***In Silico* Methods to Predict Solubility**

James L McDonagh, John BO Mitchell, David S Palmer & Rachael E Skyner

Addresses & contact details:

James L McDonagh**:** IBM Research UK, Hartree Centre, STFC Daresbury Laboratory, Sci-Tech Daresbury, Keckwick Lane, Daresbury WA4 4FS
+44-(0)1925 568346
james.mcdonagh@uk.ibm.com

John BO Mitchell: EaStCHEM School of Chemistry and Biomedical Sciences Research Complex, University of St Andrews, St Andrews, KY16 9ST, UK.
+44-(0)1334-467259
jbom@st-andrews.ac.uk

David S Palmer: Department of Pure and Applied Chemistry, University of Strathclyde, Thomas Graham Building, 295 Cathedral Street, Glasgow, Scotland G1 1XL, U.K.
david.palmer@strath.ac.uk

Rachael E Skyner**:**
Diamond Light Source Ltd., Diamond House, Harwell Science and Innovation Campus, Fermi Ave, Didcot, OX11 0DE.
Structural Genomics Consortium, University of Oxford, Old Road Research Campus, Roosevelt Drive, Oxford, OX3 7DQ.
+44-(0)1235 567537
rachael.skyner@diamond.ac.uk

IN SILICO METHODS TO PREDICT SOLUBILITY

1. Solubility: What is all the fuss about?

Solubility is a vital parameter in many different scenarios from the direct question "*will compound A dissolve in solvent B?*" to questions such as "*what is the toxicity of a substance?*" (Hutchinson *et al.*, 1979), "*what environmental impacts may a substance have?*" (Doerr-MacEwen *et al.*, 2006) and "*how sensitive are humans to a particular odour*?" (Sell, 2014). As a result, many industries have an interest in determining solubility at an early stage in molecular discovery and development, as well as in later stages of chemical and product formulation. Some may even wish to optimize a molecular design to maximize or minimize solubility depending on the use case (Leach *et al*., 2006). As a result, the computational prediction of solubility is an attractive idea as no chemicals need to be used, hence many predictions can in principle be run in parallel at minimal cost in terms of human research time, chemical usage and chemical disposal. Computational methods therefore offer the potential to dramatically shorten the time to solution, minimise the environmental impact of molecular discovery and reduce research costs. Nonetheless, the general aim in applying computational methods is not to completely replace laboratory experiments, but rather to guide investigators towards focusing their experimental resources on the most promising areas of research.

Solubility is a particularly important property in the pharmaceutical industry, playing a critical role in determining *pharmacokinetics* — the mechanism by which a substance is transported around the body and excreted; and *pharmacodynamics* — an active substance's pharmaceutical action *in vivo*.

Pharmacokinetics covers the Absorption, Distribution, Metabolism, Excretion and Toxicity (ADMET) properties, which are all influenced by a substance's *bioavailability*, *i.e.* in what concentration and how fast an active compound will be available under physiological conditions at the point of action. For orally administered pharmaceuticals, active substances must pass through gastric fluids, cell membranes, and blood in order to reach their sites of action. This covers a wide variety of environments: hydrophobic, hydrophilic, acidic stomach and slightly acidic to neutral intestinal tract. Many drug molecules are weak acids or bases meaning that ionization can occur in response to such environmental changes, affecting the probability of absorption at different sites within the gastrointestinal tract. As a result of these factors, if a substance has a low solubility, or a significant shift in its solubility profile due to these changes in environment, then the quantity of that substance available within the gastrointestinal tract and bloodstream may consequently be low. This leads to difficulties in formulation as there may be a high variance in the quantity of the active substance available in different patients (Leach *et al*., 2006).

Pharmacodynamics describes the pharmaceutical's therapeutic action. Solubility can affect how easily an active substance can bind to a target, and hence contributes to a substance's pharmacological activity. For the reasons discussed here, solubility has become a major source of attrition in the development of new pharmaceuticals, and the subject of regulatory conditions for low solubility active substances (FDA, 2014).

2. Definitions and Concepts

A solution is a homogenous mixture of solute(s) and solvent(s) in any physical state, solid, liquid or gas, where the solute is the substance dissolved in the solvent. A substance's solubility describes the extent to which the substance can be dissolved in a given solvent, resulting in a solution. Solubility is a thermodynamic property, related to the equilibrium between the solute and solvent. From these general ideas, we can begin to discuss the finer details of solubility and the process of solvation.

2.1 Solubility Data and Experimental Determinations

Solubility data, particularly in the pharmaceutical industry, tend to be of two distinct types: *equilibrium solubility* measurements, also called thermodynamic solubility, and *kinetic solubility* measurements. In the following paragraphs we provide a short overview of how solubility data can be measured. A more in-depth discussion of this topic is provided in the chapter "The role of solubility to optimize drug substances – a Medicinal Chemistry perspective" of this book.

2.1.1 Equilibrium solubility

Equilibrium solubility is the concentration of solute in equilibrium with its saturated solution. These measurements are often made at the later stages of pharmaceutical development, such as during formulation (Narasimham & Barhate, 2011; Saal & Petereit, 2012).

The classical method for determining this is the *shake flask method* (Jouyban & Fakhree, 2012), which involves mixing a sample of a solute with a buffer solution until saturation. The saturated solution is then shaken until equilibrium is achieved and a final solubility determined by the dissolved concentrations. This is often achieved using high-pressure liquid chromatography. This method relies upon long shaking times, as it is difficult to determine when the solution has reached equilibrium.

A more modern approach is the *CheqSol (chasing equilibrium solubility*) method (Stuart & Box, 2005; Llinàs *et al.*, 2008; Box *et al.*, 2009; Etherson *et al.*, 2014). The procedure involves acid-base titrations between concentrations slightly above and slightly below that where precipitation occurs. A major advantage of the CheqSol method over others is its speed. The CheqSol experiment takes approximately 20 – 80 minutes (Stuart & Box, 2005), whereas traditional shake-flask methods can take days, and other titration-based methods (Avdeef, 1998) can take up to 10 hours.

Another commonly used method is the *synthetic* method (Jouyban & Fakhree, 2012), which is particularly useful for viscous solutions. This method uses a laser and detector and determines the equilibrium point by a significant drop in laser light reaching the detector, signifying that the solute is no longer entering solution.

2.1.2 Kinetic solubility

Kinetic solubility is the solubility at which an induced precipitate is first detected. The kinetic solubility value is attributable to a metastable state, which results from a supersaturated solution (a solution in which the concentration of the solute is greater than its equilibrium

value). Hence, typically these values suggest a compound's solubility to be higher than the true equilibrium solubility. As mentioned in 2.1.1, equilibrium measurements often require a long time, and thus kinetic solubility measurements have emerged as an alternative due to their speed, allowing fast screening in the early stages of molecular discovery. Kinetic solubility values can be helpful in guiding the experimental design of a molecule towards an optimal solubility (Narasimham & Barhate, 2011; Saal & Petereit, 2012).

Kinetic solubility is often determined using *turbidimetric assays*. Turbidimetry is a process that measures the loss of transmitted light intensity due to the scattering effect of suspended particles. This usually involves mixing the solute with an organic solvent and using UV spectroscopy to detect when precipitation occurs. The solubility is then determined based on the concentration that has been added to the organic solvent. Experimental solubility determinations have been discussed in much more detail by Lipinski *et al.* (1997) and by Alsenz & Kansy (2007), and also in other chapters of the current volume.

2.2 Intrinsic Solubility

An important definition in the solubility literature is *intrinsic solubility*. The intrinsic solubility of an ionisable molecule is defined as *the equilibrium solubility of the unionised form at a given set of thermodynamic conditions* (Hörter *et al.*, 2001; Palmer *et al.*, 2008).

This is a significant quantity in many industries where it is used to indicate the bioavailability of a substance. This is important for pharmaceuticals — where bioavailability determines how effective an active ingredient can be and the dosage required — as well as in industries like agrochemicals where environmental concerns are critical for pesticide and insecticide development. Several well established models also link the intrinsic solubility to pH dependent solubility and the dissolution process through the *Noyes-Whitney equation* (Noyes & Whitney, 1897) (equation 1) and *Henderson-Hasselbalch equation* (Po & Senozan, 2001) (equations 4a & 4b), as discussed in more detail below. Intrinsic solubility is generally referred to using the notation S_0 or, for its base 10 logarithm, $\log S_0$. The solubility is most often referred to units of moles per litre (M), though when considered as an equilibrium constant it is technically unitless.

2.3 The Solvation Process and Factors Which Affect Solvation

2.3.1 Solvation: Equilibrium solubility and dissolution

The process of solvation involves two clear and distinct concepts, the first is equilibrium solubility (described in 2.1.1) and the second is *dissolution.* Dissolution is a kinetic property, and describes the rate at which molecules become available for dispersal from the solid solute into solution. Dispersal of solute molecules through a solvent continues until a constant equilibrium concentration is achieved. Solubility and dissolution are important concepts, particularly in the context of pharmaceuticals, as drug delivery is impacted by the dissolution rate whilst drug activity is impacted by a solute's equilibrium solubility (Jouyban & Fakhree, 2012). The dissolution rate can be described by the Noyes-Whitney equation (Noyes & Whitney, 1897):

$$\frac{dW}{dt} = \frac{kA(C_S - C)}{L} \quad (1)$$

Equation 1. *Noyes-Whitney equation: dW/dt is the rate of dissolution, A stands for the solute surface area that is in contact with the solvent, C represents the instantaneous solute concentration in the bulk solvent, C_s is the diffusion layer solute concentration (given from the solubility of the molecule with the assumption that the diffusion layer is saturated), k is the diffusion coefficient, and L is the diffusion layer thickness.*

2.3.2 Thermodynamic effects on solubility

There are a number of physical and chemical factors that affect a substance's solubility. As discussed above, solubility is a thermodynamic property, thus thermodynamic variables such as temperature and pressure influence a substance's solubility.

Temperature affects a substance's solubility in accordance with the second law of thermodynamics: *for an isolated system a spontaneous change will occur in the direction of increasing entropy*. The Gibbs free energy of solvation is composed of enthalpic and entropic terms:

$$\Delta G^*_{sol} = \Delta H^*_{sol} - T\Delta S^*_{sol} \quad (2)$$

where the asterisks refer to the 1 M standard state (see below). The entropic term plays an increasingly important role as the temperature increases, since molecular motion increases at higher temperatures, leading to a more disordered system. For example, a gas generally has a much greater available volume, is more dispersed, and therefore has a higher entropy than a liquid or solution. Hence, dissolving a gas in a liquid is entropically unfavourable, and a gas will generally be less soluble at a higher temperature. Therefore gas solvation must be enthalpically driven. The opposite is true of a solid, since breaking up an ordered crystal lattice is an entropically favourable process. Thus, a solid becomes more soluble at higher temperature, an observation so familiar that it can be considered common sense, with the solution representing a higher entropy state. Solid solvation can therefore be entropically driven.

The partial pressure of a gas *(the pressure that a single gas component of a mixture would have if it alone occupied the same volume at the same temperature as the mixture)* is another example of a thermodynamic variable that affects a substance's solubility. The solution is in equilibrium with the surroundings, so if the composition, temperature or pressure of the surrounding gas changes, the partial pressure also changes, and the equilibrium responds giving a change in solubility. This is the thermodynamic explanation of the process which, for example, occurs when one opens a carbonated drink and carbon dioxide escapes from the solution.

When discussing the solubility of a solute material, molecular interactions that exist between the solute molecules need to be accounted for. Where the solute is a solid, due consideration of the nature of the solid-state form is required, whether that be amorphous or crystalline.

The polymorphic form *(polymorphs are alternative 3D crystalline arrangements of molecules)* of a crystalline material must also be considered. Closer packing of molecules within a crystal lattice, and stronger interactions, lead to energetically favourable lattice energies. More enthalpy is required to dissociate the energetically favourable crystal, which makes it less soluble. Therefore, the lowest energy polymorph of a compound is the least soluble. A more weakly bound structure, such as a less stable polymorph or especially an amorphous state, is enthalpically easier to break up, and thus the solubility is higher as these structures can dissociate more easily. Polymorphism is highly relevant in the pharmaceutical industry, as polymorphic transitions can substantially change the solubility, pharmacokinetics, and other physicochemical properties of a substance. This is perhaps most famously presented pharmaceutically in the case of the HIV protease inhibitor Ritonavir (Bauer *et al.*, 2001; Chemburkar *et al.*, 2000; Neumann & van de Streek, 2018).

Ionization also plays an important role in determining a substance's solubility. Many biologically active ingredients fall into a category of being a weak acid or weak base, meaning the active molecules ionization changes, depending on environmental pH. This ionization occurs due to a reaction with the solvent, which forms an equilibrium. An example of the equilibrium states, where water is the solvent, is given in equations 3a & 3b**Error! Reference source not found.** (for an acid – 3a, and for a base – 3b):

Acid:

$$HA + H_2O \rightleftharpoons A^- + H_3O^+ \quad \therefore \quad pKa = -log_{10}\left(\frac{[A^-][H_3O^+]}{[HA]}\right) \qquad (3a)$$

Base:

$$B + H_2O \rightleftharpoons BH^+ + OH^- \quad \therefore \quad pKb = -log_{10}\left(\frac{[BH^+][OH^-]}{[B]}\right) \qquad (3b)$$

Equations 3a & 3b. *(a) Acid and (b) base equilibria and definitions of pK_a and pK_b. HA represents an acidic molecule and B represents a basic molecule, the ionized species are then represented by the charged forms.*

The strength of an acid in solution is measured by pK_a, which is the negative base 10 logarithm of the acid dissociation constant K_a. A more positive pK_a value represents a smaller extent of dissociation at a given pH, as shown by the Henderson-Hasselbalch equation – equations 4a & 4b. Generally, as ionization increases, the solubility of electrolytes (substances that dissociate upon solvation into ions and enable the solution to conduct electricity) also increases and the solubility of non-electrolytes decreases. As a result, the pH at which experimental measurements are made is an important factor when assessing a substance's solubility. The total solubility of an ionisable substance is calculable by the Henderson-Hasselbalch equation (for an acid – equation 4a, for a base – equation 4b) by consideration of the intrinsic solubility, pH of the environment and pK_a of a substance. Note that the exponent of the rightmost term differs in the case of an acidic or basic solute.

$$log_{10}S^{Acid} = log_{10}S_0 + log_{10}(1 + 10^{pH-pKa}) \qquad (4a)$$

$$log_{10}S^{Base} = log_{10}S_0 + log_{10}(1 + 10^{pKa-pH}) \quad (4b)$$

Equations 4a & 4b. *The Henderson-Hasselbalch equation. S_0 is the intrinsic solubility, pK_a is defined in Equation 3a, and pH is the acidity or basicity of the solution.*

Solubility is influenced by intermolecular interactions between molecular species. This means the interactions between the solute molecules, interactions between solvent molecules, and the cross interactions between solute and solvent molecules all need to be accounted for when performing solubility predictions. In addition to this, thermodynamic variables and ionization need to be considered. This leads to a large number of degrees of freedom, hence demonstrating the difficulty faced in making accurate predictions *in silico* (Jouyban & Fakhree, 2012; Bergström & Larsson, 2018).

3. Computational Prediction of Solubility

Methods typically applied to solubility prediction broadly fall into two categories: *first principles* calculations and *chemoinformatics*.

