202.1	35.52	(Bustamante,Peña and Barra, 1998)
			199.8	35.52	(Bustamante,Peña and Barra, 1998)
			200	34.129	(Grandelli *et al.*, 2012)
			200.55	35.85	(Sotomayor *et al.*, 2012)
			200.75	35	(Drebushchak *et al.*, 2006)
				34.54	(Nielsen *et al.*, 2001)
			200.3	36.3	(Wassvik *et al.*, 2006)
propylparaben	96.05-96.15	16.85-28.01	96.05	28.01	(Gmehling, 2018)

			96.15	27.2	(Giordano *et al.*, 1999)
			96.05	27.99	(Manzo and Ahumada, 1990)
				16.85	(Nielsen *et al.*, 2001)
saccharin	225.85-229.75	26.77-32.1	225.85	27.41	(Grady *et al.*, 1973)
			229.75	32.1	(Good and Naír, 2009)
			229.55	26.77	(Basavoju,Bostrom and Velaga, 2008)
			227.59	29.89	(Matos *et al.*, 2005)
salicylic acid	158.15-160.95	12.84-28.8	158.85	23.205	(Gmehling, 2018)
				18.2	(Raphaël Sabbah and Le, 1993, Baptiste Bouillot,Teychené and Biscans, 2013)
			158.95	26.1	(Pinto,Diogo and Minas Da Piedade, 2003)
			158.05	28.8	(German L. Perlovich,Volkova and Bauer-Brandl, 2006)
			160.95	27.1	(Good and Naír, 2009)
			159.35	23.05	(Peña *et al.*, 2009)

			158.15	24.448	(Mota *et al.*, 2008)
			158.65	24.6	(Lim *et al.*, 2013)
			160.3	12.845	(Campanella *et al.*, 2010)
			158.25	25.269	(Baptiste Bouillot,Teychené and Biscans, 2013)
				23.52	(Meltzer and Pincu, 2009)
			159.3	23.05	(Peña *et al.*, 2006)
sorbitol	93.35-106	28.05-39.53		31.6	(Siniti,Jabrane and Létoffé, 1999)
			99	28.054	(Roos, 1993)
			95	35	(Talja and Roos, 2001)
			100.7	35.705	(Nakada *et al.*, 2006)
			93.35	30.2	(Barone *et al.*, 1990)
			96.8	39.531	(Gombás *et al.*, 2003)
			106	32.4269	(Nakada *et al.*, 2006)
			96.05	30.35	(B. Tong *et al.*, 2008)
			97.05	30.5	(Paduszyński,Okuniewski and Domańska, 2015)

			101.65	31.6	(Siniti *et al.*, 1993)
succinic acid	181.85-188	31.26-53.1	188	32.946	(Gmehling, 2018)
			183.85	32.97	(Cingolani and Berchiesi, 1974)
			181.85	34	(Roux,Temprado and Chickos, 2005)
			184.85	31.259	(Booth *et al.*, 2010)
			185.5	32.72	(Y. Li *et al.*, 2014)
				53.1	(Nielsen *et al.*, 2001)
			184.85	34.46	(Khetarpal,Lal and Bhatnagar, 1980)
sulfacetamide	182-183	22.38-29.8	182	29.76	(Fleming Martínez and Gómez, 2002)
			182	29.8	(F. Martínez,Ávila and Gómez, 2003)
			182	29.76	(F. Martínez and Gómez, 2001)
			183	22.384	(Shiu Shiang Yang and Guillory, 1972)

sulfaguanidine	191-192	21.12-25.94	191	24.83	(Shiu Shiang Yang and Guillory, 1972)
			191	22.384	(Shiu Shiang Yang and Guillory, 1972)
			191	25.94	(Shiu Shiang Yang and Guillory, 1972)
			191	21.13	(Shiu Shiang Yang and Guillory, 1972)
			192	23.597	(Shiu Shiang Yang and Guillory, 1972, Basavoju,Bostrom and Velaga, 2008)
sulfathiazole	171.8-201.5	0.163-33.31	200.95	28.909	(Gmehling, 2018)
			200.15	28.89	(E. Marti, 1988)
			172.65	33.31	(E. Marti, 1988)
			195.15	23.92	(E. Marti, 1988)
			199.8	30.25	(Fleming Martínez and Gómez, 2002)
			199.8	30.3	(F. Martínez,Ávila and Gómez, 2003)

			201.5	26.13745	(Sunwoo and Eisen, 1971)
			198.7	25.543	(Martin,Wu and Velasquez, 1985)
			171.8	0.1629	(Khattab, 1983)
			197.2	24.1	(Khattab, 1983)
theophylline	269.7-277.4	28.72-36.24	277.4	36.2494	(Franceschi *et al.*, 2008)
			276.37	30.664	(Franceschi *et al.*, 2008)
			274.5	29.69385	(Adjei, 1980)
			270.8	30.4481	(Szterner,Legendre and Sghaier, 2010)
			269.7	28.71851	(Szterner,Legendre and Sghaier, 2010)
			270.7	30.1058	(Szterner,Legendre and Sghaier, 2010)
			274.7	28.01586	(Szterner,Legendre and Sghaier, 2010)
			271.31	29.034	(Hahnenkamp, 2008)
			271.51	31.824	(Liu *et al.*, 2013)
			274.5	29.5235	(Ho-Meei Lin and Nash, 1993)

urea	132.7-134.85	12.93-15.03	132.7	14.79	(Gmehling, 2018)
			134.75	15.03	(Ferloni and Gatta, 1995)
				13.9	(Kozyro,Dalidovich and Krasulin, 1986)
			133.1	13.61	(Zordan *et al.*, 1972)
			134.05	14.6	(U.S. Rai and Rai, 1999)
			132.05	13.6	(Jamróz *et al.*, 1998)
			133.55	14.6	(U. Rai and Rai, 1998)
			134.85	12.93	(Kabo *et al.*, 1990)
			132.65	13.61	(Vogel and Schuberth, 1980)
			133.35	14.79	(Della Gatta and Ferro, 1987)
			132.7	14.518	(Miller and Dittmar, 1934)
			132.65	13.9	(Qin *et al.*, 2010)
			134.5	14.8	(Reddi *et al.*, 2011)

8 Appendix Two

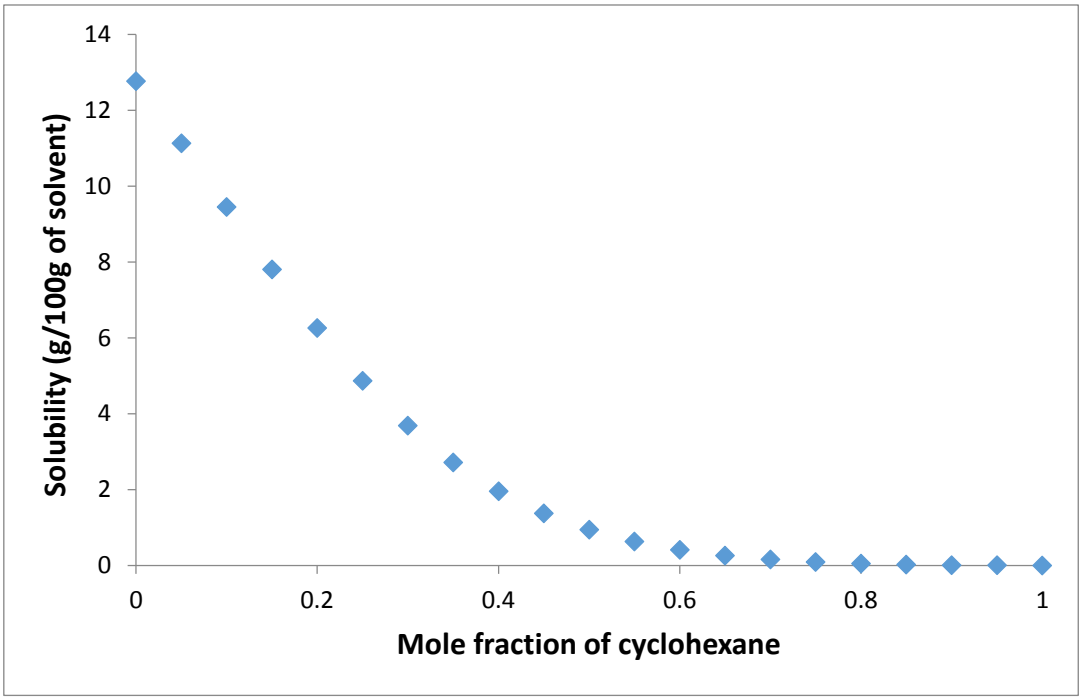


Figure 8-1 predicted solubility curve for paracetamol in 1,3-dioxane and cyclohexane at 25°C using COSMOtherm

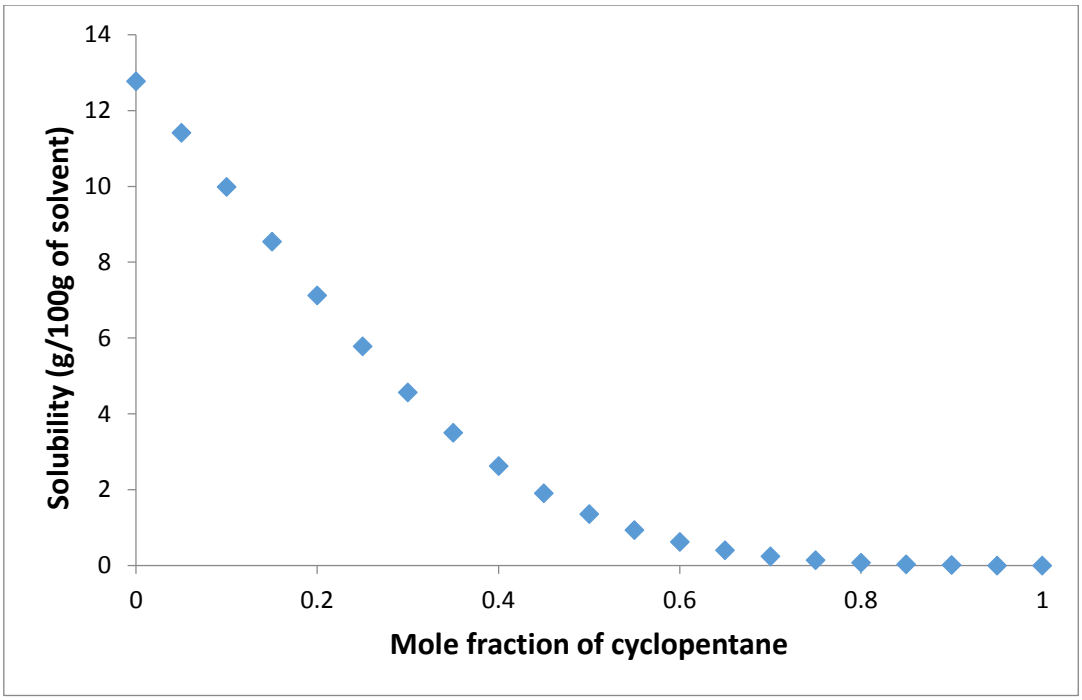


Figure 8-2 predicted solubility curve for paracetamol in 1,3-dioxane and cyclopentane at 25°C using COSMOtherm

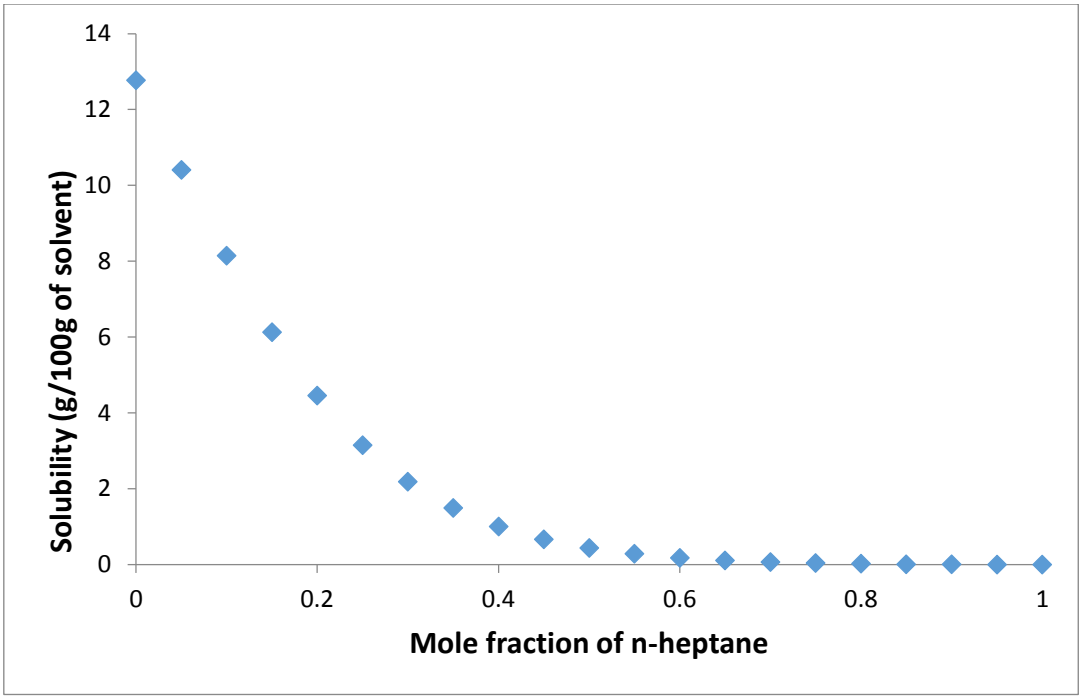


Figure 8-3 predicted solubility curve for paracetamol in 1,3-dioxane and n-heptane at 25°C using COSMOtherm

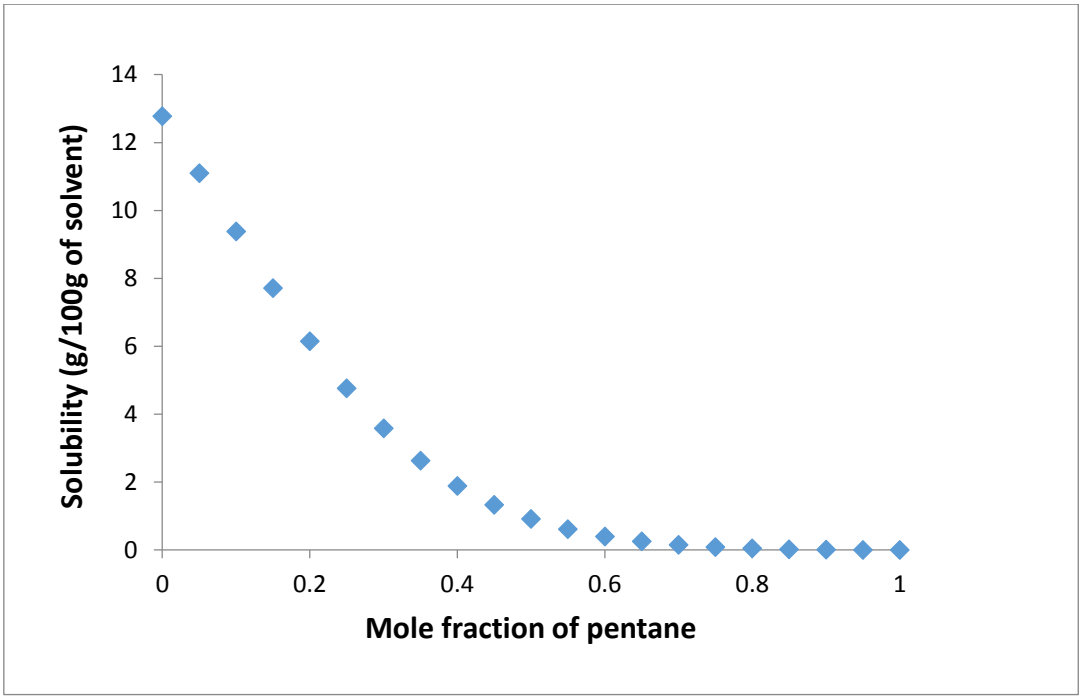


Figure 8-4 predicted solubility curve for paracetamol in 1,3-dioxane and pentane at 25°C using COSMOtherm

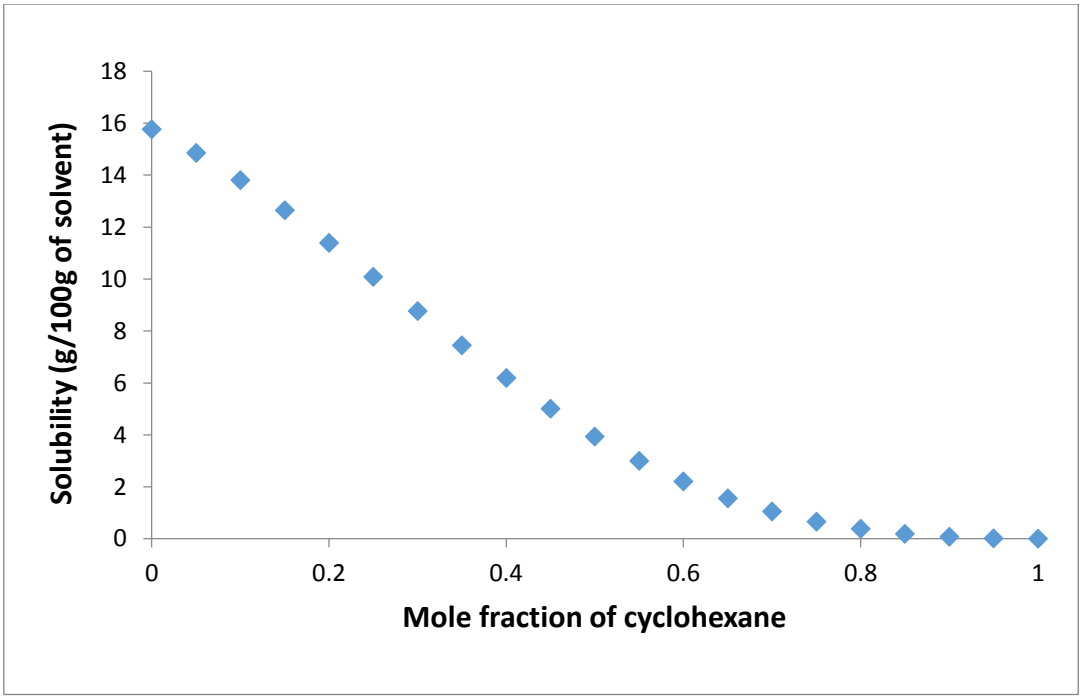


Figure 8-5 predicted solubility curve for paracetamol in 2-methoxyethanol and cyclohexane at 25°C using COSMOtherm

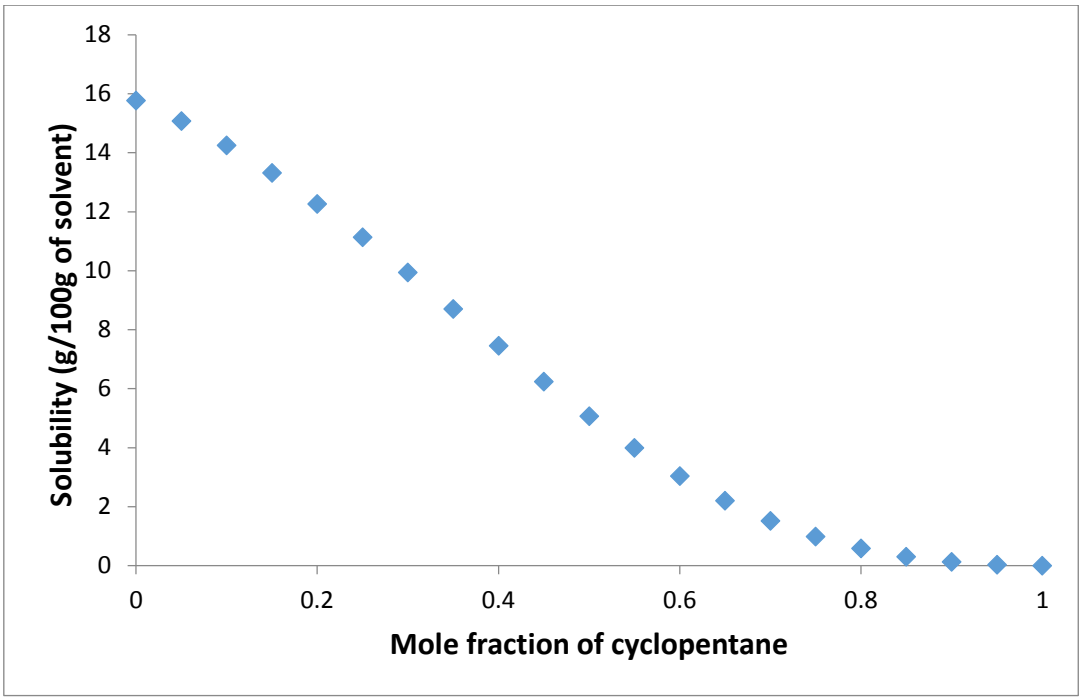


Figure 8-6 predicted solubility curve for paracetamol in 2-methoxyethanol and cyclopentane at 25°C using COSMOtherm

