Machine Learning Guided Approach for Studying Solvation Environments

Yasemin Basdogan,[†] Mitchell C. Groenenboom,[†] Ethan Henderson,[†] Sandip De,[‡] Susan Rempe,[¶] and John A. Keith[*,†]

[†]*Department of Chemical and Petroleum Engineering Swanson School of Engineering, University of Pittsburgh, Pittsburgh, USA*

[‡]*Laboratory of Computational Science and Modelling, Institute of Materials, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland*

[¶]*Department of Nanobiology, Sandia National Laboratories, Albuquerque, USA*

E-mail: jakeith@pitt.edu

Abstract

Molecular level understanding and characterization of solvation environments is often needed across chemistry, biology, and engineering. Toward practical modeling of local solvation effects of any solute in any solvent, we report a static and all-quantum mechanics based cluster-continuum approach for calculating single ion solvation free energies. This approach uses a global optimization procedure to identify low energy molecular clusters with different numbers of explicit solvent molecules and then employs the Smooth Overlap for Atomic Positions (SOAP) learning kernel to quantify the similarity between different low-energy solute environments. From these data, we use sketch-maps, a non-linear dimensionality reduction algorithm, to obtain a two-dimensional visual representation of the similarity between solute environments in differently sized microsolvated clusters. After testing this approach on different ions having charges of 2+, 1+, 1-, and 2-, we find that the solvation environment around each

ion can be seen to progressively become more similar along with each corresponding calculated single-ion solvation free energy. Without needing either dynamics simulations or an *a priori* knowledge of local solvation structure of the ions, this approach can be used to calculate solvation free energies with errors within five percent of experimental measurements for most cases, and it should be transferable for the study of other systems where dynamics simulations are not easily carried out.

1 Introduction

Solvation plays an essential role in chemical and biological processes ranging from homogeneous catalysis to ion channel transport to energy storage. In many cases, the explicit interactions between small ions with nearby solvent molecules are crucial for molecular-scale understanding of the systems. In such cases, single-ion solvation free energies can be several hundreds of kcal/mol (or greater than 10 eV), which can make accurate predictions quite challenging. Molecular dynamics (MD) or Monte Carlo (MC) simulations have been used, notably for systems that have anions and complex small molecules,[1–3] but the accuracy of these simulations depends on the availability of high-quality force field parameters. In the absence of reliable parameters, MD simulations involving quantum mechanics (QM) calculations can be accurate, but they are far more computationally laborious. Semi-empirical continuum solvation models (CSMs)[4–9] have been developed as practical means to determine absolute solvation free energies, but CSMs can sometimes result in large errors, especially with systems that have non-uniform charge distributions. Such errors can significantly impede predictions of thermodynamic properties and severely bias mechanistic predictions.[10]

A standard approach to address these problems has been to include explicit solvent molecules with the QM calculation of the solute, using cluster-continuum or mixed implicit/explicit modeling, since this often provides better solvation free energies from thermodynamic cycles. Of the methods in this classification, the cluster formulation of quasi-chemical theory (QCT), developed by Pratt, Rempe, and colleagues, is a rigorous treat-

ment that uses an electronic structure calculation on the ion with one or more solvation shells.[11–13] This approach has produced accurate predictions of solvation free energies for hydration of alkali metal ions (Li^+, Na^+, K^+, Rb^+),[14–19] alkaline earth metals (Mg^{2+}, Ca^{2+}, Sr^{2+}, Ba^{2+}),[20–23] transition metals,[24,25] halide ions (F^-, Cl^-),[26–28] small molecules (Kr, H_2, CO_2),[29–32] ion solvation from non-aqueous solvents,[22,33] and binding sites of proteins and other macromolecules,[34–40] generally to within 5% error.[23] However, the correct use of QCT requires determining an appropriate solvation shell for the system, and this can be non-trivial.[41,42]

