PCCP

PERSPECTIVE

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2015, **17**, 6174

A review of methods for the calculation of solution free energies and the modelling of systems in solution

R. E. Skyner,†[a] J. L. McDonagh,†[a] C. R. Groom,[b] T. van Mourik[a] and J. B. O. Mitchell*[a]

Received 16th January 2015,
Accepted 23rd January 2015

DOI: 10.1039/c5cp00288e

www.rsc.org/pccp

Over the past decade, pharmaceutical companies have seen a decline in the number of drug candidates successfully passing through clinical trials, though billions are still spent on drug development. Poor aqueous solubility leads to low bio-availability, reducing pharmaceutical effectiveness. The human cost of inefficient drug candidate testing is of great medical concern, with fewer drugs making it to the production line, slowing the development of new treatments. In biochemistry and biophysics, water mediated reactions and interactions within active sites and protein pockets are an active area of research, in which methods for modelling solvated systems are continually pushed to their limits. Here, we discuss a multitude of methods aimed towards solvent modelling and solubility prediction, aiming to inform the reader of the options available, and outlining the various advantages and disadvantages of each approach.

1. Introduction

Poor aqueous solubility is a major cause of attrition (failure) in the pharmaceutical development process and remains a vital property to quantify in the development of agrochemicals, and in the identification and quantification both of metabolites and of potential environmental contaminants. It is estimated that around 70% of pharmaceuticals in development are poorly soluble with 40% of those currently approved also being poorly soluble.[1,2] Solubility is determined by structural and energetic components emanating from solid phase structure and packing interactions, in addition to relevant solute–solvent interactions and structural reorganisation in solution. In this review, we focus on the methods currently available to model the solution phase and to predict solubility for a wide range of applications, including ligand binding, molecular property prediction and molecular design.[3] Readers specifically interested in solubility prediction are also referred to the solubility challenge.[4]

Accurate and timely prediction of solubility could save time and money in drug development, agrochemical development and environmental monitoring. An early-stage analysis of drug and agrochemical candidates allows organisations to focus on those molecules most likely to meet their required solubility criteria. Many models exist in this area, with differing levels of accuracy, physical interpretability, and calculation time.

Quantitative Structure Activity Relationships (QSAR) and Quantitative Structure Property Relationships (QSPR) are very successful in this field, providing good predictive results at a reasonably low computational cost. These models, however, tend to be limited to molecules similar to those used in their training set. Moreover, these models lack a full physical interpretation, although some do allow assessments of descriptor importance that can perhaps to some extent be physically interpreted.

Several fitted or derived general equations, which take only a few pieces of empirical data as arguments, have also been produced. One of the most successful is the General Solubility Equation (GSE),[5] taking the melting point and the base ten logarithm of the partition coefficient (log *P*; partition coefficient for neutral molecules in octanol and water) as empirical input.

The field has also seen the revival of old ideas as new automated data driven design protocols, such as Matched Molecular Pair Analysis (MMPA).[6] MMPA allows one to acquire previously 'unknown' data from existing data sets by exploring how a single molecular change can impact a particular property or activity of interest. We now see large scale data mining following these kinds of protocols, consortia such as SALT MINER, and programs developed by individual companies such as GSK's BioDig.[7,8]

In addition to these approaches, we see physics based models ranging from classical simulations to quantum chemical calculations being applied to solubility prediction. These methods vary greatly in complexity. Classical simulations can encompass simple

[a] *School of Chemistry, University of St Andrews, Purdie Building, North Haugh, St Andrews, Fife, KY16 9ST, UK. E-mail: jbom@st-andrews.ac.uk*
[b] *Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, CB2 1EZ, UK*
† Signifies that these authors contributed equally to this work.

Molecular Dynamics (MD), studying the interactions between solute and solvent, to more complex perturbations of solutes from the solution phase to the gas phase. Recent advances have seen a new generation of polarisable force fields emerging with a greater capacity to account for changes in the electronic charge distribution. Many of these force fields utilise multipole moments, as opposed to point charges, to capture the anisotropy of the charge distribution. Force fields such as Atomic Multipole Optimised Energetics for Biomolecular Applications (AMOEBA) have been used to study the solvation dynamics of ions.[9] Newer, polarisable force fields, such as the quantum chemical topology force field (QCTFF), use multipolar electrostatics calculated based on quantum chemical topology, supplemented with machine learning (Kriging) to model the system. This force field has been used to model amino acids with small water clusters.[10] Some force fields can be mixed with a quantum chemical core region in mixed Quantum Mechanics–Molecular Mechanics (QM/MM) approaches.

Other common models include those representing the solvent as a continuous field with no explicit solvent coordinates. In most cases, these models come at much higher computational cost than their informatics counterparts, and often at lower accuracy. However, if such a method were feasible and accurate enough to predict solubility, it would not have a domain of applicability restricted by the molecules within a training set and would also be physically interpretable. Thus, there is a continuing search for such physical methods. These methods have proven useful for modelling or approximating the solution phase, hence their applications are diverse and widespread outside of solubility prediction.

1.1 Thermodynamics and solubility

A solution is considered as an equilibrium between solute and solvent, reaching equilibrium when the number of molecules transferred from the solution to a non-solute state is equal to the transfer of molecules from a non-solute state to solution, *i.e.* when the forward rate is equal to the backward rate and both phases are in equilibrium. Solubility is a quantitative term, most simply describing the amount of a substance that will dissolve in a given amount of solvent, and is a property of thermodynamic equilibrium. A second process involved in solvation is dissolution; a kinetic term describing the rate at which a substance is transferred from a non-solute phase into solution. Solubility and dissolution are fundamental terms describing the process of solvation, and are related by the Noyes–Whitney equation;[11]

$$\frac{\mathrm{d}W}{\mathrm{d}t} = \frac{kA(C_\mathrm{s} - C)}{L} \quad (1)$$

where $\mathrm{d}W/\mathrm{d}t$ is the rate of dissolution, A is the solute surface area in contact with the solvent, C is 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.

As solubility is a thermodynamic term, it is inherently affected by factors such as temperature and pressure, as well as ionisation, solid state effects, and gaseous partial pressure for solvated gases.

pH is considered to have a significant effect on solubility, as many organic molecules can behave as weak acids or weak bases, due to ionisable basic or acidic functional groups, with polarisation of ionisable groups in solution increasing or decreasing the overall solubility. The pH of the aqueous solution in which such molecules are dissolved determines whether the molecule exists in its neutral or ionised form. The charged form of a molecule is more soluble, and thus the aqueous solubility of a substance is pH-dependent.[12] This dependence is described by the Henderson–Hasselbalch (HH) equations as follows;

$$\begin{aligned} \log S_\mathrm{total}^\mathrm{acidic} &= \log S_0 + \log\left(1 + 10^{\mathrm{pH}-\mathrm{p}K_\mathrm{a}}\right) \\ \log S_\mathrm{total}^\mathrm{basic} &= \log S_0 + \log\left(1 + 10^{\mathrm{p}K_\mathrm{a}-\mathrm{pH}}\right) \end{aligned} \quad (2)$$

where S_total is the equilibrium (thermodynamic) solubility, $\log S_0$ is the intrinsic solubility, defined as the solubility of an unionised species in a saturated solution, $\mathrm{p}K_\mathrm{a}$ is the negative logarithm of the ionisation constant of the molecule, and the final term on the right hand side is the solubility of the ionised form.[12] The HH relationship can be utilised in the prediction of pH-dependent aqueous solubility of drugs when the $\mathrm{p}K_\mathrm{a}$ and $\log S_0$ values of a compound are known.[13] The intrinsic solubility is a particularly important quantity as it can be used to find the pH dependent profile and estimate the $\mathrm{p}K_\mathrm{a}$; it is a quantity required by industry and hence the focus of several prediction methods.[14] The pH dependent profile of a drug is particularly important in pharmaceutics, as it has a direct effect on the absorption profile of a drug once it has entered the body. A basic drug-like molecule at a high pH (>2 pH units above the $\mathrm{p}K_\mathrm{a}$) will be fully unionised with solubility at a minimum (intrinsic solubility). Protonation of the base increases as pH becomes more acidic, and solubility increases. When pH and $\mathrm{p}K_\mathrm{a}$ are equal, half of the solute molecules are protonated and the solubility of the drug becomes double the intrinsic solubility. According to the HH equation, this rise in solubility increases indefinitely with decreased pH, however in practice a limit is reached at the salt solubility. Two intersecting concentration curves for the base solubility and the salt solubility can be combined to give a composite curve for base solubility as a function of pH. If any one point on this curve is known (solubility and pH at which it was measured), the whole curve can be predicted providing $\mathrm{p}K_\mathrm{a}$ and the acid solubility factor $C_{0\mathrm{A}}/C_{0\mathrm{B}}$ (the ratio of S_0 of acid to S_0 of base) are known.[15]

Intermolecular interaction strengths play an important role in the solvation of substances from the solid state. Solutes which exhibit weak intermolecular forces (*i.e.* are weakly bound) tend to have a higher solubility, as the energy cost of breaking up the lattice is lower. Polymorphic effects can also lead to complications in solubility prediction. A classically cited example of this is the case of the anti-HIV drug Ritonavir,[16,17] in which a polymorphic shift led to a significant change in solubility, leaving the drug with a greatly reduced bio-availability. This exemplifies the consideration of solubility as a property which

is dependent upon solid, solute, solvent, and solution state properties and interactions.

Two common approaches to the calculation of the Gibbs free energy of solution utilise a thermodynamic cycle approach. A first approach calculates the free energy of solution by addition of the free energy of sublimation (taking the molecule in the crystalline phase and subliming it into the gaseous phase) and free energy of solvation (taking the molecule in its gaseous phase and solvating it into aqueous solution). An example of this approach is shown in Section 5 of this review, and other examples are also cited within the literature.[14,18,19] A second approach involves calculation of the free energy of solution by addition of the free energy of fusion (taking a molecule from the crystalline state to a hypothetical supercooled liquid) and the free energy of transfer (transfer from a supercooled liquid into aqueous solution). This method is widely cited within the literature, and common GSE methods are also derived from this approach.[5] Both thermodynamic cycle approaches are depicted in Fig. 1.

The solid state is an important consideration for the initial crystalline phase calculated within thermodynamic cycle approaches. Lattice minimisation calculations and periodic DFT provide excellent tools for modelling these systems. Recent advances in these methods show promise for improving predictions, these include updated codes and improved dispersion corrections in periodic DFT.[20,21]

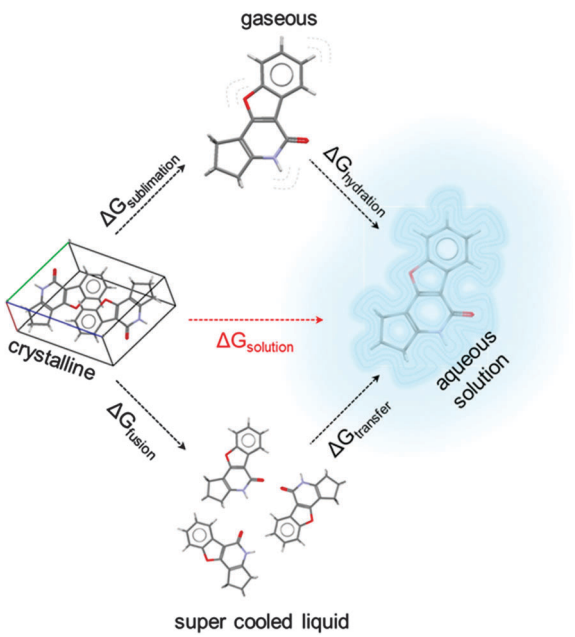

Fig. 1 Calculating the Gibbs free energy of solution is often achieved through the utilisation of thermodynamic cycles. Two routes are depicted here. The first route is shown at the top of the diagram, whereby a molecule is taken in its crystalline form and sublimed, and then hydrated. The addition of the Gibbs free energy terms of these processes gives the free energy of solution. The second thermodynamic cycle is represented at the bottom of the diagram, whereby the molecule is taken in its crystalline form and undergoes fusion into a hypothetical supercooled liquid, and then is transferred into aqueous solution. The addition of the free energy terms for these two processes also gives the Gibbs free energy of solution.