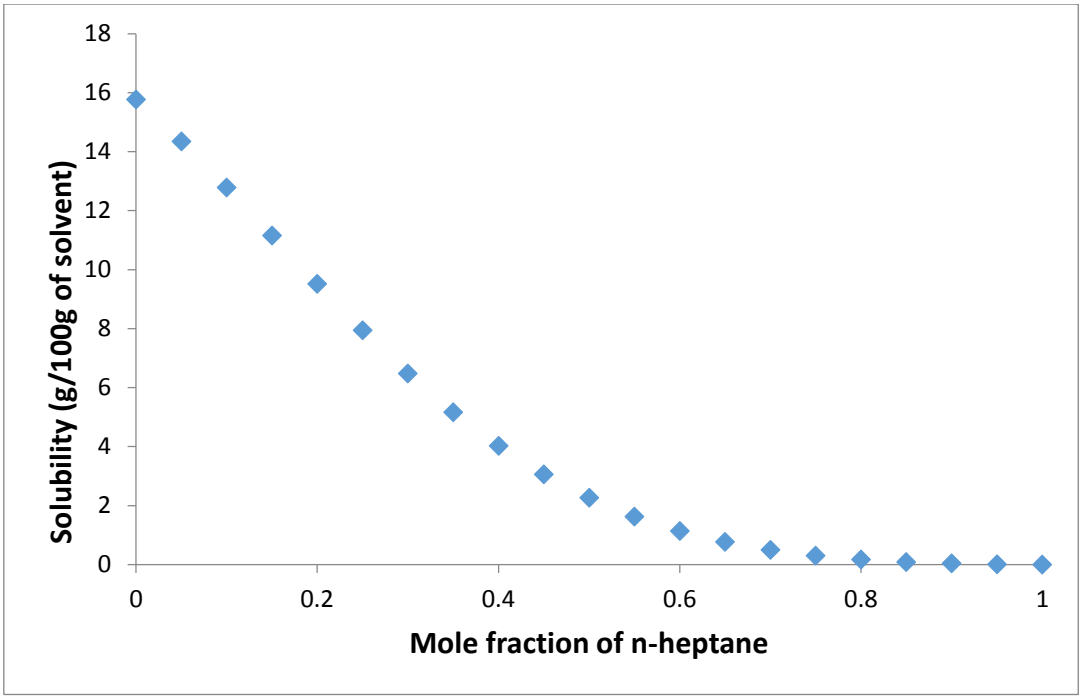


Figure 8-7 predicted solubility curve for paracetamol in 2-methoxyethanol and n-heptane at 25°C using COSMOtherm

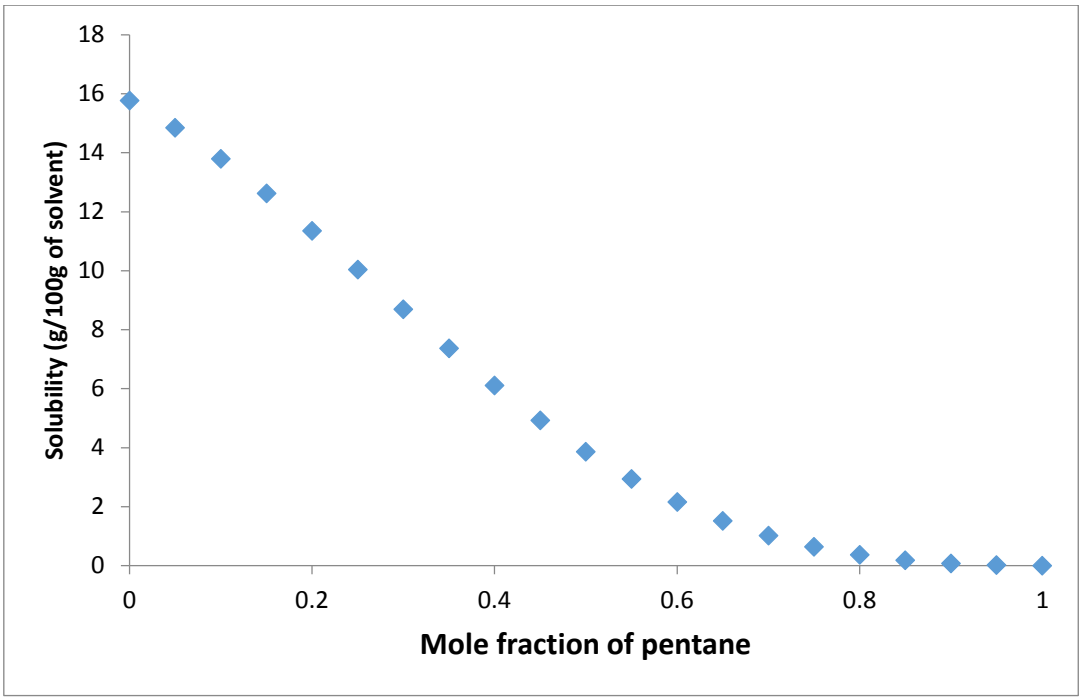


Figure 8-8 predicted solubility curve for paracetamol in 2-methoxyethanol and pentane at 25°C using COSMOtherm

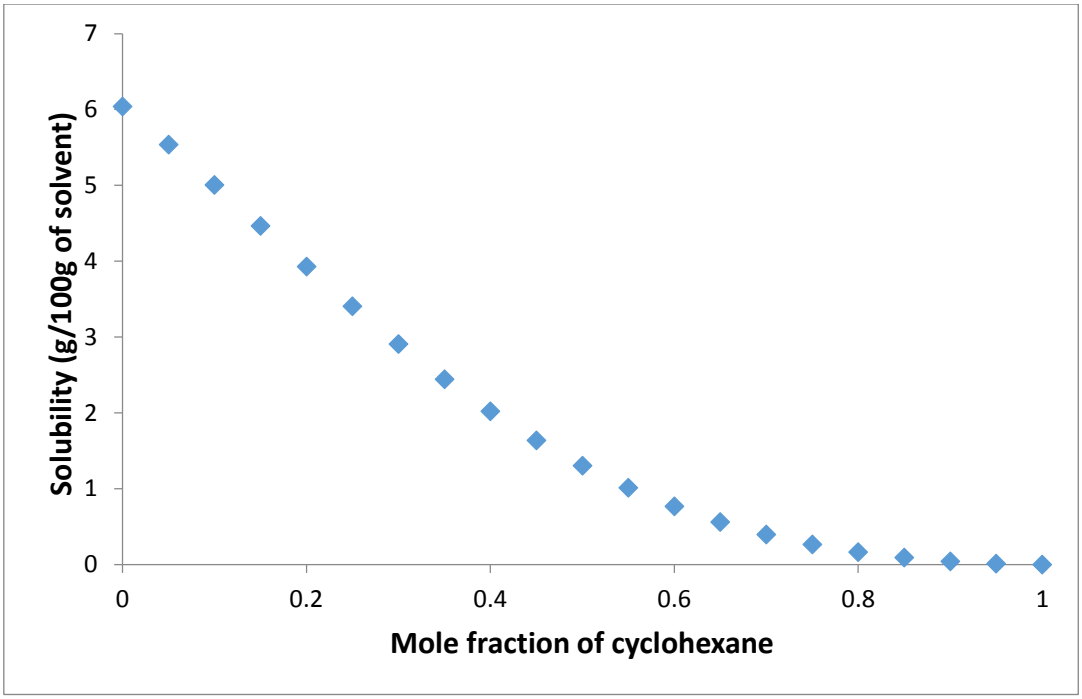


Figure 8-9 predicted solubility curve for paracetamol in 2-propanol and cyclohexane at 25°C using COSMOtherm

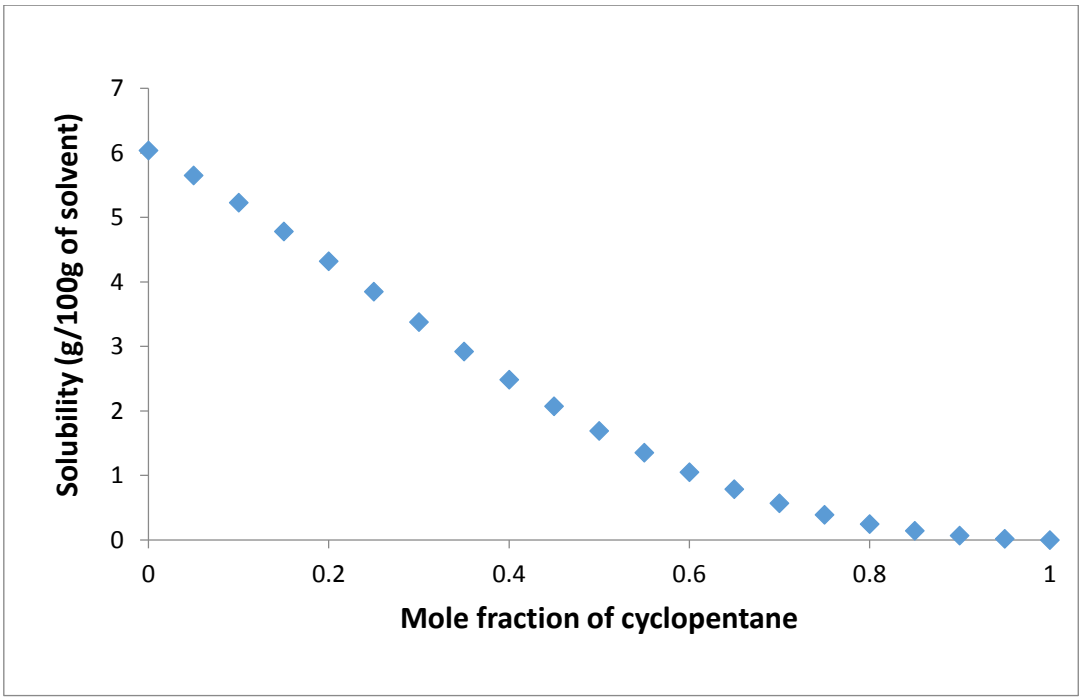


Figure 8-10 predicted solubility curve for paracetamol in 2-propanol and cyclopentane at 25°C using COSMOtherm

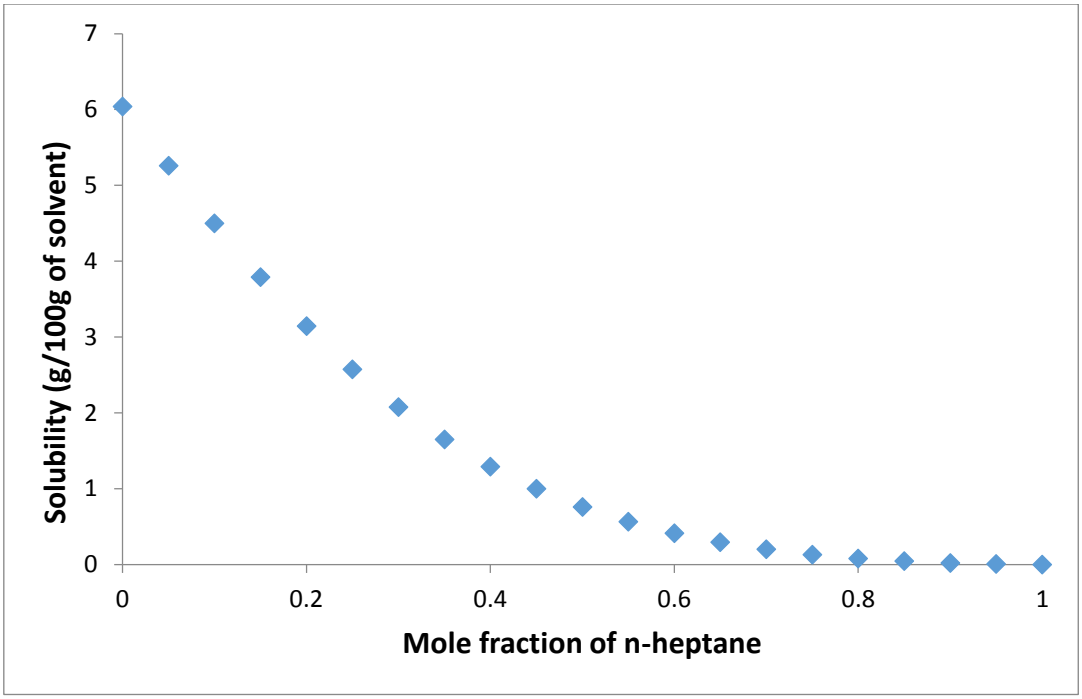


Figure 8-11 predicted solubility curve for paracetamol in 2-propanol and n-heptane at 25°C using COSMOtherm

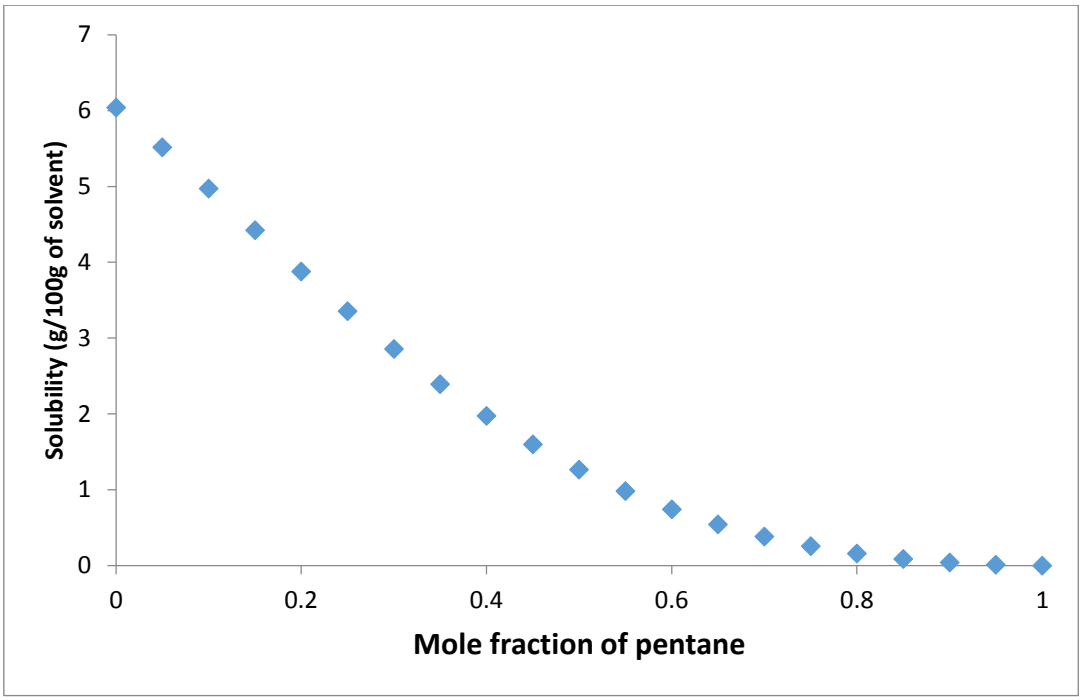


Figure 8-12 predicted solubility curve for paracetamol in 2-propanol and pentane at 25°C using COSMOtherm

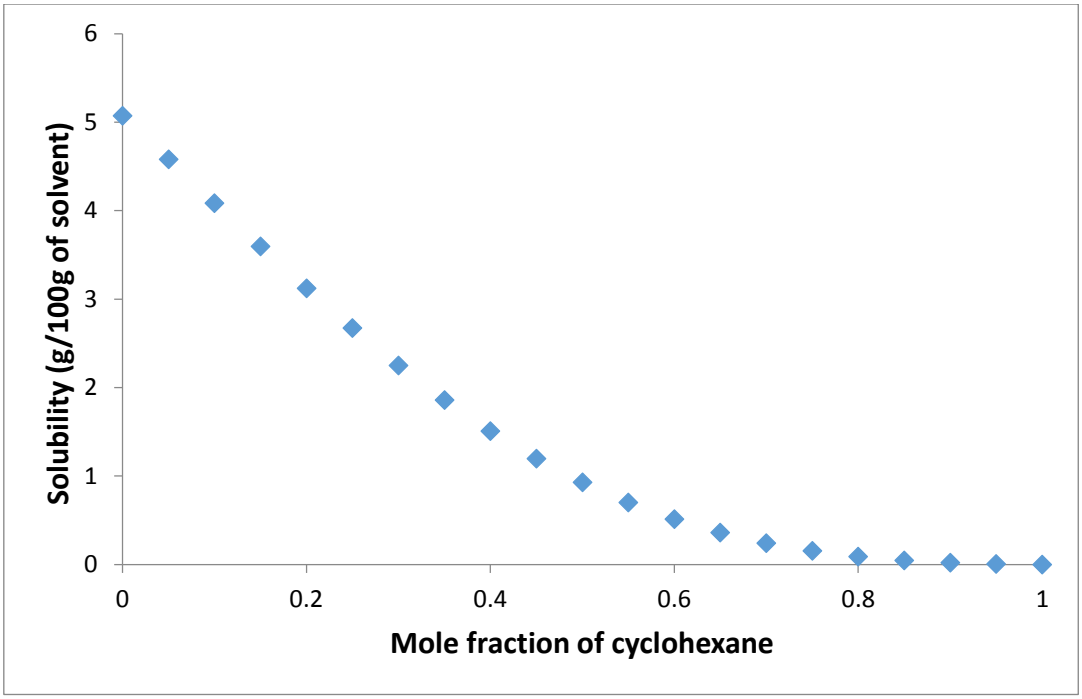


Figure 8-13 predicted solubility curve for paracetamol in 4-methyl-2-pentanone and cyclohexane at 25°C using COSMOtherm

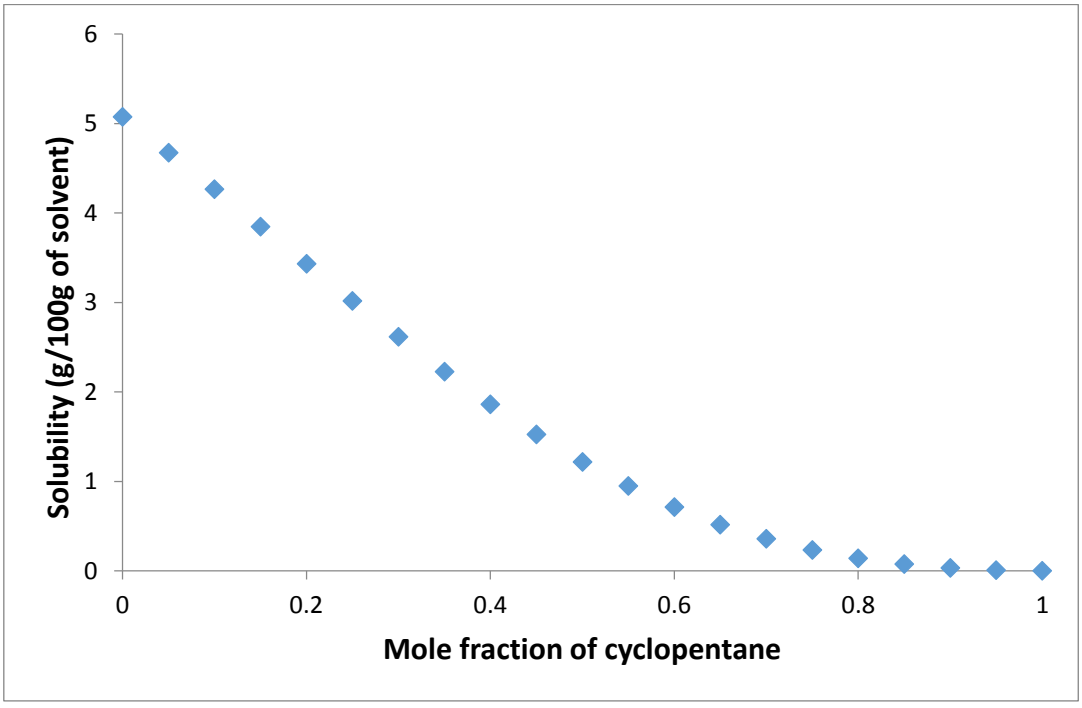


Figure 8-14 predicted solubility curve for paracetamol in 4-methyl-2-pentanone and cyclopentane at 25°C using COSMOtherm