Adding to the complexity of single-ion solvation predictions is that there are two different free energy scales that are frequently misunderstood or not acknowledged in the literature. One is often called the 'absolute' scale, while the other is called the 'real' scale. The real scale includes the phase potential[43] or surface potential,[44–46] which is the total reversible work to move an ion across the vacuum/liquid interface, whereas the absolute scale does not. The absolute scale is associated with data from Marcus, who studied and reported experimental solvation free energies for a large number of ions.[47] Those data rely on the 'classical' extra-thermodynamic assumption, referred to as the TATB hypothesis, that two specific ions of opposite charges have similar absolute free energies. That hypothesis assumes the system is independent of any interfacial potential that arises from the anisotropic distribution of the solvent molecules near the interface.[48] In a real physical system, a solvation free energy will also include a phase potential contribution that depends on the interfacial potential at the water-air interface. The real scale can be associated with data from Tissandier et al., who have extrapolated conventional free energy measurements on small ionic hydrates to obtain real solvation free energies of ions in bulk phase.[49] This idea is often referred as the cluster pair-based (CPB) approximation.

The absolute solvation free energy scale can be converted into the real solvation free

energy scale by incorporating the phase potential using the following equation:

$$\Delta G_{\text{solv}}^{\text{real}} = \Delta G_{\text{solv}}^{\text{abs}} + zF\phi \tag{1}$$

where F is the Faraday constant, z is the atomic charge and ϕ is the interfacial potential. Table S1 compares Marcus's data with data from Tissandier et al., and it highlights the phase potential contribution in real solvation free energy calculations, which is ∼-10 kcal/mole for alkali metals and ∼12 kcal/mole for halides. With two sets of experimental data to compare to, there has often been a lack of consensus on which calculation schemes result in which solvation free energy scale and why. It is generally understood that free energy calculations using periodic boundary conditions (such as MD and MC simulations) do not include the phase potential contribution, and thus represent absolute solvation data because there is no physical vacuum/liquid interface.[50]

For cluster-based calculations this is murkier. Specifically, QCT literature cites the absence of phase potentials in theoretical predictions and reports data in closest agreement with the absolute solvation data of Marcus,[23] while other computational studies using a similar thermodynamic cycle and cluster-continuum approach have reported closer agreement with the real solvation scale.[51,52] Of course, solvation energies will depend on how many solvent molecules are used and where they are placed. Kemp and Gordon demonstrated the effective fragment potential (EFP) method, coupled with Monte Carlo simulations can be used to study the solvation of F^- and Cl^- anions.[53] Their approach showed that 15 water molecules in this model were required to fully solvate a single F^- anion while 18 water molecules were required to completely solvate a Cl^- anion. Merz and coworkers also used molecular dynamics simulations to identify the first solvent shell that can be used in calculations using continuum solvation methods.[54] It would be beneficial to have a general and automatable way to model local solvation environments (of any solute in any general solvent) without using molecular dynamics simulations that can be computationally prohibitive.

This work elucidates the theory between two different thermodynamic cycles (schemes 1 and 2) and how they result in two different solvation free energy scales. To automatically generate microsolvated clusters, we used a global optimization method, called ABCluster.[55,56] We then calculated the real solvation free energies with cluster-continuum modeling using the thermodynamic cycle outlined in scheme 2. We initially hypothesized that solvation free energies will improve if we systematically add explicit water molecules around each ion while ensuring that each microsolvated state is a reasonable approximation of a thermodynamically low energy structure. A similar idea was previously studied by Bryantsev and co-workers by increasing water cluster sizes to 18 explicit solvent molecules around the Cu^{2+} ion, which significantly decreased the error compared to the CSM-computed solvation free energy.[57] Here, we introduce a way to leverage machine learning (ML) to study local solvation structures. Unlike other studies that use supervised ML to predict solvation energies with algorithms like quantitative structure-activity relationship models (QSAR),[58] random forest,[59] decision tree,[60] or artificial neural networks,[61] we are using unsupervised ML to study how similarities between microsolvated structures coincide with solvation energies so that one can learn the inherent arrangement of our data without using explicitly provided labels. We first assemble our data-set of microsolvated structures and then use dimensionality reduction algorithms to study similar patterns in the microsolvated clusters. One of the main challenges in applying ML techniques to chemistry problems is to find the best representation of the system so that it is complete and concise. In this study we use Smooth Overlap of Atomic Positions (SOAP) descriptors[62] to represent our microsolvated clusters. Next we use sketch-maps to reduce the dimensions of our feature vectors and to study similar patters between our microsolvated clusters.[63] Using the combination of SOAP and sketch-map ML algorithms we demonstrate that low energy molecular clusters produced by our procedure have structurally similar local solvation environments. This suggests that this calculation scheme can be used to quantify the number of explicit solvent molecules needed to accurately model the relevant local solvation environment of a charged solute.