Complete polymorphic screening and prediction still eludes our capabilities and hence hampers our ability to predict solubility from purely first principles.

A further consideration is that of the standard states used in the different physical states. Typically sublimation data is reported in a 1 atmosphere standard state. Solvation is typically quoted in the Ben-Naim standard state of 1 mol L^{-1} with a fixed centre of mass. The difference between the two standard states is a constant 1.89 kcal mol^{-1} (7.91 kJ mol^{-1}), calculated as $\Delta G_{atm} \rightarrow$ mol L^{-1} = $RT\ln(24.46)$, where 24.46 is the molar volume at ambient conditions.

The free energy of solution can be calculated directly by the following formula:

$$\Delta G_{solution} = -RT\ln(S_0 V_m)$$
$$\log(S_0 V_m) = \frac{-\Delta G_{solution}}{2.303RT} \quad (3)$$

where S_0 is the intrinsic solubility V_m is the crystalline molar volume, R is the gas constant and T is the temperature in Kelvin (K).

A convenient formula[19] allows the solution free energy to be calculated using the native standard states, and removes the dependence on the crystalline molar volume:

$$S_0 = \frac{-p_0}{RT}\exp\left(\frac{\Delta G_{sub}^{1\,atm} + \Delta G_{solv}^{1\,mol\,L^{-1}}}{RT}\right) \quad (4)$$

2. Informatics – 'Smart' machines in solubility prediction

Informatics is the science of information processing, storage, and data mining. There are many applications and methodologies available for this type of task. Commonly used methods in chemistry are QSAR/QSPR models which are built from known data. These models correlate structural features of molecules with physical properties of interest. A major supposition of QSPR is that molecules similar in structure will have similar physical properties, and for QSAR models, perhaps chemical or biological similarities. Therefore it is possible to train a model defining a specific relationship between structure and property/activity on a training dataset, and apply it to similar molecules to predict their properties and activities. For this reason, QSAR/QSPR models are not broadly applicable (*i.e.*, they cannot be applied to molecules differing considerably from the training set). While QSPR was once dominated by multiple linear regression, nowadays machine learning represents the state of the art. Both regression and machine learning protocols can identify these structure–property relationships by correlating structural features with experimentally determined physical data. A brief introduction to some of these methods is provided below, and for a more detailed account, see "An Introduction to Cheminformatics"[22,23] and references therein. Initially, one must represent a molecule in a machine readable format to enable the calculation of molecular descriptors. Two of the most common methods for doing this are the Simplified

Molecular Input Line Entry System (SMILES)[24] and the IUPAC International Chemical Identifier (InChI).[25]

2.1 Molecular descriptors

Descriptors represent physical, chemical, topological or energetic features of chemical structures, and can vary greatly in form and derivation. In general, a descriptor is a vector of single numerical values (features), each encoding specific information about an individual molecule.[22] This information can be a simple number, such as the molecular weight or the count of a specific atom type, or they can be a prediction of corresponding experimental quantities, such as the octanol–water partition coefficient (usually expressed as $\log P$). Alternatively, they can also be derived from semi-empirical or quantum chemistry. Clearly the cost of calculating different descriptors can vary dramatically. It is often the case that descriptors offering higher levels of refinement, and therefore more useful molecular discrimination, incur a higher computational cost.[22] There are many different molecular descriptors and numerous pieces of software to calculate them.[22]

2.2 Methods

2.2.1 Regression. Regression analysis is a fundamental tool in informatics. Simple linear regression expresses a relationship between a scalar dependent variable Y and a single explanatory independent variable X. Multiple Linear Regression (MLR) extends this to allow for multiple dependent y_i variables or explanatory independent variables x_i, expressed as;

$$y = \sum_{i}^{j} \alpha_i x_i \qquad (5)$$

These methods have seen widespread use in many fields.[26] A disadvantage of MLR is the apparent ease of over-fitting. It is suggested that a useful rule of thumb is that the number of data points should be in excess of five times the number of explanatory variables (Fig. 2).[22,23]

2.2.2 Random forest. Random Forest (RF), is a learning method based on decision trees. These are stacked sets of binary separators following a tree like graph structure. RF uses a 'forest' of these decision trees, making use of "*the wisdom of crowds*"; hence, it is considered an ensemble learning method. RF can be used for classification or regression. For application to classification problems, the binary splitting is based upon the Gini index, which is a calculation of the maximal discrimination of the data points. For regression, splitting is generally based on a minimisation of the root mean squared error (RMSE). The initial node is known as the root node, with subsequent nodes being called branch nodes. The final nodes are referred to as leaf nodes and contain molecules with similar predictions of the property or activity (Fig. 2).[14,23]

2.2.3 Support vector machines. Another commonly used machine learning method is that of Support Vector Machines (SVM). SVM supports both regression and classification tasks, and is capable of handling multiple continuous and categorical variables. Methods for handling classification tasks are based on typically non-linear kernel functions. These kernel functions allow the transformation of data points into a higher dimensional feature space (Fig. 2).

SVM training algorithms are built up of binary categorised data, whereby a particular data point belongs to one of two categories. Thus, the test set data is also categorised, producing a clear separation, which should be as wide as possible, in the feature space. Alternatively, in the case of regression, the surface behaves analogously to a regression line, providing a maximal explanation of the data within the bounds of an acceptable error margin whilst attempting to remain relatively flat to avoid overfitting.[22,23]

2.2.4 Networks. Artificial Neural Networks (ANNs) and deep learning architectures are another common form of machine learning method in chemistry. These are models conceptually based on the brain's neuron network (although a great simplification). ANNs contain an input layer which receives the molecular information, an output layer which provides the prediction to the

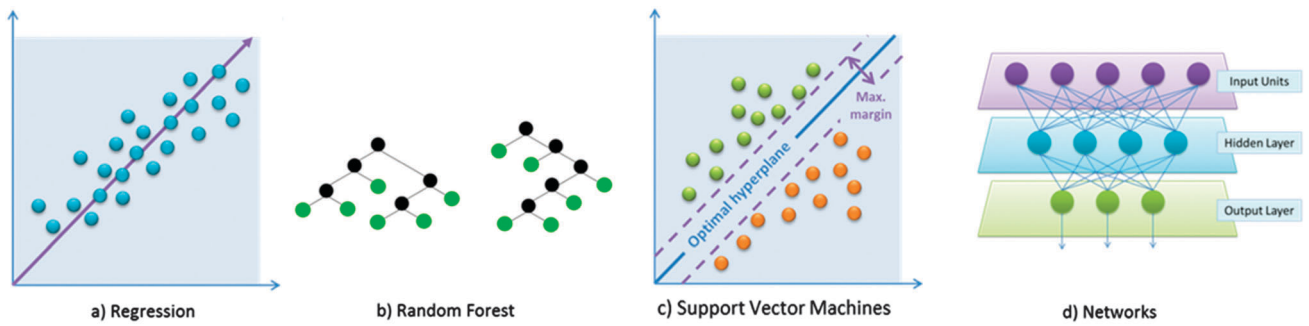

Fig. 2 Machine learning methods; (a) regression analysis aims to describe how the typical value of the dependent variable changes as the independent variables are changed. The regression function (purple arrow) characterises variation; (b) decision trees consisting of a binary separation at the nodes, leading to predictions or classifications at the leaf nodes (green circles); (c) an example of SVM separating data into distinct categories by an optimal hyperplane, which should have optimal margins either side for a clear distinction in data categorisation; (d) a typical network consists of layers of nodes. All nodes have connections with all other nodes in adjacent layers. The input units (top) do not count as a layer of nodes, as they do not carry out any typical arithmetic operations. A typical arithmetic operation is the generation of a net signal and transformation by a transfer function into an output signal. The input units distribute input values to all of the neurons in the layer below. The connections between nodes each have a different weight, representing different descriptors used in machine learning.

user, and between these at least one hidden layer which is trained using data to link the neurons of the input layer and output layer in a suitable fashion for the problem at hand. The training generally involves weighting specific paths between the neurons.[7,8,13] Deep learning architectures aim to enhance the learning capabilities of machine learning methods such as ANNs. Deep learning algorithms attempt to abstract data on a high level through model architectures comprising multiple non-linear transformations. In the case of ANNs, enhanced data abstraction can be achieved through the addition of hidden layers, capturing the interaction of many factors which contribute to the observed data.

2.3 The General Solubility Equation (GSE)

GSE (as briefly mentioned in the introduction) is a QSPR model based on the melting point and the octanol–water partition coefficient $\log P$ of a chemical substance, used to predict the aqueous solubility of non-ionisable compounds,[28] and acts as a useful guide for ionisable compounds using lipophilicity ($\log D$) at the pH of the aqueous buffer employed. The equation states that;

$$\log S = 0.5–0.01(\text{m.p.}^{\circ}\text{C} - 25) - \log P \quad (6)$$

Or in terms of $\log D$;

$$\log S_{\text{pH}(x)} = 0.5–0.01(\text{m.p.}^{\circ}\text{C} - 25) - \log D_{\text{pH}(x)} \quad (7)$$

GSE is a simple QSPR model, with powerful predictive ability (coefficient of determination (r^2) = 0.96 and root mean squared error (RMSE) = 0.53 $\log S$ units for a data set of 1026 organic molecules[29]), and the simplicity of the model means it has found wide application in the pharmaceutical industry. However, the reliance of the GSE on experimentally determined descriptors limits its applicability, and datasets sparsely populated at their limits can lead to overestimation of the model's predictive power.[30]

Ali *et al.*[30] have revisited the GSE and have attempted to relieve the reliance of the GSE on the experimentally determined melting point by replacing it with a descriptor that describes the topological polar surface area (TPSA). They demonstrate the effects of inflated predictive power of the GSE by using a subset of an initial dataset, which reduced the overall predictive power of the GSE by approximately 6.4%. TPSA was included in a revised model to account for the fact that 88.5% of poorly performing compounds contained polarisable groups. The pure GSE model employed provided r^2 = 0.818, and the TPSA replacement of melting point model provided r^2 = 0.813, showing a comparable effectiveness. The number of compounds containing polarisable groups with $\log S$ predicted within ±1 log unit of experimentally determined values was also higher for the revised TPSA model (83.2% TPSA; 79.6% GSE). A final model combining melting point, $\log P$ and TPSA was also tested, and was found to have a better predictive power than both of the previously employed models (r^2 = 0.869) with 90.8% of compounds containing polarisable groups predicted within ±1 log unit of experimentally determined values.

The work of Ali *et al.*[30] highlights the importance of reliable descriptors in improving the overall performance of QSPR models, particularly when polar or polarisable functionality is included in test sets, and when experimentally determined values are required. As such, experimentally determined values may be best suited only for comparative analysis of predictive models to experimental data as a measure of performance in many cases.