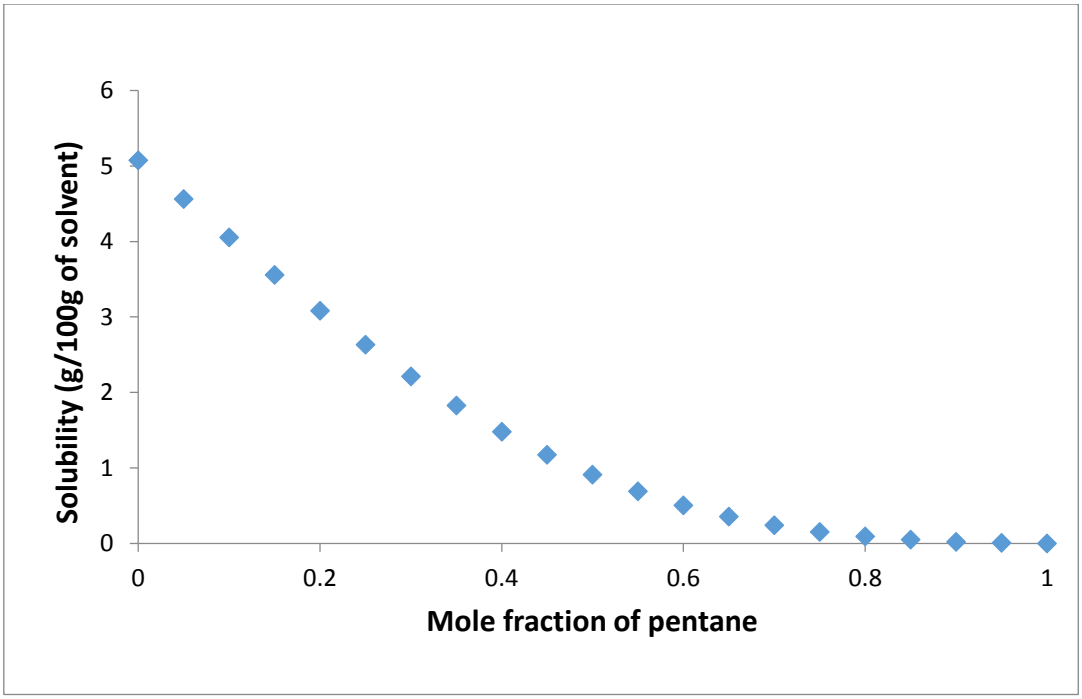


Figure 8-15 predicted solubility curve for paracetamol in 4-methyl-2-pentanone and pentane at 25°C using COSMOtherm

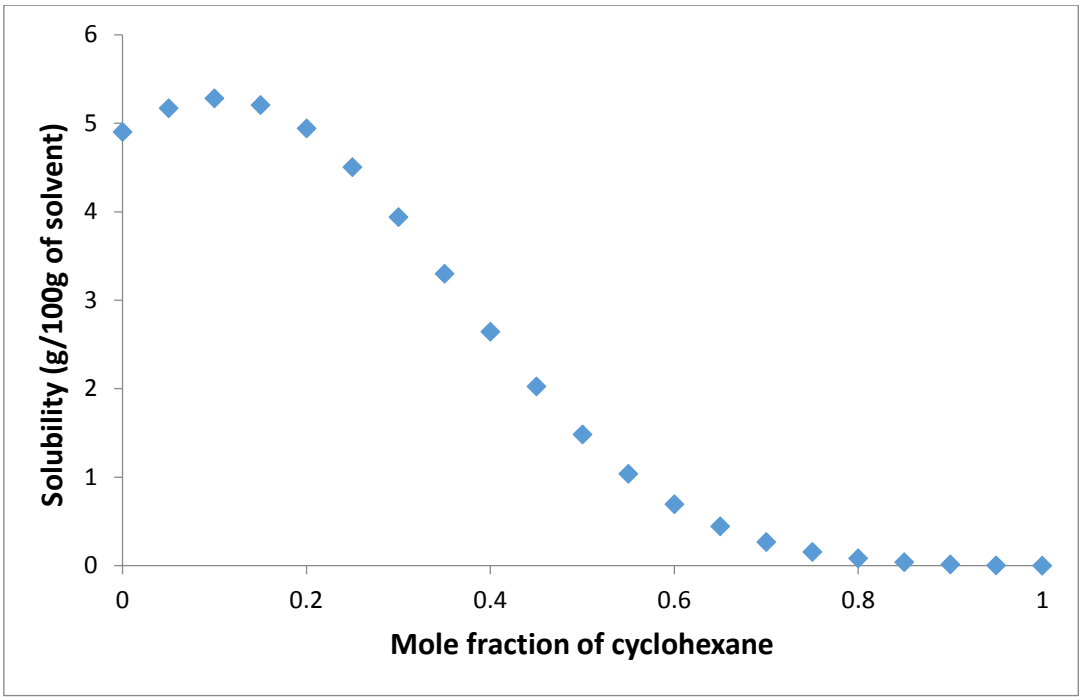


Figure 8-16 predicted solubility curve for paracetamol in acetonitrile and cyclohexane at 25°C using COSMOtherm

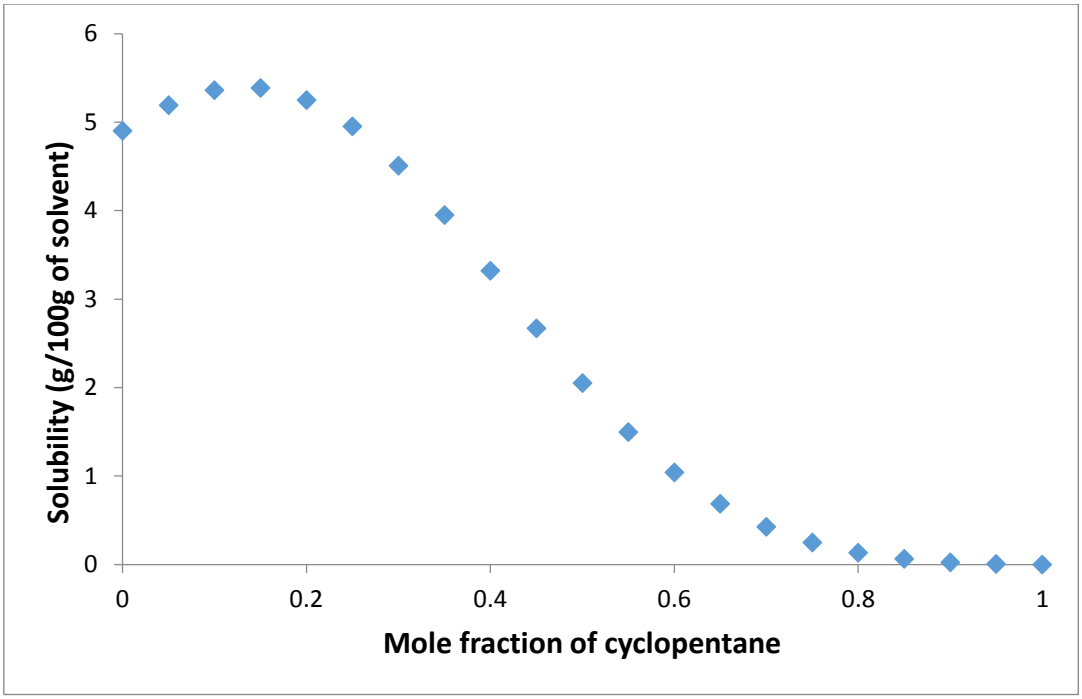


Figure 8-17 predicted solubility curve for paracetamol in acetonitrile and cyclopentane at 25°C using COSMOtherm

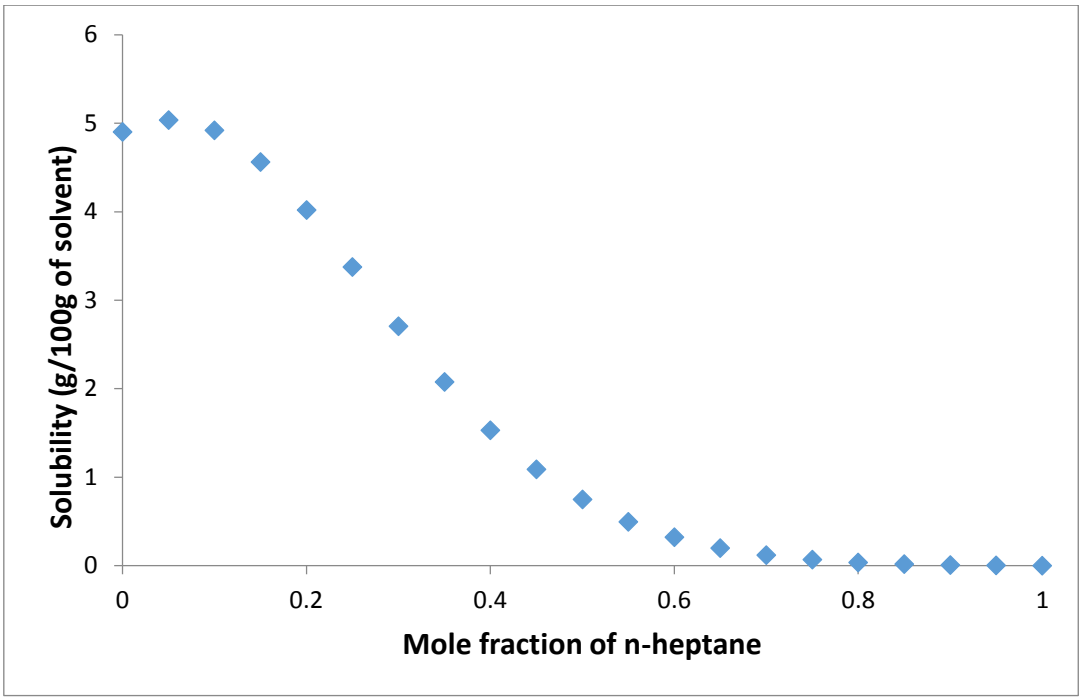


Figure 8-18 predicted solubility curve for paracetamol in acetonitrile and n-heptane at 25°C using COSMOtherm

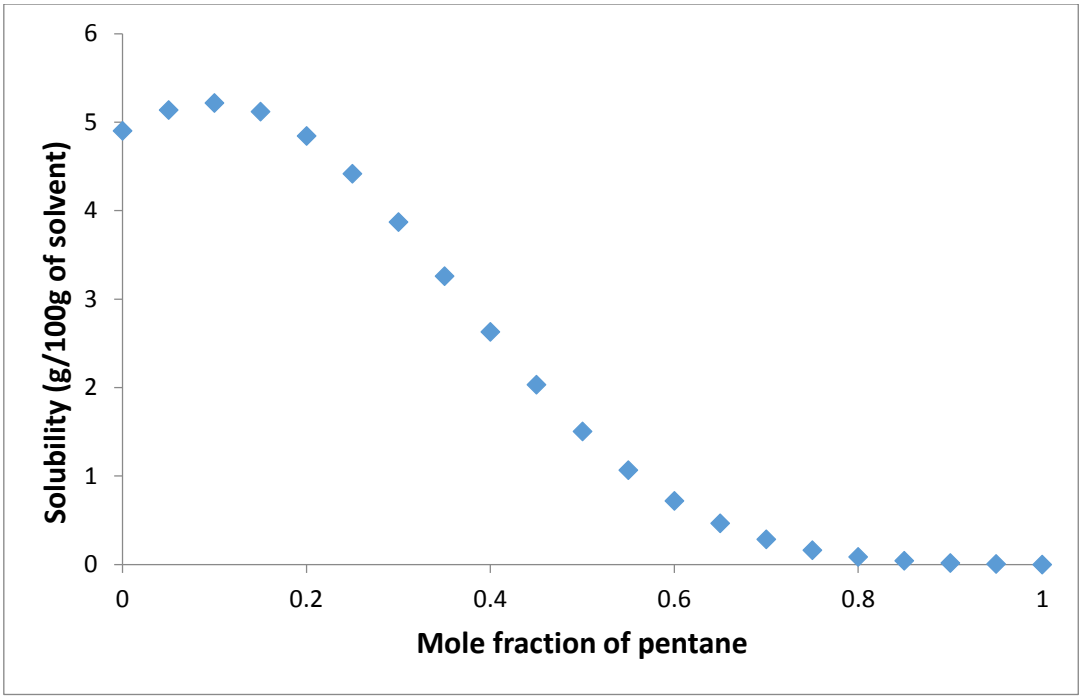


Figure 8-19 predicted solubility curve for paracetamol in acetonitrile and pentane at 25°C using COSMOtherm

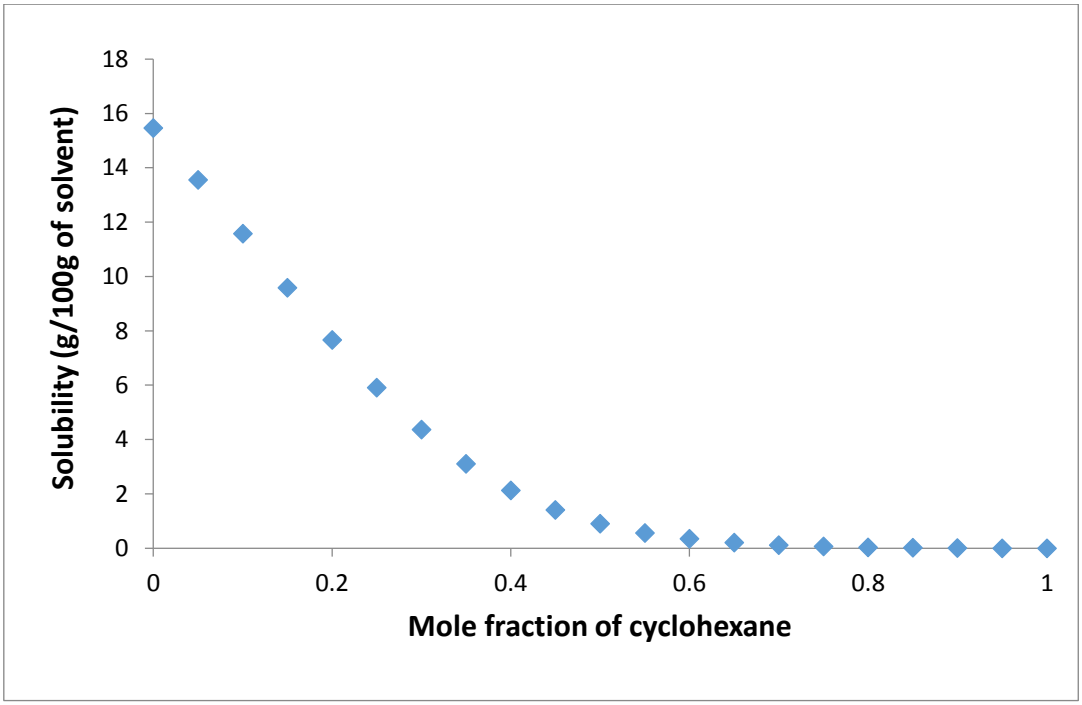


Figure 8-20 predicted solubility curve for paracetamol in dioxane and cyclohexane at 25°C using COSMOtherm

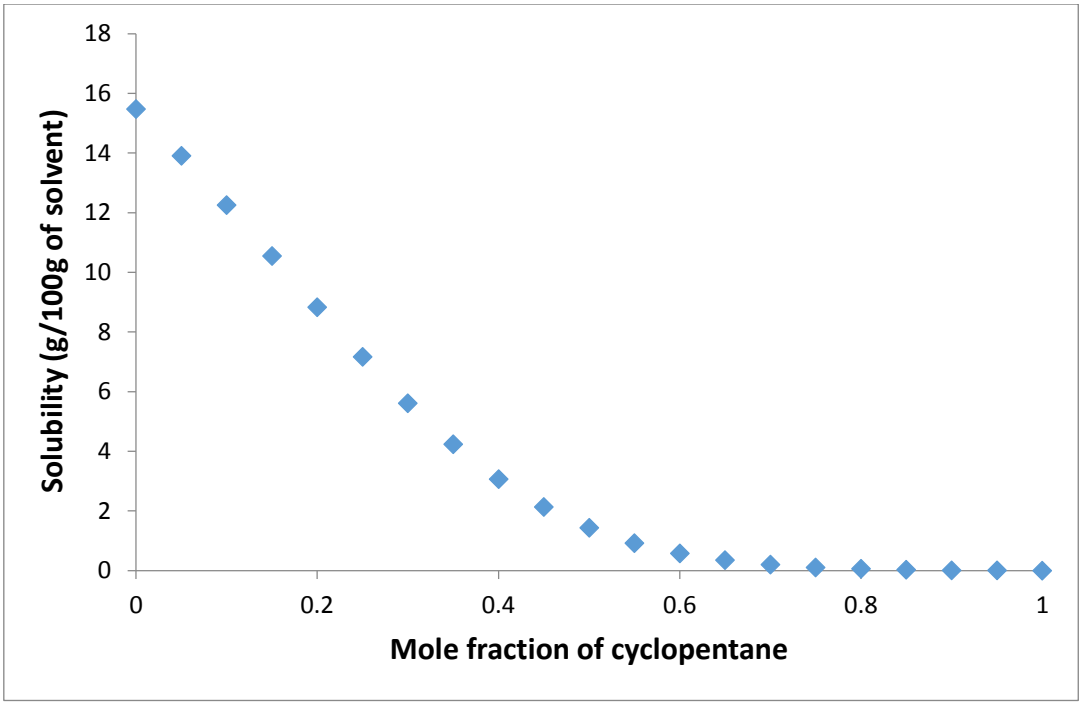


Figure 8-21 predicted solubility curve for paracetamol dioxane and cyclopentane at 25°C using COSMOtherm

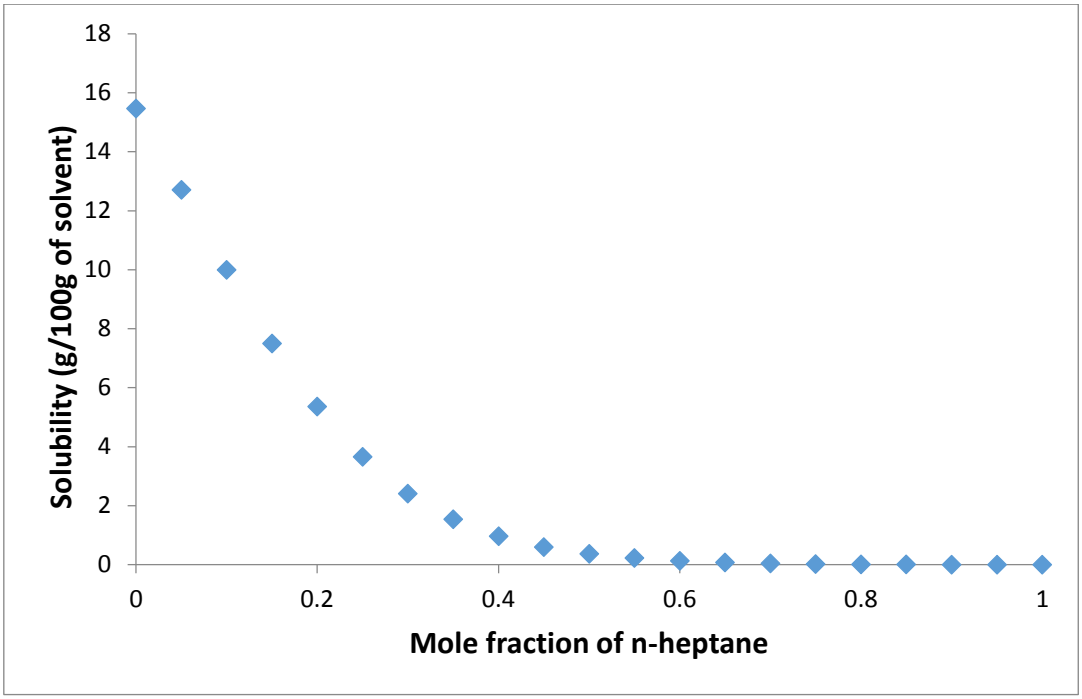


Figure 8-22 predicted solubility curve for paracetamol dioxane and n-heptane at 25°C using COSMOtherm