2 Theory

Cluster continuum modeling has been used in different formulations to calculate solvation free energies of small ions.[54,64–67] These methods involve different approximations, ranging from including a single water molecule to using MD simulations to obtain physical solvent structures at room temperatures. This hybrid approach has received further developments in the theory (e.g. by using cluster expansion treatments).[68–70] QCT is the most robust approach of these because it is based on statistical mechanics,[71] and it has been proven to be reliable in different applications.[13–16,20,23,24,27,28,33,34]

The starting point for QCT is to partition the region around the solute into inner and outer shell solvent domains. Akin to cluster-continuum modeling schemes, the inner shell is typically treated quantum mechanically, while the outer shell is treated with a dielectric continuum model. Applied to hydration of ions X with charge $m\pm$, the inner-shell reactions are given as cluster association equilibria:

$$\mathrm{X}^{m\pm} + n\mathrm{H_2O} \rightleftharpoons \mathrm{X(H_2O)}_n^{m\pm} \tag{2}$$

A clustering algorithm is applied to identify the populations of the clusters on the right side of Eq. 2. The theory then treats the cluster $\mathrm{X(H_2O)}_n^{m\pm}$ as a molecular system under analysis. A natural procedure to identify inner-shell configurations for an ion is to define waters within a distance λ from an ion as an inner-shell partner. With n water ligands in the cluster, the excess chemical potential, or hydration free energy, consists of several terms,

$$\mu_{\mathrm{X}}^{(\mathrm{ex})} = -kT \ln \left[K_n^{(0)} {\rho_{\mathrm{H_2O}}}^n \right] + kT \ln \left[p_{\mathrm{X}}(n) \right] + \left(\mu_{\mathrm{X(H_2O)}_n}^{(\mathrm{ex})} - n\mu_{\mathrm{H_2O}}^{(\mathrm{ex})} \right) \tag{3}$$

This formula is correct for any physical choices of λ and n.

The terms in Eq. 3 describe contributions to the total ion hydration free energy from the inner and outer shell solvation environments. The first term gives ion association reactions with water molecules in the inner shell taking place in an ideal gas phase. The association

reactions are scaled by the water density, ρ_{H_2O}, to account for the availability of water ligands to occupy the inner shell. The second term accounts for the thermal probability that a specific ion has n inner-shell partners in solution. The last terms describe the solvation of the $X(H_2O)_n^{m\pm}$ cluster and the de-solvation of n individual water molecules from aqueous solution in the outer-shell environment.

A judicious selection of λ and n in Eq. 3 can simplify the free energy analysis. First, by considering a specific λ, the minimum value in $kT\ln[p_X(n)]$ identifies the most probable n, denoted as $\bar{n}$. Then that contribution, associated with the work of selecting n waters for ion association, can be dropped to result in Eq. 4,

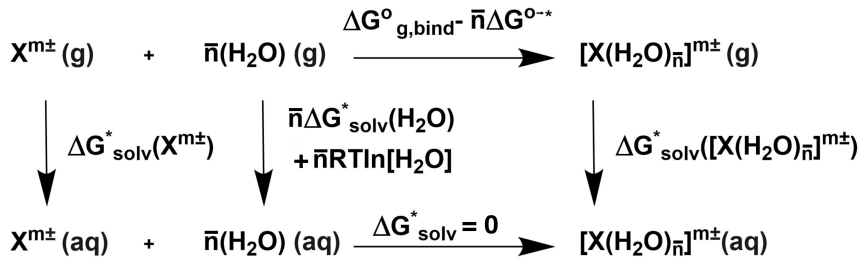
$$\mu_X^{(ex)} \approx -kT \ln\left[K_{\bar{n}}^{(0)} {\rho_{H_2O}}^{\bar{n}}\right] + \left(\mu_{X(H_2O)_{\bar{n}}}^{(ex)} - \bar{n}\mu_{H_2O}^{(ex)}\right) \tag{4}$$