2.4 Other cheminformatics applications

A recent approach to predict solubility proposed by McDonagh *et al.*[14] applied three models, exploiting both cheminformatics descriptors and theoretically derived thermodynamic properties. The initial models use theoretical chemistry and QSPR models alone, with further development combining the two approaches into a unified QSPR model. The developed models aim to calculate solubilities in agreement with experiment and in a reasonable time period. It was found that quantitatively accurate solvation free energies are unobtainable from the specific simple theoretical chemistry approach applied. The authors suggest that QSPR models are the most effective method, when both time and accuracy are considered. The machine learning methods employed, which use a modest number of cheminformatics descriptors, predict solubility values comparable to those obtained with currently available commercial software.‡ Notably, only a small improvement in accuracy was found on combining the two approaches. This suggests that the cheminformatics descriptors and the theoretically derived quantities are not very complementary, but duplicate much of the same information.[14]

Another recent approach, by Lusci *et al.*,[27] applies deep learning to the solubility prediction problem. The deep learning method is based on recursive neural networks adapted for undirected graph representations of molecules. The method produces good predictions of solubility on a number of standard datasets in the field.[27]

A further example of a cheminformatics approach is demonstrated by Shayanfar *et al.*[31] who apply a simple QSPR model to the prediction of aqueous solubility of drugs, validated by cross-validation. A training set of 220 drug-like molecules was used to build a model with MLR. Six descriptors (solute, melting point, experimental $\log P$, calculated Abraham solvation parameters, calculated $C\log P$ values and calculated melting points) were regressed against experimental aqueous solubility from the literature to develop a three-variable model, calculating aqueous solubility from Abraham solvation parameters, $C\log P$ and melting points. The three variable model was then tested with cross validation, and a final two-variable model was developed with the excess molar fraction of the compound E and $C\log P$. The two variables used gave an R^2 value of 0.934 and a standard error estimate s of 0.893. The proposed model was compared to a GSE model and a linear/solvation/energy relationship (LSER) model. Correlations between each model's computationally determined

‡ A recent machine learning method and dataset proposed by some of the authors is available from the Mitchell group web server: http://chemistry.st-andrews.ac.uk/staff/jbom/group/Informatics_Solubility.html

values of aqueous solubility with corresponding experimental values gave an $R^2 = 0.62$ for GSE, $R^2 = 0.57$ for LSER and $R^2 = 0.66$ for the proposed MLR method.

Recent work has also suggested that, contrary to popular arguments, the quality of the experimental data available is not the limiting factor for the predictive accuracy of solubility predictions obtained from cheminformatics models.[32] This work may suggest that inherent limitations within the models are responsible for the largest part predictive errors.

3. Implicit solvation – an isotropic field as a solvent representation

Continuum solvation models consider solvent as a continuous isotropic medium. An underlying assumption of implicit solvation models is that explicit solvent molecules may be removed from the model, provided that the continuous medium replacing them sufficiently represents equivalent properties.

A simplification of continuum models can be thought of in terms of a Hamiltonian as;

$$\hat{H}^{\mathrm{tot}}(r_{\mathrm{M}}) = \hat{H}^{\mathrm{M}}(r_{\mathrm{M}}) + \hat{H}^{\mathrm{MS}}(r_{\mathrm{M}}) \quad (8)$$

where M refers to a single solute molecule, S refers to the solvent, and r refers to position. Solvent coordinates do not appear within the Hamiltonian term, exemplifying the representation of solute in a continuum, rather than as definite atoms, as with explicit models. $\hat{H}^{\mathrm{MS}}$ is a sum of different interaction operators, which can be expressed in terms of solvent response functions, indicated by $Q_x(\vec{r}',\vec{r}')$, where $\vec{r}$ indicates a position vector, and x represents a contributing interaction. More in-depth discussions are available in textbooks specific to computational chemistry, such as that by Cramer,[3] and reviews by Tomasi *et al.*[15]

In a standard continuum model, generally represented by Polarisable Continuum Models (PCM), solute–solvent interaction energies can be represented by a number of Q_x operators. The free energy of M is therefore described by an expression of five terms;

$$G(M) = G_{\mathrm{cav}} + G_{\mathrm{el}} + G_{\mathrm{dis}} + G_{\mathrm{rep}} + G_{\mathrm{tm}} \quad (9)$$

with the order of terms corresponding to the best performing order of the 'charging processes', integration processes coupling a distribution function with a potential function. The terms are the free energy of cavitation, electrostatic energy, dispersion energy, repulsion energy and thermal fluctuation, respectively.

3.1 Continuum models for electrostatic interactions

PCM models are advantageous in that they can represent a statistically averaged (continuum) solvent so that meaningful results can be acquired within a single calculation. PCM models have been particularly useful in modelling reactivity and spectroscopy of various solvents with different polarities.[33]

In a solvent–solute system where atom Q (solute) has a positive charge, solvent water molecules will preferentially orientate their negative dipoles towards the solute's positive charge (Fig. 3, left). For a single water molecule, there is only a slight preference in orientation, which is smaller than that of its average thermal fluctuations. Therefore, this effect is averaged over the long range of electrostatic interactions of water in the bulk (Fig. 3, right). For an isotropic solvent with random thermal motion, the average electric field is zero at any given point. However, introduction of a solute gives a net change in orientation, introducing an overall change in electric field, known as the 'reaction field'.

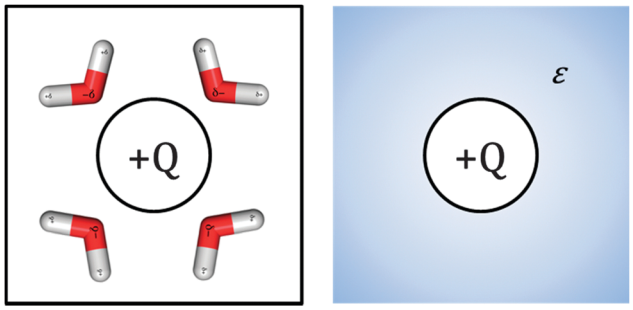

Fig. 3 (left) Water molecules reorient themselves to preferentially point the negative end of their dipole towards the positive solute charge (+Q). (right) The system is modelled with a continuous polarisable field. Polarisability is represented by the bulk dielectric constant, ε.

Accounting for the reaction field increases the solute's polarity proportionally to the solute polarisability, and the strength of the external electric field. This causes an increase in the dipole moment of Q, consequently polarising and increasing the change in orientation of the solvent to oppose the dipole moment of Q.[3]

There are energy costs associated with both the orientation and polarisation of the solvent, and the dipole moment of Q. As solvent molecules oppose the dipole moment of Q, they interact unfavourably with the reaction field. They also lose configurational freedom, with an associated free-energy cost. In a continuum model, the charge distribution of a solvent is represented as a continuous electric field, statistically averaged over all degrees of freedom at thermodynamic equilibrium. The electric field at any given point is the gradient of the electrostatic potential. The work required to create the charge distribution is determined from the interaction of solute charge density ρ with the electrostatic potential ϕ from;

$$G = \frac{1}{2}\int \rho(r)\phi(r)\mathrm{d}r \quad (10)$$

The polarisation component of G, which we call G_{P}, is the difference between charging the system in gas and solution phases; thus only the electrostatic potentials in both gas and solution phases are needed to calculate G_{P}.

PCM methods are generally applied through two models: the Poisson–Boltzmann (PB) model, and the Generalised Born (GB) model. Both models are advantageous for different systems, and the accuracy of either model is mostly dependent upon the suitability of the cavity type used to surround the solute molecule within an ideal solvent system.

3.1.1 The Poisson–Boltzmann (PB) model. The Poisson equation (eqn (11)) combines the terms for electrostatic potential and the differential form of Gauss's law to define the electrostatic

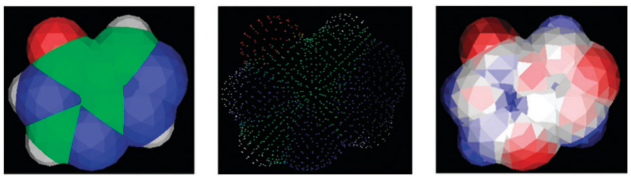

Fig. 4 The PCM cavity of allopurinol. (left) The solvent accessible surface of allopurinol from a PCM calculation. (middle) The reaction field evaluation points. (right) Surface polarisation as a result of reaction field.

potential ϕ as a function of the dielectric constant ε and charge density ρ. When a surrounding dielectric medium responds linearly to an embedded charge, Poisson's equation states that;

$$\nabla^2\phi(r) = -\frac{4\pi\rho(r)}{\varepsilon} \tag{11}$$

Continuum solvation models represent the charge distribution on the basis of two separate areas: inside (solute) and outside (solvent) of a cavity (Fig. 4). For this case, the Poisson equation states;

$$\nabla\varepsilon(r)\cdot\nabla\phi(r) = -4\pi\rho(r) \tag{12}$$

The Poisson equation as expressed above is valid only for systems under non-ionic conditions. In a real solution, dissolving a solute produces mobile electrolytes. This effect is accounted for by an expansion of the Poisson equation, known as the Poisson–Boltzmann (PB) equation;

$$\nabla\varepsilon(r)\cdot\nabla\phi(r) - \varepsilon(r)\lambda(r)\frac{8\pi q^2 I k_B T}{\varepsilon k_B T \quad q}\sinh\left[\frac{q\phi(r)}{k_B T}\right] = -4\pi\rho(r) \tag{13}$$

where q gives the magnitude of electrolyte ionic charge, λ is a function equal to 0 in areas inaccessible to electrolyte ions and 1 for accessible areas, and I indicates the ionic strength of the electrolyte system.

PB equations are best used to calculate the electrostatic potential of systems where the cavitation of solute is near-spherical or ellipsoidal (ideal cavitation), as the convergence of the predicted electrostatic component of the solvation free energy ΔG_E is computationally expensive and often inaccurate. Thus, derivations applying approximations of the Poisson equation are often used in continuum models,[33] the most common of which are Self-Consistent Reaction Field (SCRF) models, such as the Onsager model.[34]

A further limitation of PB based models is the definition of cavitation. A number of variational SCRF models have been proposed in order to optimise cavitation parameters, most commonly using tessellation (tiling) of the cavity surface to simplify and reduce iterations of the PB equation.[33]

3.1.2 The Generalised Born (GB) model. For systems in which ideal cavitation is not accurate, arbitrary cavitation can be applied. Arbitrary cavitation refers to the construction of a cavity around the solute similar to the shape represented by space-filling models generated from the overlap of atomic spheres at volumes representing van der Waals (vdW) radii. An alternative method to SCRF models involves an approximation of the Poisson equation that can be analytically solved, known as the Generalised Born (GB) approach.

A conducting sphere with charge q can be considered representative of a monatomic ion. If the surface of the sphere is assumed to be entirely smooth, the charge distribution around it will be uniform, and the charge density at any point is given by;

$$\rho(s) = \frac{q}{4\pi a^2} \tag{14}$$

where s is a point on the sphere's surface, and a is the spherical radius. Integrating over the entire outside surface and adding a term for the electrostatic potential, the energy term G, with $|r| = a$, becomes;

$$G = -\frac{1}{2}\int\left(\frac{q}{4\pi a^2}\right)\left(-\frac{q}{\varepsilon a}\right)\mathrm{d}s = \frac{q^2}{2\varepsilon a} \tag{15}$$

The Born equation for the polarisation of a monatomic ion is calculated from the difference in the required work in the gas and solution phases applied to eqn (8);

$$G_P = -\frac{1}{2}\left(1 - \frac{1}{\varepsilon}\right)\frac{q^2}{a} \tag{16}$$

The GB method extends the Born equation to polyatomic molecules to express polarisation energy as;

$$G_P = -\frac{1}{2}\left(1 - \frac{1}{\varepsilon}\right)\sum_{k,k'}^{\text{atoms}} q_k q_{k'}\gamma_{kk'} \tag{17}$$

where k and k' run over all atoms, each with a partial charge q. The determination of suitable parameters for γ for polyatomic systems involves a radial integration of the charge q to determine the interaction of atom k with the surrounding medium. γ has units of reciprocal length, thus representing an inverse Coulomb integral. γ is given a suitable functional form in order to approximate the PB equation, and has a limiting behaviour, becoming closer to the exact reciprocal length r^{-1} at large interatomic distances.

3.2 Continuum models for non-electrostatic interactions

Similarly to the electrostatic components of solvation free energy, non-electrostatic contributions to the solvation free energy are not experimentally measurable. The solubility of experimental systems may be more susceptible to some effects than others. Various neutral model systems have been developed in accordance with this.