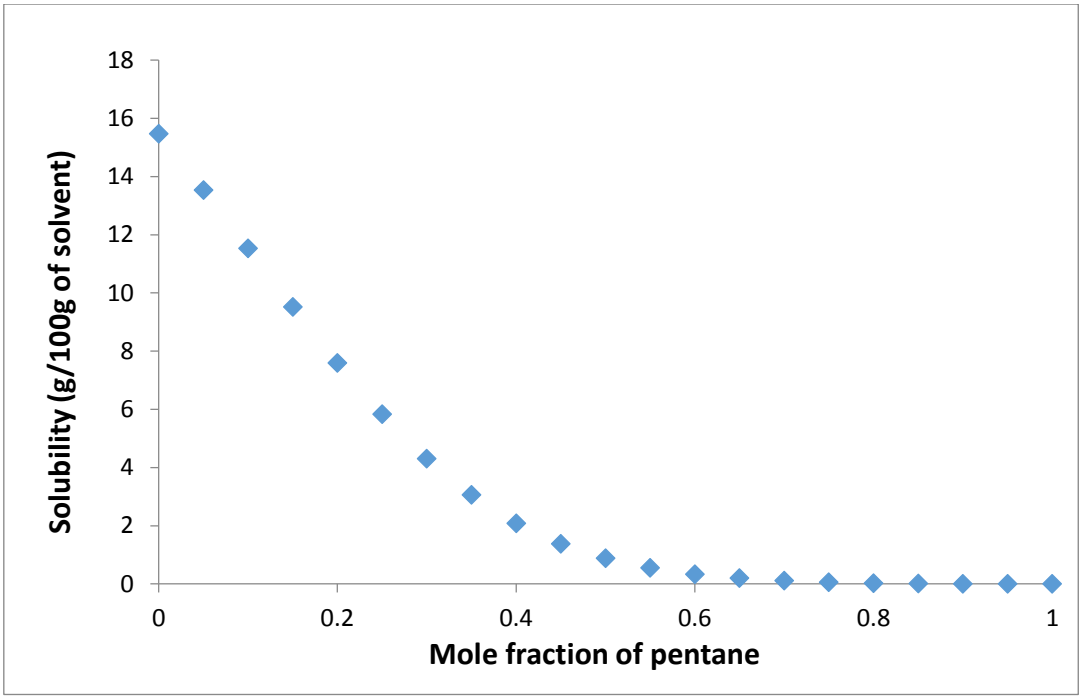


Figure 8-23 predicted solubility curve for paracetamol dioxane and pentane at 25°C using COSMOtherm

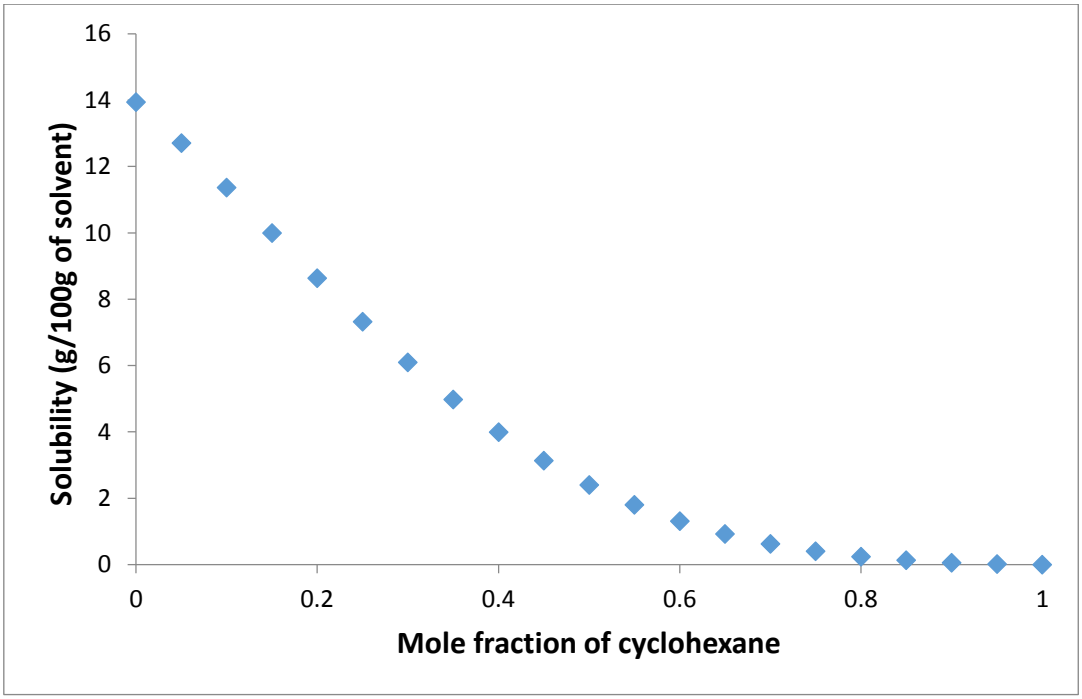


Figure 8-24 predicted solubility curve for paracetamol ethanol and cyclohexane at 25°C using COSMOtherm

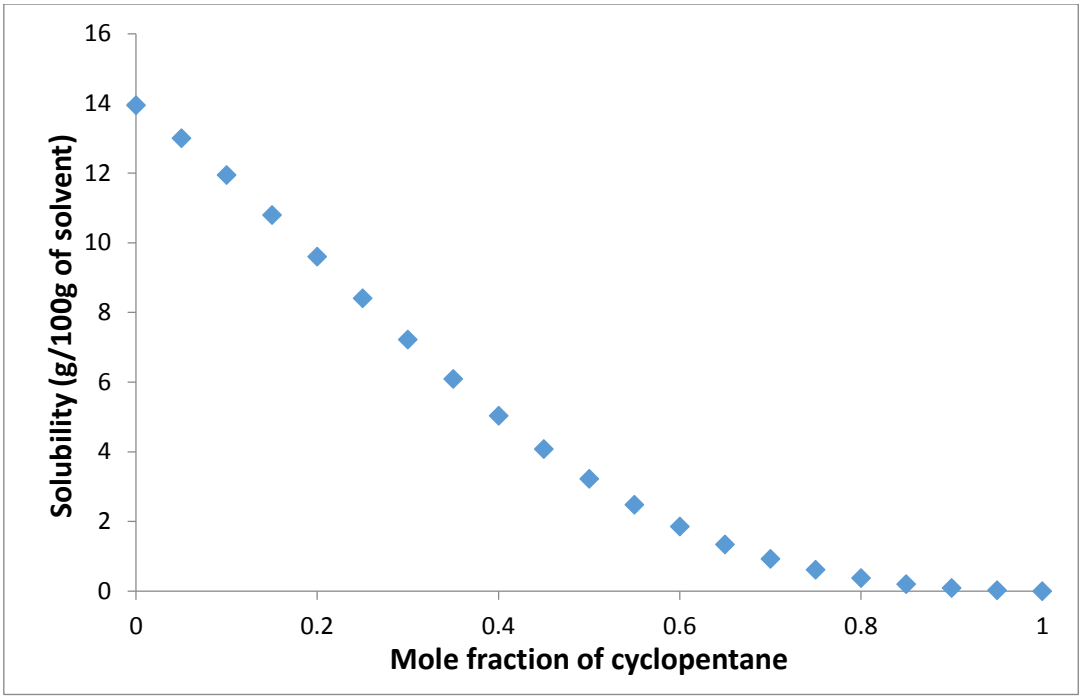


Figure 8-25 predicted solubility curve for paracetamol ethanol and cyclopentane at 25°C using COSMOtherm

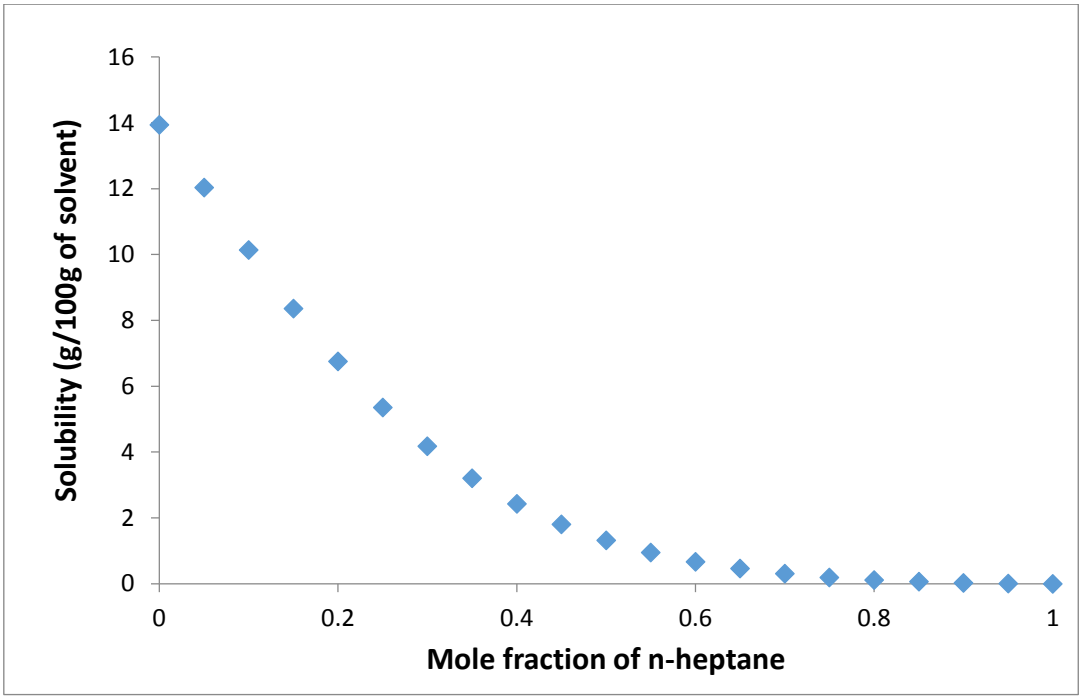


Figure 8-26 predicted solubility curve for paracetamol ethanol and n-heptane at 25°C using COSMOtherm

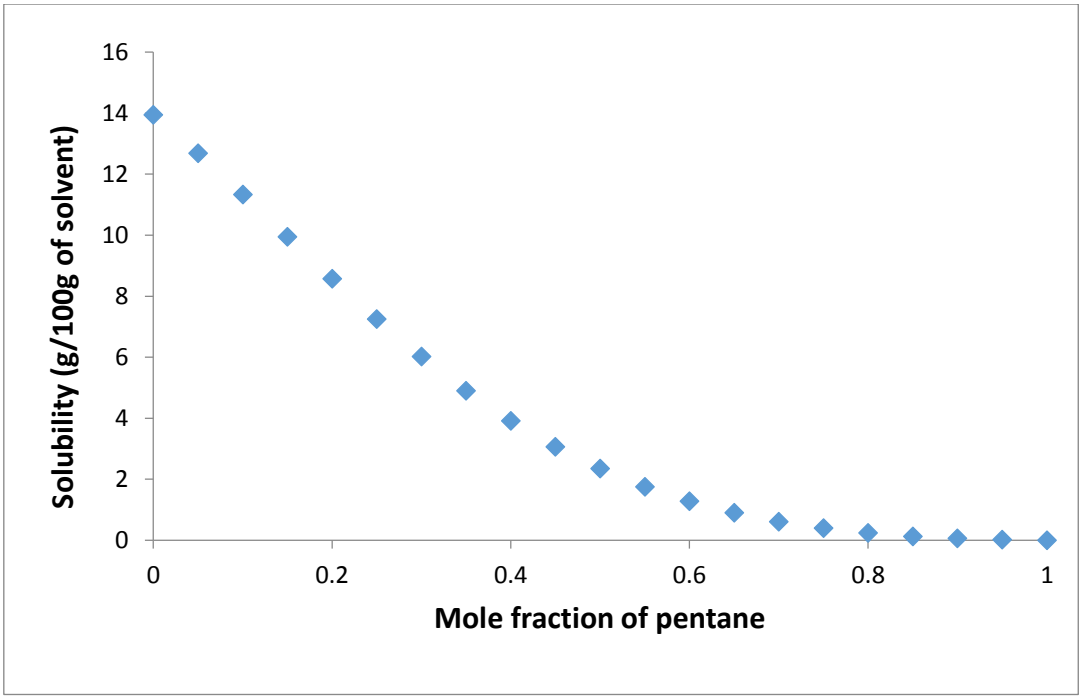


Figure 8-27 predicted solubility curve for paracetamol ethanol and pentane at 25°C using COSMOtherm

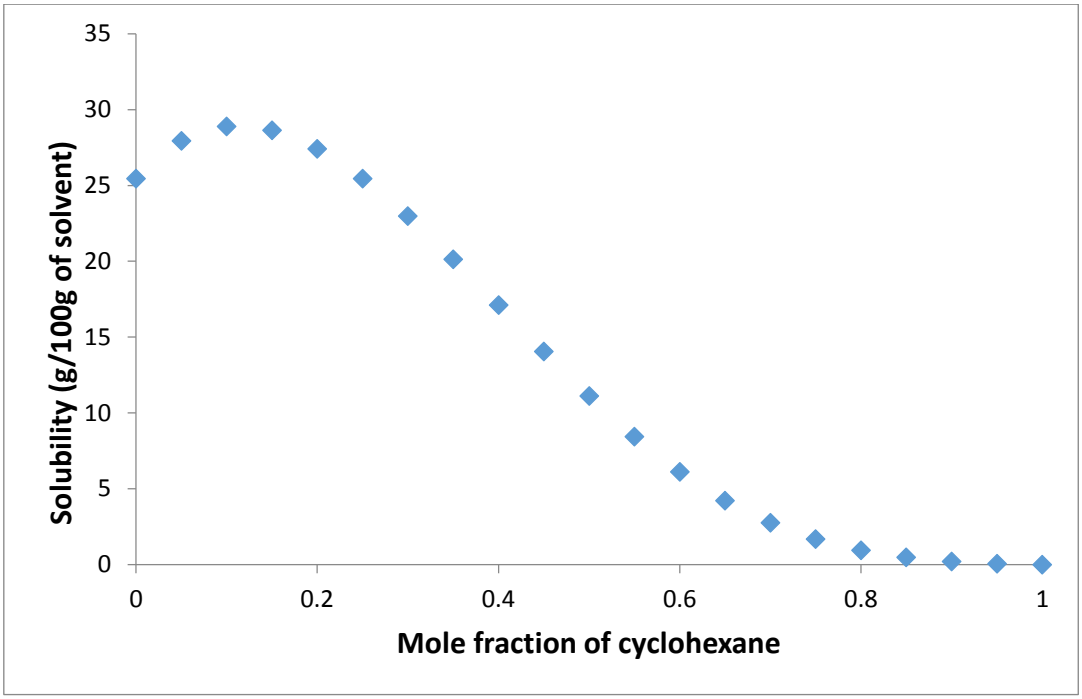


Figure 8-28 predicted solubility curve for paracetamol formic acid and cyclohexane at 25°C using COSMOtherm

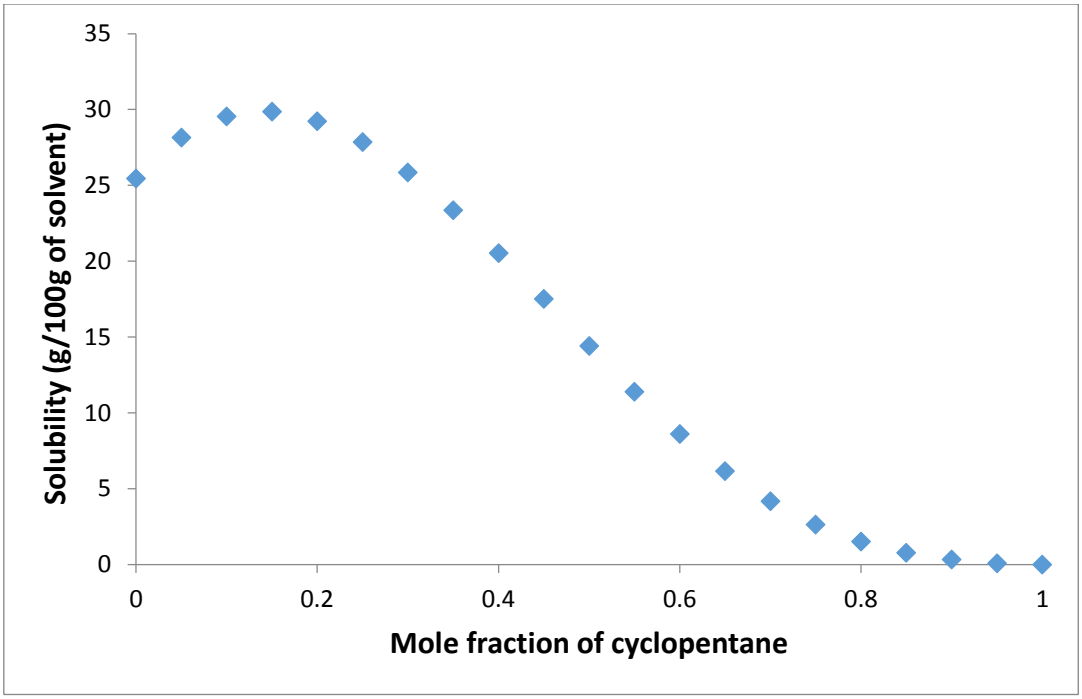


Figure 8-29 predicted solubility curve for paracetamol formic acid and cyclopentane at 25°C using COSMOtherm

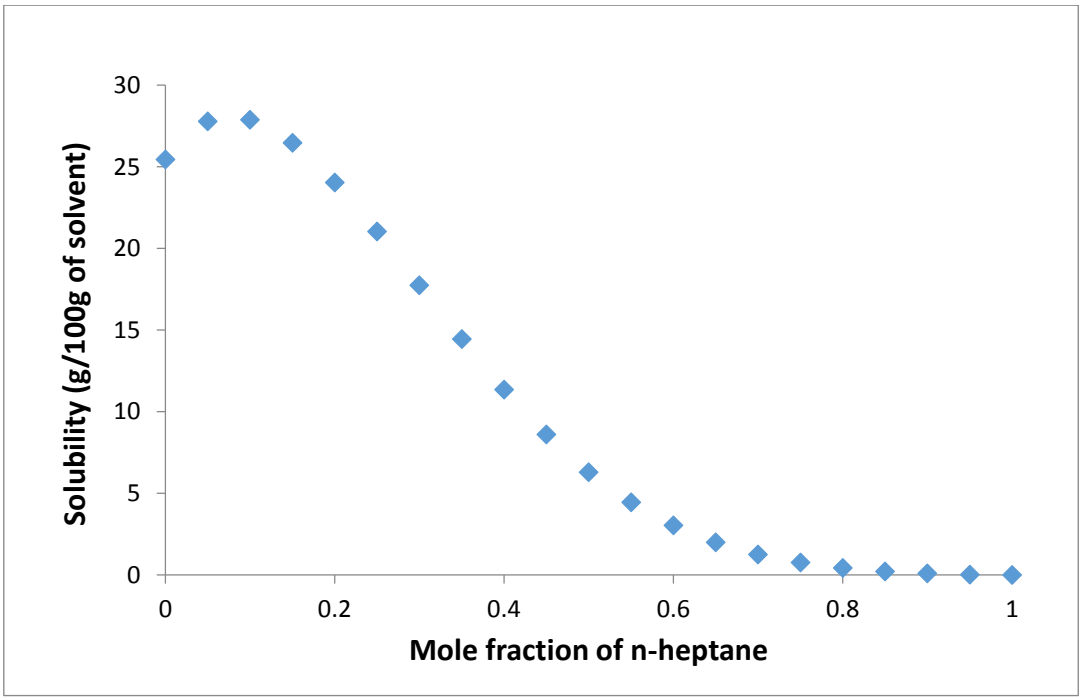


Figure 8-30 predicted solubility curve for paracetamol formic acid and n-heptane at 25°C using COSMOtherm