First principles calculations generally apply physical modelling methods such as coarse-grained simulations, molecular dynamics (MD; Jorgensen & Tirado-Rives, 1996) and quantum chemistry. While considerably closer to physics than are chemoinformatics approaches, such methods are rarely *first principles* in the sense of doing fully *ab initio* quantum chemistry on both solvent and solute. Nonetheless, these methods look to solve real physical equations to elucidate the physicochemical processes that are occurring. Chemoinformatics, in contrast, seeks methods to correlate so called *features* with a property of interest, in our current case solubility. These features range in complexity from simple counts (*e.g.*, the number of carbon atoms in a molecule), to more complex descriptors such as those representing the topology of a molecule, *e.g.* shape indices (Sharma *et al.*, 1997; Hall & Kier, 2007).

Both first principles and chemoinformatics approaches have their advantages and disadvantages. Chemoinformatics is a data driven discipline whose models have been shown to provide accurate results quickly once they have been suitably trained. However, these methods usually provide little in the way of phenomenological or mechanistic information and can perform poorly outside of the domain they were trained for. First principles calculations have also been shown to provide good results, but are usually a little less accurate and require much more computational time and hence higher cost. Such methods do however provide chemical and physical insights into the phenomena and mechanisms that physically transpire.

In first principles and some cheminformatics models, it is common to apply a thermodynamic cycle to predict the solubility of a molecule. There are two frequently employed cycles: the *fusion cycle* and the *sublimation cycle*. These are shown in Figure 1. The fusion cycle describes the transition from the solid to solution state through an intermediate supercooled liquid state. From a physical perspective, this provides two free energy changes: the free energy of fusion and the free energy of transfer. The sublimation cycle couples the solid state to the solution state through a gaseous state, which again

consists of two free energy changes: free energy of sublimation and free energy of solvation or hydration. These cycles, or variations on them, have been used for first principles models (Lüder *et al.*, 2007a; Palmer *et al.*, 2008; Palmer *et al.*, 2012; Paluch *et al.*, 2010; Moučka *et al.*, 2015; Li *et al.*, 2017), chemoinformatics models (McDonagh *et al.*, 2014) and model equations (Ran & Yalkowsky, 2001).

3.1 Standard state conventions

Sublimation energies are typically quoted or calculated in the *1 atm standard state*. In contrast to this, solvation free energies are often given using the *1M or Ben-Naim* standard state (Ben-Naim, 1978; Ben-Naim & Marcus, 1984). This must be accounted for in modelling solubility by a thermodynamic cycle. In this chapter, ΔG° will be used to indicate the 1 atm standard state. ΔG* will be used to indicate the 1M standard state (Ben-Naim, 1978).

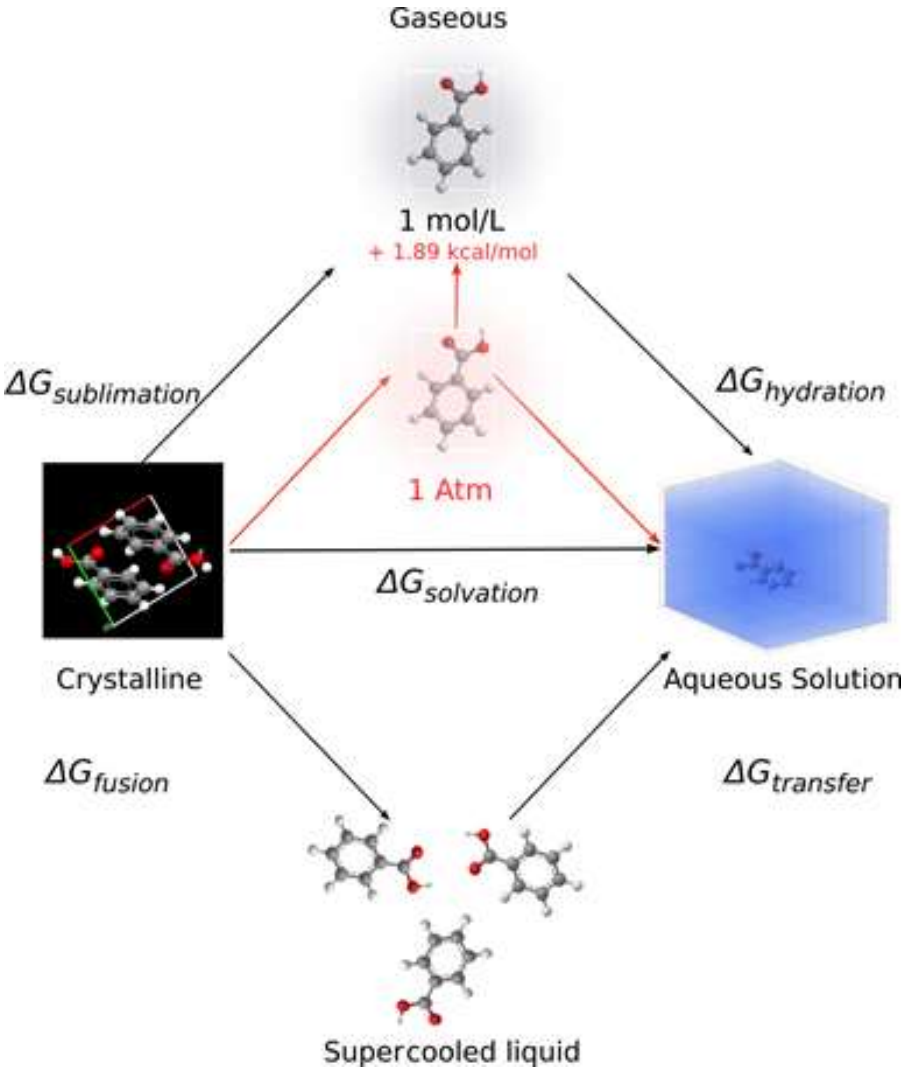

Figure 1. *Sublimation and fusion cycles used to predict solubility. Representation of the 1 atm and 1M standard states with their difference in energies calculated at 298 K.*

The energetic difference across these two standard states is calculated from the work for isothermal expansion or compression of a gas between its initial volume V_i and final volume V_f as:

$$\Delta G = RT\ ln\left(\frac{V_f}{V_i}\right) \quad (5)$$

Taking the initial condition to be 1M and the final as 1 atm, this becomes RT ln(24.46) at 298 K, since the molar volume of an ideal gas at 1 atm is 24.46 litres. The corresponding energy difference at 298K is therefore 1.89 kcal/mol or 7.91 kJ/mol. When calculating the sublimation energy, the energetic correction is positive in the conversion from 1atm to 1 M and negative in the conversion 1 M to 1 atm, this is shown diagrammatically in Figure 1.

3.2 Solubility from First Principles

It is very striking that a wide diversity of methods exist for computing solubility on the basis of chemical and physical theory. We will refer to such methods as *first principles* approaches, notwithstanding the fact that to varying degrees they all depend in one way or another on empirical parameters.

3.2.1 Computational models of the Solvent for First Principles Calculations

Across computational chemistry, it has been shown that the chemical environment often needs to be modelled in order to accurately reproduce the chemical and physical characteristics of a system (Luccarelli *et al.*, 2010). In some cases the environment takes an active role in chemical processes, meaning that the environment needs to be explicitly accounted for when modelling the system (Jorgensen *et al.*, 1983; Bhat *et* al., 1994; Jorgensen & Tirado-Reves, 2005; Luccarelli *et al.*, 2010).

Models for solvents can by and large be split into two distinct categories: *explicit* and *implicit*, although some hybrid models (Skyner *et al*., 2015; Ratkova *et al*., 2015) have also been generated. Implicit models are computationally more efficient, but lack the explicit insights into the role of the solvent. Implicit models are particularly common in quantum chemistry (Cramer & Truhlar, 1999; Tomasi *et al*., 2005; Klamt 2011). Depending on the property one is considering, the additional detail afforded by explicit models may not be strictly required. Implicit models tend to provide bulk response properties such as polarization, but do not explicitly represent discrete solvent molecules. Implicit models model a solvent as a continuous field, which interacts with a bounding surface around the solute. The reaction field, induced by the charge distribution of the solute polarizing the solvent field, is then evaluated on this surface. This is shown in Figure 2. Notable versions of implicit solvation models include the Polarizable Continuum Model (PCM; Miertuš *et al.,* 2001; Cammi & Tomasi 1995), the Solvation Model based on Density (SMD; Marenich *et al.,* 2009) and the Conductor-like Screening Model (COSMO; Klamt & Schüürmann, 1993).

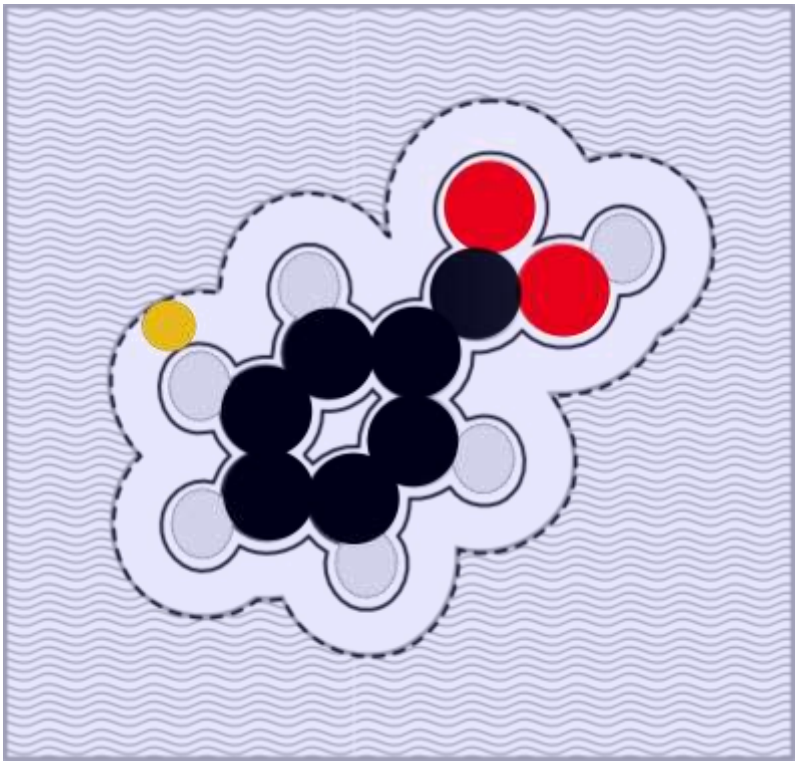

Figure 2. *Implicit solvation by a continuum model. A probe shown in yellow is used to trace a surface known as the solvent accessible surface defining a surface area which is accessible for solvent interactions with the solute. A reaction field is then induced as a result of the solute molecule's electrostatic potential polarizing the continuum field.*

Explicit models directly model solvent molecules, thereby accounting for spatial and orientational degrees of freedom of the solvent. Explicit models pay additional overheads for accounting directly for all the degrees of freedom, but provide greater insight into the role a solvent plays in a chemical system. Explicit models are most common in Molecular Dynamics (Jorgensen & Tirado-Rives, 1996) — a simulation method for studying the physical motion of atoms and molecules, where the trajectories of atoms and molecules are determined from Newton's equations. Explicit models for various solvents have been generated over decades, which has allowed the tuning of efficient parameterized models to represent the physical processes that are occurring. Genuinely first principles quantum chemical methods have also advanced to a state at which a limited number of explicit solvent molecules can be included in a calculation (Huan *et al.*, 2016; Hogan *et al.*, 2016), in these examples only one or two water molecules. In addition, coupled multi-scale methods such as Quantum Mechanics / Molecular Mechanics (QM/MM; Warshel & Levitt, 1976) have allowed quantum chemical calculations of reactions in the presence of a larger solvent environment — quantum mechanics is used to treat the area in which chemical processes take place at an appropriate level of quantum theory, and molecular mechanics is used to model the rest of the system with a force field. Such advances have enabled computational studies of enzyme reactions (Mulholland, 2005; Senn *et al*., 2005; Alderson *et al.*, 2012).

There are some emerging hybrid models which offer averaged spatial distribution information without the extra computational cost of explicitly representing the solvent molecules, an important example being 3D-RISM (3-Dimensional Reference Site Interaction Model: Chandler *et al.*, 1986; Kovalenko & Hirata, 1999; Genheden *et al*., 2010; Palmer *et al*. 2010; Ratkova *et al*., 2015). 3D-RISM stands at a level intermediate between the kind of explicit representation of solvent molecules used in QM simulations and the implicit solvation of semi-empirical quantum mechanics (methods that rely on the same formalisms as QM methods, but include some information from experimental or empirical sources). 3D-RISM uses three-dimensional solvent density distributions, but not spatial coordinates of individual solvent molecules. Although 3D-RISM is based on rigorous statistical mechanics, carefully chosen 3D-RISM functionals are required to obtain numerically accurate results.

3.2.2 Modelling Water

Aqueous solution is one of the most critical chemical environments, and the subject of much computational as well as experimental research. Therefore, many models have been generated to represent water in various computational methods. In this section we outline some of the common water models. An exhaustive list of water models in computational chemistry is outside the scope of this text, but we provide references to extended discussions on water models (Cramer & Truhlar, 1999; Mark & Nilsson, 2001; Tomasi *et al*., 2005; Skyner *et al*., 2015; Ratkova *et al*., 2015).

The water molecule is composed of two hydrogen atoms covalently bound to a single oxygen atom in a distorted tetrahedral geometry. The oxygen lone pairs above and below the plane of the molecule cause a deviation from the ideal 109.5° tetrahedral bond angle, the experimentally determined value being 104.5°. Water is a polar molecule, having point group C_{2v} and a permanent dipole moment along its principal axis. Water can also act as both a hydrogen bond donor and acceptor, and is able to form up to four hydrogen bonds with neighboring molecules. This leads to highly ordered structures being formed; particularly in its crystalline solid state, ice, which has many different polymorphic forms that differ in the regular 3D arrangements of the water molecules and can be formed under different conditions of temperature and pressure. One of the notable anomalous properties of water is that water's maximum density is not in the solid phase, as is the case with most substances. In fact, water's maximum density occurs as a liquid around 4°C.

Generally, explicit water models can be categorised through three characteristics: 1) the number of interaction sites each molecule has — the most commonly employed models place fixed interaction sites at physically relevant locations, 2) whether the molecule is treated as a rigid or a flexible body, and 3) whether polarization effects are included or not.

The most minimal models in common use are the three site models. These models generally have a rigid structure and place interaction sites at the atomic locations within the water molecule. The hydrogen sites are typically represented by a point charge. The oxygen site is represented by a point charge and a Lennard-Jones repulsion-dispersion potential. Regularly utilized models of this type are: the *Simple-Point-Charge* (*SPC*) model (Berendsen *et al*., 1981), *SPC-Extended* (*SPC/E*; Berendsen *et al*., 1987) and *Transferable-Intermolecular-Potential with 3 Points* (*TIP3P*; Jorgensen *et al.*, 1983). The SPC model was the first of these

models to be published, followed by TIP3P. Both of these models are rigid with the SPC model having an ideal tetrahedral HOH angle of 109.5°, while the TIP3P HOH angle is the experimental 104.5° (Berendsen *et al*., 1981; Jorgensen *et al.*, 1983). The SPC/E model is an extension, adding an average polarization correction to the SPC model. A further addition to the SPC model is the flexible-SPC model, which has been shown to be a very accurate three-site model (Praprotnik *et al*., 2004; Toukan & Rahman, 1985; Wu *et al.*, 2006). The flexible-SPC model enables O-H bond distance and angle variations during a simulation. Three site models are very computationally efficient owing to their relative simplicity.

A two-site model also exists, and is based on the SPC model, maintaining the dipole moment, size and charge separation of the three-site SPC model. This model has been shown to capture the solvation properties of water well for apolar solutes and the bulk solvent (Dyer *et al.,* 2009).