3.2.1 Specific component models. Pierotti[35] developed a model formula, based on scaled particle theory, for the calculation of cavitation free energy through the observation of the solvation energy for noble gases. Scaled particle theory is a statistical-mechanical theory of fluids derived from exact radial distribution functions, to give an expression for the work required to place a spherical particle into a fluid of spherical particles. Noble gas atoms do not exhibit permanent electrical moments, thus their transfer into solution is considered to be the most analogous example of perfect cavitation.

The experimental data from Pierotti's work has been complemented by simulation data,[36] including free energy of formation data of molecular-sized cavities in 12 common solvents obtained from free energy perturbation simulations. Pierotti's formula has since been expanded for molecular cavities by Colominas *et al.*[37]

A further, specific contributing factor to solvation free energy is dispersion. A somewhat simplistic explanation of dispersion is as follows. The average electron cloud of an atom is spherically symmetrical, but at any instantaneous time point there may be a polarisation of charge causing an instantaneous dipole moment. This dipole moment interacts with neighbouring atoms, inducing a second instantaneous dipole, and so on, and an interaction occurs between these. The in-phase correlation of instantaneous and induced dipoles mean the overall interaction energy does not average to zero over time.[3] The average interaction energy falls off (largely) proportionally to r^{-6} (where r is the distance between interacting particles). The multipole expansion of the dispersion interaction is written;

$$V(r) = -\frac{C_6}{r^6} - \frac{C_8}{r^8} - \frac{C_{10}}{r^{10}} \cdots \quad (18)$$

where C_6, C_8 and C_{10} are dispersion coefficients dependent on the atomic species. This is normally evaluated as a sum over all pairs of atoms in different interacting molecules.

3.2.2 Atomic surface tensions. Another approach for the evaluation of the non-electrostatic components of solvation free energy assumes the non-electrostatic component to be atom or group specific, and proportional to atomic surface area. A recent review by Wang *et al.*[38] (2009) considers four QSPR aqueous solubility models developed on the principle of weighted atom type counts and Solvent Accessible Surface Areas (SASA). They note that models considering SASA are often developed with small test-sets, and are therefore, in common with QSAR/QSPR models, poor performers for test molecules dissimilar to the original training set. The authors found that SASA descriptors did not enhance model performance any further than weighted atom type counts. This suggests the influences upon the non-electrostatic components of solvation free energy may be more complex than simple surface area considerations.

A further notable feature of continuum models based on surface tension is the neglect of any other contribution; that is, the development of these models assumes surface area as the sole determinant of solvation free energy, and that electrostatic components are implicit within the calculation parameters used.[33]

3.3 The current state of continuum models

There are a large number of available continuum solvent models, all with relative merits and shortcomings. The following is a brief description of those most commonly applied.

Integral Equation Formalism PCM (IEFPCM) is the current version of PCM applied in common quantum chemistry packages. IEFPCM is a reformulation of dielectric PCM (DPCM) in terms of the integral equation formalism. One of the biggest challenges to PCM methods is that they are all derived assuming the solute charge density is entirely encapsulated in the cavity. This is often not the case, as the electron distributions often extend beyond the cavity. IEFPCM has been shown to cope well with this effect when compared to other PCM based methods.[33]

A further variation of PCM is the conductor-like polarisable continuum model (CPCM), which is often considered one of the most successful solvation models.[39] The conductor-like screening model/conductor-like screening model for real solvents (COSMO/COSMO-RS)[40] is a variation on Poisson–Boltzmann PCM and CPCM. In COSMO the dielectric permittivity (ε) is set to infinity ($\varepsilon = \infty$). This defines the solvent as a conductor, which is suggested as a more realistic approximation for strong dielectric media such as water, with the first version of COSMO[40] having values of the dielectric constant with a relative error of less than $\frac{1}{2}\varepsilon^{-1}$. COSMO has been shown to be a reliable and readily available method for calculations on the liquid and solution phases. The use of a boundary condition for the calculation of total potential in place of a traditional dielectric boundary condition for the electric field found values within 10% of the exact results obtained from dielectric boundary condition methods.[41] COSMO-RS extends the COSMO code to also define the ability of the solvent to screen the surface charge on the cavity of the solute. Parametrisation of COSMO and COSMO-RS performed by the software developers tested 217 small to medium neutral molecules, spanning a vast functionality of H, C, N, O and Cl. An overall accuracy of 0.4 (rms) kcal mol^{-1} for chemical potential differences was achieved.[41]

A recent addition is the solvation model based on density (SMD). This model applies the IEFPCM protocol, solving the non-homogeneous Poisson equation using a set of optimised atomic Coulomb radii. The non-electrostatic contributions are calculated on the basis of a parameterised function which includes terms for atomic and molecular surface tensions as well as the solvent accessible surface area.[42]

A recent investigation of gas to solution phase standard state Gibbs free energies of solution compares energies obtained for six combustion gas flue compounds at the G4 level of theory using IEFPCM, CPCM and SMD implicit solvent models for 178 organic solvents. It is found that IEFPCM and CPCM produce similar ΔG_S values for all six flue compounds, with maximum absolute intra-solvent deviations of <1.6 kJ mol^{-1}. Intra-solvent deviations between the IEFPCM and SMD models up to 45.5 kJ mol^{-1} were observed. IEFPCM and CPCM also showed strong correlation between calculated solvent ε and ΔG_S for all solvents, whereas SMD showed a much more varied relationship.[43]

4. Explicit solvation models

Explicit solvation models are the primary choice of solubility models where solvent-specific effects are considered. The explicit treatment of water should, in principle, provide the most descriptive and realistic model for the investigation of solvation,[44] however it intrinsically requires a large number of degrees of freedom and thus is associated with a phase space of high dimensionality. This requires statistical averaging over the entire phase space, particularly

when extracting specific underlying physical behaviour, such as thermodynamic properties.

Statistical thermodynamics relates all observable thermodynamic properties to the partition function, Q. The partition function is summarised as;

$$Q = \iint e^{-\frac{E(q,p)}{k_B T}} dq\, dp \quad (19)$$

where Q is the classical formulation integrated over all phase space of all spatial q and momentum p coordinates.

Explicit models consider solvation in terms of free energy calculations, with different models for water available, as discussed below.

4.1 Free Energy Calculations – Monte Carlo (MC) and Molecular Dynamics (MD) simulations

Free energy considerations are distinctly different for intramolecular and intermolecular degrees of freedom. For intramolecular components, free energy contributions rely on vibrational and librational motions on an intramolecular energy surface.[45] For well-defined energy-minima, the free energy is easily accessible from the partition function (eqn (19)) from vibrational frequencies treated with the harmonic approximation. The harmonic approximation estimates the nuclear potential of a molecular system in its equilibrium geometry at a potential energy surface minimum in terms of normal vibrational modes, each governed by a 1D harmonic potential. Anharmonic effects are accounted for with MC or MD simulations for the calculation of entropy on the intramolecular energy surface.[45] Due to diffusion, the particles of a solution system do not exhibit motion definable by harmonic approximations. Thus, conventional MC and MD methods do not involve the direct determination of Q, and exhibit an extremely slow convergence for densities of typical chemical systems, due to the exponential dependence of the Boltzmann factor on the occupation of available energy levels at a given temperature.

4.1.1 Free Energy Perturbation (FEP) methods. Free Energy Perturbation (FEP) methods were first introduced by Zwanzig[46] in 1954, who related the thermodynamics of two different systems, in order to evaluate differences in intermolecular potentials. Zwanzig notes that at high temperatures, the forces of repulsion between molecules determine the equation of state of a gas, and that at lower temperatures the equation of state should be determinable by considering forces of attraction as perturbations on the forces of repulsion. The energy change from state A to state B is calculated by;

$$\begin{aligned}\Delta G(\mathrm{A} \rightarrow \mathrm{B}) &= G_\mathrm{B} - G_\mathrm{A} \\ &= -k_\mathrm{B} T \ln\left\langle \exp\left(-\frac{E_\mathrm{B} - E_\mathrm{A}}{k_\mathrm{B} T}\right)\right\rangle_\mathrm{A}\end{aligned} \quad (20)$$

where T is temperature, and the triangular brackets indicate an average over the simulation runs for A. A normal simulation run for A coincides with a new energy state of B on each optimisation run. The energy difference between A and B is either between the atoms in each state, or in an isomeric difference, for example A may be the *cis*-isomer of a structure, and B the *trans*-isomer, with A and B in different energy states due to different intra- and/or intermolecular interaction. For isomeric differences, the free energy map is calculated along reaction coordinates. The convergence of FEP calculations is only reliable for a small difference between A and B, thus traditional perturbation theory only holds true for systems which remain similar upon dissolution.

More recent derivations of Zwanzig's model allow the division of perturbations into smaller calculations, allowing parallelisation. These models involve breaking the reaction pathway down into a series of intermediate transition state steps, allowing better convergence between the initial and final structures investigated.[47] However, FEP calculations remain one of the most computationally expensive methods for calculating free energy differences.

An example of this is shown by Lüder *et al.*[48] who have investigated the effectiveness of FEP methods for the calculation of free energy of solvation in pure melts for 46 drug molecules. Simulations were performed in two stages, scaling down the Coulomb and Lennard-Jones (LJ) interactions independently. Results were interpreted under the assumption that the free energy of the vapour to liquid process ΔG_{vl} can be calculated from the sum of the free energy term for cavitation ΔG_{cav} and the energy associated with LJ interactions and half of the Coulomb interaction term. ΔG_{cav} is obtained from hard-body theories. Interaction energies and molar volumes for each of the 64 drug molecules were compared for systems comprising 260 molecules. Deviations between systems were found to be an average of 2.9% for intermolecular interaction energy, and 1.4% for molar volume, suggesting the dataset selected would provide reliable results. Predicted and simulated ΔG_{cav} values are found to be systematically underestimated by approximately 15%. An overall average deviation of calculated ΔG_{vl} values in comparison to experiment is -1.8 kJ mol^{-1}, with reasonable errors expected in the range -1 to 1 kJ mol^{-1}. This investigation suggests that overall, FEP methods require more work at the theory level, particularly due to systematic errors that occur in phase space relationships between reference and perturbed systems.

An alternative approach to calculating the free energy difference from one state to another is to treat the change from A to B as a transformation, rather than to calculate free energies of independent structures, and calculate an energetic difference, as in traditional FEP methods.[3]

A recent application of this method, derived from FEP, has been demonstrated by Liu *et al.*[49] for the calculation of the solubility of gases in ionic liquids. The Bennett acceptance ratio (BAR) method utilises the method of transferring between states instead of treating each state as an individual structure. The Coulomb and LJ terms are calculated separately. It is found that simulated solubilities are found in good agreement with Henry's law constants. However, comparison to experimental data finds poorly soluble gases to have larger errors, with underestimated and overestimated gas solubilities found with similar calculation methods in complementary studies.

4.1.2 Enthalpy–entropy decomposition. A further offshoot of free energy calculations is the decomposition of the free energy term into enthalpic and entropic components. Entropy

and enthalpy complement free energy as they provide interpretive information to link molecular perturbations and thermodynamic changes. Two solutes may have similar hydration free energies (HFE), but may have solubilities dependent on distinct chemical functional groups.[44] As both enthalpy and entropy are experimentally measurable, the difference between theory and experiment is ascertainable, and may be applied as benchmarks for force field optimisations,[44] and give insight into the mechanism of solvation. Levy and Gallicchio have reviewed a variety of different approaches to the thermodynamic decomposition of free energies.[44]

Wyczalkowski *et al.*[50] recently proposed two new methods for the estimation of entropy and enthalpy decomposition of free energy calculations, evaluated for the solvation of *N*-methylacetamide (NMA). The methods investigated found thermodynamic contributions to be in disagreement with experimental data, highlighting the difficulty in obtaining decompositions comparable in quality to free energy estimates, with thermodynamic decomposition of computational Helmholtz free energies of solvation (ΔF at fixed volume) values yielding errors approximately two orders of magnitude larger than the initial ΔF values found. It is noted that ΔF values are statistically reliable and can be used for quantitative comparison to experimental data. The calculation of entropic and enthalpic contributions is also extremely computationally demanding, as every temperature point of a simulation requires recalculation of the overall free energy.[3] The authors highlight that where calculation of free energies of solvation has advanced so that computational errors are on par with experimental ones, thermodynamic decomposition calculations suffer from statistical errors 10–100 times larger than free energy of solvation calculations.