9 Appendix Three

Table 9-1 Scripts available for download

Name of Script	**Description of Script**
input_file_only	start COSMO*therm* in command line using only input file
1_solubility_curve	Produces data points for a solubility curve
2_linear_regression_model_doe	Produces nine points for solute/solvent combinations changing the enthalpy of fusion, temperature and melting temperature for linear regression
3_solvent_screen	Solvent screen for solutes
4_solute_binary_solvent	Produces data points for binary solvent system
5_ternary_phase_diagram	Produces data points for ternary phase diagram: one solute and two solvents
6_solid_liquid_extraction	produces data points two solutes one solvents
7_miscibility_solvent_anti_solvents	Miscibility calculated using number of moles
8_salt_solubility_curve	Produces data points for salt solubility curve
9_salt_solvent_screen	Solvent screens for salts
10_DHfus_Melttemp_change_solubility_automation	Produces data points by changing enthalpy of fusion and melting temperature across a range
11_liquid-liquid_equilibrium_ternary	Produces data points for a ternary solvent system
12_DHfus_Melttemp_extremes	Produces data points by changing enthalpy of fusion and melting temperature across a range
13_Job_queuer	Multiple jobs can be queued at one time for solubility curves
13_solubility_list	Produces COSMO*therm* data points specified in the job-queuer
14_new_entity_model_generator	Produces data points for one solute/solvent for linear regression
15_two_solute_two_solvents	Produces data points for two solutes and two solvents
16_miscibility_mole_fraction	Automated miscibility using mole fraction instead of number of moles
17_ternary_solvents_miscibility	Produces data points for ternary solvents at specific mole fraction

18_project_pure_solvents	Produces single solubility points for a specific temperature and solute/solvent combination
19_project_pure_solvents_cp_heat	Produces solubility points using heat capacity estimate
__init__	Informs the main scripts where to find functions and variables
functions	Script containing functions called by main scripts
lists	Contains lists of data required by main scripts
paths	Contains file addresses required by main scripts
3_wash solvent_selector	Produces solubility data for API and Impurities showing differences in solubility for crystallisation and wash solvent selection
3_wash solvent_summary	Produces tables for the selection of wash solvent
1_Table_Grapher	Produces solubility curve graphs
1_VantHoff_Grapher	Produces Van't Hoff solubility graphs
2_linear_regression_model_doe_tabulator	Produces an excel file for results of linear regression modelling of COSMO*therm*
2_linear_regression_model_doe_txt_file_creator	Produces a text file for results of linear regression modelling of COSMO*therm*
3_solvent_screen_ranker	Produces a graph for the solvent screen of a solute
4_solute_binary_solvent_writer	Produces a table for a solvent binary system with solute
5_ternary_reader_for_origin	Produces an excel sheet for use in ORIGIN
6_solid_liquid_extraction_tabler_grapher	Produces table and a graph for solute/solvent solid/liquid extraction
7_miscibility_reader	Produces a table of solvent miscibility
8_Salt_Solubility_Table_Grapher	Produces table and a graph of salt solubility curves
9_salt_solvent_screen_ranker	Produces a table and graph for salt solvent screens
10_Table_Grapher_for_DHfus_changer	Produces a table and a graph for solubility points that have different enthalpy of fusion and melting temperatures
11_LLE-equilibrium_ternary_reader_for_origin	Produces an excel sheet for use in ORIGIN

12_Table_Grapher_for_DHfus_range	Produces a table and a graph for solubility points that have different enthalpy of fusion and melting temperatures
13_grapher_and_error_assessment	Produces a table and graph for solubility points and compare with experimental solubility points
17_miscibility_ternary_reader	Produces miscibility table for three solvents
18_reader_pure_solvents	Produces a table and graph for solubility points and compare with experimental solubility points
dripfeed_analysis	Analysis of the drip-feed model
remove_by_solute_g_100g_error	solute-Fold model machine learning for error prediction using g/100g as units
remove_by_solute_g_100g_solubility_prediction_only	solute-Fold model machine learning for solubility prediction using g/100g as units
remove_by_solute_log_error	solute-Fold model machine learning for error prediction using log g/100g as units
remove_by_solute_log_solubility_prediction_only	solute-Fold model machine learning for solubility prediction using log g/100g as units
RF_kfoldcv_g_100g_error	k-Fold model machine learning for error prediction using g/100g as units
RF_kfoldcv_g_100g_solubility_prediction	k-Fold model machine learning for solubility prediction using g/100g as units
RF_kfoldcv_log_error	k-Fold model machine learning for error prediction using log g/100g as units
RF_kfoldcv_log_solubility_prediction	k-Fold model machine learning for solubility prediction using log g/100g as units
dripfeed_RF_loop	drip-feed model script

10 References

Abramov, Y.A. (2018) 'Rational Solvent Selection for Pharmaceutical Impurity Purge'. *Crystal Growth and Design,* 18 (2), pp. 1208-1214.

Acevedo, I.L., Pedrosa, G.C. and Katz, M. (1993) 'Dynamic viscosities of non-electrolyte mixtures'. *Phys. Chem. Liq., ,* 26 (2), pp. 99-106.

Aceves-Hernandez, J.M. *et al.* (2009) 'Indomethacin polymorphs: Experimental and conformational analysis'. *Journal of Pharmaceutical Sciences,* 98 (7), pp. 2448-2463.

Acree, W.E. (1991) 'Thermodynamic properties of organic compounds: enthalpy of fusion and melting point temperature compilation'. *Thermochimica Acta,* 189 (1), pp. 37-56.

Adjei, A. (1980) 'Solubility of Xanthine Derivatives in Polar and Nonpolar Solvents'. *University of Texas, Austin, thesis,* 1-209.

Adriano Antunes Souza, A. *et al.* (2010) 'Determination of the melting temperature, heat of fusion, and purity analysis of different samples of zidovudine (AZT) using DSC'. *Brazilian Journal of Pharmaceutical Sciences,* 46 (1), pp. 37-43.

Allan, J.R. *et al.* (1989) 'Thermal analysis studies on pyridine carboxylic acid complexes of zinc(II)'. *Thermochimica Acta,* 153 249-256.

Almeida, A.R.R.P., Oliveira, J.A.S.A. and Monte, M.J.S. (2015) 'Thermodynamic study of nicotinamide, N- methylnicotinamide and N, N- dimethylnicotinamide: Vapour pressures, phase diagrams, and hydrogen bonds'. *The Journal of Chemical Thermodynamics,* 82 108-115.

Almeida, A.R.R.P. *et al.* (2015) 'Thermodynamic properties of sublimation of the ortho and meta isomers of acetoxy and acetamido benzoic acids'. *The Journal of Chemical Thermodynamics,* 86 6-12.

Andrews, D.H., Lynn, G. and Johnston, J. (1926) 'The heat capacities and heat of crystallization of some isomeric aromatic compounds'. *Journal of the American Chemical Society,* 48 (5), pp. 1274-1287.

Arias, M.J., Moyano, J.R. and Ginés, J.M. (1998) 'Study by DSC and HSM of the oxazepam–PEG 6000 and oxazepam– D- mannitol systems: Application to the preparation of solid dispersions'. *Thermochimica Acta,* 321 (1), pp. 33-41.

Armstrong, N.A., James, K.C. and Wong, C.K. (1979) 'Inter- relationships between solubilities, distribution coefficients and melting points of some substituted benzoic and phenylacetic acids'. *The Journal of pharmacy and pharmacology,* 31 (9), pp. 627-631.

Awwad, A.M., Jabara, K.A. and Salman, M.A. (1988) 'Viscosities of ethylbenzene + and 1,3,5-trimethylbenzene + n-alkanes at 298.15 K'. *J. Pet. Res., ,* 7 (1), pp. 157-168.

Babinkov, A.G. *et al.* (1979) 'Thermodynamic properties of aclipic acid'. *Termodin. Org. Soedin.,* 8 28.

Bardavid, S. *et al.* (1996) *Excess molar volumes and excess viscosities for the n - hexane+dichloromethane+tetrahydrofuran ternary system at 25°C.*

Bardyshev, I.I. (1948) 'The viscosity of terpene hydrocarbons'. *Zh. Prikl. Khim., ,* 21 1019-1024.

Barone, G. *et al.* (1990) 'Enthalpies and entropies of sublimation, vaporization and fusion of nine polyhydric alcohols'. *Journal of the Chemical Society, Faraday Transactions,* 86 (1), pp. 75-79.

Barrio, M. *et al.* (2017) 'The Pressure-Temperature Phase Diagram of Metacetamol and Its Comparison to the Phase Diagram of Paracetamol'. *Journal of Pharmaceutical Sciences,* 106 (6), pp. 1538-1544.

Basavoju, S., Bostrom, D. and Velaga, S.P. (2008) 'Indomethacin--Saccharin Cocrystal: Design, Synthesis and Preliminary Pharmaceutical Characterization.(Author abstract)(Report)'. *Pharmaceutical Research,* 25 (3), pp. 530.

Ben Abacha, A. *et al.* (2015) 'Text mining for pharmacovigilance: Using machine learning for drug name recognition and drug–drug interaction extraction and classification'. *Journal of Biomedical Informatics,* 58 122-132.

Benazzouz, A. *et al.* (2014) 'Hansen approach versus COSMO-RS for predicting the solubility of an organic UV filter in cosmetic solvents'. *Colloids And Surfaces A-Physicochemical And Engineering Aspects, 2014 Sep 20, Vol.458, pp.101-109,* 458.

Berchiesi, G., Cingolani, A. and Leonesi, D. (1974) 'Thermodynamic properties of organic compounds'. *Journal of thermal analysis,* 6 (1), pp. 91-99.

Beret, S. and Prausnitz, J.M. (1975) 'Perturbed hard-chain theory: An equation of state for fluids containing small or large molecules'. *AIChE Journal,* 21 (6), pp. 1123-1132.

Bhardwaj, R.M. *et al.* (2015) 'A random forest model for predicting the crystallisability of organic molecules'.

BIOVIA, D.S. (2017) *Pipeline Pilot v 17.2.*

Blokhina, S.V. *et al.* (2015) 'Solution thermodynamics of pyrazinamide, isoniazid, and p-aminobenzoic acid in buffers and octanol'. *The Journal of Chemical Thermodynamics,* 91 396-403.

Boobier, S., Osbourn, A. and Mitchell, J.B.O. (2017) 'Can human experts predict solubility better than computers?'. *Journal of Cheminformatics,* 9 (1), pp.

Booss, H.J. and Hauschildt, K.R. (1972) 'Die Schmelzenthalpie des Benzils und 4-Nitrophenols'. *Fresenius' Zeitschrift für analytische Chemie,* 261 (1), pp. 32-32.

Booth, A. *et al.* (2010) 'Solid state and sub- cooled liquid vapour pressures of substituted dicarboxylic acids using Knudsen Effusion Mass Spectrometry (KEMS) and Differential Scanning Calorimetry'. *Atmospheric Chemistry and Physics,* 10 (10), pp. 4879.

Bothe, H. and Cammenga, H.K. (1980) 'Composition, Properties, Stability and Thermal Dehydration of Crystalline Caffeine Hydrate'. *Thermochim. Acta,* 40 29-39.

Bouillot, B. (2011) 'Approches Thermodynamiques pour la Prediction de la Solubilite de Molecules d'Interet Pharmaceutique'. *Universite de Toulouse thesis,* 1-272.

Bouillot, B. *et al.* (2017) 'Solubility of pharmaceuticals: A comparison between SciPharma, a PC-SAFT-based approach, and NRTL-SAC'. *European Physical Journal: Special Topics,* 226 (5), pp. 913-929.

Bouillot, B., Teychené, S. and Biscans, B. (2013) 'An Evaluation of COSMO-SAC Model and Its Evolutions for the Prediction of Drug-Like Molecule Solubility: Part 1'. *Industrial & Engineering Chemistry Research,* 52 (26), pp. 9276-9284.

Breiman, L. (2001) 'Random forests'. *MACH LEARN,* 45 (1), pp. 5-32.

Bret-Dibat, P. and Lichanot, A. (1989) 'Thermodynamic properties of positional isomers of disubstituted benzene in condensed phase'. *Thermochim. Acta,* 147 261-271.

Brittain, H.G. (2009) 'Vibrational spectroscopic studies of cocrystals and salts. 2. The benzylamine-benzoic acid system'. *Crystal Growth and Design,* 9 (8), pp. 3497-3503.

Brown, C.J. *et al.* (2018) 'Enabling precision manufacturing of active pharmaceutical ingredients : workflow for seeded cooling continuous crystallisation'.

Bruner, L. (1894) 'Heats of fusion of organic compounds'. *Ber. Dtsch. Chem. Ges.,* 27 2102-2107.

Bryant, M.J. *et al.* (2019) 'The CSD Drug Subset: The Changing Chemistry and Crystallography of Small Molecule Pharmaceuticals'. *Journal of Pharmaceutical Sciences,* 108 (5), pp. 1655-1662.

Bustamante, P., Peña, M.A. and Barra, J. (1998) 'Partial solubility parameters of piroxicam and niflumic acid'. *International Journal of Pharmaceutics,* 174 (1), pp. 141-150.

Bustamante, P. *et al.* (1998) 'Enthalpy– entropy compensation for the solubility of drugs in solvent mixtures: Paracetamol, acetanilide, and nalidixic acid in dioxane– water'. *Journal of Pharmaceutical Sciences,* 87 (12), pp. 1590-1596.

Bustamante, P., Romero, S. and Reillo, A. (1995) 'Thermodynamics of Paracetamol in Amphiprotic and Amphiprotic—aprotic Solvent Mixtures'. *Pharmacy and Pharmacology Communications,* 1 (11), pp. 505-507.

Campanella, L. *et al.* (2010) 'Solid-- liquid phase diagrams of binary mixturesAcetylsalicylic acid(1) + E(2) (E = salicylic acid, polyethylene glycol 4000, d-mannitol)'. *Journal of Thermal Analysis and Calorimetry,* 99 (3), pp. 887-892.

Campbell, A.N. and Campbell, A.J.R. (1941) 'The system naphthalene-p-nitrophenol: an experimental investigation of all the variables in an equation of the freezing point curve'. *Can. J. Res,* B19 73-79.

Cares-Pacheco, M.G. *et al.* (2014) 'Physicochemical characterization of d- mannitol polymorphs: The challenging surface energy determination by inverse gas chromatography in the infinite dilution region'. *International Journal of Pharmaceutics,* 475 (1-2), pp. 69-81.

Cesàro, A. (1980) 'Thermodynamic properties of caffeine crystal forms'. *The Journal of Physical Chemistry,* 84 (11), pp. 1345-1346.

Chen, C.C. (1993) 'A segment- based local composition model for the gibbs energy of polymer solutions'. *Fluid Phase Equilibria,* 83 301-312.

Chen, C.C. and Song, Y.H. (2004) 'Solubility modeling with a nonrandom two-liquid segment activity coefficient model'. *Industrial & Engineering Chemistry Research, 2004, Vol.43(26), pp.8354-8362,* 43 (26), pp.

Chen, H. *et al.* (2018) 'The rise of deep learning in drug discovery'. *Drug Discovery Today,* 23 (6), pp. 1241-1250.

Chen, Y.P., Tang, M. and Kuo, J.C. (2005) 'Solid- liquid equilibria for binary mixtures of N-phenylacetamide with 4- aminoacetophenone, 3- hydroxyacetophenone and 4-hydroxyacetophenone'. *Fluid Phase Equilibria,* 232 (1-2), pp. 182-188.

Cheng, Y.S. *et al.* (2010) 'Workflow for managing impurities in an integrated crystallization process'. *AIChE Journal,* 56 (3), pp. 633-649.

Cilurzo, F. *et al.* (2010) 'Effect of drug chirality on the skin permeability of ibuprofen'. *International Journal of Pharmaceutics,* 386 (1), pp. 71-76.

Cingolani, A. and Berchiesi, G. (1974) *Thermodynamic properties of organic compounds - Note I. A DSC study of phase transitions in aliphatic dicarboxylic acids.*

Connett, J.E. (1979) *CERTIFIED REFERENCE MATERIALS FOR THE CALIBRATION OF THERMAL ANALYSIS APPARATUS.*

Danilin, V.N. *et al.* (2001) *Phase diagrams of binary systems formed by saturated fatty acids.*

David, D.J. (1964) 'Determination of Specific Heat and Heat of Fusion by Differential Thermal Analysis: Study of Theory and Operating Parameters'. *Analytical Chemistry,* 36 (11), pp. 2162-2166.

de la Iglesia, D. *et al.* (2014) 'A Machine Learning Approach to Identify Clinical Trials Involving Nanodrugs and Nanodevices from ClinicalTrials.gov.(Research Article)'. *PLoS ONE,* 9 (10), pp.

Dean, J.A. (1992) *Landolt-Boernstein.* Springer,Berlin.

Della Gatta, G. and Ferro, D. (1987) 'Enthalpies of fusion and solid-to-solid transition, entropies of fusion for urea and twelve alkylureas'. *Thermochim. Acta,* 122 143-152.

Detherm (2016) '*Detherm'*.

Diedrichs, A. (2005) 'Optimization of a dynamic differential scanning calorimeter for the experimental determination of heat capacity'. *Masters thesis*.

Domalski, E. and Hearing, E. (1996) '*Heat capacities and entropies of organic compounds in the condensed phase, vol 3'. J. Phys. Chem. Ref. Data.*

Domańska, U. *et al.* (2009) 'pKa and Solubility of Drugs in Water, Ethanol, and 1-Octanol'. *The Journal of Physical Chemistry B,* 113 (26), pp. 8941-8947.

Domańska, U. *et al.* (2010) 'Solubility and pKa of select pharmaceuticals in water, ethanol, and 1-octanol'. *The Journal of Chemical Thermodynamics,* 42 (12), pp. 1465-1472.

Domingos, P. (2012) 'A few useful things to know about machine learning'. *Communications of the ACM,* 55 (10), pp. 78-87.

Dong, J.-X. *et al.* (2007) 'The standard molar enthalpy of formation, molar heat capacities, and thermal stability of anhydrous caffeine'. *The Journal of Chemical Thermodynamics,* 39 (1), pp. 108-114.

Donnelly, J.R. *et al.* (1990) 'Purity and heat of fusion data for environmental standards as determined by differential scanning calorimetry'. *Thermochimica Acta,* 167 (2), pp. 155-187.

Draucker, L.C. *et al.* 'Experimental determination and model prediction of solid solubility of multifunctional compounds in pure and mixed nonelectrolyte solvents'. *Industrial & Engineering Chemistry Research, 2007, Vol.46(7), pp.2198-2204,* 46 (7), pp.

Drebushchak, V.A. *et al.* (2006) 'Thermoanalytical investigation of drug- excipient interaction:Part I. Piroxicam, cellulose and chitosan as starting materials'. *Journal of Thermal Analysis and Calorimetry,* 84 (3), pp. 643-649.

DrugBank (2018) '*DrugBank'*.