Beyond these models are the four-site models. These models add a fictitious negatively charged interaction site along the dipole of the water molecule. The additional site better represents the electrostatic charge distribution over the water molecule. Of the four-site models, TIP4P (Jorgensen *et al.*, 1983) is the most commonly applied, and is used in a variety of forms that have been optimized for different scenarios. One of the most notable optimizations of the TIP4P is the TIP4P/Ew model (Horn *et al*., 2004) that has been optimized to account for changes in thermodynamic, structural, and electrostatic properties of liquids by the application of Ewald summation for electrostatics, and more-precise long-range Lennard-Jones interactions. The TIP4P/Ew model has been shown to improve the parameters it was designed to optimize, and does so significantly in comparison to the TIP4P model (Horn *et al*., 2004). TIP4P/Ice (Abascal *et al*., 2005) is a model optimized for simulating and recovering properties of ice structures, with Ewald and Lennard-Jones parameters used similarly to the TIP4P/Ew model, and fitting equations of state for different forms of ice. TIP4P/2005 (Abascal & Vega, 2005) is a general-purpose parameterization for the condensed phases of water which shows promising reproductions of experimental properties.

Further models have been constructed adding additional sites to the water molecule with five-site models typically adding interaction sites for the oxygen lone pairs and removing the fictitious site from the four-site models. TIP5P is an example of this type of model (Mahoney & Jorgensen, 2000; Jorgensen & Tirado-Rives, 2005), and has been shown to provide some improvements in reproducing the structure of water clusters.

The majority of the models mentioned in the preceding paragraphs are rigid, non-polarizable models, although many have extensions that aim to add parameters for polarization (*e.g*., TIP3P/Fw and SPC/Fw – Wu *et al.,* 2006). However, when introduced as a solvent in biomolecular simulations, or for dilute solutions, their performance decreases. Polarizable water models aim to account for the non-equivalence of water molecules in solution — that is, each water molecule in solution is inequivalent as they differ in their exact geometry, vibrations, and charge distribution — rather than trying to produce an 'average' representation of water.

Generally, rigid models tend to over-stabilize the water dimer, in comparison to polarizable models (Baranyai & Bartók, 2007). A model that includes polarizable interactions is the Drude oscillator model, where a classical charged Drude particle is attached to the water

oxygen by a harmonic spring, (SWM4; Lamoureux *et al.,* 2003). This model represents the permanent charge distribution of water by three point-charges; two on the hydrogen atoms, and an additional point at the HOH bisector. There are five charged sites in total. Other charge-on-spring (COS; Straatsma & McCammon, 1990) models include the COS/B1-B2 (Yu *et al.,* 2003), COS/G2-G3 (Yu & van Gunsteren, 2004) — based upon the TIP4P geometry, and COS/D (Bachmann & van Gunsteren, 2014) models. COS/G2 and COS/G3 models are very similar in performance, and arguably better than the COS/D model.

BK3 (Kiss & Baranyai, 2013) is a polarizable model with Gaussian spatial distributions of atomic charges. Multipolar representations of water molecules have been generated, which can be used in conjunction with MD. These methods go beyond point charges to improve the representation of the anisotropy of the molecular charge distribution (Stone, 1981; Ren & Ponder, 2003; Handley *et al.*, 2009). Electronic coarse graining now enables very accurate water models to be applied to research problems (Jones *et al.,* 2013).

Coarse-grained techniques have also been produced. These methods enable larger simulations to be carried out using simplified models. Methods such as coarse-grained MD and Dissipative Particle Dynamics (DPD) are now becoming common in industrial and academic settings. Some efforts are now underway to construct optimal coarse-grained models and investigate the transferability of such models (Yesylevskyy *et al*., 2010; Bejagam *et al*., 2018).

3.2.3 Computing Sublimation Energies

Those methodologies using the sublimation cycle (Figure 1) require the computation of the sublimation free energy. There are three main approaches to this which are used either in solubility calculations, or in Crystal Structure Prediction (CSP). These are firstly the ψ_{mol} approach, secondly the ψ_{crys} method, and thirdly the Einstein Crystal technique.

In chemistry and materials science, for systems where a compound's crystal structure is unknown, prediction methods are widely used. These typically involve generating possible crystal packings and identifying those with the most favourable lattice energies. These methods fall under the category of CSP. Conveniently, lattice energy corresponds almost exactly to the sublimation energy described in Figure 1, albeit with a sign reversal. Thus, techniques originally developed for CSP can be incorporated into first principles solubility computation.

One approach (Price *et al*., 2010), which has been popular in CSP, is to obtain the lattice energy with a model potential for the repulsion and dispersion terms, plus Distributed Multipole Analysis (DMA; Stone, 1981) for the electrostatics. This approach to CSP or lattice energy computation is sometimes known as the ψ_{mol} approach. The name indicates that the wavefunction, or charge density for Density Functional Theory (DFT) methods, is calculated explicitly only for the isolated molecule. From this charge density, one obtains the DMA, giving an atom-atom anisotropic description of the electrostatic interactions. The lattice energy is then minimised by relaxing the lattice parameters and the positions and orientations of the molecules within the unit cell in the lattice minimisation program DMACRYS (Price *et al.*, 2010). This approach is used in simulation-free solubility calculation (Palmer *et al*., 2008; Palmer *et al*., 2012; Buchholz *et al.*, 2017), with the entropic

components then being approximated using statistical thermodynamics (McQuarrie & Simon, 1997).

A second possibility is to compute the lattice energy by quantum chemistry, which approaches genuine first principles accuracy. The state of the art is periodic DFT, which however requires some correction for the missing dispersion energy. Hoja & Tkatchenko (2018) used the PBE0 functional (Adamo, 1999) along with a many body dispersion correction and mode-by-mode analysis of the vibrational contributions for their most accurate CSP results. The approach is known as ψ_{crys} since it obtains the wavefunction or charge density for the periodic crystalline system. Curtis *et al*. (2018) similarly used a dispersion-corrected periodic DFT ψ_{crys} approach as part of their genetic algorithm approach to CSP.

Buchholz *et al.* (2017) obtained results for both the ψ_{mol} and ψ_{crys} approaches, and also both the melt and sublimation cycles, in their simulation-free study of the relative solubilities of racemic and enantiopure organic crystals. They also discussed the benefits of going beyond the statistical thermodynamics approximation of the vibrational contributions by explicitly computing the contribution of each phonon or molecular vibrational mode. Iuzzolino *et al*. (2018) used ψ_{crys} and ψ_{mol} approaches successively in their CSP protocol.

The performance of modern CSP methods is covered in the literature describing blind tests held periodically by the Cambridge Crystallographic Data Centre (CCDC). There have been six such tests to date, covering a multitude of different models from various researchers (Reilly *et al.* 2016; Bardwell *et al.,* 2011; Day *et al.,* 2009, Day *et al.*, 2005, Motherwell *et al.,* 2002, Lommerse *et al.,* 2000).

A third approach to the sublimation energy, more popular in solubility prediction than in CSP, is to obtain the free energy difference between an Einstein crystal, which is a simple hypothetical model of a solid, and the real crystal by simulation. This approach has been used by Li *et al.* (2017) and by Sanz & Vega (2007). Simulation-based approaches to this problem are discussed at greater length in Moustafa *et al*. (2013).

3.2.4 Computing Hydration Energies

The computation of hydration energies has been reviewed at length by Skyner *et al.* (2015) and will only be summarised here. There exist a wide diversity of approaches. Explicit water molecules can be used in an MD simulation, as by Mobley *et al*. (2009). Greater accuracy, at significantly increased expense, can be obtained from Free Energy Perturbation calculations, as described by Westergren *et al.* (2007). QM/MM approaches (Mulholland, 2005) are also possible. Moving down the scale of accuracy and cost, we can consider solvent density without the need for simulations of explicit water molecules by utilising an integral equation theory model with a suitable free energy functional, *e.g.* 3D-RISM with the UC, PC, or PC+ functional (Misin *et al.* 2015; Misin *et al.* 2016; Sergiievskyi *et al.*, 2015; Palmer *et* al. 2010), or classical molecular density functional theory (Wu, 2017), which describes the density of molecules in a fluid rather than the more familiar use of DFT to describe the density of electrons in a molecule. This can provide an excellent compromise between cost and precision. Hydration energies can also be calculated at relatively low computational cost

using implicit solvent continuum models (Sulea *et al*., 2009; Marenich *et al*., 2009). Finally, the COSMO-RS model can also generate good results (Klamt *et al.*, 2009; Klamt, 2011).

3.2.5 First Principles Routes to Solubility

Figure 3 shows the variety of different thermodynamic cycles used to obtain the chemical potential or free energy difference between crystalline solute and aqueous solution. As discussed earlier, a simple cycle where the total solid-to-solution free energy change is obtained from considering a route *via* the gas phase is known as a sublimation cycle. The archetypical sublimation cycle is the simple crystal-gas-solution route shown in yellow in Figure 3, and as the top half of Figure 1. A cycle proceeding *via* a liquid, often a notional supercooled liquid at room temperature, is known as a melt cycle and corresponds to the bottom half of Figure 1. The General Solubility Equation (GSE) method of Yalkowsky's group, (Ran & Yalkowsky, 2001) uses a cycle of this general kind, though also considering solution in octanol, and corresponds to the red line in Figure 3.

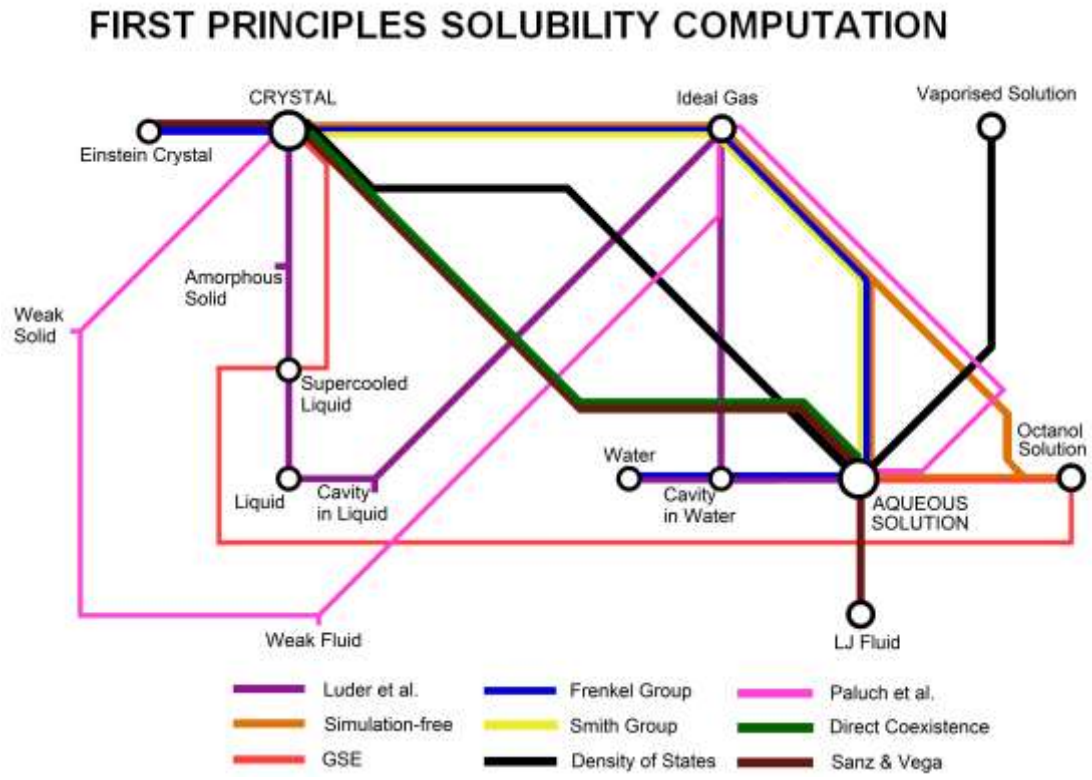

Figure 3: *States considered in a sample of nine first principles approaches to solubility calculation. All such methods require a suitable route, or thermodynamic cycle, linking the crystal and aqueous solution. This route must permit the change in free energy between crystal and aqueous solution, and therefore the equilibrium constant describing solubility, to be deduced from contributions which can all be computed with available techniques and models. States are included here if they are considered explicitly, visited in simulations, or used as a reference for the calculation of chemical potential or free energy. Some states correspond to real systems, others are hypothetical.*

3.2.5.1 Direct Coexistence

One approach is to run an MD or similar simulation of a solute and solvent until equilibrium is reached. This equilibrium will then represent the solubility limit of this system, and the concentration can then be obtained simply by counting the number of solvent and solute particles in the solution phase in the simulation. Kolafa (2016) used both the SPC/E model of water (Berendsen *et al.*, 1987) and the BK3 model (Kiss & Baranyai, 2013) to simulate a slab of crystalline NaCl in direct contact with brine, and obtained solubilities around 3.7 mol/kg, rather less than the experimental 6.1 mol/kg. Such direct coexistence methods tend to be expensive and on occasion are described rather harshly as "brute force" approaches, despite the substantial technical and scientific expertise involved.

3.2.5.2 Chemical Potentials from Simulation

Another popular approach is based on the equality of the chemical potentials of the crystalline solid and aqueous solution phases at equilibrium:

$$\mu_{aq}^{solute} = \mu_{solid}^{solute} \tag{6}$$

where the chemical potentials (μ) are functions of temperature, pressure and, in solution, concentration (Paluch *et al.*, 2010; Bergström & Larsson, 2018). If the absolute chemical potentials of both phases can be calculated at a given temperature and pressure, then the problem reduces to finding the solution chemical potential as a function of concentration. The concentration at which the two chemical potentials become equal is the limiting equilibrium or thermodynamic solubility. Computing the absolute chemical potentials of these phases requires comparison with a reference state, for example an ideal gas as in a sublimation cycle, whose absolute chemical potential can be obtained. Alternatively, absolute values are not required if both crystal and solution phases are referred to the same reference state, since the requirement of equation (6) is simply that the *difference* in their chemical potentials is zero.

Recent work in the Frenkel group (Li *et al.,* 2017; Li *et al.,* 2018) calculated the chemical potential of the crystalline phase by using MD to simulate thermodynamically reversible paths between the Einstein crystal and the full crystalline solute. The solution phase chemical potential was obtained by MD simulation of the processes of growing a cavity in SPC water (Berendsen *et al*., 1981), inserting a solute molecule into the cavity, and then finally shrinking the cavity away to leave the molecule in aqueous solution (Bergström & Larsson, 2018). Although Li *et al.* in fact simplified their analysis by assuming high dilution, such that only the insertion of a single isolated solute molecule into pure water need be considered, their method is fairly easily extensible to higher solubilities. They applied their approach to naphthalene and obtained $\log S_0$ = -5.32, almost identical to the experimental value of -5.36 (Li *et al.,* 2017). While they suggest that this essentially perfect agreement may be somewhat fortuitous, it provides a tantalising hint that their methodology may be highly effective. More recently (Li *et* al., 2018), they calculated the solubility of phenanthrene as $\log S_0$ = -6.51 compared with the experimental -6.96, an error of only 0.45 $\log S_0$ units. However, their calculations underestimated the solubility of caffeine by approximately two orders of magnitude. To investigate the accuracy of their method

properly, a much larger set of predictions on several tens of druglike compounds would be required.