A recent study by Ahmed and Sandler[51] uses the decomposition of free energies of hydration and self-solvation of low polarity nitrotoluenes to consider an array of thermodynamic terms and physiochemical properties. These include: solid-phase vapour pressures, solubilities, Henry's law constants, hydration and self-solvation entropies, enthalpies, heat capacities and enthalpies of vaporisation or sublimation. Their study focuses on the temperature-dependence of various terms. Decomposition of hydration free energies into enthalpic and entropic contributions is performed by a method utilising polynomial fitting of temperature-dependent self-solvation free energies (with respect to temperature). The use of fitting increases the sensitivity of derived values of hydration free energies. Self-solvation enthalpy (ΔH_{self}) values and entropy ($T\Delta S_{self}$) values are calculated within approximately 2 kcal mol^{-1} of experimentally determined values.

4.2 Combined Quantum Mechanical/Molecular Mechanical Methodologies (QM/MM)

Explicit solvation models are often developed with respect to biological systems, due to the role of water in catalytic mechanisms, protein folding and protein–DNA recognition, to name but a few, which all require the specific detail of explicit water–substrate interactions to hold descriptive meaning. Of particular interest are combined QM/MM models, with QM describing electronic system changes (where precise system description is needed) and the rest of the system (where less precision is required) being described by a MM force field.[3] Applications of QM/MM combined models are discussed in a recent review.[52]

The foundational concepts involve the partitioning of a desired system into two subsystems: the QM subsystem, containing a small number of atoms and described by QM, with the remainder of the system described by a suitable MM force field. The Hamiltonian of the whole system is simply written;

$$H = H_{QM} + H_{MM} + H_{QM/MM} \quad (21)$$

where H_{QM} is a QM Hamiltonian, H_{MM} is an empirical force field and $H_{QM/MM}$ describes interactions at the QM/MM interface. The energy of the system is also described as the sum of QM, MM and QM/MM contributions. This model is often referred to as a two-layered approach (Fig. 5, left). A derivative of this model involves adding a third "layer" as a continuum solvent representation around the MM region, and is known as a three-layered approach (Fig. 5, right).

Theoretically, any desired level of accuracy can be used within the QM region of the simulated system, within the scope of available methods. However, more accurate methods are susceptible to high computational cost. Thus, careful consideration is required by the user as to what level of accuracy is required, and at what cost. A succinct overview of different available QM methods is provided by Friesner and Guallar[52] for QM/MM methods applied to enzymatic catalysis, with descriptions, advantages and disadvantages of respective QM methods available in textbooks such as the one by Cramer.[3]

A primary consideration when selecting a QM/MM method is the interactions at the QM/MM interface. Two aspects must be considered; (i) the presence of covalent bonds across the interface – a particular concern for large (*e.g.*, biomolecular) molecules, (ii) the influence of the MM solvent region on the QM region – electrostatic and van der Waals interaction terms must be included.

In order to treat covalent bonds at the interface, it is possible to introduce "link atoms". Link atoms are QM hydrogen atoms that fill free valencies of QM atoms connected to MM atoms. A disadvantage of this method is the debate about inclusion of

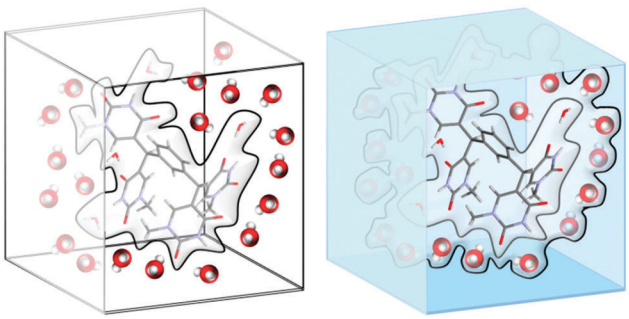

Fig. 5 (left) Two-layered approach to the QM/MM method. The solute molecule and a few water molecules are treated with QM (centre) and the rest of the solvent system is represented by MM up to a user-defined distance. (right) Three-layered approach – an additional layer surrounds the MM region and uses a continuum approach to describe the long range solvent in the bulk.

Coulombic interaction terms for the link atoms. Other methods developed in order to avoid the use of link atoms include the Local Self-Consistent Field (LSCF) method, which applies a mixture of hybrid and atomic orbitals to represent the QM system, and the "connection atom" method, where MM and QM interface atoms are described as QM methyl groups with a free sp^3 valence.

A recent three-layered approach aiming to tackle the issues associated with the QM/MM interface and the interaction terms for MM solvent effects has been proposed by Steindal *et al.*[53] This approach is described as the fully polarisable QM/MM/PCM method (see Section 3 for a description of PCM), and is designed for the effective inclusion of a medium in a QM calculation. Short range solvent electrostatic potentials are described by an atomistic model (QM/MM) whilst the long range potentials are described by a continuum. The method is implemented in combination with linear response techniques with a non-equilibrium formulation of environmental response. The authors find a faster convergence with respect to system size for QM/MM/PCM than for QM/MM methods. This approach allows for reduction of the MM part of the calculation with PCM, allowing less demanding calculations, and reduced sampling. However, three-layered approaches such as this often require much more user input and method manipulation, for example, considerations for MM/PCM interactions have to be considered in addition to QM/MM interactions, and so such methods are suited only to advanced users.

4.3 Explicit representations of water atoms

When solvent is represented explicitly, solvent molecules usually greatly outnumber solute molecules. Thus, in order for a model to be efficient, it is advantageous to use the simplest possible solvent representation.[44] Water is often considered the most useful solvent system, and thus is the solvent most widely used in explicit solvent models. The macroscopic properties are well established, yet the microscopic forces that determine water structure are not fully understood.

The treatment of water can be rigid or flexible. Rigid models often include a fictitious H–H bond to constrain bond angles in the water monomer.[3] Three of the most common rigid models for water are the TIP3P (transferable intermolecular potential 3P), SPC (simple point charge) and SPC/E (simple point charge extended) models, and their modified counterparts. These three models are effectively rigid pair potentials comprising LJ and Coulombic terms. However, the terms used differ in each model, and give rise to different calculated bulk properties for water.[54] Values for various properties of water obtained with different rigid models of water are shown below, in Table 1.

Table 1 Model *vs.* experimental (exp.) values for bulk properties of water under standard conditions (298 K; 1 bar), including dipole μ, density ρ, static dielectric constant ε_0 and heat capacity C_p

Property	TIP3P[55,56]	TIP4PEw[57]	SPC/E[56,58]	Exp.[56]
μ (D)	2.348	2.32	2.352	2.5–3.0
ρ (g cm^{-3})	0.980	0.995	0.994	0.997
ε_0	94	63.90	68	78.4
C_P (cal K^{-1} mol^{-1})	18.74	19.2	20.7	18

MD calculations require the integration of Newton's equations of motion for all atoms, which is achieved through the evaluation of all atomic forces at each time step. Non-bonded interactions, especially long-range electrostatic interactions, dominate computationally, requiring extensive CPU time. In order to minimise this to an acceptable level, approximations are necessary. Boundaries are introduced into water models to restrain the system to a finite size, which almost always leads to artefacts in the obtainable data.[54] The most commonly utilised method for cost-effective solute computations is the application of a spherical cut-off, limiting the number of pairwise interactions to those within a specified radius.[54] The use of cut-offs for non-bonded interactions can have undesirable effects. LJ interactions are susceptible to small energetic effects, and large pressure effects induced by cut-offs. Pressure scaling can be used to correct for pressure related cut-off effects, usually to the order of several hundred bar. Cut-off effects for systems with dipolar electrostatic interactions are more prominent, with cut-offs selected within the parameters of experimental radial distribution functions up to ~1.0 nm. However, computer simulations have shown ordering within water up to ~1.4 nm, so the full structure of water is not typically accounted for, resulting in a poor description of dielectric properties. A further, and the most prominent, effect of cut-offs occurs in systems with full charges, where accumulation of the charge occurs at the cut-off boundary.[59]

Spoel *et al.*[59] (1998) investigated the effectiveness of TIP3P, TIP4P, SPC, and SPC/E models in describing the density and energy, dynamic, dielectric and structural properties of water. All simulations and analyses were identical for each model investigated, allowing the evaluation of simulation methodology independent of the model. It was found that system size, cut-off length and reaction fields had comparable effects on the overall calculated structural properties of water.

System size effects are considered through the comparison of systems comprising a small (216) and a large (820) number of molecules. The average thermodynamic properties (ρ, E_{pot}, T, P) are the same regardless of system size. Fluctuations in thermodynamic properties are known to be proportional to the square root of the system size, which is confirmed within the study. However, differences between large and small systems are observed, particularly for the dielectric constant, which is higher for all systems with a large number of molecules. The diffusion constant for large systems is also higher, attributed to periodic boundary conditions (PBC).

Cut-off effects are considered by the use of two different cut-off lengths (0.9 nm and 1.2 nm) for the large systems. It is found that density increases with an increased cut-off length, and energy decreases. There is no effect on dielectric behaviour.

In all simulations density decreased by approximately 1 kJ mol^{-1} on application of a reaction field. The self-diffusion constant D, and rotational correlation times were found to increase, indicating that the reaction field affects both the translational and rotational mobility of molecules.

Quantum chemical MD simulations of water are often developed with Density Functional Theory (DFT) methods, using either plane wave or atom-centred basis sets, to determine the electronic structure and forces. These methods offer

reasonable estimates of the structural and dynamic properties of water when compared to experimental measurements. However, problems exist in the description of electronic gradient corrections, and equilibrium pressure. The interatomic forces of early quantum simulations, including DFT based methods, were originally parameterised with classical mechanics, leading to an unsatisfactory agreement between quantum and experimental results. DFT models also tend to calculate liquid structure with too much order, and underestimate equilibrium density. This is often attributed to the inability of local functionals to describe dispersion effects.

A recent approach to water simulation has claimed to provide a model, called the electronically coarse-grained model, capable of accounting for the shortcomings of both existing classical and quantum models.[60] Jones *et al.*[60] (2013) base their method on the replacement of valence electrons of an atom with an embedded Quantum Drude oscillator (QDO). QDO treatment of water is based upon the TIP4P classical rigid model of water, with the three water atoms supplemented by a dummy atom with a negative charge, added along the $\angle$HOH bisector to create an additional interaction point. The QDO parameters aim to reproduce the isotropic parts of the dipole, polarisability, and the dispersion coefficient. The dispersion interaction is then adjusted by scaling, whilst preserving polarisability. The baseline unadjusted model produces a realistic, but over-structured liquid with a density that is too low by up to 20%, attributed to its underestimation of dispersion. Note also that the value of the enthalpy of vaporisation (at ambient pressure) Δh_{vap} was found at 40 $\pm$ 2 kJ mol^{-1}, close to the experimental value of 43.91 kJ mol^{-1}. Scaling the dispersion term results in an increased equilibrium density for increased dispersion. This induces a weakening effect on the H-bonding network of water, bringing the overall structure closer to agreement with benchmark data. However, the calculated Δh_{vap} increases to 46 $\pm$ 2 kJ mol^{-1}, which is 4% higher than the experimental value. It is also found that the H-bond network is sensitive to changing polarisation at fixed dispersion, affirming the independent importance of both polarisation and dispersion effects on an overall explicit model.