Dufal, S. *et al.* (2014) 'Prediction of thermodynamic properties and phase behavior of fluids and mixtures with the SAFT- γ mie group-contribution equation of state'. *Journal of Chemical and Engineering Data,* 59 (10), pp. 3272-3288.

Eckert, F. (2015) *COSMOtherm Reference Manual.* Leverkusen.

Eckert, F. and Klamt, A. (2002) 'Fast solvent screening via quantium chemistry: COSMO-RS approach'. 48 369.

El Moussaoui, A., Chauvet, A. and Masse, J. (1993) *Study of the solid state reaction of the binary system Nordazepam III/nicotinic acid.*

Espitalier, F., Biscans, B. and Laguérie, C. (1995) 'Physicochemical Data on Ketoprofen in Solutions'. *Journal of Chemical and Engineering Data,* 40 (6), pp. 1222-1224.

Eykman, J.F. (1889) 'Zur kryoskopischen Molekulargewichtsbestimmung'. *Z. Physik. Chem.,* 4 497-519.

Fall, D.J. and Luks, K.D. (1965) *The Coal Tar Data Book.* The Coal Tar Research Association.

Ferloni, P. and Gatta, G.D. (1995) 'Heat capacities of urea, N-methylurea, N-ethylurea, N-(n)propylurea, and N-(n)butylurea in the range 200 to 360 K'. *Thermochimica Acta,* 266 (C), pp. 203-212.

Ferro, D. *et al.* (1987) 'Vapor pressures and sublimation enthalpies of urea and some of its derivatives'. *J. Chem. Thermodyn.,* 19 915-923.

Fiebig, A., Jones, M.J. and Ulrich, J. (2007) 'Predicting the Effect of Impurity Adsorption on Crystal Morphology'. *Crystal Growth & Design,* 7 (9), pp. 1623-1627.

Fischer, J. and Weiss, A. (1986) 'Transport Properties of Liquids. V. Self Diffusion, Viscosity, and Mass Density of Ellipsoidal Shaped Molecules in the Pure Liquid Phase'. *Berichte der Bunsengesellschaft für physikalische Chemie,* 90 (10), pp. 896-905.

Franceschi, E. *et al.* (2008) 'Phase behavior and process parameters effects on the characteristics of precipitated theophylline using carbon dioxide as antisolvent'. *Journal of Supercritical Fluids,* 44 (1), pp. 8-20.

Fredenslund, A., Jones, R.L. and Prausnitz, J.M. (1975) 'GROUP-CONTRIBUTION ESTIMATION OF ACTIVITY COEFFICIENTS IN NONIDEAL LIQUID MIXTURES'. *AIChE Journal, November 1975, Vol.21(6), pp.1086-1099,* 21 (6), pp.

Gartenmeister, R. (1890) 'The viscosity of liquid hydrocarbon compounds and its relation to the chemical constitution'. *Z. Phys. Chem. ,* 524-551.

Giordano, F. *et al.* (1999) 'Physical properties of parabens and their mixtures: Solubility in water, thermal behavior, and crystal structures'. *Journal of Pharmaceutical Sciences,* 88 (11), pp. 1210-1216.

Gmehling, J. (1998) 'Present status of group-contribution methods for the synthesis and design of chemical processes'. *Fluid Phase Equilibria,* 144 (1-2), pp. 37-47.

Gmehling, J. (2018) '*Dortmund Data Bank'*.

Gombás, Á. *et al.* (2003) 'Study of thermal behaviour of sugar alcohols'. *Journal of Thermal Analysis and Calorimetry,* 73 (2), pp. 615-621.

Goncalves, E.M. and Da Piedade, M.E.M. (2012) 'Solubility of nicotinic acid in water, ethanol, acetone, diethyl ether, acetonitrile, and dimethyl sulfoxide.(Report)'. *The Journal of Chemical Thermodynamics,* 47 362.

Gonçalves, E.M., Rego, T.S. and Minas da Piedade, M.E. (2011) 'Thermochemistry of aqueous pyridine-3-carboxylic acid (nicotinic acid)'. *The Journal of Chemical Thermodynamics,* 43 (6), pp. 974-979.

Good, D.J. and Naír, R.H. (2009) 'Solubility advantage of pharmaceutical cocrystals'. *Crystal Growth and Design,* 9 (5), pp. 2252-2264.

Gorniak, A. *et al.* (2011) 'Phase diagram and dissolution studies of the fenofibrate-- acetylsalicylic acid system.(Report)'. *Journal of Thermal Analysis and Calorimetry,* 104 (3), pp. 1195.

Gracin, S., Brinck, T. and Rasmuson, Å.C. (2002) 'Prediction of solubility of solid organic compounds in solvents by UNIFAC'. *Industrial and Engineering Chemistry Research,* 41 (20), pp. 5114-5124.

Gracin, S. and Rasmuson, Å.C. (2002) 'Solubility of phenylacetic acid, p- hydroxyphenylacetic acid, p-aminophenylacetic acid, p- hydroxybenzoic acid, and ibuprofen in pure solvents'. *Journal of Chemical and Engineering Data,* 47 (6), pp. 1379-1383.

Grady, L.T. *et al.* (1973) 'Drug Purity Profiles'. *Journal of Pharmaceutical Sciences,* 62 (3), pp. 456-464.

Granberg, R.A. and Rasmuson, A.C. (1999) 'Solubility of paracetamol in pure solvents'. *Journal of Chemical and Engineering Data,* 44 (6), pp. 1391-1395.

Granberg, R.A. and Rasmuson, Å.C. (2000) 'Solubility of paracetamol in binary and ternary mixtures of water + acetone + toluene'. *Journal of Chemical and Engineering Data,* 45 (3), pp. 478-483.

Grandelli, H.E. *et al.* (2012) 'Melting point depression of Piroxicam in carbon dioxide+co-solvent mixtures and inclusion complex formation with β-cyclodextrin'. *The Journal of Supercritical Fluids,* 71 19-25.

Grant, D.J.W. (1990) *Solubility behavior of organic compounds.* New York: New York : Wiley.

Graubner, G. (2008) 'Experimental determination and prediction of solubility of drugs for the crystallization of solutions'. *Universität Oldenburg thesis*.

Grigor'ev, A.A. *et al.* (1994) 'Thermochemical investigation of enthalpies of fusion and solution of adducts'. *Zh. Obshch. Khim.,* 64 (4), pp. 564-570.

Gross, J. and Sadowski, G. (2001) 'Perturbed-chain SAFT: An equation of state based on a perturbation theory for chain molecules'. *Industrial & Engineering Chemistry Research, 2001, Vol.40(4), pp.1244-1260,* 40 (4), pp.

Guo, K. *et al.* (2010) 'Co-crystallization in the caffeine/ maleic acid system: Lessons from phase equilibria'. *Crystal Growth and Design,* 10 (1), pp. 268-273.

Gupta, P. *et al.* (2012) 'Phase equilibria and molecular interaction studies on (naphthols + vanillin) systems'. *The Journal of Chemical Thermodynamics,* 48 291-299.

Gupta, R.K. and Singh, R.A. (2004) 'Thermochemical and microstructural studies on binary organic eutectics and complexes'. *J. Cryst. Growth,* 267 340-347.

Hahnenkamp, I. (2008) 'Experimental and theoretical studies on the solubility of drugs in solvents'. *Universität Oldenburg thesis,* 1-151.

Hala, S. *et al.* (1979) 'Temperature dependence of heat of vaporization of saturated hydrocarbons C5-C8. Experimental data and an estimation method'. *Collect. Czech. Chem. Commun., ,* 44 (3), pp. 637-651.

Hamdi, N. *et al.* (2004) 'Solvates of indomethacin'. *Journal of Thermal Analysis and Calorimetry,* 76 (3), pp. 985-1001.

Hammer, B. (2014) '*Neural Networks*'.

Hanaya, M. *et al.* (2002) 'Low- temperature adiabatic calorimetry of salol and benzophenone and microscopic observation of their crystallization: Finding of homogeneous- nucleation- based crystallization'. *Journal of Chemical Thermodynamics,* 34 (8), pp. 1173-1193.

Hancock, B.C. and Parks, M. (2000) 'What is the true solubility advantage for amorphous pharmaceuticals?'. *Pharmaceutical research,* 17 (4), pp. 397-404.

Heath, E.A., Singh, P. and Ebisuzaki, Y. (1992) 'Structure of p - hydroxybenzoic acid and p - hydroxybenzoic acid- acetone complex (2/1'. *Acta Crystallographica Section C,* 48 (11), pp. 1960-1965.

Hendriksen, B.A. and Grant, D.J.W. (1995) 'The effect of structurally related substances on the nucleation kinetics of paracetamol (acetaminophen)'. *Journal of Crystal Growth,* 156 (3), pp. 252-260.

Hendriksen, B.A. *et al.* (1998) *Crystallization of Paracetamol (Acetaminophen) in the Presence of Structurally Related Substances.*

Hirotaka Ikeda, K.C., Atushi Kanou and Noriaki Hirayama (2005) 'Prediction of solubility of Drugs by Conductor-Like Screening Model for Real Solvents'. *Chem. Pharm. Bull.,* 53 (2), pp. 253-255.

Hong, J. *et al.* (2010) 'Solid- liquid- gas equilibrium of the ternaries ibuprofen + myristic acid + CO 2 and ibuprofen + tripalmitin + CO 2'. *Journal of Chemical and Engineering Data,* 55 (1), pp. 297-302.

Hopfe, D. (1990) 'Thermophysical data of Pure Substances'. *Data Compilation of FIZ CHEMIE Germany,* 1.

Hrynakowski, K. and Smoczkiewiczowa, A. (1937) 'Thermal analysis. II. Measurement of the fusion heat of solid substances'. *Rocz. Chem.,* 17 165-168.

Hulse, W.L., Grimsey, I.M. and De Matas, M. (2008) 'The impact of low-level inorganic impurities on key physicochemical properties of paracetamol'. *International Journal of Pharmaceutics,* 349 (1), pp. 61-65.

ICH (2015) 'Q3D Elemental Impurities,'.

Irany, E.P. (1943) 'The Viscosity Function. IV. Non-ideal Systems'. *Journal of the American Chemical Society,* 65 (7), pp. 1392-1398.

Jain, A. and Yalkowsky, S.H. (2006) 'Estimation of Melting Points of Organic Compounds-II'. *Journal of Pharmaceutical Sciences,* 95 (12), pp. 2562-2618.

Jain, A., Yang, G. and Yalkowsky, S.H. (2004) '*Estimation of melting points of organic compounds*'.

Jamróz, M.E. *et al.* (1998) 'The urea– phenol(s) systems 1 Paper presented at the International Conference on Applied Physical Chemistry, Warsaw, 13–15 November 1996. 1'. *Fluid Phase Equilibria,* 152 (2), pp. 307-326.

Jia, Q. *et al.* (2007) 'Solid- Liquid Equilibria of Benzoic Acid Derivatives in 1- Octanol * * Supported by the National Natural Science Foundation of China (No.20676101) and the Natural Science Foundation of Tianjin University of Science & Technology (No.20050207)'. *Chinese Journal of Chemical Engineering,* 15 (5), pp. 710-714.

Joback, K.G. and Reid, R.C. (1987) 'Estimation of Pure-component properties from group contributions'. *Chemical engineering Comm,* 57 233-243.

Johnston A, J.B., Kennedy AR, Florence AJ (2008) 'Targeted crystallisation of novel carbamazepine solvates based on a retrospective Random Forest classification.'. *CrystEngComm,* 10 (1), pp. 23-35.

Jorgensen, W.L. and Duffy, E.M. (2002) 'Prediction of drug solubility from structure'. *Advanced Drug Delivery Reviews,* 54 (3), pp. 355-366.

Joseph, A., Bernardes, C.E.S. and Minas Da Piedade, M.E. (2012) 'Heat capacity and thermodynamics of solid and liquid pyridine- 3- carboxylic acid (nicotinic acid) over the temperature range 296 K to 531 K'. *The Journal of Chemical Thermodynamics,* 55 23-28.

Jozwiakowski, M.J. *et al.* (1996) 'Solubility behavior of lamivudine crystal forms in recrystallization solvents'. *Journal of Pharmaceutical Sciences,* 85 (2), pp. 193-199.

Kabo, G.Y. *et al.* (1990) 'Thermochemistry of alkyl derivatives of carbamide'. *Izv. Akad. Nauk SSSR Ser. Khim,* 750-755.

Kan, A.T. (1996) 'UNIFAC prediction of aqueous and nonaqueous solubilities'. *Environmental Science & Technology,* 30 (4), pp. 1369-1377.

Kant, S., Rai, U.S. and Rai, R.N. (2012) 'Thermal and physico- chemical studies on binary organic eutectic systems 4- Aminoacetophenone with benzoin and 4- nitrophenol'. *Journal of Thermal Analysis and Calorimetry,* 110 (2), pp. 551-557.

Katritzky, A.R. *et al.* (1998) 'QSPR studies on vapor pressure, aqueous solubility, and the prediction of water-air partition coefficients'. *Journal of Chemical Information and Computer Sciences,* 38 (4), pp. 720-725.

Keerrthega, M.C. and Thenmozhi, D. (2016) 'Identifying Disease -Treatment Relations Using Machine Learning Approach'. *Procedia Computer Science,* 87 306-315.

Kennedy, J.H. and Carr, P.W. (1973) 'The effect of inert solids on the differential scanning calorimetric behavior of benzoic acid'. *Thermochimica Acta,* 7 (4), pp. 325-329.

Khattab, F.I. (1983) 'Thermal analysis of pharmaceutical compounds. V. The use of differential scanning calorimetry in the analysis of certain pharmaceuticals'. *Thermochimica Acta,* 61 (3), pp. 253-268.

Khetarpal, S.C., Lal, K. and Bhatnagar, H.L. (1980) 'Thermophysical data'. *Indian J. Chem. Sect. A,* 19 516-519.

Kikic, I. *et al.* (2010) 'Solubility estimation of drugs in ternary systems of interest for the antisolvent precipitation processes'. *The Journal of Supercritical Fluids,* 55 (2), pp. 616-622.

Kirklin, D.R. (2000) 'Enthalpy of combustion of acetylsalicylic acid'. *The Journal of Chemical Thermodynamics,* 32 (6), pp. 701-709.

Kishi, H. and Hashimoto, Y. (1989) 'Evaluation of the procedures for the measurement of water solubility and n-octanol/water partition coefficient of chemicals results of a ring test in Japan'. *Chemosphere,* 18 (9), pp. 1749-1759.

Klamt, A. (2012) '*Comment on "comparison of the a priori COSMO- RS models and group contribution methods: Original UNIFAC, modified UNIFAC(Do), and modified UNIFAC(Do) consortium"*'.

Klamt, A. (2015) '*COSMOlogic*'.

Klamt, A. and Eckert, F. (2000) 'COSMO- RS: a novel and efficient method for the a priori prediction of thermophysical data of liquids'. *Fluid Phase Equilibria, 2000 Jun 28, Vol.172(1), pp.43-72,* 172 (1), pp.

Kleineberg, H. (2009) 'Experimental and theoretical determination of the solubility of drugs'. *Universdität Oldenburg Thesis*.

Klous, M.G. *et al.* (2005) 'Pharmaceutical heroin for inhalation: Thermal analysis and recovery experiments after volatilisation'. *Journal of Pharmaceutical and Biomedical Analysis,* 39 (5), pp. 944-950.

Klímová, K. and Leitner, J. (2012) 'DSC study and phase diagrams calculation of binary systems of paracetamol'. *Thermochimica Acta,* 550 59-64.

Kohn, W., Meir, Y. and Makarov, D.E. (1998) 'van der Waals Energies in Density Functional Theory'. *Physical Review Letters,* 80 (19), pp. 4153-4156.

Kokkonen, P. and Nissema, A. (1979) 'The thermodynamic properties of binary and ternary systems. Excess volumes and logarithmic Excess viscosities of the system triethylamine + mesitylene'. *Finn. Chem. Lett., ,* 3 69-71.

Kozyro, A.A., Dalidovich, S.V. and Krasulin, A.P. (1986) *SPECIFIC HEAT, HEAT OF FUSION, AND THERMODYNAMIC PROPERTIES OF UREA.*

Krogh, A. (2008) 'What are artificial neural networks?'. *Nature biotechnology,* 26 (2), pp. 195-197.

Lahav, M. and Leiserowitz, L. (2001) *The Effect of Solvent on Crystal Growth and Morphology.*

Lakshmana Prabu, S. and Timmakondu, S. (2010) *Impurities and its importance in pharmacy.*

Lazerges, M. *et al.* (2010) 'Thermodynamic studies of mixtures for topical anesthesia: Lidocaine- salol binary phase diagram'. *Thermochimica Acta,* 497 (1 2), pp. 124.

Leach, A.R. (1996) *Molecular Modelling: Principles and Applications.* Addison Wesley Longman Limited.

Lebedeva, N.D. (1964) 'Heats of combustion of monocarboxylic acids'. *Russ. J. Phys. Chem.(Engl. Transl.),* 38 1435-1437.

Lebedeva, N.D. *et al.* (1976) 'Enthalpies of formation of polysubstituted derivatives of benzene'. *Termodin. Org. Soedin.,* 12-29.
Lebedeva, N.D., Ryadnenko, V.L. and Kuznetsova, I.N. (1971) 'Heats of combustion and enthalpies of formation of certain aromatic nitro-derivatives'. *Russ. J. Phys. Chem. (Engl. Transl.),* 45 549.
Legendre, B. and Feutelais, Y. (2004) 'Polymorphic and Thermodynamic Study of Indomethacin'. *Journal of Thermal Analysis and Calorimetry,* 76 (1), pp. 255-264.
Lerdkanchanaporn, S., Dollimore, D. and Evans, S.J. (2001) 'Phase diagram for the mixtures of ibuprofen and stearic acid'. *Thermochimica Acta,* 367 (368), pp. 1-8.
Li, J., Zhou, J.K. and Huang, S.X. (1991) 'An investigation into the use of the eutectic mixture sodium acetate trihydrate - tartaric acid for latent heat storage'. *Thermochim. Acta,* 188 (1), pp. 17-23.
Li, S., Varadarajan, G.S. and Hartland, S. (1991) 'Solubilities of theobromine and caffeine in supercritical carbon dioxide: correlation with density-based models'. *Fluid Phase Equilibria,* 68 (C), pp. 263-280.
Li, Y. *et al.* (2014) 'Solubilities of adipic acid and succinic acid in a glutaric acid + acetone or n-butanol mixture'. *Journal of Chemical and Engineering Data,* 59 (12), pp. 4062-4069.
Li, Z.J. *et al.* (1999) 'Characterization of racemic species of chiral drugs using thermal analysis, thermodynamic calculation, and structural studies'. *Journal of Pharmaceutical Sciences,* 88 (3), pp. 337-346.
Lim, J. *et al.* (2013) 'Solubility of salicylic acid in pure alcohols at different temperatures'. *The Journal of Chemical Thermodynamics,* 57 295-300.
Lin, H.M. and Nash, R.A. (1993) 'An experimental method for determining the hildebrand solubility parameter of organic nonelectrolytes'. *Journal of Pharmaceutical Sciences,* 82 (10), pp. 1018-1026.
Lin, S.T. and Sandler, S.I. (2002) 'A priori phase equilibrium prediction from a segment contribution solvation model'. *Industrial & Engineering Chemistry Research, 2002, Vol.41(5), pp.899-913,* 41 (5), pp.
Lipinski, C.A. *et al.* (2001) 'Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings1PII of original article: S0169-409X(96)00423-1. The article was originally published in Advanced Drug Delivery Reviews 23 (1997) 3–25.1'. *Advanced Drug Delivery Reviews,* 46 (1), pp. 3-26.
Liu, C. *et al.* (2013) 'Facile method for the prediction of anhydrate/ hydrate transformation point'. *Industrial and Engineering Chemistry Research,* 52 (46), pp. 16506-16512.
Loschen, C. and Klamt, A. (2012) 'COSMOquick: A Novel Interface for Fast sigma-Profile Composition and Its Application to COSMO- RS Solvent Screening Using Multiple Reference Solvents'. *Industrial & Engineering Chemistry Research, 2012, Vol.51(43), pp.14303-14308,* 51 (43), pp.
Luther, H. and Waechter, G. (1949) 'Preparation and physical measured values of alkyl-substituted naphthalene'. *Chem. Ber., ,* 82 (2), pp. 161-176.
Luts, J. *et al.* (2010) 'A tutorial on support vector machine-based methods for classification problems in chemometrics'. *Analytica Chimica Acta,* 665 (2), pp. 129-145.
Lutskii, A.E., Obukhova, E.M. and Sidorov, I.A. (1958) 'Properties of binary mixtures of organic compounds as a function of association and concentration'. *J. Gen. Chem. USSR,* 28 (9), pp. 2423-2432.
Lux, A. and Stockhausen, M. (1993) 'A Dielectric Relaxation Study of Some Liquid Dihydric Alcohols and Their Mixtures with Water'. *Physics and Chemistry of Liquids,* 26 (1), pp. 67-83.
Lvova, M.S., Garber, N.I. and Kozlov, E.I. (1988) 'Use of differential scanning calorimetry to analyse the quality of certain vitamin drugs'. *Journal of thermal analysis,* 33 (4), pp. 1231-1234.
Magalhães, A. (2014) 'Gaussian-Type Orbitals versus Slater-Type Orbitals: A Comparison'. *Journal of Chemical Education,* 91 (12), pp. 2124.
Magoń, A. *et al.* (2014) 'Heat capacity and transition behavior of sucrose by standard, fast scanning and temperature- modulated calorimetry'. *Thermochimica Acta,* 589 183-196.