The Smith group's work, Moučka *et al.* (2015), similarly seeks conditions where equation (6) is satisfied. They used an Osmotic Ensemble Monte Carlo approach to calculate the chemical potential of NaCl solutions as a function of concentration, carrying out a series of Monte Carlo (MC) simulations with different numbers of ion pairs solvated in water. They also used MC simulations to compute the chemical potential of solid NaCl. Since both solid and solution chemical potentials were calculated relative to the ideal gas reference state, this work is also an example of the so-called sublimation cycle; see Figure 1. Like any approach intended for highly soluble compounds, their methodology is required to operate without any assumptions of high dilution. Like most NaCl solubility calculations, their results both depended quite strongly on the force field used and were underestimated relative to experiment. Their best result was a solubility of 3.6 mol/kg using a potential function from Joung & Cheatham (2008), a factor of nearly two smaller than the experimental value of 6.1 mol/kg and similar to the results of Kolafa (2016). Other choices of force field led to values in the range 0.8 to 1.0 mol/kg. The inadequacies of potential energy functions are likely to be a major cause of underestimation of NaCl solubility in simulation studies.

Paluch *et al.* (2010) also computed the solubility of NaCl. Their work contains a particularly useful discussion of the role of reference states in calculating both relative and absolute chemical potentials. They also obtain μ_{aq}^{solute} as a function of concentration, and seek the conditions where it is equal to μ_{solid}^{solute}. The solid chemical potential is found by performing MD simulations along a pathway in which the crystal is transformed *via* two hypothetical states, a weakly interacting ordered solid and a weakly interacting liquid, into an ideal gas. Path integration is performed along this transformation pathway, so that the solid chemical potential can be found relative to the ideal gas reference state. For the solution, they carry out an analogous process using what they term an Expanded Ensemble method, linking together the ideal gas and the solution in SPC/E water (Berendsen *et al*., 1987) *via* solutions of varying concentrations, all simulated with an MC method. Like Moučka *et al.* (2015), they underestimate the NaCl solubility significantly, obtaining 0.8 mol/kg.

Sanz & Vega (2007) had earlier used MC simulations and thermodynamic integration to link the solid chemical potential to that of an Einstein crystal, and the solution chemical potential to a hypothetical Lennard-Jones fluid. Both were also related to an ideal gas reference state. They found the solid and solution chemical potentials of NaCl to be equal at a concentration of 5.4 mol/kg in SPC/E water (Berendsen *et al*., 1987), which translates to 4.8 M. Their result is significantly closer to the experimental value of 6.1 mol/kg, or 5.4 M, than many other simulation-derived solubilities for NaCl. Possible reasons for this are discussed at some length by Paluch *et al.* (2010).

3.2.5.3 Free energy change via amorphous phases

In a series of four publications (Westergren *et al*., 2007; Lüder *et al*., 2007a; Lüder *et al*., 2007b; Lüder *et al*., 2009), a Swedish team computed the solubility of druglike molecules using simulations *via* an elaborately planned route that visited crystalline and amorphous solids, supercooled liquid, liquid melt, and ideal gas on its way to the aqueous solution.

Rather than explicitly modelling the crystalline phase, they linked crystalline to amorphous solubility by the empirical relationship

$$S_0^{amorph} \approx S_0^{crys} \exp\left(\frac{\Delta S_m}{R} \ln\left(\frac{T_m}{T}\right)\right) \quad (7)$$

where T_m is the melting point and ΔS_m is the entropy of melting. Their work is notable for its use of a simple linear response approximation, whereby the free energy required to transfer a single molecule from the vapour into an amorphous phase is:

$$\Delta G_{va} = \Delta G_{cav} + E_{LJ} + \frac{E_{QQ}}{2} \quad (8)$$

This is the sum of the free energy required to form a cavity in TIP4P water (ΔG_{cav}) (Jorgensen *et al.*, 1983), the Lennard-Jones energy of interaction of the molecule with the amorphous phase (E_{LJ})and only half the Coulombic interaction energy (E_{QQ}), since the other half is assumed to be cancelled out by a corresponding entropy change. Their most accurate free energies were obtained by an expensive Free Energy Perturbation method, but they demonstrated that good results could also be achieved by simpler and cheaper models which greatly reduced the cost of the simulations.

3.2.5.4 Simulation-free approaches

An alternative to simulation is to calculate the free energy changes in a static manner, without use of dynamics. Two related publications (Palmer *et al.* 2008; Palmer *et al.*, 2012) illustrate this approach, both computing solubility as an equilibrium constant derived from free energy changes calculated under standard conditions. This contrasts with the approach based on equation (6) and used by many simulation methods, which seeks the non-standard conditions under which the change in chemical potential or free energy on solvation is zero. In Palmer *et al*.'s work a sublimation cycle is used, the crystal-gas leg being computed by the kind of ψ_{mol}-based lattice energy minimisation common in crystal structure prediction (Price *et al.*, 2010). In the 2008 publication, the gas-solution leg was computed *via* a simple semi-empirical quantum chemical model of the solution state. The authors looked both at a direct gas-aqueous route, and also at one *via* octanol solution. The ease of computing logP, and hence the relevant partition equilibrium constant P, facilitated accurate modelling of transfer between the two solution environments. The authors found that significant improvements in accuracy could be obtained by allowing the contributions to the free energy change to be scaled by parameters fitted from training data. A regression model including three descriptors — first principles lattice energy, estimated logP, and the number of rotatable bonds — achieved an excellent Root Mean Squared Error (RMSE) of 0.71 $\log S_0$ units over an unseen test set of 26 druglike molecules (Palmer *et al*., 2008). Adding the computed hydration energy to the model was found not to improve the regression statistics.

Modest accuracy in the hydration energy was to be expected, since continuum solvation models contain many approximations. Solvent structure features from the solvation shell structure are missing in continuum models, and non-electrostatic energy terms are not represented in a first principles manner. In Palmer *et al*. (2012), the more sophisticated 3D-RISM/UC model of Palmer *et al*. (2010), parameterised specifically to yield numerically accurate hydration free energies, was used for the hydration leg.

Combining the two legs of the cycle, the solubility S_0 can be obtained from the free energy of dissolving the solid into aqueous solution, which is given by the sum

$$\Delta G^*_{solu} = \Delta G^*_{sub} + \Delta G^*_{hyd} = -RT \ln(S_0 V_m) \quad (9)$$

where S_0 is the intrinsic solubility in moles per litre and V_m is the crystal's molar volume. The molar volume V_m (in litres) appears due to the use of the Ben-Naim approach with the molecular centres of mass fixed and a standard state (denoted by *) of 1 M concentration. $S_0 V_m$ is the ratio of the compound's concentrations in aqueous solution and in the crystal, which Ben-Naim & Marcus (1984) treat like a partition coefficient. A useful formula (Palmer *et al.* 2012; Skyner *et al*., 2015) allows one to calculate the sublimation free energy at its usual 1 atm standard state and the solvation term similarly at the common 1 M, while eliminating the crystalline molar volume:

$$S_0 = -\frac{p_0}{RT} \exp\left(\frac{\Delta G_{sub}^{1\ atm} + \Delta G_{solv}^{1 mol/L}}{RT}\right) \quad (10)$$

Palmer *et al.* (2012) achieved a RMSE of 1.45 $\log S_0$ units from first principles across a set of 25 druglike molecules, good enough to be a useful prediction though with a larger error than is typical of informatics approaches to similar problems.

Also borrowing from CSP, Buchholz *et al.* (2017) considered both ψ_{mol} and ψ_{crys} approaches to the sublimation energy in their investigation of the relative solubilities of racemic and enantiopure crystals of druglike organic molecules. For the gas-to-solution leg, they utilised the COSMO-RS model (Klamt, 2011). Their objective was to evaluate the feasibility of using differential solubility for enantiomer separation.

3.2.5.5 Solubility from Density of States

A rather different method has been described by the Anwar group (Boothroyd *et. al.,* 2018). This involves computation of the density of states of the solution, which is carried out by means of MC simulations on NaCl solutions of different concentrations in SPC/E water (Berendsen *et al*., 1987), using the force field from Joung & Cheatham (2008). These simulations visit both solution and gaseous states, in order to facilitate particle insertion. The solid phase can be simulated similarly to other methods discussed above, or one can simply import an externally simulated value of μ_{solid}^{NaCl}. Restricting themselves to conditions at which the solid and solution chemical potentials were equal, the authors then looked at the computed density of states, in the form of the probability distribution function. If the density of states is known, scanning the probability distribution function in temperature at a given pressure determines the phase coexistence condition, and *vice versa* for pressure. The single component probability distribution function contains two peaks, one corresponding to pure solute and the other corresponding to the saturated solution; the mole fraction of this peak is the limiting solubility. Like other simulated-based studies, they found a NaCl solubility around half the experimental value, and also noted a counterfactual decrease of computed solubility with increasing temperature.

3.3 Solubility from Informatics

Informatics methods, in contrast to first principles ones, are designed with the simple objective of accurate numerical prediction. One seeks any method that will link the inputs – in this case, representations of molecular structures, to the required outputs – accurate estimates of experimental solubility, without any requirement to incorporate real-world physics or chemistry. Any interpretability or mechanistic insight from the model would be a secondary consideration at best. For a property where we believe that the underlying processes are unknown, difficult or expensive to compute accurately, this approach of letting the data speak for themselves has much to commend it. Properties like bioactivity, logP and melting point are therefore generally computed with informatics methods. On the other hand, some other properties, such as dipole moments and IR frequencies, are relatively easy to compute fairly accurately from first principles, and not generally predicted by chemoinformatics. Solubility is an intermediate property, where first principles computation is tractable, but currently informatics provides both much faster and also more numerically accurate predictions.

3.3.1 Test sets, comparison and experimental design

The traditional experimental design in informatics-based property prediction is to train the model on a training set. Optimisation of a model on this training set would ultimately lead to overfitting, and thus any model must be validated on an independent external test set which has not in any way been used in the model's construction. Alternatively, multiple models are built using different subsets of the data for training and for testing. Techniques of this kind include cross-validation, bootstrapping, and jack-knifing, and again depend on the test data for a particular model being independent of its training. Where an algorithm has variable parameters, these may be optimised using an internal validation set, once more this needs to be entirely independent of the final external testing. For small datasets, tuning model parameters against a single validation set may induce model bias, which can sometimes be avoided by using cross-validation rather than simple training and validation to optimise these parameters, see Figure 4a. For instance, in a 10-fold cross-validation exercise with an internal validation fold, the roles of the 8 training, one internal and one external validation set are permuted cyclically and 10 separate models generated, as shown in Figure 4b.

Validation methods, such as in the cyclical process described above, are often used for model selection. In these cases, the validation method needs to be sensitive enough to estimate differences between models, as discussed by Gütlein *et al.* (2013). For example, if the validation method has a bias in calculating predictivity, the method can still be applied, providing the systematic error applies across all models. Validation is also sensitive to dataset size, and the distribution of the data. The data in the training set will only produce a model that is good for predicting values from the same distribution. Validation estimates therefore only hold true for unseen data that fall within the same distribution as the training data. Sample selection bias (using data biased towards a particular distribution) can be reduced or avoided by using larger datasets for testing, by repeating the splitting of data, or by using stratified splitting where the folds of data are designed to have a similar distribution of property values, or sometimes coverage of chemical space, to the overall dataset. For example, if the dataset contains a few large hydrophobic and probably insoluble

compounds, the splitting is designed as far as possible to put them in different folds so that each fold is representative of the distribution of these selected groups in the complete dataset.

Unfortunately, almost every published study uses somewhat different data. This makes it almost impossible to compare the quality of prediction between separate studies. Although the relative performance and ranking of methods within a given study may be meaningful, the examples below will show that these relative rankings are in fact often not maintained when studies are compared. Although a number of standard datasets exist, such as the Huuskonen set (Huuskonen, 2000) and the DLS-100 set (Mitchell *et al.*, 2017), typically studies adopt a pick-and-mix approach to dataset construction. It is well known that some molecules are harder to predict solubility for than others (Hopfinger *et al.*, 2009; Boobier, *et al*., 2017), and by extension predictivity cannot easily be compared between datasets drawn from different regions of chemical space or different degrees of molecular diversity. This means that prediction statistics from different studies can provide only a rough guideline for comparison.

A rare attempt to standardise solubility prediction was the Solubility Challenge (Llinàs *et al.*, 2008; Hopfinger *et al.*, 2009) where the idea was that 100 accurate solubilities would be provided as a training set and a further 32 would be held back as a test set. Research groups around the world were challenged to predict intrinsic aqueous solubilities for these 32 compounds. Of these, 94 training and 28 test compounds had usable numerical solubility data. The other compounds were either too soluble for CheqSol to detect any precipitate, in which case no numerical solubility is available, or found to decompose or react during the experiment; as in the cases of aspirin and, it was later discovered (Comer *et al*., 2014), indomethacin. There were 99 entries received, each of these being a set of predicted solubilities submitted by one participating research group. The best were high quality models, while some of the others were not predictively useful. The organisers reported only R^2 on the test set, and number of answers correct to within the exacting margin of ± 0.5 $\log S_0$ units. On these criteria, only 18 entries obtained an R^2 above 0.5 on the 28 test compounds, and only 19 predicted at least half of the solubilities correctly. Fuller reporting of results, and especially a description of the methods employed by each entrant, would have made the Solubility Challenge even more valuable to the academic and industrial communities interested in solubility prediction. Nonetheless, these 122 compounds and their solubility data provide a standard training and test set than can be used to evaluate new and existing methods.

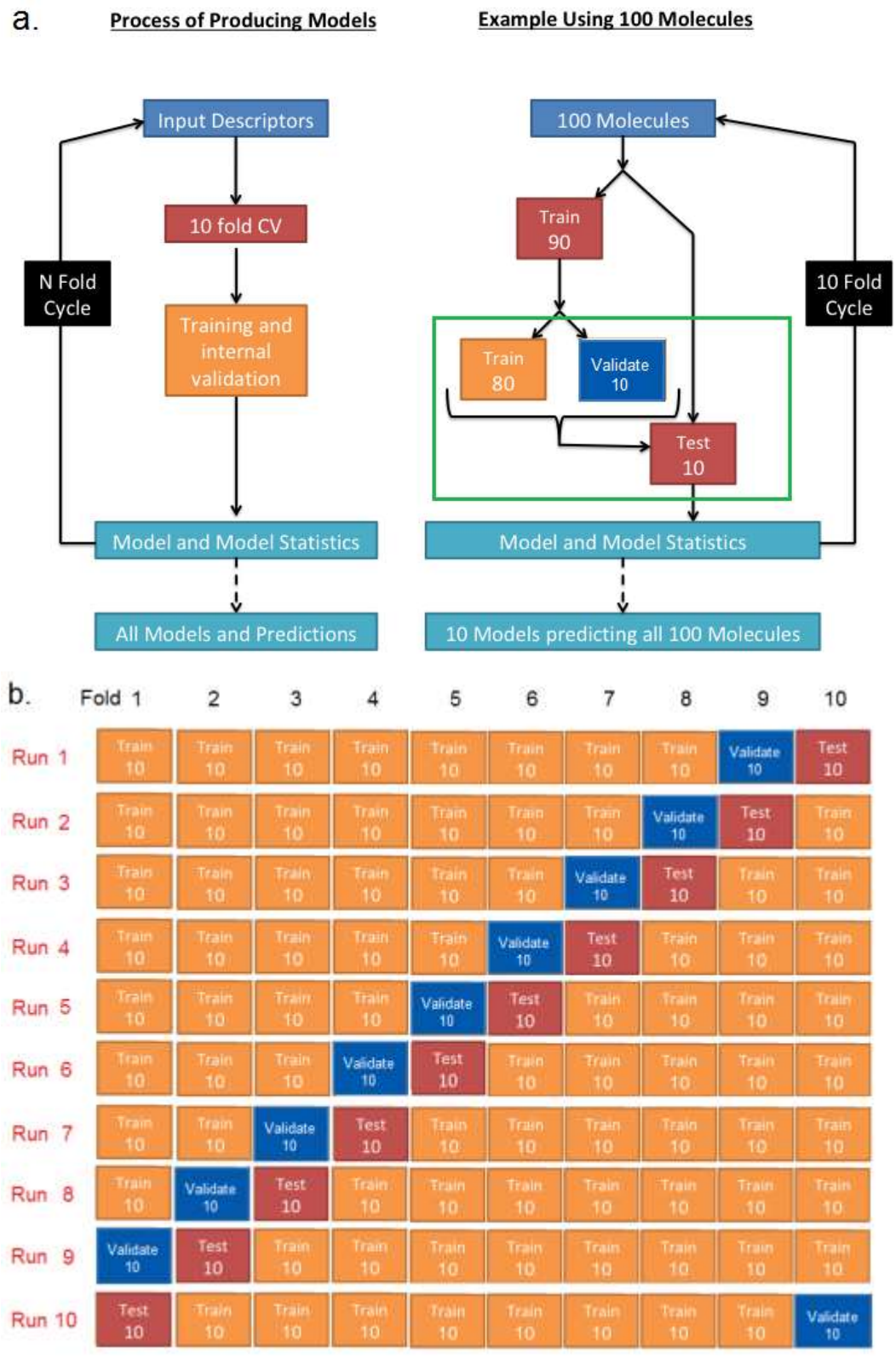

Figure 4: *(a) Structure of a 10-fold cross-validation experiment. Each model makes unseen predictions of solubility for 10 molecules, such that each molecule is predicted once by the ensemble of 10 models. (b) A detailed view of the 10-fold cross-validation experiment, showing the roles of each fold in the 10 different runs. This corresponds to the region outlined in green in Figure 4a.*