5. Efficient hybrid models – statistical mechanics

Within an aqueous solution phase, single snapshot images of structure are of limited use. Water is one of the few single component liquids for which there are highly competitive interactions at short range (hydrogen bonding), capable of damping the effects of repulsion. For this reason, ensemble averaging is required to identify the most probable geometric configurations which most heavily contribute to the system's interactions. This idea has already been introduced within explicit models of solvation using ensembles taking snapshots at specific time periods. However, the cost of calculating the many configurations accessible in a solution is enormous, hence, in this section we focus on statistical mechanics methods which enable a more efficient calculation process.

5.1 Correlation functions

From a chemical point of view, a solution is a highly mobile system in which the dynamics are a vital contribution to the system's properties and behaviour. Therefore, mathematically we wish to capture this. Attempting to quantify dynamics with static properties is not sufficient; we must therefore provide averages or probabilities of interactions occurring at given distances. For this reason a natural choice is to represent the solvent using Pair Correlation Functions (PCF), or equivalently Radial Distribution Functions (RDF). These functions allow us to determine a probabilistic structure of the solvent.

PCF can be interpreted as showing the probability against distance of there being an atom of interest at that distance from the atom under study. For example the first large blue peak in Fig. 6 would correspond to either a water H at a distance from an O atom under study or *vice versa*. These functions are experimentally determinable from scattering experiments. We would expect that the PCF/RDF would go to a constant value of 1 at large values of r (*i.e.* it would become isotropic, like a continuum model, as there are no solute interactions to perturb the system). However, at small values of r we would not expect this. At very small values (less than the van der Waals radii of the solute atoms) we expect zero as only one particle can occupy the space at a time. Just outside this distance we see sharp non-uniform behaviour as solvent in the space interacts favourably with the solute holding a more rigid form. This leads to troughs in the PCF/RDF just behind the peaks, thus deviating from the value of 1 for a uniform solvent (Fig. 6).

5.1.1 Computational use and determination of correlation functions. The starting point for the use and determination of these functions for solvation modelling in statistical mechanics is integral equation theory (IET). In this theory a molecule is fully described by a six-dimensional vector (three degrees of freedom relate to position x, y, z and three degrees of freedom determine the orientation ψ, θ, φ). To refer to these two sets of variables collectively, we will use the following symbols $r = \{x, y, z\}$ and $\Theta = \{\psi, \theta, \varphi\}$. These variables are conveniently incorporated into the fundamental 6D integral equation, the

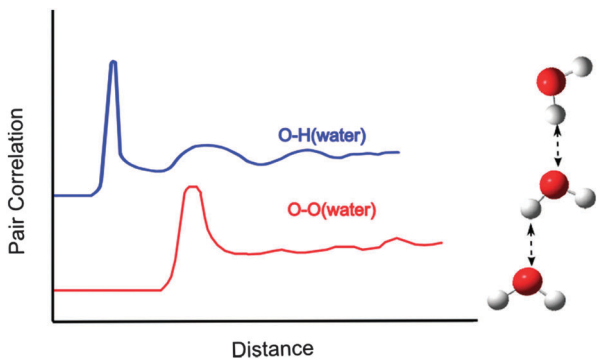

Fig. 6 A schematic representation of PCF for liquid water; water oxygen – water hydrogen (blue) and water oxygen – water oxygen (red).

Molecular Ornstein–Zernike equation (MOZ). This equation utilises PCF/RDF between the various constituents of the liquid, $g(r_1, r_2, \Theta_1, \Theta_2)$. This simplifies for homogeneous solution to relative positions and orientation of the constituents, $g(r_1 - r_2, \Theta_1 - \Theta_2)$. This can most conveniently be written with reference to the total correlation function $h(r, \Theta)$.[61]

$$h_{ij}(r_1 - r_2, \Theta_1 - \Theta_2) = g_{ij}(r_1 - r_2, \Theta_1 - \Theta_2) - 1 \quad (22)$$

We can simplify this equation by assuming spherical symmetry of molecules, hence removing consideration of orientational degrees of freedom by treating each water molecule as a hard sphere. We can now further separate the contributions to the total correlation function into direct and indirect components. To do this we must introduce the direct correlation function $c(r)$. We can now re-write the MOZ equation assuming spherical symmetry as follows:

$$h(r_{1,2}) = c(r_{1,2}) + \int \mathrm{d}r_3 c(r_{1,3})\rho(r_3)h(r_{2,3}) \quad (23)$$

Two effects contribute to the total correlation function (eqn (22)): (i) the direct correlation between r_1 and r_2, and (ii) an indirect correlation *via* a third body, r_3. The indirect correlation *via* r_3 is weighted by the density at r_3, and thus allows the consideration of all possible positions of the third body (Fig. 7).[61]

To solve this equation, $h(r)$ and $c(r)$ need to be found. As we have only a single equation and two unknown functions, $h(r)$ and $c(r)$, another equation is required; a closure relation must be introduced. There are several such equations available from statistical mechanics. The exact closure relation is as follows:

$$g(r) = \mathrm{e}^{-\beta U(r)+h(r)-c(r)+B(r)} \Rightarrow \mathrm{e}^{-\beta U(r)+T(r)+B(r)} \quad (24)$$

where β is equal to $1/k_\mathrm{B}T$ and $U(r)$ is the interaction potential which is often of the following form:

$$U(r) = 4\varepsilon\left[\left(\frac{\sigma_{ab}}{r}\right)^{12} - \left(\frac{\sigma_{ab}}{r}\right)^{6}\right] + \frac{q_a q_b}{r} \quad (25)$$

$T(r)$ is known as the indirect correlation function as it is the difference between the total and direct correlation functions, and quantifies the indirect contribution. $B(r)$ is the bridge function, which comes from graph theory – its exact form is not known. Several approximate closure relations exist; some will be discussed here, although others are available. Originally the HyperNetted-Chain (HNC) approximate closure was used:

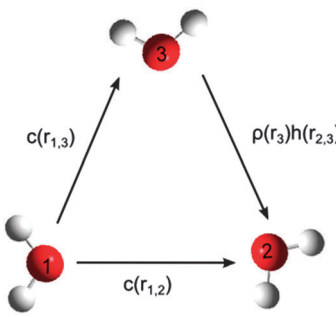

Fig. 7 Illustration of the contributions, both direct and indirect, to the total correlation function.

$$h(r) = \mathrm{e}^{(-\beta U(r)+T(r))} - 1 \quad (26)$$

This closure works in principle for charged systems but neglects the bridge function term completely, assuming it to be zero. This can lead to poor convergence due to uncontrolled growth in the argument of the exponent. An alternative is the Partially Linearised HyperNetted Chain (PLHNC). This closure linearises the HNC once a cut off value (C) is exceeded:[62]

$$\Lambda = -\beta U(r) + T(r)$$
$$h(r) = \begin{cases} \mathrm{e}^{(-\beta U(r)+T(r))} - 1 & \text{when } \Lambda \leq C \\ -\beta U(r) + T(r) + \mathrm{e}^{C} - C - 1 & \text{when } \Lambda > C \end{cases} \quad (27)$$

This improves the convergence of the equations and is now regularly used in many applications for a variety of systems.

Due to the spherical symmetry approximation, the MOZ can only be applied to simple solutions. Additionally, due to the high dimensionality of the full equation, before the spherical symmetry approximation was invoked, it was practically incomputable. For this reason a number of approximations have been developed which are collectively referred to as Reference Interaction Site Models (RISM).

5.2 3D-RISM: a hybrid solvation model

As we have seen, the explicit treatment of solvent is considered to be a necessary step in the understanding of solvent structure. However, this naturally carries high computational costs.[3] The alternative continuum treatment of solvents lacks the ability to account for the underlying physical theory; energy contributions from solvation shell features are computable, but not transferable. Solvent structure features from the first and second solvation shells are lost in continuum models, and non-electrostatic energy terms are not described from first principles, thus are not transferable to more complex models.[63]

The 3D derivation of RISM (3D-RISM)[64,65] is a 3D molecular theory of solvation, applied through solvent distributions, rather than explicit solvent molecules, and conceives solvation structure and dynamics from the first principles of statistical mechanics.

3D-RISM is derived from a partial integration over the orientational degrees of freedom; this leaves a set of 3D integral equations (one equation per solvent site; N_{solvent}). This method utilises solvent site – solute total correlation functions and direct correlation functions in the solution of the RISM equations. The 3D-RISM equations take the following form:[62]

$$h(\alpha) = \sum_{\xi}^{N_{\mathrm{solvent}}} \int_{R^3} c_\xi(r_1 - r_2)\chi_{\xi,\alpha}(|r_2|)\mathrm{d}r_2 \quad (28)$$

here $\chi_{\xi,\alpha}$ labels the solvent susceptibility function. This function models the bulk solvent mutual correlations. For the example of water, this function models the intermolecular correlation between water oxygen and water hydrogen. This function can be calculated from the intramolecular solvent correlation function ($\omega_{\xi\gamma}^{\mathrm{solvent}}(r)$), the radial site to site total

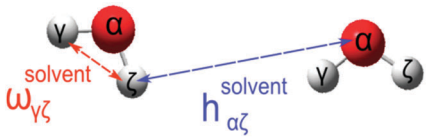

Fig. 8 Illustration of the contributions to the solvent susceptibility function.

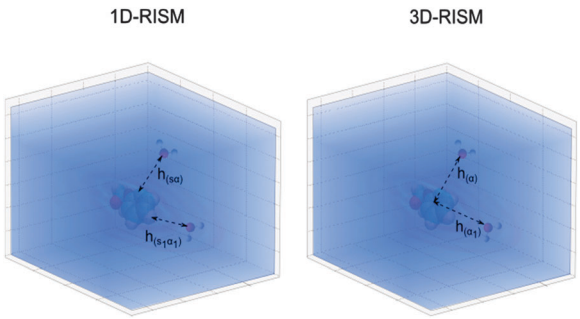

Fig. 9 A schematic representation of 1D-RISM and 3D-RISM. The conceptual difference in the models is that the total correlation functions are calculated considering the solute as a set of sites (1D-RISM) or as a single site (3D-RISM). α labels the solvent site in both models, s labels the solute site in the 1D-RISM case.

correlation functions ($h_{\zeta\alpha}^{\text{solvent}}(r)$) and the number density at each solvent site (ρ_α) (Fig. 8):

$$\chi_{\xi,\alpha}(r) = \omega_{\zeta\gamma}^{\text{solvent}}(r) + \rho_\alpha(h_{\zeta\alpha}^{\text{solvent}}(r)) \tag{29}$$

3D RISM can reliably account for the spatial correlation of the solvent density around the solute. As displayed above, the solvent molecules are modelled as a set of atomic sites, with 3D structure described by intramolecular correlation functions (Fig. 9).[62,66]

5.3 1D-RISM: a high throughput solvation model

Another RISM method is 1D RISM, which separates the solute into a set of sites (generally the atoms) and utilises solvent site – solute site total correlation functions and direct correlation functions. This leads to a set of ($N_{\text{solute site}} \times N_{\text{solvent site}}$) closure relations. 1D RISM is extremely quick to calculate but does not account properly for spatial correlations of the solvent density around the solute:

$$h_{s'\alpha}(r) = \sum_{s'=1}^{N_{\text{solute}}} \sum_{\zeta=1}^{N_{\text{solvent}}} \int_{R^3}\int_{R^3} \omega_{ss'}(|r_1 - r'|) c_{s'\zeta}(|r' - r''|)\chi_{\xi,\alpha}(|r'' - r_2|)\mathrm{d}r'\mathrm{d}r'' \tag{30}$$

N_{solute} is the number of sites in the solute and N_{solvent} is number of sites in the solvent molecule. $\omega_{ss'}$ are the intramolecular correlation functions representing the solute molecule.[66]

Implementations of both 1D- and 3D-RISM are available in well-known computational packages such as AMBER. There are also implementations in some quantum chemistry codes such as ADF.