Alternatively, the magnitude of the contribution, $kT\ln[p_X(n)]$ from Eq. 3 can be estimated from molecular simulation results for any n. Second, CSMs can be used to determine outer shell contributions. With most CSMs, the external boundary of the model cavity is defined by spheres centered on each of the atoms. Typically, CSM results are sensitive to the radii of the spheres that define the solute cavity, but when the ion is surrounded by a full shell of solvating ligands, the sensitivity is lessened (when the radii for the ligands are adequate), and this results in a fortuitous error cancellation in the last terms of (Eq. 3 and 4). Third, selecting clusters with small n generally results in stronger solute-solvent interactions, which helps ensure that vibrational motions are characterized by small displacements from equilibrium, which is required when assuming a harmonic potential energy surface for the analysis of a free energy. Prior work suggests that anharmonic vibrational motions become prominent with clusters as small as n=5 for Na^+ and K^+ ions in clusters with water molecules,[37] and they can be even smaller sizes for anion-water clusters.[28]

As an aside, the solvation energy represented in Eq. 4 can also be equivalently represented using the thermodynamic cycle shown in Scheme 1 which is mathematically expressed (us-

ing different notation) with Eq 5. This alternative notation is based on the observable macroscopic quantities coming from thermodynamics and is often used in chemistry and engineering communities (e.g. see Refs.[66] and[57]). We note that when Eq 5 and Eq 6 are written in per mole basis so they are equivalent to chemical potentials.

Scheme 1: Monomer cycle for calculating an absolute solvation free energy.



$$\Delta G^{*}_{\mathrm{solv}}(\mathrm{X}^{m\pm}) = \Delta G^{\circ}_{\mathrm{g,bind}} - \bar{n}\Delta G^{\circ\to*} + \Delta G^{*}_{\mathrm{solv}}(\mathrm{X(H_2O)}^{m\pm}_{\bar{n}}) - \bar{n}\Delta G^{*}_{\mathrm{solv}}(\mathrm{H_2O}) - \bar{n}RT\ln[\mathrm{H_2O}] \quad (5)$$

Scheme 1 is sometimes called the "monomer cycle" since it involves individual water monomers rather than a water cluster. The free energy of binding a gas phase cluster is expressed as $\Delta G^{\circ}_{\mathrm{g,bind}}$, where the circle denotes a the free energy difference at a gas phase standard state of 1 bar. The solvation free energies ($\Delta G^{*}_{\mathrm{solv}}$) are labeled with asterisks to denote energies conventionally expressed at an aqueous standard state of 1 M, and they are calculated here using SMD implicit solvation. Additional corrections ($\Delta G^{\circ\to*}$, each having a magnitude of 1.9 kcal/mol or 0.08 eV) are needed to account for the change from a gas phase standard state to an aqueous phase standard state. Here, we use scheme 1 to map the QCT theory on macroscopic variables. Solvation energies calculated using Scheme 1 are comparable to the Marcus scale and thus will be comparable to MD/free energy perturbation or QCT studies that do not have the phase potential contribution.[72]

Of the different cluster continuum procedures that do not require dynamics, the procedure by Bryantsev et al. is promising since it appears to yield solvation free energies that agrees

well with the experimental data for both the proton and Cu^{2+}, and with results that appear to match the real solvation scale. Their cycle, outlined in Scheme 2, is similar to the monomer cycle in Scheme 1, but it involves pre-formed water clusters containing n interacting water molecules that have been optimized at 0 K and free energy contributions are obtained using standard ideal gas, rigid rotor, and harmonic oscillator approximations. The single ion solvation free energy from the Scheme 2 cluster cycle is calculated with Eq. 6:

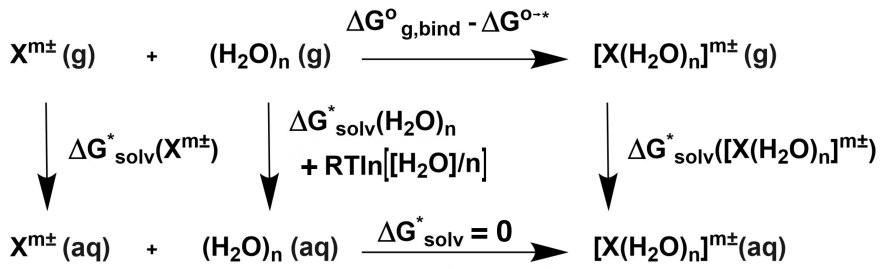
$$\Delta G^*_{\rm solv}(X^{m\pm}) = \Delta G^\circ_{\rm g,bind} - \Delta G^{\circ\to *} + \Delta G^*_{\rm solv}(X({\rm H_2O})^{m\pm}_n) - \Delta G^*_{\rm solv}({\rm H_2O})_n - {\rm RTln}([{\rm H_2O}]/n) \tag{6}$$

Scheme 2 also evaluates the same QCT theory shown in Eq. 3, but by applying QCT to both the water dehydration problem ($\mu^{\rm (ex)}_{\rm H_2O}$) as well as the ion hydration problem ($\mu^{\rm (ex)}_{\rm X}$). This dual QCT approach has advantages due to anticipated error cancellations.[32] Successful use of CSM models is known to require that CSM parameters are properly chosen since results can vary greatly with surface type, cavity size, and continuum model used.[73] However, by using similar sizes of clusters for ion hydration and water dehydration, the boundary λ between inner and outer shells is approximately balanced on both sides of the equation, leading to a cancellation of errors to the outer-shell solvation contribution from a CSM model.

The same balance in cluster sizes may also lead to an approximate cancellation of anharmonic contributions in the inner-shell contributions to the solvation free energy. Eq. 3 depends on using the most probable n to eliminate the $kT\ln[p_{\rm X}(n)]$ term (as done in Eq.4), or requires molecular simulations to explicitly evaluate that term. It also needs a filled inner-shell occupancy so that the CSM model is minimally dependent on specific radii used to compute the outer shell contribution to hydration free energy. Scheme 2 approximately eliminates these constraints by error cancellations. With Scheme 2, large n values can be used; however, care must be applied when the cavity radius is around 6 Å. In such length scales and above, the surface or phase potential contributions to the solvation free energy, ϕ,

should be included in the calculation.[74–78] In the analysis here, outer-shell contributions to the solvation free energy go to zero as cluster size increases,[20,23] and then the phase potential enters into the calculation and then is accounted for naturally. This explains why results from Scheme 2 agree better with the real solvation scale than the absolute solvation scale. We acknowledge there are limitations of Scheme 2, for instance the use of 0 K optimized water clusters would not physically represent real water droplets at ambient temperatures, but by using similar sizes of clusters for ion hydration and water dehydration, one might take advantage of fortuitous error cancellation from this scheme.

Scheme 2: Cluster cycle for the calculation of real solvation free energy.



As mentioned before, a limitation with using Scheme 2 is to know how many solvent molecules should be used. To identify the degree of solvation necessary we employ ML algorithms first to assemble our micrsolvated structures then to study the similarities between the structures. The main challenge in using ML based algorithms is to come up with an appropriate representation that will give a complete description of your system. SOAP algorithm makes use of Gaussian functions to find a best match between pairs of atomic environments. Previously, it has been used to study different geometries of fullerene, amorphous silicon, pentacene, and ice structures.[63,79–82] We believe SOAP is a good representation for our systems because it is invariant to rotations, translations, and permutations. Assembling techniques like SOAP identify an agglomeration of data in chemical space, but they do not offer a Euclidan relationship between different structures. Such relations can be determined and represented using maps that represent geometrically similar structures as data points

that are adjacent to each other. Here, we use sketch-map non-linear dimensionality reduction techniques to construct a two-dimensional representation of the free energy surface. Given a certain cut-off radius, this algorithm identifies structures that show similar local solvation motifs, e.g. if two points are close to each other on the sketch-map then the local solvation environments from the two data points are relatively similar.