3.3.2 General Solubility Equation

Before discussing purely informatics approaches to solubility prediction, we should mention one method that falls somewhere between informatics and first principles. Ran & Yalkowsky (2001) give a theoretical justification of their General Solubility Equation (GSE) in terms of a version of the melt cycle, where the melting point determines the ratio of the solubilities of the solid and liquid phases, while the octanol-water partition coefficient logP describes the difference between solubility in an ideal solvent and in water. This leads to an equation

$$logS_0 = 0.5 - 0.01(MP - 25°C) - logP \quad (11)$$

where MP is the melting point in degrees Celsius, with the melting point term set to zero for compounds which are liquid at room temperature. Notwithstanding its theory-related derivation and its parameters being conveniently round numbers, this equation shares much in common with purely empirical relationships. Where experimental melting points and partition coefficients are available, the GSE can be quite accurate. Ran & Yalkowsky obtained a best root mean squared error (RMSE) of 0.52 $logS_0$ units for a dataset including both druglike and smaller organic molecules. Yalkowsky's group later reported RMSEs between 0.53 and 0.86 on a variety of test sets (Sanghvi *et al.*, 2003). McDonagh *et al.* (2015) found an RMSE of 0.77 for 30 druglike molecules using a version of the GSE that incorporated experimental melting points and *in silico* predicted logP values, and 0.86 when both melting points and logP were predicted computationally. Although melting point is hard to predict accurately (Nigsch *et al*, 2006; Bhat *et al*., 2008; Hughes *et al*., 2008), each 1 K error in the melting point affects the GSE's predicted solubility by only 0.01 $logS_0$ units, so even modest quality predicted melting points are acceptable for this purpose. Ali *et al*. (2012) fitted a version of the GSE using training data; the coefficients, as expected, changed upon fitting and the RMSE therefore is slightly better than that of the Ran & Yalkowsky GSE on the same dataset. Ali *et al*. also looked at models involving topographical polar surface area (TPSA) as an alternative to melting point that is more easily predictable for unsynthesised virtual library compounds; this change has very little effect on the RMS error. Alternatively TPSA can be used as well as melting point, in which case the error falls from 0.71 to 0.61 $logS_0$ units with the inclusion of the one extra parameter.

3.3.3 Quantitative Structure-Property Relationships (QSPR)

The idea that there exist property-to-property and structure-to-property correlations amongst molecules goes back to at least the 1860s (Dearden, 2006; Yousefinejad & Hemmateenejad, 2015). As early as 1863, Cros related solubility to toxicity in a PhD thesis (Cros, 1863). Such relationships are predicated on molecular structure being both physically meaningful and capable of being represented, notions only gaining widespread acceptance in the mid-nineteenth century. Edinburgh chemist Alexander Crum Brown was an early advocate of the representation of molecular structure as bonds between atoms, using deliberately two dimensional topological diagrams showing interatomic connections and bond orders. His 1868 paper (Crum Brown & Fraser, 1868) linked "the mutual relations of the atoms in the substance" to its physiological effect. QSPRs relating structure specifically to solubility included the work of Führer (1924) and Erickson (1952), both of whom observed that adding extra CH_2 units reduces solubility by an approximately constant factor.

QSAR pioneer Hansch demonstrated that solubility could be predicted by assuming a linear relationship between $\log S_0$ and logP (Hansch *et al.*,1968).

3.4 Specific Techniques in QSPR and Machine Learning

A substantial number of different mathematical and computational techniques have been used in the construction of QSPR models of solubility and other physicochemical properties (Mitchell, 2014). The more sophisticated and typically non-linear algorithms that have more recently become mainstream in the field are generally categorised as supervised Machine Learning methods. That is, the computer learns the relationship between chemical structure and solubility by being trained on available data and generates a predictive model. Below we describe some of the interesting and important techniques, though this list is not exhaustive.

3.4.1 Linear Techniques

Many historically important and mathematically compact approaches to the QSPR problem are linear. At its simplest, such a model uses a linear combination of input properties describing chemical structure and multiplies them by fitted coefficients to predict the value of the output property, here solubility. This can be simply visualised as an extension of the idea of interpolation from a line of best fit, although the number of variables used in the model is usually rather greater than one.

3.4.1.1 Group contribution and multi-linear regression methods

The constant effect of a CH_2 moiety described in Section 3.3.3 is a rudimentary example of a group contribution. This concept can be used to build models where each molecule is broken down into fragments, often similar to conventional functional groups, and by using a suitable training set each such group is assigned a numerical parameter defining a transferable contribution to solubility which it is presumed to make whenever it occurs in a molecule. For each test molecule, $\log S_0$ is calculated as the sum of the contributions from each group in the compound (Klopman *et al.*, 1992; Klopman & Zhu, 2001; Marrero & Gani, 2002) – this is known as an additive group contribution method. Klopman *et al.* (1992) carried out several predictive tests on different models; their most generally applicable model gave an RMSE of 1.25 on 21 organic molecules. Hou *et al*. (2004) defined contributions per atom, rather than per molecule, and obtained an RMSE of 0.79 $\log S_0$ units over 120 test compounds. Wang *et al.* (2007) modified the group contribution idea by calculating an accessible surface area associated with each fragment, rather than just counting the occurrences, and also added descriptors for other key properties into their model; they validated their model thoroughly and obtained an RMSE of 0.705 on their 120 molecule test set, compared with between 1.23 and 2.06 on external databases of druglike compounds. The UNIFAC (UNIversal quasi-chemical Functional-group Activity Coefficients) approach is effectively a group contribution method but proceeds *via* estimation of activity coefficients. Its use to predict organic solubilities was described by Gracin *et al*. (2002) for nine different organic solutes in a variety of polar and non-polar solvents. Abraham's group (Abraham & Le, 1999) similarly used linear regression, but the quantities in their equation were designed to have specific physicochemical meanings: refractivity, polarizability, hydrogen bond acidity and basicity, and a characteristic volume. Their work gave an

impressive RMSE of 0.5 over 65 test compounds. The same concepts can be extended to parameterise solvents, both allowing solubility to be predicted in different solvents and also permitting similarities between solvents to be identified (Bradley *et al.*, 2015).

Multi-linear regression (MLR) is also applicable based on molecular, rather than groupwise, descriptors. For instance, Hewitt *et al.* (2009) obtained an RMSE of 0.95 for the 28 usable test compounds of the Solubility Challenge (Llinàs *et al.*, 2008; Hopfinger *et al.*, 2009) using an MLR model based on only three descriptors. Catana *et al.* (2005) implemented an MLR model, amongst other linear and non-linear methods, and obtained an impressive RMSE of 0.57 over 177 test compounds.

3.4.1.2 Partial Least Squares

Partial Least Squares (PLS), or Projection to Latent Structures, is a linear regression method that uses latent variables to project both the input and output variables into a new space. PLS is essentially modelling covariance, seeking the linear combination of input features that explains the maximum proportion of the variance in the output variable, here solubility. Since it is effectively seeking a single, maximally explanatory, direction in the input space, PLS is robust against redundancy and mutual correlation amongst the input variables. It can be used similarly to multilinear regression, but without the need for prior aggressive feature selection. Catana *et al*. (2005) implemented PLS and achieved RMSE values of around 0.5 over 177 test compounds for their PLS models. Hughes *et al.* (2008) found that PLS was almost as good as support-vector machines (SVM) and a little better than random forest (RF) in terms of RMSE over 87 test compounds, with RMSE values around 0.95 depending on the exact descriptor set used. However, Palmer *et al.* (2007) had previously observed that PLS was slightly less effective than either SVM or RF with an RMSE of 0.773 for 330 test molecules. Boobier *et al*. (2017) had PLS as seventh best of ten machine learning algorithms, with an RMSE of 1.265 for 25 molecules. Cao *et al*. (2010) got an RMSE of 0.769 on 45 test set molecules, marginally better than artificial neural networks (ANN) but behind SVM.

3.4.2 Non-linear machine learning methods

While powerful enough often to obtain good empirical fits to experimental solubility data, the methods discussed above are essentially limited to linear relationships between descriptors and solubility, although Abraham & Le (1999) did consider taking products of two descriptors as a pragmatic if inelegant means of capturing inter-descriptor interactions. However, machine learning methods provide a more natural way of accounting for more complicated, non-linear, QSPRs. In fact, although chemistry has often been rather slow to adopt techniques from computer science, several different machine learning approaches have now found application in solubility prediction and related QSPR problems.

3.4.2.1 Artificial Neural Networks

Neural networks have been studied since the 1950s, with the invention of backpropagation by Werbos (1975) being a major breakthrough. An artificial neural network (ANN) is a mathematical approach to pattern recognition and machine learning problems. While the ANN is inspired by the structure of the human brain, it is not realistically attempting to

simulate or reproduce the way the brain works. Indeed the ANN is typically smaller even than the 302-neuron brain of the nematode worm *C. elegans* (Connors & Long, 2004). The ANN consists of nodes known by biological analogy as neurons, which are joined together by weight-carrying connections and arranged in an input layer, a hidden layer or layers, and an output layer. The mathematical weights are varied during training, and adjusted through backpropagation. Although there is a significant risk of overfitting if ANN training is not stopped at an appropriate stage, the approach is capable of producing good results.

ANNs have been used to predict solubility by a number of groups. Hewitt *et al.* (2009) implemented a number of different models, including ANN, as part of their participation in the Solubility Challenge (Llinàs *et al.*, 2008; Hopfinger *et al.*, 2009). They limited their models to no more than five input descriptors, albeit chosen by a genetic algorithm from an available pool of 426. Their best ANN model was a multilayer perceptron with only two input descriptors. The first was logP and the second a hard-to-interpret size-and-connectivity feature known as R2e+, or more fully as the R maximal autocorrelation of lag 2 weighted by atomic Sanderson electronegativities. Rather disappointingly, their ANN performed worse than multi-linear regression, with an RMSE of 1.51 $\log S_0$ units on the 28-compound Solubility Challenge test set. Catana *et al.* (2005) found an RMSE of 0.608 over 130 test molecules using a multi-layer perceptron (MLP), a variety of ANN, with a single hidden layer. Louis *et al.* (2010) used a backpropagation network to obtain an RMSE of 0.738 on a small 14-compound test set; Cao *et al.* (2010) used the same methodology to obtain an RMSE of 0.789 on 45 molecules. Boobier *et al.* (2017) also found an MLP to do very well, being the best of 10 assorted machine learning methods with an RMSE of 0.985 over a challenging test set of 25 druglike molecules. Erić *et al.* (2012) used a counter-propagation neural network designed for interpretability, obtaining an RMSE of 0.679 on a 94-compound test set and interpreting their model in terms of the computed importances of the seven input descriptors of which logP was the most significant. Palmer *et al.* (2007) reported that ANN did a little less well than RF in their study, with an RMSE of 0.751 over 330 test set molecules. Bhat *et al*. (2008) designed an ensemble technique with 50 neural networks and implemented it to predict the melting points of organic molecules. This approach could be combined with logP prediction to obtain solubility *via* the GSE.

3.4.2.2 Random Forest

Random Forest (RF) creates an ensemble of many diverse decision or regression trees, based on distinct samples from the same pool of data (Breiman 2001; Svetnik *et al.*, 2003). For numerical prediction of solubility, regression rather than decision trees are used. Each tree is grown by recursive partitioning of the training set compounds, based on their descriptor representations. The resulting forest of regression trees is described as random for two reasons. Firstly, each tree is built from a new bootstrap sample of the training data, a sample of *N* out of *N* compounds chosen with replacement. Secondly, at each node a tree must make its partition by considering only a random subset of the descriptors, the size of which subset is a parameter known as m_{try}. A fresh set of m_{try} descriptors is selected for decision making at each node as the tree is grown. A Gini-optimal (Raileanu &, Stoffel, 2004) split of the training data is made at each mode, so that the compounds are grouped into increasingly homogeneous sets down the tree. Thus the collection of molecules assigned to

each terminal leaf node will share similar values of solubility, or of whatever other property is being predicted. Once built, the Random Forest consisting of a total of n_{tree} regression trees can be used to predict solubilities of previously unseen test compounds. The consensus of the different trees that forms the overall prediction of the forest is based simply on the mean of the individual trees' predictions. The probability of a given molecule not being selected for the bootstrap sample of a particular tree is $(1 - 1/N)^N$, which tends to the limiting value of 1/e as N becomes large. This means that for each tree approximately 37% of the training data are unused, and these can be adopted as a so-called out-of-bag validation set. RF tends to cope well with the presence of correlated descriptors and is generally robust against overfitting (Svetnik *et al.*, 2003). Indeed, many different descriptors will play at least some role in an RF model, given that only a modestly sized sample is available to be selected at any one node of any one tree.

RF has been applied to various aqueous solubility datasets by different authors. Schroeter *et al.* (2007) obtained an RMSE of 0.855 on an external test set of 536 compounds. Palmer *et al.* (2007) reported the RMSE of 0.690 for 330 test molecules, and subsequently (Palmer & Mitchell, 2014) compared RF models built firstly on literature solubilities and secondly on new CheqSol (Box *et al.*, 2009) experiments for 80 compounds. Perhaps surprisingly, they found that the new experiments, despite having a consistent methodology and reporting scheme for all 80 compounds, generated a model with an almost identical cross-validated RMSE to that obtained from solubilities harvested from a variety of literature sources and methodologies, around 0.88 $\log S_0$ units. RF methods have been applied by Hughes *et al.* (2008), Kovdienko *et al.* (2010), McDonagh *et al.* (2014), and also by Boobier *et al.* (2017) who found that it was the joint second best amongst 10 machine learning predictors tested and of similar quality to the second best of a panel of 22 human predictors. Kew *et al.* (2015) observed RF to generate an RMSE of 1.02 in a 10-fold cross-validation using 262 molecules from Hughes *et al*. (2008), and thus to be essentially joint best alongside Support Vector Machine of 15 methods for solubility prediction.