Majer, V. and Svoboda, V. (1985) 'Thermophysical data'. 1.
Man, Y.B.C. and Tan, C.P. (2002) 'Comparative differential scanning calorimetric analysis of vegetable oils: II. Effects of cooling rate variation'. *Phytochemical Analysis,* 13 (3), pp. 142-151.
Manic, M.S. *et al.* (2012) 'Solubility of high-value compounds in ethyl lactate: Measurements and modeling'. *J. Chem. Thermodyn.,* 48 93-100.
Manin, A.N., Voronin, A.P. and Perlovich, G.L. (2014) 'Acetamidobenzoic acid isomers: Studying sublimation and fusion processes and their relation with crystal structures'. *Thermochimica Acta,* 583 72-77.
Mantheni, D.R. *et al.* (2012) 'Solid state studies of drugs and chemicals by dielectric and calorimetric analysis.(Report)'. *Journal of Thermal Analysis and Calorimetry,* 108 (1), pp. 227.
Manzo, R.H. and Ahumada, A.A. (1990) 'Effects of solvent medium on solubility. V: Enthalpic and entropic contributions to the free energy changes of Di- substituted benzene derivatives in ethanol: Water and ethanol: Cyclohexane mixtures'. *Journal of Pharmaceutical Sciences,* 79 (12), pp. 1109-1115.
Mao, Z. *et al.* (2009) 'Measurement and Correlation of Solubilities of Adipic Acid in Different Solvents'. *Chinese Journal of Chemical Engineering,* 17 (3), pp. 473-477.
Marti, E. (1988) 'Applied chemical thermodynamics and kinetics on pharmaceutical compounds'. *Journal of Thermal Analysis and Calorimetry,* 33 (1), pp. 37-45.
Marti, E.E. (1972) 'Purity determination by differential scanning calorimetry'. *Thermochim. Acta,* 5 173-220.
Martin, A., Wu, P.L. and Velasquez, T. (1985) 'Extended hildebrand solubility approach: Sulfonamides in binary and ternary solvents'. *Journal of Pharmaceutical Sciences,* 74 (3), pp. 277-282.
Martinez, J.C. *et al.* (2018) 'Enhanced Quality Control in Pharmaceutical Applications by Combining Raman Spectroscopy and Machine Learning Techniques'. *International Journal of Thermophysics,* 39 (6), pp. 1-13.
Martínez, F. and Gómez, A. (2001) 'Thermodynamic study of the solubility of some sulfonamides in octanol, water, and the mutually saturated solvents'. *Journal of Solution Chemistry,* 30 (10), pp. 909-923.
Martínez, F. and Gómez, A. (2002) 'Estimation of the Solubility of Sulfonamides in Aqueous Media from Partition Coefficients and Entropies of Fusion'. *Physics and Chemistry of Liquids,* 40 (4), pp. 411-420.
Martínez, F., Ávila, C.M. and Gómez, A. (2003) 'Thermodynamic Study of the Solubility of Some Sulfonamides in Cyclohexane'. *Journal of the Brazilian Chemical Society,* 14 (5), pp. 803-808.
Matos, R. *et al.* (2005) 'Saccharin: a combined experimental and computational thermochemical investigation of a sweetener and sulfonamide'. *Molecular Physics,* 103 (2-3), pp. 221-228.
Matsuda, H. *et al.* (2015) 'Determination and prediction of solubilities of active pharmaceutical ingredients in selected organic solvents'. *Fluid Phase Equilibria,* 406 116-123.
Meenan, P. (2001) 'From Molecules to CrystallizersAn Introduction to Crystallization Roger Davey and John Garside. Oxford University Press, New York. 2000. ISBN 0198504896'. *Crystal Growth & Design,* 1 (1), pp. 101-101.
Meltzer, V. and Pincu, E. (2009) *A dsc study for binary mixture of 2-chlorobenzoic acid with salicylic acid.*
Meltzer, V. and Pincu, E. (2012) 'Thermodynamic study of binary mixture of citric acid and tartaric acid'. *Central European Journal of Chemistry,* 10 (5), pp. 1584-1589.
Miller, F.W. and Dittmar, H.R. (1934) 'The Solubility of Urea in Water. The Heat of Fusion of Urea'. *Journal of the American Chemical Society,* 56 (4), pp. 848-849.
MOE, M.O.E. (2018) 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7: 2013.08;Chemical Computing Group ULC.
Monte, M. *et al.* (2010) 'Vapour pressures, enthalpies and entropies of sublimation of para substituted benzoic acids'. *J Therm Anal Calorim,* 100 (2), pp. 465-474.
Montgomery, D.C. (2008) *Design and analysis of experiments.* London: John Wiley & Sons:.

Moore, D.J. *et al.* (2007) 'Infrared spectroscopy and differential scanning calorimetry studies of binary combinations of cis- 6- octadecenoic acid and octadecanoic acid'. *Chemistry and Physics of Lipids,* 150 (1), pp. 109-115.
Moradabadi, A. (2017) *Theoretical study of charge transport in Li-based batteries.*
Moreno, E. *et al.* (2007) 'Polymorphism of even saturated carboxylic acids from n- decanoic to n- eicosanoic acid'. *New Journal of Chemistry,* 31 (6), pp. 947-957.
Moriwaki, H. *et al.* (2018) 'Mordred: A molecular descriptor calculator'. *Journal of Cheminformatics,* 10 (1), pp. 1-14.
Mota, F.L. *et al.* (2009) 'Temperature and solvent effects in the solubility of some pharmaceutical compounds: Measurements and modeling'. *European Journal of Pharmaceutical Sciences,* 37 (3), pp. 499-507.
Mota, F.L. *et al.* (2008) 'Aqueous solubility of some natural phenolic compounds'. *Industrial and Engineering Chemistry Research,* 47 (15), pp. 5182-5189.
Moura Ramos, J.J., Correia, N.T. and Diogo, H.P. (2004) 'Vitrification, nucleation and crystallization in phenyl- 2- hydroxybenzoate (salol) studied by Differential Scanning Calorimetry (DSC) and Thermally Stimulated Depolarisation Currents (TSDC)'. *Physical Chemistry Chemical Physics,* 6 (4), pp. 793-798.
Mu, T.C., Rarey, J. and Gmehling, J. (2007) 'Performance of COSMO- RS with sigma profiles from different model chemistries'. *Industrial & Engineering Chemistry Research, 2007 Sep 26, Vol.46(20), pp.6612-6629,* 46 (20), pp.
Muller, F.L., Fielding, M. and Black, S. (2009) 'A practical approach for using solubility to design cooling crystallisations'. *Organic Process Research and Development,* 13 (6), pp. 1315-1321.
Mumford, S.A. and Phillips, J.W.C. (1950) '19. The physical properties of some aliphatic compounds'. *Journal of the Chemical Society (Resumed),* (0), pp. 75-84.
Murdande, S.B. *et al.* (2010) 'Solubility advantage of amorphous pharmaceuticals: I. A thermodynamic analysis'. *Journal of Pharmaceutical Sciences,* 99 (3), pp. 1254-1264.
Murray, J.P., Cavell, K.J. and Hill, J.O. (1980) 'A DSC study of benzoic acid: a suggested calibrant compound'. *Thermochimica Acta,* 36 (1), pp. 97-101.
Murthy, S.S.N., Paikaray, A. and Arya, N. (1995) 'Molecular relaxation and excess entropy in liquids and their connection to the structure of glass'. *The Journal of Chemical Physics,* 102 (20), pp. 8213-8220.
Musuc, A.M., Razus, D. and Oancea, D. (2002) *Analele Universitatii Bucuresti Chimie,* 11 (2), pp. 147.
Myrdal, P., Manka, A. and Yalkowsky, S. (1995) 'AQUAFAC 3: Aqueous functional group activity coefficients; application to the estimation of aqueous solubility'. *Chemosphere,* 30 (9), pp. 1619-1637.
Nakada, H. *et al.* (2006) 'Estimation of Melting Point/Latent Heat of Binary Mixtures of Phase Change Material of Erythritol and Polyalcohol'. *KAGAKU KOGAKU RONBUNSHU,* 32 (5), pp. 429-434.
Naoki, M. *et al.* (1999) 'A new phase of hydroquinone and its thermodynamic properties'. *Journal of Physical Chemistry B,* 103 (30), pp. 6309-6313.
Neau, S.H., Bhandarkar, S.V. and Hellmuth, E.W. (1997) 'Differential molar heat capacities to test ideal solubility estimations'. *Pharmaceutical research,* 14 (5), pp. 601-605.
Negoro, H. *et al.* (1960) *Yakugaku Zasshi,,* 80 670.
Nicoli, S. *et al.* (2008) 'Ethyl-paraben and nicotinamide mixtures: Apparent solubility, thermal behavior and X- ray structure of the 1: 1 co-crystal'. *Journal of Pharmaceutical Sciences,* 97 (11), pp. 4830-4839.
Nielsen, T.L. *et al.* (2001) 'The CAPEC database'. *Journal of Chemical and Engineering Data,* 46 (5), pp. 1041-1044.
Nigam, R.K. and Dhillon, M.S. (1970) 'Thermodynamic Properties of Binary Mixtures in the Condensed Phases: Entropy of Fusion of Molecular Compounds'. *Indian J. Chem.,* 8 821-825.
Nikolova, N. and Jaworska, J. (2003) 'Approaches to Measure Chemical Similarity – a Review'. *QSAR & Combinatorial Science,* 22 (9-10), pp. 1006-1026.

Nordström, F. (2008) '*Solid-Liquid Phase Equilibria and Crystallization of Disubstituted Benzene Derivatives'*.
Nordström, F.L. and Rasmuson, Å.C. (2006) 'Phase equilibria and thermodynamics of p - hydroxybenzoic acid'. *Journal of Pharmaceutical Sciences,* 95 (4), pp. 748-760.
Nti-Gyabaah, J. *et al.* (2008) 'Solubility of lovastatin in a family of six alcohols: Ethanol, 1-propanol, 1-butanol, 1-pentanol, 1-hexanol, and 1-octanol'. *International Journal of Pharmaceutics,* 359 (1), pp. 111-117.
Ottoboni, S. (2018) '*Thesis: Developing strategies and equipment for continuous isolation of active pharmaceutical ingredients (APIs)by filtration, washing and drying'*.
Paduszynski, K., Okuniewski, M. and Domanska, U. (2013) '"Sweet-in-Green" Systems Based on Sugars and Ionic Liquids: New Solubility Data and Thermodynamic Analysis'. *Industrial & Engineering Chemistry Research,* 52 (51), pp. 18482-18491-18482-18491.
Paduszyński, K., Okuniewski, M. and Domańska, U. (2015) 'Solid– liquid phase equilibria in binary mixtures of functionalized ionic liquids with sugar alcohols: New experimental data and modelling'. *Fluid Phase Equilibria,* 403 167-175.
Palmer, D. *et al.* (2007) 'Random Forest Models To Predict Aqueous Solubility'. *Journal of Chemical Information and Modeling,* 47 (1), pp. 150.
Palmer, D. and Mitchell, J. (2014) 'Is experimental data quality the limiting factor in predicting the aqueous solubility of druglike molecules?'.
Papaioannou, V. *et al.* (2014) 'Prediction of Thermodynamic Properties and Phase Behavior of Fluids and Mixtures with the SAFT-gamma Mie Group-Contribution Equation of State'. *J CHEM ENG DATA,* 59 (10), pp. 3272-3288.
Parks, W.G., LeBaron, I.M. and Molloy, E.W. (1941) 'The viscosity of formamide - dioxane solutions at 5, 25 and 40 °C'. *J. Am. Chem. Soc,* 63 3331-3336.
Pasquali, I., Bettini, R. and Giordano, F. (2007) 'Thermal behaviour of diclofenac, diclofenac sodium and sodium bicarbonate compositions'. *Journal of Thermal Analysis and Calorimetry,* 90 (3), pp. 903-907.
Paus, R. *et al.* (2015) 'Dissolution of Crystalline Pharmaceuticals: Experimental Investigation and Thermodynamic Modeling'. *Industrial & Engineering Chemistry Research,* 54 (2), pp. 731-742.
Pedley, J.B., Naylor, R.D. and Kirby, S.P. (1986) *Thermochemical data of organic compounds.* London; New York: Chapman and Hall.
Perisanu, S. *et al.* (2006) *The structure and thermochemistry of 3: 4, 5: 6-dibenzo-2-hydroxymethylene-cyclohepta-3, 5-dienenone (1) and some related compounds.*
Perlovich, G. and Bauer-Brandl, A. (2001) *The Melting Process of Acetylsalicylic Acid Single Crystals.*
Perlovich, G.L., Volkova, T.V. and Bauer-Brandl, A. (2006) '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 (7), pp. 1448-1458.
Peña, M.A. *et al.* (2009) 'Thermodynamics of Cosolvent Action: Phenacetin, Salicylic Acid and Probenecid'. *Journal of Pharmaceutical Sciences,* 98 (3), pp. 1129-1135.
Peña, M.A. *et al.* (2006) 'Solubility parameter of drugs for predicting the solubility profile type within a wide polarity range in solvent mixtures'. *International Journal of Pharmaceutics,* 321 (1), pp. 155-161.
Pinto, S.S., Diogo, H.n.P. and Minas Da Piedade, M.E. (2003) 'Enthalpy of formation of monoclinic 2-hydroxybenzoic acid'. *The Journal of Chemical Thermodynamics,* 35 (1), pp. 177-188.
Pinto, S.S. and Diogo, H.P. (2006) 'Thermochemical study of two anhydrous polymorphs of caffeine'. *The Journal of Chemical Thermodynamics,* 38 (12), pp. 1515-1522.
Poeti, G., Fanelli, E. and Braghetti, M. (1982) 'A differential scanning calorimetric study of some phenol derivatives'. *J. Therm. Anal.,* 24 273-279.

Ponomarenko, S.M. *et al.* (1995) 'Physicochemical properties and electronic structure of some aprotic solvents'. *Zh. Obshch. Khim.,* 65 (2), pp. 190-198.
Prakash, D.J. *et al.* (1996) 'Density and Viscosity of Methanol + 1,2-Dichloroethane, Methanol + 1,1,1-Trichloro Ethane, and Methanol + 1,1,2,2-Tetrachloroethane Mixtures'. *Physics and Chemistry of Liquids,* 33 (4), pp. 249-254.
Prasad, K.V.R. *et al.* (2001) 'Crystallization of paracetamol from solution in the presence and absence of impurity'. *International Journal of Pharmaceutics,* 215 (1), pp. 29-44.
Qin, F. *et al.* (2010) 'Study of the heat of absorption of co 2 in aqueous ammonia: Comparison between experimental data and model predictions'. *Industrial and Engineering Chemistry Research,* 49 (8), pp. 3776-3784.
Qiu, J. and Albrecht, J. (2018) 'Solubility Correlations of Common Organic Solvents'. *Organic Process Research & Development,* 22 (7), pp. 829-835.
Rai, U. and Rai, R. (1998) 'Physical Chemistry of Organic Eutectics'. *Journal of Thermal Analysis and Calorimetry,* 53 (3), pp. 883-893.
Rai, U.S. and Mandal, K. (2011) *Some physicochemical studies on organic eutectics and 1:1 addition compound; p-phenylenediamine – benzoic acid system.*
Rai, U.S. and Mandal, K.D. (1990) 'Chemistry of organic eutectics: p-phenylenediamine-m-nitrobenzoic acid system involving the 1:2 addition compound'. *Bull. Chem. Soc. Jpn.,* 63 1496-1502.
Rai, U.S. and Rai, R.N. (1999) 'Some Physicochemical Studies on Organic Eutectics and Molecular Complex: Urea -- p-nitrophenol System'. *J. Mater. Res,* 14 1299-1305.
Reddi, R.S.B. *et al.* (2011) 'Phase equilibria, crystallization, thermal and microstructural studies on organic monotectic analog of nonmetal– nonmetal system; urea– 4- bromo- 2- nitroaniline'. *Fluid Phase Equilibria,* 313.
Rehberg, C.E. and Dixon, M.B. (1950) 'n-Alkyl lactates and their acetates'. *J. Am. Chem. Soc,* 72 1918-1922.
Renon, H. and Prausnitz, J.M. (1968) 'Local compositions in thermodynamic excess functions for liquid mixtures'. *AIChE Journal,* 14 (1), pp. 135-144.
Riggio, R. *et al.* (2011) *Mixtures of methyl isobutyl ketone with three butanols at various temperatures.*
Rodríguez, S. *et al.* (1997) 'Excess volumes and excess viscosities of binary mixtures of cyclic ethers with bromobenzene'. *Journal of Solution Chemistry,* 26 (2), pp. 207-215.
Rojas-Aguilar, A. *et al.* (2004) 'Thermochemistry of benzoquinones'. *Journal of Chemical Thermodynamics,* 36 (6), pp. 453-463.
Romero, S. *et al.* (2004) 'Characterization of the solid phases of paracetamol and fenamates at equilibrium in saturated solutions'. *Journal of Thermal Analysis and Calorimetry,* 77 (2), pp. 541-554.
Roos, Y. (1993) 'Melting and glass transitions of low molecular weight carbohydrates'. *Carbohydrate research,* 238 39-48.
Ross, T.J. (2010) *Fuzzy Logic with Engineering Applications.*
Rotich, M.K., Glass, B.D. and Brown, M.E. (2001) 'Thermal Studies on Some Substituted Aminobenzoic Acids'. *Journal of Thermal Analysis and Calorimetry,* 64 (2), pp. 681-688.
Roux, M.V., Temprado, M. and Chickos, J.S. (2005) 'Vaporization, fusion and sublimation enthalpies of the dicarboxylic acids from C 4 to C 14 and C 16'. *The Journal of Chemical Thermodynamics,* 37 (9), pp. 941-953.
Roy, S., Riga, A.T. and Alexander, K.S. (2002) 'Experimental design aids the development of a differential scanning calorimetry standard test procedure for pharmaceuticals'. *Thermochimica Acta,* 392 399-404.
RSC (2019) '*RSC*'.

Rubin, T.R., Levedahl, B.H. and Yost, D.M. (1944) 'The Heat Capacity, Heat of Transition, Vaporization, Vapor Pressure and Entropy of 1,1,1-Trichloroethane'. *Journal of the American Chemical Society,* 66 (2), pp. 279-282.

Ruether, F. and Sadowski, G. (2009) 'Modeling the Solubility of Pharmaceuticals in Pure Solvents and Solvent Mixtures for Drug Process Design'. *Journal of Pharmaceutical Sciences,* 98 (11), pp. 4205-4215.

Ruslim, F. *et al.* (2009) 'Evaluation of pathways for washing soluble solids'. *Chemical Engineering Research and Design,* 87 (8), pp. 1075-1084.