5.4 RISM corrections and derivations

5.4.1 Correction schemes. A well-known error in both 1D and 3D-RISM occurs due to accounting for the cavitation term in the solution phase incorrectly. Other limitations also exist, associated with the use of approximations. Several schemes to correct these errors have been developed for 3D-RISM, however these are beyond the scope of this review, and thus are discussed in minimal detail.

Many studies have been conducted over the last two decades with a view to improving the accuracy of 3D-RISM for a variety of applications. Modifications to the original equations have included cavity corrections,[67] parallelisation with fast Fourier transforms[68] and MD modifications,[63] amongst others.

The universal correction (UC)[69] given in eqn (31) is a two parameter correction derived by regression. $\Delta G_{\text{hydration}}^{\text{GF}}$ refers to the Gaussian fluctuation hydration free energy (HFE) functional discussed below, a and b are regression coefficients ($a = -3.2217$ and $b = 0.5783$), and ρV is the dimensionless partial molar volume as calculated by 3D-RISM.

$$\begin{aligned} \Delta G_{\text{hydration}}^{\text{3D-RISMUC}} &= \Delta G_{\text{hydration}}^{\text{GF}} + a(\rho V) + b \\ \text{UC} &= a(\rho V) + b \end{aligned} \tag{31}$$

A second scheme known as cavity corrected 3D-RISM fits a single parameter calculated on the basis of a solution composed of spheres which interact exclusively by LJ type interactions.[70] A very recent addition offers a theoretical justification for such schemes; applying a thermodynamic-ensemble partial molar volume correction.[71]

Correction schemes for 1D-RISM also exist. These correction schemes must correct for additional approximations from the 1D RISM theory. A recent addition is the Structural Descriptor Correction (SDC). This applies QSPR methods and group contributions to correct 1D-RISM.[66]

A primary concern in the improvement of 3D-RISM remains its ability to describe the thermodynamic properties of solvation. One view adopted by Palmer *et al.*[69] is that solubility calculations should be considered in terms of a simple thermodynamic cycle, calculating the solvation free energy from summation of the free energy of sublimation, and the free energy of hydration, as illustrated in Fig. 10.

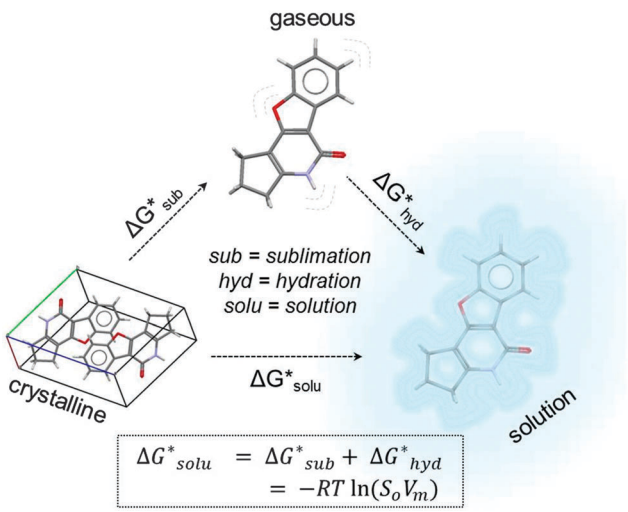

Fig. 10 Solubility prediction *via* a thermodynamic cycle. The free energy change from crystalline to aqueous phase is calculated from the summation of the free energy change of sublimation and the free energy change of hydration.

A recent investigation by Palmer *et al.*[62] implements the thermodynamic cycle approach to the calculation of solubility, with sublimation free energies calculated from crystal lattice minimisation and HFEs calculated with 3D-RISM. Crystal lattice calculations are performed on known crystal structures.

The authors highlight a plethora of existing approximate functionals which can provide HFE values from the solvent site-solute total correlation functions and direct correlations of 3D-RISM. However, the functionals investigated previously to Palmer *et al.*'s work often provide HFEs with RMSE errors higher than the standard deviation of experimental data, and worse than those reported in QSPR models.

The investigation[62] implementing the thermodynamic cycle approach to the calculation of solubility applied the previous work of Palmer *et al.*[61] and found that the thermodynamic cycle approach predicted HFEs in good agreement with experiment (R = 0.94, σ = 0.99 kcal mol^{-1}). However, the predictions did not perform as well as purely empirical approaches, and this was mostly attributed to a lack of parameterisation against experimental data.

5.4.2 Hydration free energy functionals. In order to calculate HFEs a HFE functional must be applied to the RISM output. There are a number of such functionals which vary in accuracy. Some of the correction schemes above recommend a specific HFE functional for use (UC recommends the Gaussian fluctuation HFE functional[72]). It is suggested to the user that where possible several functionals are tested for accuracy. Where this is not possible, the guidance given for the selection of a HFE functional for specific schemes should be followed, as these are generally well documented by the developer groups.

5.4.3 RISM and quantum chemical applications. RISM has also been applied to quantum chemical applications. RISM was extended for applications to quantum chemistry – this extension is called RISM-SCF. This theory provides the following definition of the Helmholtz free energy of the system:

$$A = E_{\text{solute}} + \Delta\mu \tag{32}$$

where A is the total Helmholtz free energy, E_{solute} is the solute energy and $\Delta\mu$ is the solvation free energy from the RISM equations. A is functionally connected to both the site-to-site density correlation functions and the wavefunction of the solute, hence mutual solution of E_{solute} and $\Delta\mu$ provide the joint system's equilibrium energies.[33] 3D-RISM has been combined with Kohn–Sham DFT, offering an alternative to continuum solvents and *ab initio* MD.[73] These calculations have been extended to higher levels of quantum mechanical theory (multi-reference methods) which are currently unaffordable at the QM/MM level.[33]

5.5 Other hybrid models

Combined implicit–explicit hybrid models work on a common framework; the central part of the system contains explicit solute and a few explicit solvent molecules, and the rest of the system is treated as a dielectric continuum.

The improvement associated with the insertion of explicit water molecules within a dielectric continuum has been demonstrated by Kelly, Cramer and Truhlar,[74] who use the calculation of aqueous acid dissociation constants to demonstrate the effects of inserting a single explicit solvent molecule into a continuum solvent representation. Along with previous work,[75] the authors show that in many cases an implicit solvation method is sufficient for the calculation of pK_a values. However, when strong and specific solute–solvent hydrogen bonding interactions are expected to contribute significantly to the aqueous phase, a single explicit molecule inserted to the continuum significantly improves pK_a calculation. Using their own implicit continuum model (SM6), it is found that addition of further explicit waters, up to three, significantly increases the accuracy of the calculation. However, the use of alternative continuum models, namely SM5.43R and PCM, finds a worsening of results when an increasing number of explicit atoms are added. This exemplifies the importance of choosing a suitable continuum representation in implicit–explicit hybrid models.

Zhu and Krilov[76] discussed two flexible boundary hybrid solvation models for biomolecular systems, based upon the traditional hybrid model with both explicit and implicit solvent regions. The proposed models aim to account for short-range solvent effects *via* elimination of PBC by limiting the number of explicit solvent molecules to two or three solvation shells. The first model, the dynamic boundary model, imposes a confining potential on the solvent, which responds dynamically to fluctuations in solvent distribution and solute conformation. The second model, the exchange boundary solvation model, allows pairwise exchanges between the explicit and implicit regions of the system, maintaining a uniform hydration of the solute. Comparison of the two methods with traditional PBC methods shows good agreement between calculated energies, and the two models are found to improve computational efficiency by up to two orders of magnitude, attributed to the reduced number of explicit solvent molecules in comparison to other models.

Chaudhury *et al.*[77] recently discussed the discrepancies between explicit and implicit methods for solvation models of biological systems such as proteins, and consequently investigate a Hybrid Replica Exchange Molecular Dynamics (REMD) method for protein solvation. Temperature-based REMD involves running multiple simultaneous simulations at a wide range of temperatures, while allowing temperature exchange between simulation steps. This relates the relative probability of finding each conformation at a given temperature to conformational energy. Traditional REMD successfully models small peptides and proteins, but becomes more cost-constrained for larger systems. In order to account for discrepancies between implicit and explicit methods, the authors propose a hybrid implicit–explicit method with each simulation step run exclusively in explicit solvent. During exchange between time steps, the entire solvent system is replaced with an implicit solvent model. Finally, the explicit solvent is re-inserted for the next simulation step. The use of an implicit solvent model during exchange significantly reduces computational cost. Where implicit and explicit models give different behaviours, the hybrid method gives mixed results in terms of thermodynamic and structural descriptions. However, the explicit model of solvent molecules describes solvent-specific features of energy landscapes well.

A further emerging method that similarly attempts to reduce the cost-constraints of explicit methods is Grid Cell Theory

(GCT).[78] GCT spatially resolves the enthalpic and entropic components of hydration on a 3D grid, covering a volume of space around a solute. The grid can be non-uniform and unevenly spaced. The solute is constrained to adopt a single conformation, speeding up convergence by only allowing rigid body translations and rotations of water molecules. A second benefit of GCT is that graphical analysis of a calculated grid is possible. A drawback of GCT method development emanates from the fact that there does not exist a unique method of partitioning a free energy into a sum of contributions, as contributions are susceptible to coupling. Gerogiokas *et al.*[78] have recently proposed a GCT method, and evaluate the enthalpic and entropic contributions to hydration, making visualisation of hydration thermodynamics possible. GCT is a slower method than other thermodynamic integration methods, but such alternative methods are not as descriptive in terms of thermodynamic contributions.

6. Outlook and conclusions

The aim of this review is to introduce the multitude of available methods and concepts for the calculation of solution free energies, and the modelling of systems in solution. Through the highlighting of many traditional and emerging methods within explicit, implicit, informatics and hybrid methods, it has become clear that each modelling category has its own advantages and disadvantages. The trade-off between the inaccuracies of implicit solvent models and the computational cost-constraints of explicit models is a prominent issue, and has conceived a number of hybrid solvation methods, each of which aims to provide a model of reasonable accuracy at an appropriate cost. The plethora of such available methods exemplifies the importance of accurate solvation models.

We have placed particular emphasis on 3D-RISM and its derived counterparts, as we believe that RISM based methods are a strong contender in the challenge of finding a computationally viable solubility prediction method which is also descriptive enough for the theoretical study of a system's thermodynamics. However, it is also noted that such methods are a long way from perfection, and require further refinements of solute–solvent correlation functions.

With the increase of computing power, as described by Moore's law, it is hard to predict how much of an issue computational costs associated with solvation modelling will be over the coming years. However, increases in computing power will inevitably allow more accurate methods to be employed within a faster timeframe. We predict the emergence of hybrid models which describe the theoretical and physical components of solvation at an ever increasing rate, with the need to trade off accuracy over time becoming less as computing power increases.

Although future prospects for solvation modelling are bright, we are also aware that there is a very present need for good models. We would like to note that the best choice in model for solvation is entirely dependent on the requirements of the user. For high-throughput screening of molecules of similar structural features, we suggest QSPR/QSAR as a suitable and reliable approach for thermodynamic property calculation (*e.g.*, solvation free energy). However, where specific physical and mechanistic meaning is desired, it is best to employ either explicit solvent representations, suitable for relatively small solute sizes, or where larger solutes are used, hybrid models. The choice of hybrid models for such investigations is not intuitively obvious, as highlighted within this review, as some systems are described sufficiently with addition of a single solute molecule, whereas for other systems it is necessary to add enough explicit solvent molecules to describe full solvation shells. Thus, it is often necessary to consider whether solvent behaviour is a significant contributor to the property of interest. If so, explicit/hybrid methods are advisable, dependent upon available computing resources. Otherwise, continuum models could offer sufficient physical description of the solvent environment. Of course, where sufficient and trustworthy experimental data are available, several models should be tested and evaluated for correlation with available experimental data.