3.4.2.3 Support Vector Machine

Support Vector Machine (SVM: Vapnik, 1998; Noble, 2006) is a popular machine learning method which transforms the input data into a high dimensional space, by means of a typically non-linear kernel function. For binary classification, SVM seeks a hyperplane which optimally separates the data between the two classes, ideally such that they lie almost entirely on opposite sides of it. This is achieved by maximizing the margin between the closest points, known as support vectors, and the hyperplane. For solubility, however, it is usual to generate quantitative predictions, rather than a binary soluble-insoluble classification, and thus SVM is adapted for regression with the hyperplane now playing the role of a regression line.

Numerous studies have addressed solubility in this way. Lind *et al.* (2003) reported cross-validated RMSEs between 0.57 and 0.77 on different datasets. Palmer *et al.* (2007) found that SVM obtained an RMSE of 0.72 on a 330 compound test set, slightly worse than RF. In contrast Hughes *et al.* (2008) in the same research group found SVM to somewhat outperform RF over 87 molecules. Comparison of these two studies emphasises that the performance of these machine learning methods is similar and that there is unlikely to be a

universally best-performing algorithm. Louis *et al.* (2010) found SVM to do a little better than ANN on a small test set of 14 molecules with an RMSE of 0.832. Cao *et al.* (2010) found a better RMSE, 0.731 over 45 compounds, with SVM than with two other machine learning methods. Kew *et al.* (2015) observed SVM to get an RMSE of 1.01 in a 10-fold cross-validation over 262 compounds taken from Hughes *et al*. (2008), and thus to be essentially joint best with RF of 15 methods for solubility prediction. Boobier *et al*. (2017), however, found SVM to be only the eighth best out of 10 methods for a 75-25 training-test split of the DLS-100 dataset (Mitchell *et al*., 2017) with an RMSE of 1.280 for 25 test compounds.

3.4.2.4 k-Nearest Neighbours

In k-Nearest Neighbours (kNN), the solubility of a query molecule is predicted from the solubilities of its k nearest neighbours in chemical space amongst the available dataset. In general, k is a small integer so that only the local variation of the property in chemical space affects the kNN prediction. This distinguishes it from other machine learning approaches, which attempt to generate global models. Although the predictivity is essentially local, the coverage of diverse chemical structures is as broad as the composition of the dataset. A quantitative prediction is made by averaging the solubilities of the k neighbours; in some applications the average can be weighted by distance (Nigsch *et al*., 2006). The methodology requires a robust measure of distance in chemical space, so descriptors should be appropriately scaled. Hughes *et al.* (2008) found that kNN was less effective than SVM, PLS or RF over 87 test compounds, with RMSE values around 1.10 depending on the descriptor set used. Kew *et al.* (2015) found kNN to be only 12th best of 15 methods for solubility prediction. Boobier *et al*. (2017), in contrast, found kNN to be the fourth best of ten machine learning algorithms, slightly ahead of PLS and SVM amongst others, with an RMSE of 1.204 for 25 molecules. Kühne *et al.* (2006) employed kNN in a more cryptic way, using it to select the most appropriate model for each compound rather than directly to predict solubility.

3.4.2.5 Gaussian Processes

Gaussian processes are Bayesian methods commonly applied in various fields of machine learning. A Gaussian process is a stochastic process, which extends a multivariate normal distribution to an infinite number of random variables. As a result, a Gaussian process effectively represents a probability distribution over functions. The Gaussian process is completely specified by a mean and covariance function (Rasmussen & Williams, 2006). The covariance function can be selected to represent previous knowledge about the data such as periodic patterns or how sharply a function may change between points. Common choices of covariance function include the squared exponential and Martérn covariance functions (Rasmussen & Williams, 2006). Gaussian processes have over the past decade begun to be applied to a range of chemical problems with success (Obrezanova *et al*., 2007; Popelier, 2016; Pyzer-Knapp *et al*., 2016; McDonagh *et al*., 2018).

Gaussian processes have been used effectively to predict ADMET properties and solubility. Obrezanova & Seagall (2010) applied Gaussian process classification to build QSPR models of a variety of ADMET and bioactivity properties, including blood-brain barrier penetration and hERG inhibition. They compared the results to a range of other machine learning classification methods, including RF and SVM, and found that, whilst no method was notably

more successful than the others, Gaussian processes were often the best performing. Schwaighofer *et al*. (2007) developed QSPR solubility models for a range of datasets of electrolyte molecules, achieving promising results and outperforming many of the commercially available solubility prediction packages.

3.2.4.6 Deep Learning

The notion of deep learning (LeCun *et al*., 2015) in the machine learning field comes from the use of large multi-layer neural networks, which can perform abstractions from data and hence intrinsically define features, as opposed to traditional ANNs which correlate fixed feature representations with a given property. The hidden layers are where the decisions are made in a network. Deep learning for pharmaceutical property prediction has seen widening use over the recent years. Deep learning specifically for solubility prediction has, however, been relatively underexplored, though one notable example is the work of Lusci *et al* (2013). In this paper the authors produced a novel two-step approach, using a first ANN to determine the optimal molecular representation encoding chemical structure *in silico*, and a second ANN to find the best mapping function between this representation and solubility. They achieved good predictive statistics over a range of datasets, including cross-validated RMSEs of 0.60 to 0.92 $\log S_0$ units on the 1026-molecule Huuskonen dataset (Huuskonen, 2000) and RMSEs between 1.00 and 1.41 on the Solubility Challenge set (Llinàs *et al.*, 2008; Hopfinger *et al.*, 2009) for variants of their method.

3.2.4.7 Consensus Methods

There is potential to improve the performance of machine learning by taking a consensus of different predictors. RF is itself by design a consensus of different trees, while Bhat *et al.* (2008) created an analogous ensemble of ANNs. Going beyond this, it is equally possible to combine different machine learning approaches to generate a single prediction. Kew *et al*. (2015) showed that a Greedy Ensemble incorporating several other machine learning predictors performed similarly to the best individual algorithm with an RMSE of 0.83 $\log S_0$ units on the Solubility Challenge dataset and 1.13 on the 262 molecule Hughes dataset. Boobier *et al*. (2017) similarly observed that a median-based consensus was essentially as good as the best single method, whose identity would not be apparent in advance. Thus there is scope to leverage the power of diverse algorithms and take advantage of the wisdom of crowds (Galton, 1907; Surowiecki, 2004) with consensus approaches.

4. Conclusions

This chapter has covered a wide range of methods from physical simulations to cheminformatics. Currently, informatics methods provide the best quantitative predictions of solubility in terms of the lowest RMSE and highest R^2. Nonetheless, they are very limited in the physical insight they can offer. Some methods, such as RF, have inbuilt measures of descriptor importance. However, since descriptors are often correlated with one another, the list of the most numerically significant descriptors for predictivity may not be a good guide to physical or chemical importance. Further, these kinds of models demonstrate correlation not causation, and thus do not directly inform us about the physicochemical reasons or mechanisms for some molecules being more soluble than others.

There is no clear or repeatable pattern as to which machine learning methods are most effective at solubility prediction. Indeed, looking at the various references cited herein, RF, SVM and ANN are all competitive with one another, but it is not possible to identify in advance which will be most accurate for any given dataset. However, all of the informatics methods presented herein are data driven and thus have some dependence on the accuracy and reliability of the experimental measurements used to acquire the data. To this end, reliable and carefully curated datasets for solubility and related quantities are of great value to the computational and modelling community. More laboratory data enabling better estimation of the appropriate error bars for experimental solubility values would help us to understand whether informatics methods are now close their limiting accuracy, or whether there remains substantial scope for improvement from new machine learning algorithms, from novel descriptors encoding different information, or from better feature selection.

In terms of physical models, we have presented methods which make good solubility predictions and provide varying levels of physical insight. Much work by those predicting solubility has focused on modelling of the solvent and solution. This has provided a rich choice of models which can be selected for a particular problem. The best choice of model depends upon the physics one is interested in representing. Implicit models are efficient, but provide little information on solvent structure; hybrid methods such as RISM are also efficient and provide limited statistical information on the solvent structure; explicit solvent representation can be expensive but will provide maximal information on the solvent's structure and response to a solute.

Solid state modelling has not received the same level of attention by those predicting solubility as solution phase modelling has, but there is considerable expertise and experience available from CSP. There are many options here also, which range from lattice simulation methods to periodic quantum mechanical calculations. Again, the choice depends on balancing compute time against the level of detail needed to describe the physics of interest. There is much scope in this area to investigate the effects of different solid state modelling techniques on solubility predictions.

Both the solid and solution states require suitable representation for accurate predictions *via* thermodynamic cycles. Predictions from both the fusion cycle and the sublimation cycles have shown similar levels of accuracy, as described by the references presented herein. Over recent years, physics based methods have shown improvements and begun to offer much improved solubility predictions. Among the most pressing priorities is for these methods to be tested on datasets of sufficient size to assess their accuracy and compare their quantitative performance with that of chemoinformatics.

Overall, we foresee a bright future for solubility prediction using both physics based models and informatics methods. Informatics provides excellent opportunities for fast virtual screening, with deep learning offering new opportunities for automated feature extraction. Physical models will continue to provide deeper insights into the chemical and physical phenomena which define a substance's solubility, enabling us to discover rules for designing molecules intelligently to enhance or reduce solubility. The combination of these methods could yield a particularly powerful tool in the future.

References

Abascal, J. L. F.; Sanz, E.; García Fernández, R.; Vega, C. A Potential Model for the Study of Ices and Amorphous Water: TIP4P/Ice. J. Chem. Phys. 2005; 122(23): 234511

Abascal, J. L., & Vega, C. A General Purpose Model for the Condensed Phases of Water: TIP4P/2005. J. Chem. Phys. 2005; 123(23):234505.

Abraham MH, Le J. The correlation and prediction of the solubility of compounds in water using an amended solvation energy relationship. Journal of Pharmaceutical Sciences. 1999; 88:868-880. http://dx.doi.org/10.1021/js9901007

Adamo C. Toward reliable density functional methods without adjustable parameters: The PBE0 model. Journal of Chemical Physics 1999; 110:6158 https://doi.org/10.1063/1.478522

Alderson RG, De Ferrari L, Mavridis L, McDonagh JL, Mitchell JBO, Nath N. Enzyme Informatics. Curr. Top. Med. Chem. 2012; 12(17):1911–1923.

Ali J, Camilleri P, Brown MB, Hutt AJ, Kirton SB. Revisiting the General Solubility Equation: In Silico Prediction of Aqueous Solubility Incorporating the Effect of Topographical Polar Surface Area. J Chem Inf Model 2012; 52:420-428. http://dx.doi.org/10.1021/ci200387c

Alsenz J, Kansy M. High Throughput Solubility Measurement in Drug Discovery and Development. Adv. Drug Deliv. Rev. 2007; 59(7): 546–567.

Avdeef, A. pH-metric Solubility. 1. Solubility-pH Profiles from Bjerrum Plots. Gibbs Buffer and pK_a in the Solid State, *Pharm. Pharmacol. Commun.* 1998, *4(3)*, 165-178.

Bachmann S.J., van Gunsteren W.J. An improved simple polarisable water model for use in biomolecular simulation, J. Chem. Phys., 2014; 141; 22D515

Baranyai A, Bartók A. Classical interaction model for the water molecule.J Chem Phys. 2007; 126(18):184508. https://dx.doi.org/10.1063/1.2730510

Bardwell DA *et al.* Towards crystal structure prediction of complex organic compounds – a report on the fifth blind test. Acta Cryst. B, 2011; B67; 535-551

Baranyai S, Bartók A. Classical interaction model for the water model, J. Chem. Phys. 2007; 126; 184508

Bauer J, Spanton S, Henry R, Quick J, Dziki W, Porter W, Morris J. Ritonavir: An Extraordinary Example of Conformational Polymorphism. Pharm. Res. 2001; 18(6):859–866.

Bejagam KK, Singh S, An Y, Berry C, Deshmukh SA. PSO-Assisted Development of New Transferable Coarse-Grained Water Models. J. Phys. Chem. B 2018; 122(6):1958–1971.

Ben-Naim A. Standard Thermodynamics of Transfer. Uses and Misuses. J. Phys. Chem. 1978; 82 (7):792–803.

Ben-Naim A, Marcus Y. Solvation thermodynamics of nonionic solutes. Journal of Chemical Physics, 1984; 81(4): 2016-2027 http://dx.doi.org/10.1063/1.447824

Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. Interaction Models for Water in Relation to Protein Hydration. In; Springer, Dordrecht, 1981; pp 331–342

Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. The Missing Term in Effective Pair Potentials. J. Phys. Chem. 1987; 91(24):6269–6271

Bergström CAS, Larsson P. Computational prediction of drug solubility in water-based systems: Qualitative and quantitative approaches used in the current drug discovery and development setting. International Journal of Pharmaceutics 2018; 540:185-193 https://dx.doi.org/10.1016/j.ijpharm.2018.01.044

Bhat AU, Merchant SS, Bhagwat SS. Prediction of Melting Points of Organic Compounds Using Extreme Learning Machines. Industrial and Engineering Chemistry Research 2008; 47:920-925. http://dx.doi.org/10.1021/ie0704647

Bhat, T. N.; Bentley, G. A.; Boulot, G.; Greene, M. I.; Tello, D.; Dall'Acqua, W.; Souchon, H.; Schwarz, F. P.; Mariuzza, R. A.; Poljak, R. J. Bound Water Molecules and Conformational Stabilization Help Mediate an Antigen-Antibody Association. Proc. Natl. Acad. Sci. U. S. A. 1994, 91 (3), 1089–1093.

Boobier S, Osbourn A, Mitchell JBO. Can human experts predict solubility better than computers? Journal of Cheminformatics. 2017; 9:63. http://dx.doi.org/10.1186/s13321-017-0250-y

Boothroyd S., Kerridge A., Broo A., Buttar D., Anwar J., Solubility prediction from first principles: a density of states approach, Phys. Chem. Chem. Phys., 2018; 20; 20981-20987 http://dx.doi.org/ 10.1039/C8CP01786G

Box K, Comer JE, Gravestock T, Stuart M. New Ideas about the Solubility of Drugs. Chemistry & Biodiversity 2009; 6(11):1767-1788. http://dx.doi.org/10.1002/cbdv.200900164

Bradley JC, Abraham MH, Acree WE, Lang A. Predicting Abraham model solvent coefficients. Chemistry Central Journal 2015; 9:12. http://dx.doi.org/10.1186/s13065-015-0085-4

Breiman L. Random Forests, Mach. Learn. 2001; 45:5-32. http://dx.doi.org/10.1023/a:1010933404324

Buchholz, H. K., Hylton, R. K., Brandenburg, J. G., Seidel-Morgenstern, A., Lorenz, H., Stein, M., Price, S. L. Thermochemistry of racemic and enantiopure organic crystals for predicting enantiomer separation. Crystal Growth & Design, 2017; 17(9):4676-4686 http://dx.doi.org/10.1021/acs.cgd.7b00582

Cammi R, Tomasi J. Remarks on the use of the apparent surface charges (ASC) methods in solvation problems: Iterative versus matrix-inversion procedures and the renormalization of the apparent charges. Journal of Computational Chemistry 1995; 16(12):1449-1458 http://dx.doi.org/10.1002/jcc.540161202

Cao DS, Xu QS, Liang YZ, Chen X, Li HD. Prediction of aqueous solubility of druglike organic compounds using partial least squares, back-propagation network and support vector machine. Journal of Chemometrics 2010; 24:584-595 http://dx.doi.org/10.1002/cem.1321

Catana C, Gao H, Orrenius C, Stouten PFW. Linear and Nonlinear Methods in Modeling the Aqueous Solubility of Organic Compounds. J Chem Inf Model. 2005; 45:170-176. http://dx.doi.org/10.1021/ci049797u

Chandler D, McCoy JD, Singer SJ. Density functional theory of nonuniform polyatomic systems. II. Rational closures for integral equations. J. Chem. Phys., 1986; 85: 5977-5982 https://doi.org/10.1063/1.451511

Chemburkar, S. R.; Bauer, J.; Deming, K.; Spiwek, H.; Patel, K.; Morris, J.; Henry, R.; Spanton, S.; Dziki, W.; Porter, W.; Quick, J.; Bauer, P.; Donaubauer, J.; Narayanan, B. A.; Soldani, M.; Riley, D.; McFarland, K. Dealing with the Impact of Ritonavir Polymorphs on the Late Stages of Bulk Drug Process Development. Org. Process Res. Dev. 2000; 4(5):413–417.