Ruslim, F. *et al.* (2007) 'Optimization of the wash liquor flow rate to improve washing of pre-deliquored filter cakes'. *Chemical Engineering Science,* 62 (15), pp. 3951-3961.

Sabbah, R. and Antipine, I. (1987) *Mise au point d'un appareil d'analyse thermique dans l'intervalle 300<t<600 k et son utilisation pour la mesure de la purete de substances, de la temperature de leur point triple et de leur enthalpie de fusion.*

Sabbah, R. and Buluku, E.N.L.E. (1991) 'Étude thermodynamique des trois isomères du dihydroxybenzène'. *Canadian Journal of Chemistry,* 69 (3), pp. 481-488.

Sabbah, R. and Gouali, M. (1996) 'Energétique liaisons inter at intramoléeulairas les trois isomères de l'aminophénol'. *Canadian Journal of Chemistry,* 74 (4), pp. 500-507.

Sabbah, R. and Ider, S. (1999) 'Thermodynamics of intermolecular and intramolecular bonds in three carboxypyridinic acids (picolinic, nicotinic and isonicotinic acids)'. *CANADIAN JOURNAL OF CHEMISTRY-REVUE CANADIENNE DE CHIMIE,* 77 (2), pp. 249-257.

Sabbah, R. and Le, T.H.D. (1993) 'Étude thermodynamique des trois isomères de l'acide hydroxybenzoïque'. *Canadian Journal of Chemistry,* 71 (9), pp. 1378-1383.

Sabbah, R. and Perez, L. (1999) 'Thermodynamic study of phthalic, isophthalic, and terephthalic acids'. *CANADIAN JOURNAL OF CHEMISTRY-REVUE CANADIENNE DE CHIMIE* 77 (9), pp. 1508-1513.

Sacchetti, M. (2001) 'Thermodynamic analysis of DSC data for acetaminophen polymorphs'. *Journal of Thermal Analysis and Calorimetry,* 63 (2), pp. 345-350.

Saini, M.K. and Murthy, S.S.N. (2014) 'Study of glass transition phenomena in the supercooled liquid phase of methocarbamol, acetaminophen and mephenesin'. *Thermochimica Acta,* 575 195-205.

Saleemi, A., Onyemelukwe, I.I. and Nagy, Z. (2013) 'Effects of a structurally related substance on the crystallization of paracetamol'. *Frontiers of Chemical Science and Engineering,* 7 (1), pp. 79-87.

Sato, K. *et al.* (1990) 'Structure and transformation in polymorphism of petroselinic acid (cis-ω- 12-octadecenoic acid)'. *Journal of Physical Chemistry,* 94 (7), pp. 3180-3185.

Schaake, R.C.F., van Miltenburg, J.C. and De Kruif, C.G. (1982) 'Thermodynamic properties of the normal alkanoic acids. II. Molar heat capacities of seven even-numbered normal alkanoic acids'. *J. Chem. Thermodyn,* 14 771-778.

Schaathun, H.G. (2012) 'Support Vector Machines'. 179-196.

Schmack, G., Rother, M. and Bittrich, H.J. (1973) 'Viscosity measurements in the systems tetrachloromethane - aromatics'. *Z. Phys. Chem. Leipzig,* 253 401-405.

Schmidt, C. and Ulrich, J. (2012) 'Predicting Crystal Morphology Grown from Solution'. *Chemical Engineering & Technology,* 35 (6), pp. 1009-1012.

Schroder, B. *et al.* (2010) 'Prediction of aqueous solubilities of solid carboxylic acids with COSMO-RS'. *Fluid Phase Equilibria, 2010 Mar 15, Vol.289(2), pp.140-147,* 289 (2), pp.

Schroeter, T. *et al.* (2007) 'Estimating the domain of applicability for machine learning QSAR models: a study on aqueous solubility of drug discovery molecules'. *Journal of Computer - Aided Molecular Design,* 21 (12), pp. 651-664.

Sharma, B.L., Sharma, N.K. and Rambal, M. (1992) *Excess Thermodynamic Functions: GE and SE of Binary Organic Eutectic Systems.*

Singh, J. and Singh, N.B. (2015) 'Phase equilibrium, crystallization behavior and thermodynamic studies of (m- dinitrobenzene + vanillin) eutectic system'. *The Journal of Chemical Thermodynamics,* 89 197-204.

Singh, M. *et al.* (2013) 'Solid–liquid equilibrium, thermal, and physicochemical studies on salicylamide–4-nitrophenol and 2-cyanoacetamide–4-aminoacetophenone organic eutectic systems'. *Journal of Thermal Analysis and Calorimetry,* 113 (2), pp. 977-983.

Singh, N.B. and Kumar, P. (1986) 'Solidification Behavior of the Cinnamic Acid- p - Nitrophenol Eutectic System'. *Journal of Chemical and Engineering Data,* 31 (4), pp. 406-408.

Singleton, W.S., Ward, T.L. and Dollear, F.G. (1950) ' Physical properties of fatty acids. I. Some dilatometric and thermal properties of stearic acid in two polymorphic forms'. *J. Am. Oil Chem. Soc.,* 27 143-146.

Siniti, M. *et al.* (1993) *Etude du comportement thermique des hexitols.*

Siniti, M., Jabrane, S. and Létoffé, J.M. (1999) 'Study of the respective binary phase diagrams of sorbitol with mannitol, maltitol and water'. *Thermochimica Acta,* 325 (2), pp. 171-180.

Smyth, C.P. and Stoops, W.N. (1928) 'The dielectric polarization of liquids. 3. The polarization of the isomers of heptane'. *J. Am. Chem. Soc., ,* 50 1883-1890.

Song, C.-C.C.a.Y. (2004) 'Solubility Modeling with a Nonrandom Two-liquid Segment Activity Coefficient Model'. *Ind. Eng. Chem. Res.,* 43 8354-8362.

Sotomayor, R.G. *et al.* (2012) 'Extended Hildebrand solubility approach applied to piroxicam in ethanol + water mixtures'. *Journal of Molecular Liquids,* 180.

Spaght, M.E., Thomas, S.B. and Parks, G.S. (1932) 'Some heat- capacity data on organic compounds, obtained with a radiation calorimeter'. *Journal of Physical Chemistry,* 36 (3), pp. 882-888.

Steinbeck, C. *et al.* (2003) 'The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics'. *Journal of chemical information and computer sciences,* 43 (2), pp. 493-500.

Stewart, B., Hylton, D.J. and Ravi, N. (2013) 'A systematic approach for understanding Slater-Gaussian functions in computational chemistry'. *Journal of Chemical Education,* 90 (5), pp. 609-612.

Subrahmanyam, C.V.S. and Sarasija, S. (1997) 'Solubility Behaviour of Carbamazepine in Binary Solvents: Extended Hildebrand Solubility Approach to obtain Solubility and other Parameters'. *Pharmazie,* 52 (12), pp. 939-942.

Sumarokova, T.N. and Nurmakova, A.K. (1960) 'Electrical Conductivity, Viscosity, and Density of the System SnBr4 - C2H5COOH; SnBR4 - C3H7COOH, SnBR4 - C5H11COOH'. *J. Gen. Chem. USSR,* 30 (1), pp. 29-36.

Sunwoo, C. and Eisen, H. (1971) 'Solubility parameter of selected sulfonamides'. *Journal of Pharmaceutical Sciences,* 60 (2), pp. 238-244.

Surov, A.O. *et al.* (2009) 'Thermodynamic and structural aspects of some fenamate molecular crystals'. *Crystal Growth and Design,* 9 (7), pp. 3265-3272.

Surwase, S. *et al.* (2013) 'Indomethacin: New Polymorphs of an Old Drug'. *Molecular Pharmaceutics,* 10 (12), pp. 4472-4480-4472-4480.

Svoboda, V. *et al.* (1982) 'Determination of heats of vaporization and some other thermodynamic properties for four substituted hydrocarbons'. *Collect. Czech. Chem. Commun.,* 47 (2), pp. 543-549.

Swarts, F. (1931) 'Study of the viscosity - The viscosity of organic compounds'. *J. Chim. Phys. ,* 28 622-650.

Szterner, P., Legendre, B. and Sghaier, M. (2010) 'Thermodynamic properties of polymorphic forms of theophylline. Part I: DSC, TG, X- ray study.(Report)'. *Journal of Thermal Analysis and Calorimetry,* 99 (1), pp. 325.

Talja, R.A. and Roos, Y.H. (2001) 'Phase and state transition effects on dielectric, mechanical, and thermal properties of polyols'. *Thermochimica Acta,* 380 (2), pp. 109-121.

Tanimoto, T.T. (1958) *An elementary mathematical theory of classification and prediction.* New York: International Business Machines Corporation.

Teixeira, A.C.T., Gonçalves Da Silva, A.M.P.S. and Fernandes, A.C. (2006) 'Phase behaviour of stearic acid– stearonitrile mixtures: A thermodynamic study in bulk and at the air–water interface'. *Chemistry and Physics of Lipids,* 144 (2), pp. 160-171.

Temprado, M., Roux, M. and Chickos, J. (2008) 'Some thermophysical properties of several solid aldehydes'. *Journal of Thermal Analysis and Calorimetry,* 94 (1), pp. 257-262.

Thakur, S.S. *et al.* (2012) 'Solid- state mechanical properties of crystalline drugs and excipients; New data substantiate discovered dielectric viscoelastic characteristics.(Report)'. *Journal of Thermal Analysis and Calorimetry,* 108 (1), pp. 283.

'Thermal Analysis T127: Application to Medical and Pharmaceutical Products (Melting Point and Fusion Heat)'. (2018) *DMF-Recovery and Purification,* 1-4.

'Thermophysical data'. (1936) *Physikalisch-Chemische Tabellen 5.Auflage Verlag Springer Berlin 3.Teil,* 1.

Thompson, C. *et al.* (2004) 'The effects of additives on the growth and morphology of paracetamol (acetaminophen) crystals'. *International Journal of Pharmaceutics,* 280 (1), pp. 137-150.

Timmermans, J. and Hennaut-Roland (1937) 'Work of the International Bureau of Physico-Chemical Standards VIII. Physical Constants of 20 Organic Compounds'. *J. Chim. Phys,* 34 693-739.

Titani, T. (1927) 'THE VISCOSITY OF LIQUIDS ABOVE THEIR BOILING POINTS. PART I'. *Bulletin of the Chemical Society of Japan,* 2 (4), pp. 95-105.

Tong, B. *et al.* (2010) 'Heat Capacities and Nonisothermal Thermal Decomposition Reaction Kinetics of d-Mannitol'. *Journal of Chemical & Engineering Data,* 55 (1), pp. 119-124.

Tong, B. *et al.* (2008) 'Thermodynamic investigation of several natural polyols (II); Heat capacities and thermodynamic properties of sorbitol.(Author abstract)(Report)'. *Journal of Thermal Analysis and Calorimetry,* 91 (2), pp. 463.

Toropov, A.P. and Kim, L.P. (1961) 'Influence of viscosity increase of components to the viscosity isotherms in normal systems'. *Uzb. Khim. Zh., ,* 2 51-55.

Tsakalotos, D.E. (1908) 'About the hydrates of fatty acids'. *C. R. Hebd. Seances Acad. Sci,* 146 1272-1274.

Tulsyan, A., Garvin, C. and Ündey, C. (2018) 'Advances in industrial biopharmaceutical batch process monitoring: Machine- learning methods for small data problems'. *Biotechnology and Bioengineering,* 115 (8), pp. 1915-1924.

Türk, M. and Kraska, T. (2009) 'Experimental and theoretical investigation of the phase behavior of naproxen in supercritical CO 2'. *Journal of Chemical and Engineering Data,* 54 (5), pp. 1592-1597.

Umnahanant, P. and Chickos, J. (2012) 'Vaporization and sublimation enthalpies of acetanilide and several derivatives by correlation gas chromatography'. *Journal of Chemical and Engineering Data,* 57 (4), pp. 1331-1337.

Valavi, M., Svärd, M. and Rasmuson, A.C. (2016) 'Prediction of the Solubility of Medium-Sized Pharmaceutical Compounds Using a Temperature-Dependent NRTL-SAC Model'. *Industrial and Engineering Chemistry Research,* 55 (42), pp. 11150-11159.

Van Norman, G.A. (2016) 'Drugs, Devices, and the FDA: Part 1: An Overview of Approval Processes for Drugs'. *JACC: Basic to Translational Science,* 1 (3), pp. 170-179.

Vecchio, S. *et al.* (2004) 'Thermal analysis study on vaporization of some analgesics. Acetanilide and derivatives'. *Thermochimica Acta,* 420 (1), pp. 99-104.

Vecchio, S. and Tomassetti, M. (2009) 'Vapor pressures and standard molar enthalpies, entropies and Gibbs energies of sublimation of three 4- substituted acetanilide derivatives'. *Fluid Phase Equilibria,* 279 (1), pp. 64-72.

Verevkin, S.P. (1999) 'Substituent effects on the benzene ring. Prediction of the thermochemical properties of alkyl substituted hydroquinones'. *Physical Chemistry Chemical Physics,* 1 (1), pp. 127-131.

Verevkin, S.P. and Kozlova, S.A. (2008) 'Di-hydroxybenzenes: Catechol, resorcinol, and hydroquinone: Enthalpies of phase transitions revisited'. *Thermochimica Acta,* 471 (1), pp. 33-42.
Vogel, L. and Schuberth, H. (1980) 'Some physicochemical data of urea near the melting point'. *Chem. Tech. (Leipzig),* 32 143.
Vold, M.J. (1949) 'Differential Thermal Analysis'. *Analytical Chemistry,* 21 (6), pp. 683-688.
Vorländer, D. and Walter, R. (1925) 'The mechanically forced birefringence of amorphous liquids in connection with their molecular shape'. *Z. Phys. Chem. ,* 118 1-30.
Walden, P. (1906) 'On organic solvents and ionization means. III. Part: Internal friction and their connection with the conductivity'. *Z. Phys. Chem. NF,* 55U 207-249.
Wang, B. *et al.* (2012) 'Measurement and correlation for solubility of dexibuprofen in different solvents from 263.15 to 293.15K'. *Thermochimica Acta,* 540 91-97.
Wang, L.-C. and Wang, F.-A. (2004) 'Solubilities of niacin in sulfuric acid + water and 3- picoline + sulfuric acid + water from (292.65– 361.35) K'. *Fluid Phase Equilibria,* 226 (1-2), pp. 289-293.
Wang, S.X. *et al.* (2004) 'Calorimetric study and thermal analysis of crystalline nicotinic acid'. *Journal of Thermal Analysis and Calorimetry,* 76 (1), pp. 335-342.
Wang, T.-C., Lai, T.-Y. and Chen, Y.-P. (2010) *Solid–Liquid Equilibria for Hexanedioic Acid + Benzoic Acid, Benzoic Acid + Pentanedioic Acid, and Hexanedioic Acid + Pentanedioic Acid.*
Wang, T.C., Li, Y.J. and Chen, Y.P. (2012) 'Solid- liquid equilibria for six binary mixtures involving heptanedioic acid, pentanedioic acid, hexanedioic acid, 2,3- dimethylbutanedioic acid, 2,2-dimethylbutanedioic acid, and 3- methylheptanedioic acid'. *Journal of Chemical and Engineering Data,* 57 (12), pp. 3519-3524.
Wassvik, C.M. *et al.* (2006) 'Contribution of solid- state properties to the aqueous solubility of drugs'. *European Journal of Pharmaceutical Sciences,* 29 (3-4), pp. 294-305.
Wassvik, C.M. *et al.* (2008) 'Molecular characteristics for solid- state limited solubility'. *Journal of medicinal chemistry,* 51 (10), pp. 3035-3039.
Watterson, S. *et al.* (2014) 'Thermodynamics of fenofibrate and solubility in pure organic solvents'. *Fluid Phase Equilibria,* 367 143-150.
Wei, D. *et al.* (2009) 'Measurement and Correlation of Solid- Liquid Equilibria of Phenyl Salicylate with C 4 Alcohols'. *Chinese Journal of Chemical Engineering,* 17 (1), pp. 140-144.
Wei, D.W., Li, F.S. and Zhang, W.X. (2004) 'Determination of Solid-Liquid Equilibria Data of Dihydroxyphenols'. *Hua Hsueh Kung Yeh Yu Kung Cheng,* 21 (3), pp. 227-230.
Weinstein, D.I., Leffler, A.J. and Currie, J.A. (1984) 'Phase Transitions In Bicyclic Compounds'. *Molecular Crystals and Liquid Crystals,* 104 (1-2), pp. 95-109.
Wildman, S.A. and Crippen, G.M. (1999) 'Prediction of Physicochemical Parameters by Atomic Contributions'. *Journal of Chemical Information and Computer Sciences,* 39 (5), pp. 868-873.
Witschi, C. and Doelker, E. (1997) 'Residual solvents in pharmaceutical products: acceptable limits, influences on physicochemical properties, analytical methods and documented values'. *European Journal of Pharmaceutics and Biopharmaceutics,* 43 (3), pp. 215-242.
Wu, H., Dang, L. and Wei, H. (2014) 'Solid- Liquid Phase Equilibrium of Nicotinamide in Different Pure Solvents: Measurements and Thermodynamic Modeling'. *Industrial & Engineering Chemistry Research,* 53 (4), pp. 1707-1711-1707-1711.
'*www.solvay.us'*. (2019).
Wyrzykowska-Stankiewicz, D. and Szafranski, A. (1975) Published. 'Determination of Purity and Heat of Fusion of Organic Compounds by Differential Scanning Calorimetry'. 1975.
Xu, F. *et al.* (2004a) *Wuji Huaxue Xuebao,* 20 (1), pp. 50.
Xu, F. *et al.* (2004b) 'Thermodynamic study of ibuprofen by adiabatic calorimetry and thermal analysis'. *Thermochimica Acta,* 412 (1-2), pp. 33-37.
Xu, F. *et al.* (2006) 'Adiabatic calorimetry and thermal analysis on acetaminophen'. *Journal of Thermal Analysis and Calorimetry,* 83 (1), pp. 187-191.

Xue, Z., Mu, T.C. and Gmehling, J. (2012) 'Comparison of the a Priori COSMO-RS Models and Group Contribution Methods: Original UNIFAC, Modified UNIFAC(Do), and Modified UNIFAC(Do) Consortium'. *Industrial & Engineering Chemistry Research, 2012 Sep 12, Vol.51(36), pp.11809-11817,* 51 (36), pp.

Yang, H., Thati, J. and Rasmuson, Å.C. (2012) 'Thermodynamics of molecular solids in organic solvents'. *The Journal of Chemical Thermodynamics,* 48 150-159.

Yang, S.S. and Guillory, J.K. (1972) 'Polymorphism in sulfonamides'. *Journal of Pharmaceutical Sciences,* 61 (1), pp. 26-40.

Ying, Y. *et al.* (2017) 'Polymorph formation in fenofibrate in the absence and presence of polymer stabilizers: a low wavenumber Raman and differential scanning calorimetry study'. *Journal of Raman Spectroscopy,* 48 (5), pp. 750-757.

Yu, S. *et al.* (2000) *Taiyangneng Xuebao,* 21 171.

Zheng, J., Xu, X. and Truhlar, D.G. (2011) 'Minimally augmented Karlsruhe basis sets.(Report)'. *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta),* 128 (3), pp. 295.

Zhou, D. *et al.* (2002) 'Physical Stability of Amorphous Pharmaceuticals: Importance of Configurational Thermodynamic Quantities and Molecular Mobility'. *Journal of Pharmaceutical Sciences,* 91 (8), pp. 1863-1872.

Zimmerman, J.H.K. (1952) 'The Experimental Determination of Solubilities'. *Chemical Reviews,* 51 (1), pp. 25-65.

Zordan, T.A. *et al.* (1972) 'Enthalpies and entropies of melting from differential scanning calorimetry and freezing point depressions: urea, methylurea, 1,1-dimethylurea, 1,3-dimethylurea, tetramethylurea, and thiourea'. *Thermochimica Acta,* 5 (1), pp. 21-24.