Acknowledgements

We are grateful for useful discussions with colleagues including the groups of Professor Maxim Fedorov and Dr David Palmer. JLMcD and JBOM are grateful to SULSA for funding; RES and JBOM thank the University of St Andrews, EPSRC (grant EP/L505079/1) and CCDC for funding.

Notes and references

1 S. Basavaraj and G. V. Betageri, *Acta Pharm. Sin. B*, 2014, **4**, 3–17.
2 H. D. Williams, N. L. Trevaskis, S. A. Charman, R. M. Shanker, W. N. Charman, C. W. Pouton and C. J. H. Porter, *Pharmacol. Rev.*, 2013, **65**, 315–499.
3 C. J. Cramer, *Essentials of Computational Chemistry: Theories and Models*, John Wiley & Sons, Chichester, 2013.
4 A. Llinàs, R. C. Glen and J. M. Goodman, *J. Chem. Inf. Model.*, 2008, **48**, 1289–1303.
5 S. H. Yalkowsky and S. C. Valvani, *J. Pharm. Sci.*, 1980, **69**, 912–922.
6 A. G. Leach, H. D. Jones, D. A. Cosgrove, P. W. Kenny, L. Ruston, P. MacFaul, J. M. Wood, N. Colclough and B. Law, *J. Med. Chem.*, 2006, **49**, 6672–6682.
7 Medchemica, SALT MINER(TM), www.medchemica.com/salt.html, 2014.
8 J. Hussain, BioDig ADME: Using matched molecular pairs to find structural changes to improve your compound's properties, UK QSAR Spring Meeting 2011, Manchester, www.ukqsar.org/slides/JameedHussain_2011.pdf.
9 J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. DiStasio, M. Head-Gordon, G. N. I. Clark, M. E. Johnson and T. Head-Gordon, *J. Phys. Chem. B*, 2010, **114**, 2549–2564.

10 S. Y. Liem and P. L. A. Popelier, *Phys. Chem. Chem. Phys.*, 2014, **16**, 4122–4134.
11 A. A. Noyes and W. R. Whitney, *J. Am. Chem. Soc.*, 1897, **19**, 930–934.
12 A. Jouyban and M. A. A. Fakhree, *Toxicity and Drug Testing*, InTech, 2012.
13 G. Völgyi, E. Baka, K. J. Box, J. E. A. Comer and K. Takács-Novák, *Anal. Chim. Acta*, 2010, **673**, 40–46.
14 J. L. McDonagh, N. Nath, L. De Ferrari, T. van Mourik and J. B. O. Mitchell, *J. Chem. Inf. Model.*, 2014, **54**, 844–856.
15 B. A. Hendriksen, M. V. S. Felix and M. B. Bolger, *AAPS PharmSci*, 2003, **5**, 35–49.
16 J. Bauer, S. Spanton, R. Henry, J. Quick, W. Dziki, W. Porter and J. Morris, *Pharm. Res.*, 2001, **18**, 859–866.
17 S. R. Chemburkar, J. Bauer, K. Deming, H. Spiwek, K. Patel, J. Morris, R. Henry, S. Spanton, W. Dziki, W. Porter, J. Quick, P. Bauer, J. Donaubauer, B. A. Narayanan, M. Soldani, D. Riley and K. Mcfarland, *Org. Process Res. Dev.*, 2000, **4**, 413–417.
18 D. S. Palmer, A. Llinàs, I. Morao, G. M. Day, J. M. Goodman, R. C. Glen and J. B. O. Mitchell, *Mol. Pharm.*, 2008, **5**, 266–279.
19 D. S. Palmer, J. L. McDonagh, J. B. O. Mitchell, T. van Mourik and M. V. Fedorov, *J. Chem. Theory Comput.*, 2012, 3322–3337.
20 S. L. Price, M. Leslie, G. W. A. Welch, M. Habgood, L. S. Price, P. G. Karamertzanis and G. M. Day, *Phys. Chem. Chem. Phys.*, 2010, **12**, 8478–8490.
21 A. M. Reilly and A. Tkatchenko, *J. Phys. Chem. Lett.*, 2013, **4**, 1028–1033.
22 A. R. Leach and V. J. Gillet, *An Introduction to Cheminformatics*, Springer, Dordrecht, 2007.
23 J. B. O. Mitchell, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2014, **4**, 468–481.
24 D. Weininger, *J. Chem. Inf. Comput. Sci.*, 1988, **28**, 31–36.
25 S. Heller, A. McNaught, S. Stein, D. Tchekhovskoi and I. Pletnev, *J. Cheminf.*, 2013, **5**, 7.
26 L. D. Hughes, D. S. Palmer, F. Nigsch and J. B. O. Mitchell, *J. Chem. Inf. Model.*, 2008, **48**, 220–232.
27 A. Lusci, G. Pollastri and P. Baldi, *J. Chem. Inf. Model.*, 2013, **53**, 1563–1575.
28 Y. Ran and S. H. Yalkowsky, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 354–357.
29 Y. Ran, Y. He, G. Yang, J. L. H. Johnson and S. H. Yalkowsky, *Chemosphere*, 2002, **48**, 487–509.
30 J. Ali, P. Camilleri, M. B. Brown, A. J. Hutt and S. B. Kirton, *J. Chem. Inf. Model.*, 2012, **52**, 420–428.
31 A. Shayanfar, M. A. A. Fakhree and A. Jouyban, *J. Drug Delivery Sci. Technol.*, 2010, **20**, 467–476.
32 D. S. Palmer and J. B. O. Mitchell, *Mol. Pharm.*, 2014, **11**, 2962–2972.
33 J. Tomasi, B. Mennucci and R. Cammi, *Chem. Rev.*, 2005, **105**, 2999–3093.
34 L. Onsager, *J. Am. Chem. Soc.*, 1936, **58**, 1486–1493.
35 R. A. Pierotti, *Chem. Rev.*, 1975, **76**, 717–726.
36 S. Höfinger and F. Zerbetto, *J. Phys. Chem. A*, 2003, **107**, 11253–11257.
37 C. Colominas, F. J. Luque and M. Orozco, *Chem. Phys.*, 1999, **240**, 254–264.
38 J. Wang, T. Hou and X. Xu, *J. Chem. Inf. Model.*, 2009, **49**, 571–581.
39 Y. Takano and K. N. Houk, *J. Chem. Theory Comput.*, 2005, **1**, 70–77.
40 A. Klamt and G. Schuurmann, *J. Chem. Soc., Perkin Trans. 2*, 1993, 799.
41 A. Klamt, V. Jonas, T. Bürger and J. C. W. Lohrenz, *J. Phys. Chem. A*, 1998, **102**, 5074–5085.
42 A. V. Marenich, C. J. Cramer and D. G. Truhlar, *J. Phys. Chem. B*, 2009, **113**, 6378–6396.
43 S. Rayne and K. Forest, *Nature Precedings*, 2010, DOI: 10.1038/npre.2010.4864.1.
44 R. M. Levy and E. Gallicchio, *Annu. Rev. Phys. Chem.*, 1998, **49**, 531–567.
45 D. L. Beveridge and F. M. DiCapua, *Annu. Rev. Biophys. Biophys. Chem.*, 1989, **18**, 431–492.
46 R. W. Zwanzig, *J. Chem. Phys.*, 1954, **22**, 1420.
47 C. Chipot and A. Pohorille, *Free Energy Calculations: Theory and Applications in Chemistry and Biology*, Springer, Berlin, Heidelberg, 2007, pp. 33–75.
48 K. Lüder, L. Lindfors, J. Westergren, S. Nordholm and R. Kjellander, *J. Phys. Chem. B*, 2007, **111**, 1883–1892.
49 H. Liu, S. Dai and D. Jiang, *J. Phys. Chem. B*, 2014, **118**, 2719–2725.
50 M. A. Wyczalkowski, A. Vitalis and R. V. Pappu, *J. Phys. Chem. B*, 2010, **114**, 8166–8180.
51 A. Ahmed and S. I. Sandler, *J. Chem. Eng. Data*, 2015, **60**, 16–27.
52 R. A. Friesner and V. Guallar, *Annu. Rev. Phys. Chem.*, 2005, **56**, 389–427.
53 A. H. Steindal, K. Ruud, L. Frediani, K. Aidas and J. Kongsted, *J. Phys. Chem. B*, 2011, **115**, 3027–3037.
54 P. Mark and L. Nilsson, *J. Phys. Chem. A*, 2001, **105**, 9954–9960.
55 M. W. Mahoney and W. L. Jorgensen, *J. Chem. Phys.*, 2000, **112**, 8910.
56 C. Vega and J. L. F. Abascal, *Phys. Chem. Chem. Phys.*, 2011, **13**, 19663–19688.
57 H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura and T. Head-Gordon, *J. Chem. Phys.*, 2004, **120**, 9665–9678.
58 L. Wang, T. J. Martinez and V. S. Pande, *J. Phys. Chem. Lett.*, 2014, **5**, 1885–1891.
59 D. van der Spoel, P. J. van Maaren and H. J. C. Berendsen, *J. Chem. Phys.*, 1998, **108**, 10220.
60 A. Jones, F. Cipcigan, V. P. Sokhan, J. Crain and G. J. Martyna, *Phys. Rev. Lett.*, 2013, **110**, 227801.
61 D. S. Palmer, A. I. Frolov, E. L. Ratkova and M. V. Fedorov, *J. Phys.: Condens. Matter*, 2010, **22**, 492101.
62 D. S. Palmer, J. L. Mcdonagh, J. B. O. Mitchell, T. van Mourik and M. V. Fedorov, *J. Chem. Theory Comput.*, 2012, 3322–3337.
63 T. Luchko, S. Gusarov, D. R. Roe, C. Simmerling, D. A. Case, J. Tuszynski and A. Kovalenko, *J. Chem. Theory Comput.*, 2010, **6**, 607–624.

64 D. Chandler, J. D. McCoy and S. J. Singer, *J. Chem. Phys.*, 1986, **85**, 5971.
65 A. Kovalenko and F. Hirata, *J. Chem. Phys.*, 1999, **110**, 10095.
66 E. L. Ratkova and M. V. Fedorov, *J. Chem. Theory Comput.*, 2011, **7**, 1450–1457.
67 J.-F. Truchon, B. M. Pettitt and P. Labute, *J. Chem. Theory Comput.*, 2014, **10**, 934–941.
68 Y. Maruyama, N. Yoshida, H. Tadano, D. Takahashi, M. Sato and F. Hirata, *J. Comput. Chem.*, 2014, **35**, 1347–1355.
69 D. S. Palmer, A. I. Frolov, E. L. Ratkova and M. V. Fedorov, *Mol. Pharm.*, 2011, **8**, 1423–1429.
70 J.-F. Truchon, B. M. Pettitt and P. Labute, *J. Chem. Theory Comput.*, 2014, 934–941.
71 V. P. Sergiievskyi, G. Jeanmairet, M. Levesque and D. Borgis, *J. Phys. Chem. Lett.*, 2014, **5**, 1935–1942.
72 F. Hirata, *Molecular theory of solvation*, Springer, 2003.
73 A. Kovalenko and F. Hirata, *J. Mol. Liq.*, 2001, **90**, 215–224.
74 C. P. Kelly, C. J. Cramer and D. G. Truhlar, *J. Phys. Chem. A*, 2006, **110**, 2493–2499.
75 C. P. Kelly, C. J. Cramer and D. G. Truhlar, *J. Chem. Theory Comput.*, 2005, **1**, 1133–1152.
76 W. Zhu and G. Krilov, *THEOCHEM*, 2008, **864**, 31–41.
77 S. Chaudhury, M. A. Olson, G. Tawa, A. Wallqvist and M. S. Lee, *J. Chem. Theory Comput.*, 2012, **8**, 677–687.
78 G. Gerogiokas, G. Calabro, R. H. Henchman, M. W. Y. Southey, R. J. Law and J. Michel, *J. Chem. Theory Comput.*, 2014, **10**, 35–48.