Comer J, Judge S, Matthews D, Towers L, Falcone B, Goodman J, Dearden J. The intrinsic aqueous solubility of indomethacin. ADMET & DMPK. 2014 http://dx.doi.org/10.5599/admet.2.1.33

Connors BW, Long MA. Electrical synapses in the mammalian brain. Annu Rev Neurosci. 2004, 27:393-418

Cramer, C. J.; Truhlar, D. G. Implicit Solvation Models: Equilibria, Structure, Spectra, and Dynamics. Chem. Rev. 1999; 99(8):2161–2200.

Cros AFA. *Action de l'alcool amylique sur l'organisme*. PhD Thesis, University of Strasbourg, 1863.

Crum Brown A, Fraser TR. On the connection between chemical constitution and physiological action. Part. I.—On the physiological action of the salts of the ammonium bases, derived from strychnia, brucia, thebaia, codeia, morphia, and nicotia, Trans. R. Soc. Edinb. 1868; 25:151–203, http://dx.doi.org/10.1017/S0080456800028155

Curtis F, Rose T, Marom N. Evolutionary Niching in the GAtor Genetic Algorithm for Molecular Crystal Structure Prediction. Faraday Discussions, 2018; 211:61-77 http://dx.doi.org/10.1039/C8FD00067K

Day G *et al.,* Significant progress in predicting the crystal structures of small organic molecules – a report on the fourth blind test. Acta Cryst. B., 2009; B65; 107-125

Day G *et al.,* A third blind test of crystal structure prediction Acta Cryst. B., 2005; B61; 511-527

Dearden JC. In silico prediction of aqueous solubility. Expert Opin Drug Disc 2006; 1:31-52.

Doerr-MacEwen, N. A.; Haight, M. E. Expert Stakeholders' Views on the Management of

Human Pharmaceuticals in the Environment. Environ. Manage. 2006; 38(5):853–866.

Dyer KD, Perkyns JS, Stell G, Pettit MB. Site-renormalised molecular fluid theory: on the utility of a two-site model of water. Mol. Phys., 2009; 107; 423-431

Erić S, Kalinić M, Popović A, Zloh M, Kuzmanovski I. Prediction of aqueous solubility of drug-like molecules using a novel algorithm for automatic adjustment of relative importance of descriptors implemented in counter-propagation artificial neural networks. International Journal of Pharmaceutics 2012; 437:232-241. http://dx.doi.org/10.1016/j.ijpharm.2012.08.022

Erickson L. The solubility of homologous series of organic compounds. Naturwiss 1952; 39:41-42

Etherson, K.; Halbert, G.; Elliott, M. Determination of Excipient Based Solubility Increases Using the CheqSol Method. Int. J. Pharm. 2014; 465 (1–2): 202–209.

FDA (US Food and Drug Administration), The Biopharmaceutics Classification System (BCS) Guidance, 2014.

Fühner H. Water-solubility in homologous series. Ber. Dtsch. Chem. Ges. 1924; 57B:510-515.

Galton F. Vox populi. Nature 1907; 75: 450-451 https://dx.doi.org/10.1038/075450a0

Genheden S, Luchko T, Gusarov S, Kovalenko A, Ryde U. An MM/3D-RISM Approach for Ligand Binding Affinities. J Phys Chem B. 2010; 114(25): 8505-8516 http://dx.doi.org/10.1021/jp101461s

Gracin S, Brinck T, Rasmuson ÃC. Prediction of Solubility of Solid Organic Compounds in Solvents by UNIFAC. Industrial & Engineering Chemistry Research 2002; 41 :5114-5124 http://dx.doi.org/10.1021/ie011014w

Gütlein M, Helma C, Karwath A, Kramer S. A Large-Scale Empirical Evaluation of Cross-Validation and External Test Set Validation in (Q)SAR. Molecular Informatics 2013; 32(5-6):516-528 http://dx.doi.org/10.1002/minf.201200134

Hall LH, Kier LB. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling. Rev. Comput. Chem. 2007; 2:367–422.

Handley, C. M., Hawe, G. I., Kell, D. B., Popelier, P. L. Optimal Construction of a Fast and Accurate Polarisable Water Potential Based on Multipole Moments Trained by Machine Learning. Phys. Chem. Chem. Phys. 2009; 11(30):6365–6376

Hansch C, Quinlan JE, Lawrence GL. Linear free-energy relationship between partition coefficients and the aqueous solubility of organic liquids. The Journal of Organic Chemistry 1968; 33:347-350. http://dx.doi.org/10.1021/jo01265a071

Hewitt M, Cronin MTD, Enoch SJ, Madden JC, Roberts DW, Dearden JC. In Silico Prediction of Aqueous Solubility: The Solubility Challenge. J. Chem. Inf. Model. 2009; 49:2572-2587

https://dx.doi.org/10.1021/ci900286s

Hogan SW, van Mourik T. Competition between hydrogen and halogen bonding in halogenated 1-methyluracil: Water systems. Journal of computational chemistry. 2016 Mar 30;37(8):763-70. https://doi.org/10.1002/jcc.24264

Hoja J, Tkatchenko A. First-principles stability ranking of molecular crystal polymorphs with the DFT + MBD approach. Faraday Discussions, 2018; 211:253-274 http://dx.doi.org/10.1039/C8FD00066B

Hopfinger AJ, Esposito EX, Llinàs A, Glen RC, Goodman JM. Findings of the Challenge to Predict Aqueous Solubility. Journal of Chemical Information and Modeling, 2009; 49:1-5 https://dx.doi.org/10.1021/ci800436c

Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. Development of an Improved Four-Site Water Model for Biomolecular Simulations: TIP4P-Ew. J. Chem. Phys. 2004; 120(20):9665–9678

Hörter, D.; Dressman, J. . Influence of Physicochemical Properties on Dissolution of Drugs in the Gastrointestinal Tract. Adv. Drug Deliv. Rev. 2001; 46 (1–3):75–87.

Hou TJ, Xia K, Zhang W, Xu XJ. ADME Evaluation in Drug Discovery. 4. Prediction of Aqueous Solubility Based on Atom Contribution Approach. Journal of Chemical Information and Computer Sciences 2004; 44:266-275 http://dx.doi.org/10.1021/ci034184n

Huan G, Xu T, Momen R, Wang L, Ping Y, Kirk SR, Jenkins S, van Mourik T. A QTAIM exploration of the competition between hydrogen and halogen bonding in halogenated 1-methyluracil: Water systems. Chemical Physics Letters. 2016; 662:67-72. https://doi.org/10.1016/j.cplett.2016.09.031

Hughes LD, Palmer DS, Nigsch F, Mitchell JBO. Why Are Some Properties More Difficult To Predict than Others? A Study of QSPR Models of Solubility, Melting Point, and Log P. Journal of Chemical Information and Modeling 2008; 48:220-232. http://dx.doi.org/10.1021/ci700307p

Hutchinson, T. C.; Hellebust, J. A.; Mackay, D.; Tarn, D.; Kauss, P. Relationship Of Hydrocarbon Solubility To Toxicity In Algae And Cellular Membrane Effects. Int. Oil Spill Conf. Proc. 1979; 1:541–547.

Huuskonen J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. Journal of Chemical Information and Computer Sciences 2000; 40:773-777 http://dx.doi.org/10.1021/ci9901338

Iuzzolino L, McCabe P, Price SL, Brandenburg JG. Crystal structure prediction of flexible pharmaceutical-like molecules: density functional tight-binding as an intermediate optimisation method and for free energy estimation Faraday Discussions 2018 https://dx.doi.org/10.1039/c8fd00010g

Jones, A.; Cipcigan, F.; Sokhan, V. P.; Crain, J.; Martyna, G. J. Electronically Coarse-Grained

Model for Water. Phys. Rev. Lett. 2013; 110 (22):227801.

Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. J. Chem. Phys. 1983; 79(2):926–935.

Jorgensen WL, Tirado-Rives J. Monte Carlo vs Molecular Dynamics for Conformational Sampling. Journal of Physical Chemistry 1996; 100(34):14508-14513 http://dx.doi.org/10.1021/jp960880x

Jorgensen, W. L.; Tirado-Rives, J. Molecular Modeling of Organic and Biomolecular Systems Using BOSS and MCPRO. J. Comput. Chem. 2005; 26(16):1689–1700. https://dx.doi.org/10.1002/jcc.20297

Joung IS, Cheatham TE. Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. Journal of Physical Chemistry B 2008; 112: 9020-9041. https://doi.org/10.1021/jp8001614

Jouyban A, Fakhree M. Experimental and Computational Methods Pertaining to Drug Solubility, in *Toxicity and Drug Testing*, Acree W Ed, Intech Open, London, 2012.

Kew W, Mitchell JBO. Greedy and Linear Ensembles of Machine Learning Methods Outperform Single Approaches for QSPR Regression Problems. Mol. Inf. 2015, 34:634-647 (2015) http://dx.doi.org/10.1002/minf.201400122

Kiss PT, Baranyai A. A systematic development of a polarizable potential of water. J Chem Phys 2013; 138(20):204507 https://dx.doi.org/10.1063/1.4807600

Klamt A., Schüürmann G. COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. J. Chem. Soc. Perkin Trans. 2, 1993; 0(5); 799-805

Klamt A. The COSMO and COSMO-RS solvation models. WIREs Comput Mol Sci. 2011; 1(5):699-709 http://dx.doi.org/10.1002/wcms.56

Klamt A, Eckert F, Diedenhofen M. Prediction of the Free Energy of Hydration of a Challenging Set of Pesticide-Like Compounds. J Phys Chem B. 2009; 113(14):4508-4510 http://dx.doi.org/10.1021/jp805853y

Klopman G, Wang S, Balthasar DM. Estimation of aqueous solubility of organic molecules by the group contribution approach. Application to the study of biodegradation. Journal of Chemical Information and Computer Sciences 1992; 32:474-482 http://dx.doi.org/10.1021/ci00009a013

Klopman G, Zhu H. Estimation of the Aqueous Solubility of Organic Molecules by the Group Contribution Approach. J Chem Inf Comput Sci. 2001; 41:439-445 http://dx.doi.org/10.1021/ci000152d

Kolafa J. Solubility of NaCl in water and its melting point by molecular dynamics in the slab geometry and a new BK3-compatible force field. Journal of Chemical Physics 2016; 145: 204509. http://dx.doi.org/10.1063/1.4968045

Kovalenko A, Hirata F. Self-consistent description of a metal–water interface by the Kohn–Sham density functional theory and the three-dimensional reference interaction site model. Journal of Chemical Physics. 1999; 110(20): 10095-10112 http://dx.doi.org/10.1063/1.478883

Kovdienko NA, Polishchuk PG, Muratov EN, Artemenko AG, Kuz'min VE, Gorb L, *et al*. Application of Random Forest and Multiple Linear Regression Techniques to QSPR Prediction of an Aqueous Solubility for Military Compounds. Molecular Informatics 2010; 29:394-406. http://dx.doi.org/10.1002/minf.201000001

Kühne R, Ebert RU, Schuurmann G. Model Selection Based on Structural Similarity-Method Description and Application to Water Solubility Prediction. J Chem Inf Model. 2006; 46:636-641 http://dx.doi.org/10.1021/ci0503762

Lamoureux G., MacKerrel Jr. A.D., Roux B., A simple polarizable model of water based on classical Drude oscillators. Journal of Chemical Physics 2003; 119(10):5185–5197

Leach AG, Jones HD, Cosgrove DA, Kenny PW, Ruston L, MacFaul P, et al. Matched Molecular Pairs as a Guide in the Optimization of Pharmaceutical Properties; a Study of Aqueous Solubility, Plasma Protein Binding and Oral Exposure. Journal of Medicinal Chemistry 2006; 49(23):6672-6682 http://dx.doi.org/10.1021/jm0605233

LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015; 521(7553):436-444 http://dx.doi.org/10.1038/nature14539

Li L, Totton T, Frenkel D. Computational methodology for solubility prediction: Application to the sparingly soluble solutes. Journal of Chemical Physics 2017; 146:214110 http://dx.doi.org/10.1063/1.4983754

Li L, Totton T, Frenkel D. Computational methodology for solubility prediction: Application to sparingly soluble organic/inorganic materials. Journal of Chemical Physics 2018; 149:054102 https://dx.doi.org/10.1063/1.5040366

Lind P, Maltseva T. Support Vector Machines for the Estimation of Aqueous Solubility. J Chem Inf Comput Sci. 2003; 43:1855-1859. http://dx.doi.org/10.1021/ci034107s

Lipinski, C. A; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Developmental Settings. Adv. Drug Deliv. Rev. 1997; 23:3–25.

Llinàs A, Glen RC, Goodman JM. Solubility Challenge: Can You Predict Solubilities of 32 Molecules Using a Database of 100 Reliable Measurements? J Chem Inf Model. 2008; 48:1289-1303 http://dx.doi.org/10.1021/ci800058v

Lommerse, JP *et al.* A test of crystal structure prediction of small organic molecules, Acta Cryst. B., 2000; B56; 687-714

Louis B, Agrawal VK, Khadikar PV. Prediction of intrinsic solubility of generic drugs using MLR, ANN and SVM analyses. European Journal of Medicinal Chemistry 2010; 45:4018-4025. http://dx.doi.org/10.1016/j.ejmech.2010.05.059

Luccarelli, J.; Michel, J.; Tirado-Rives, J.; Jorgensen, W. L. Effects of Water Placement on Predictions of Binding Affinities for P38α MAP Kinase Inhibitors. J. Chem. Theory Comput. 2010; 6 (12):3850–3856.

Lüder K, Lindfors L, Westergren J, Nordholm S, Kjellander R. In Silico Prediction of Drug Solubility: 2. Free Energy of Solvation in Pure Melts. Journal of Physical Chemistry B. 2007a; 111(7): 1883-1892 http://dx.doi.org/10.1021/jp0642239

Lüder K, Lindfors L, Westergren J, Nordholm S, Kjellander R. In Silico Prediction of Drug Solubility. 3. Free Energy of Solvation in Pure Amorphous Matter. Journal of Physical Chemistry B. 2007b; 111(25): 7303-7311 http://dx.doi.org/10.1021/jp071687d

Lüder K, Lindfors L, Westergren J, Nordholm S, Persson R, Pedersen M. In silico prediction of drug solubility: 4. Will simple potentials suffice? Journal of Computational Chemistry 2009; 30(12):859-1871 http://dx.doi.org/10.1002/jcc.21173

Lusci A, Pollastri G, Baldi P. Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. J. Chem. Inf. Model, 2013; 53(7):1563-1575 http://dx.doi.org/10.1021/ci400187y

Mahoney, M. W.; Jorgensen, W. L. A Five-Site Model for Liquid Water and the Reproduction of the Density Anomaly by Rigid, Nonpolarizable Potential Functions. J. Chem. Phys. 2000; 112 (20):8910

Marenich AV, Cramer CJ, Truhlar DG. Performance of SM6, SM8, and SMD on the SAMPL1 Test Set for the Prediction of Small-Molecule Solvation Free Energies. Journal of Physical Chemistry B. 2009; 113(14):4538-4543 http://dx.doi.org/10.1021/jp809094y

Marenich AV, Cramer CJ, Truhlar DG. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions, J. Phys. Chem. B. 2009; 113(18); 6378-6396

Mark P, Nilsson L. Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. Journal of Physical Chemistry A, 2001; 105(43):9954-9960 http://dx.doi.org/10.1021/jp003020w

Marrero J, Gani R. Group-Contribution-Based Estimation of Octanol/Water Partition Coefficient and Aqueous Solubility. Industrial & Engineering Chemistry Research 2002; 41:6623-6633 http://dx.doi.org/10.1021/ie0205290

McDonagh JL, Nath N, De Ferrari L, van Mourik T, Mitchell JBO. Uniting Cheminformatics and Chemical Theory to Predict the Intrinsic Aqueous Solubility of Crystalline Druglike Molecules. J Chem Inf Model. 2014; 54:844-856 http://dx.doi.org/10.1021/ci4005805

McDonagh JL, Silva AF, Vincent MA, Popelier PLA. Machine Learning of Dynamic Electron Correlation Energies from Topological Atoms. Journal of Chemical Theory and Computation 2018; 14(1):216-224. http://dx.doi.org/10.1021/acs.jctc.7b01157

McDonagh JL, van Mourik T, Mitchell JBO. Predicting Melting Points of Organic Molecules: Applications to Aqueous Solubility Prediction Using the General Solubility Equation. Mol Inf. 2015; 34:715-724. http://dx.doi.org/10.1002/minf.201500052

McQuarrie DA, Simon JD. Physical Chemistry: A Molecular Approach. University Science Books, Sausalito, California 1997.

Miertuš S., Scrocco E., Tomasi J. Electrostatic interaction of a solute with a continuum. A direct utilization of ab initio molecular potentials for the prevision of solvent effects; Chem. Phys. 1981; 55(1); 117-129

Misin, M., Fedorov M. V., Palmer D. S. Accurate Hydration Free Energies at a wide range of temperatures from 3D RISM; J. Chem. Phys. 2015; 142; 091105. http://dx.doi.org/10.1063/1.4914315

Misin, M., Fedorov M. V., Palmer D. S. Hydration Free Energies of Molecular Ions from Theory and Simulation J. Phys. Chem. B 2016; 120; 975–983. http:/dx.doi.org/10.1021/acs.jpcb.5b10809

Mitchell JBO. Machine learning methods in chemoinformatics. WIREs Comput Mol Sci. 2014; 4(5):468-481. http://dx.doi.org/10.1002/wcms.1183

Mitchell JBO, McDonagh JD, Boobier S. DLS-100 solubility dataset, University of St Andrews 2017. http://dx.doi.org/10.17630/3a3a5abc-8458-4924-8e6c-b804347605e8

Mobley DL, Bayly CI, Cooper MD, Dill KA. Predictions of Hydration Free Energies from All-Atom Molecular Dynamics Simulations. Journal of Physical Chemistry B. 2009; 113(14):4533-4537 http://dx.doi.org/10.1021/jp806838b

Motherwell WD *et al.,* Crystal structure prediction of small organic molecules: a second blind test, Acta Cryst. B., 2002; B58; 647-661

Moučka F, Nezbeda I, Smith WR. Chemical Potentials, Activity Coefficients, and Solubility in Aqueous NaCl Solutions: Prediction by Polarizable Force Fields. Journal of Chemical Theory and Computation 2015; 11(4): 1756-1764 http://dx.doi.org/10.1021/acs.jctc.5b00018

Mulholland, A. J. Modelling Enzyme Reaction Mechanisms, Specificity and Catalysis. Drug Discov. Today 2005; 10 (20):1393–1402.

Moustafa SG, Schultz AJ, Kofke DA. A comparative study of methods to compute the free energy of an ordered assembly by molecular simulation. Journal of Chemical Physics 2013;

139:084105 https://doi.org/10.1063/1.4818990

Narasimham, L. Y. S.; Barhate, V. D. Kinetic and Intrinsic Solubility Determination of Some Beta-Blockers and Antidiabetics by Potentiometry. J. Pharm. Res. 2011; 4 (2):532–536.

Neumann M, van de Streek J. How many Ritonavir cases are there still out there? Faraday Discussions, 2018, 211:441-458 https://dx.doi.org/10.1039/C8FD00069G

Nigsch F, Bender A, Buuren B, Tissen J, Nigsch E, Mitchell JBO. Melting Point Prediction Employing k-Nearest Neighbor Algorithms and Genetic Parameter Optimization J. Chem. Inf. Model. 2006; 46:2412-2422
https://dx.doi.org/10.1021/ci060149f

Noble WS. What is a support vector machine? Nat. Biotech. 2006; 24:1565-1567
http://dx.doi.org/10.1038/nbt1206-1565

Noyes, A. A; Whitney, W. R. The Rate of Solution of Solid Substances in Their Own Solutions. J. Am. Chem. Soc. 1897; 19 (12):930–934.

Obrezanova O, Csányi G, Gola JMR, Segall MD. Gaussian Processes: A Method for Automatic QSAR Modeling of ADME Properties. Journal of Chemical Information and Modeling 2007; 47(5):1847–1857 http://dx.doi.org/10.1021/ci7000633

Obrezanova O, Segall MD. Gaussian Processes for Classification: QSAR Modeling of ADMET and Target Activity. Journal of Chemical Information and Modeling 2010; 50(6):1053-1061.
http://dx.doi.org/10.1021/ci900406x

Palmer DS, Frolov AI, Ratkova EL, Fedorov MV. Towards a universal method for calculating hydration free energies: a 3D reference interaction site model with partial molar volume correction. Journal of Physics: Condensed Matter 2010; 22(49): 492101
http://dx.doi.org/10.1021/jp111271c

Palmer DS, Llinàs A, Morao I, Day GM, Goodman JM, Glen RC, Mitchell JBO. Predicting Intrinsic Aqueous Solubility by a Thermodynamic Cycle. Mol Pharmaceutics. 2008; 5(2): 266-279 http://dx.doi.org/10.1021/mp7000878

Palmer DS, McDonagh JL, Mitchell JBO, van Mourik T, Fedorov MV. First-Principles Calculation of the Intrinsic Aqueous Solubility of Crystalline Druglike Molecules. J Chem Theory Comput. 2012; 8(9): 3322-3337 http://dx.doi.org/10.1021/ct300345m

Palmer DS, Mitchell JBO. Is Experimental Data Quality the Limiting Factor in Predicting the Aqueous Solubility of Druglike Molecules? Molecular Pharmaceutics 2014; 11:2962-2972.
http://dx.doi.org/10.1021/mp500103r

Palmer DS, O'Boyle NM, Glen RC, Mitchell JBO. Random Forest Models To Predict Aqueous Solubility. Journal of Chemical Information and Modeling. 2007; 47:150-158.
http://dx.doi.org/10.1021/ci060164k

Paluch AS, Jayaraman S, Shah JK, Maginn EJ. A method for computing the solubility limit of solids: Application to sodium chloride in water and alcohols. Journal of Chemical Physics 2010; 133(12): 124504 http://dx.doi.org/10.1063/1.3478539

Paricaud P., Predota M., Chialvo A.A., Cummings P.T. From dimer to condensed phases at extreme conditions: accurate predictions of the properties of water by a Gaussian charge polarizable model. J. Chem. Phys. 2005; 122(24); 24451

Po HN, Senozan NM. The Henderson-Hasselbalch Equation: Its History and Limitations. J. Chem. Educ. 2001; 78 (11):1499.

Popelier PLA. Molecular simulation by knowledgeable quantum atoms. Physica Scripta 2016; 91:033007 http://dx.doi.org/10.1088/0031-8949/91/3/033007

Praprotnik M, Janezic D, Mavri J. Temperature Dependence of Water Vibrational Spectrum: A Molecular Dynamics Simulation Study. J Phys Chem A 2004; 108(50):11056-11062 http://dx.doi.org/10.1021/jp046158d

Price SL, Leslie M, Welch GWA, Habgood M, Price LS, Karamertzanis PG, Day GM. Modelling organic crystal structures using distributed multipole and polarizability-based model intermolecular potentials. Phys Chem Chem Phys 2010; 12(30): 8478-8490 http://dx.doi.org/10.1039/c004164e

Pyzer-Knapp EO, Simm GN, Aspuru Guzik A. A Bayesian approach to calibrating high-throughput virtual screening results and application to organic photovoltaic materials. Mater Horizons 2016; 3:226–233 http://dx.doi.org/10.1039/C5MH00282F

Raileanu L, Stoffel K. Theoretical Comparison between the Gini Index and Information Gain Criteria. Annals of Mathematics and Artificial Intelligence 2004; 41:77-93. https://doi.org/10.1023/B:AMAI.0000018580.96245.c6

Ran Y, Yalkowsky SH. Prediction of Drug Solubility by the General Solubility Equation (GSE). J Chem Inf Comput Sci. 2001; 41: 354-357. http://dx.doi.org/10.1021/ci000338c

Rasmussen CE, Williams CKI. Gaussian Processes for Machine Learning, MIT Press, 2006, ISBN 026218253X. http://www.gaussianprocess.org/gpml/chapters/RW.pdf

Ratkova, E. L.; Palmer, D. S.; Fedorov, M. V. Solvation Thermodynamics of Organic Molecules by the Molecular Integral Equation Theory: Approaching Chemical Accuracy. Chem. Rev. 2015; 115(13):6312–6356.

Reilly A.M. *et al.,* Report on the sixth blind test of crystal structure prediction methods, Acta. Cryst. B., 2016; B72; 439-459

Ren P, Ponder JW. Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation. J. Phys. Chem. B 2003; 107(24):5933–5947.

Saal, C.; Petereit, A. C. Optimizing Solubility: Kinetic versus Thermodynamic Solubility Temptations and Risks. Eur. J. Pharm. Sci. 2012; 47 (3):589–595.

Sanghvi T, Jain N, Yang G, Yalkowsky S. Estimation of Aqueous Solubility by the General Solubility Equation (GSE) The Easy Way. QSAR & Combinatorial Science 2003; 22:258-262 http://dx.doi.org/10.1002/qsar.200390020

Sanz E, Vega C. Solubility of KF and NaCl in water by molecular simulation. Journal of Chemical Physics 2007; 126(1):014507 http://dx.doi.org/10.1063/1.2397683

Schroeter TS, Schwaighofer A, Mika S, Ter Laak A, Suelzle D, Ganzer U, *et al*. 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, 2007; 21:485-498. http://dx.doi.org/10.1007/s10822-007-9125-z

Schwaighofer A, Schroeter T, Mika S, Laub J, ter Laak A, Sülzle D, *et al*. Accurate Solubility Prediction with Error Bars for Electrolytes: A Machine Learning Approach. Journal of Chemical Information and Modeling 2007; 47(2):407-424. http://dx.doi.org/10.1021/ci600205g

Sell, C. S. The Mechanism of Olfaction. In *Chemistry and the Sense of Smell*; John Wiley & Sons, Inc., 2014; pp 32–187.

Senn HM, O'Hagan D, Thiel W. Insight into Enzymatic C-F Bond Formation from QM and QM/MM Calculations, J. Am. Chem. Soc. 2005; 127:13643-13655 https://dx.doi.org/10.1021/ja053875s

Sergiievskyi, V., Jeanmairet, G., Levesque, M., Borgis, D. Solvation Free-Energy Pressure Corrections in the Three Dimensional Reference Interaction Site Model. J. Chem. Phys. 2015; 143; 184116. https://doi.org/10.1063/1.4935065

Sharma, V.; Goswami, R.; Madan, A. K. Eccentric Connectivity Index: A Novel Highly Discriminating Topological Descriptor for Structure−Property and Structure−Activity Studies. J. Chem. Inf. Comput. Sci. 1997; 37 (2):273–282.

Skyner, R. E.; McDonagh, J. L.; Groom, C. R.; van Mourik, T.; Mitchell, J. B. O. A Review of Methods for the Calculation of Solution Free Energies and the Modelling of Systems in Solution. Phys. Chem. Chem. Phys. 2015; 17(9):6174–6191. https://dx.doi.org/10.1039/c5cp00288e

Stone AJ. Distributed Multipole Analysis, or how to describe a molecular charge-distribution, Chemical Physics Letters 1981; 83(2):233-239 https://dx.doi.org/10.1016/0009-2614(81)85452-8

Straatsma TP, McCammon JA. Molecular dynamics simulations with interaction potentials including polarization development of a noniterative method and application to water. Molecular Simulation, 1990; 5; 181-192

Stuart M, Box K. Chasing Equilibrium: Measuring the Intrinsic Solubility of Weak Acids and Bases, Anal. Chem, 2005; 77, 983–990

Sulea T, Wanapun D, Dennis S, Purisima EO. Prediction of SAMPL-1 Hydration Free Energies Using a Continuum Electrostatics-Dispersion Model. J. Phys. Chem. B, 2009; 113(14):4511–4520 http://dx.doi.org/10.1021/jp8061477

Surowiecki J. The wisdom of crowds: why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations. Doubleday, New York 2004

Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. J. Chem. Inf. Comput. Sci. 2003; 43:1947-1958. http://dx.doi.org/10.1021/ci034160g

Tomasi J, Mennucci B, Cammi R. Quantum Mechanical Continuum Solvation Models. Chem. Rev. 2005; 105(8):2999–3093.

Toukan K, Rahman A. Molecular-dynamics study of atomic motions in water. Phys. Rev. B 1985; 31(5):2643 https://dx.doi.org/10.1103/PhysRevB.31.2643

Vapnik VN. Statistical Learning Theory. Wiley, New York, 1998.

Wang J, Krudy G, Hou T, Zhang W, Holland G, Xu X. Development of Reliable Aqueous Solubility Models and Their Application in Druglike Analysis. Journal of Chemical Information and Modeling. 2007; 47:1395-1404 http://dx.doi.org/10.1021/ci700096r

Warshel A, Levitt M. Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. J. Mol. Biol. 1976; 2(15); 227-249

Werbos PJ. Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences, PhD thesis, Harvard University, 1975.

Westergren J, Lindfors L, Höglund T, Lüder K, Nordholm S, Kjellander R. In Silico Prediction of Drug Solubility: 1. Free Energy of Hydration. Journal of Physical Chemistry B 2007; 111(7): 1872-1882 http://dx.doi.org/10.1021/jp064220w

Wu J. Classical Density Functional Theory for Molecular Systems. In: Wu J. (eds) Variational Methods in Molecular Modeling. Molecular Modeling and Simulation (Applications and Perspectives). Springer, Singapore, 2017.

Wu, Y., Tepper H.L., Voth, G.G. Flexible simple point-charge water model with improved liquid state properties. J. Chem. Phys. 2006; 124; 024503

Yesylevskyy, S. O., Schäfer, L. V., Sengupta, D., & Marrink, S. J. Polarizable Water Model for the Coarse-Grained MARTINI Force Field. PLoS Computational Biology. 2010; 6(6):e1000810.

Yousefinejad S, Hemmateenejad B. Chemometrics tools in QSAR/QSPR studies: A historical perspective. Chemometrics and Intelligent Laboratory Systems 2015; 149:177-204. http://dx.doi.org/10.1016/j.chemolab.2015.06.016

Yu H., Hansson T. van Gunsteren W.F, Development of a simple self-consistent polarizable model for water. J. Chem. Phys, 2003; 118; 221-234

Yu H., Hansson T. van Gunsteren W.F, Charge-on-spring polarizable water models revisited: From water clusters to liquid water to ice. J. Chem. Phys., 2004; 121; 9549–9564