In this study, we will apply Scheme 2 to systematically model microsolvated ions and water clusters using $n = 1-20$ water molecules. Below, we will show a modeling scheme that involves modern tools such as ABCluster, dispersion-corrected Kohn-Sham density functional theory, and the SOAP algorithm to analyze this thermodynamic cycle to quantify energy contributions, assess likely causes for errors, and understand the local structures of water molecules in these solvation environments. While more calculations are required for Scheme 2 than would be needed for Scheme 1, we find that the former scheme provides reasonably accurate single ion solvation free energies while also eschewing the need for *a priori* knowledge of the solvation environment. Thus, calculations from such cycles should be generalizable and easily automatable for any solute in any solvent environment.

3 Computational Methodology

We generated microsolvated solutes using the rigid molecular optimizer module of the ABCluster program.[55,56] We generated 1,000 low energy candidates using CHARMM forcefield parameters from MacKerell's CGenFF website together with TIP4P water parameters.[83] All force filed parameters used in this study are reported in Table S8. Five lowest energy structures obtained from the CHARMM forcefield optimization were then further optimized at the BP86[84,85]-D3BJ[86]/def2-SVP[87] or B3LYP[88]-D3BJ[86]/def2-SVP[87] level of theory, as implemented in ORCA[89] using the RI-J or RIJCOSX approximations. To study the solvation effects we have used SMD implicit solvation with both geometry optimizations and single point energy calculations. Free energy contributions were calculated using the ideal gas, rigid

rotor, and harmonic oscillator approximations at the same level of theory as the geometry optimizations. Finally, to assess the significance of higher levels of theory, we calculated single point energies on fully optimized geometries at the B3LYP[88]-D3BJ[86]/def2-TZVP[87] and ωB97X-D3[90]/def2-TZVP[87] levels of theory. Every energy reported in this manuscript is the Boltzmann-weighted average of the five low energy structures identified with a global optimization code (ABCluster). It is our experience that ABCluster with appropriate sampling is reliable at identifying low energy candidates as demonstrated by our comparisons to identifying globally optimized water clusters, but caution is still recommended in future work to ensure that meaningful structures are obtained. The thermodynamic cycle reported in Scheme 2 requires calculations on water clusters. To generate the water clusters, we followed the same procedure outlined above using with TIP4P parameters for the water molecules.[91] Cluster geometries were then optimized at the same level of QM theory as the solute-solvent clusters, as discussed above.

Finally, to compare our microsolvated structures with structures obtained from MD trajectories we performed simulations using the AMEOBA force field[92] with the TINKER[93] software package. First, we performed NPT simulations with a water box of 500 water molecules and equilibrated for 200 picoseconds at 298.15K and 1 atm. Next, we inserted an ion into the system (while removing one water molecule) and performed NVT simulations for another 200 picoseconds. Finally, we extracted 100 structures from the NVT trajectory and compared them with the structures generated with our clustering approach.

4 Results and Discussion

We first benchmarked low energy water clusters generated from ABCluster compared to global minimum energy water clusters from the Cambridge Cluster Database that also used the TIP4P forcefield.[94] In all cases (n = 1, 2, 4, 8, 12, 16, 20), the energy differences between our structures and the reference structures from the database were all at most +1.2 kcal/mol

(Table S2). This agreement demonstrates that ABCluster is an effective tool for identifying low energy structures and that our water clusters are comparable to well-established and globally optimized water cluster structures.

Next, we calculated solvation free energies of $(H_2O)_n$ clusters using the thermodynamic cycle outlined in Scheme S1 and compared it with solvation energy calculations using the SMD solvation model. Table S3 shows solvation free energies for water clusters derived using the thermodynamic cycle shown in Scheme S1, solvation free energies calculated directly using a CSM, and the difference between the two calculation schemes. For cases where $n = 2, 4, 8$, the difference in the two sets of solvation free energies is under 5 kcal/mol. However, for larger clusters ($n = 12, 16, 20$), the difference in free energies from these two calculation schemes significantly increases by as much as 30 kcal/mol. This trend shows that when relatively large clusters of water are solvated with a CSM, the model seems to introduce significant errors that would then make them less reliable if used for calculations with Scheme 2. The observation also in part justifies the use of QCT methods that use Scheme 1 and relatively small cluster sizes. The lowest error arises with $n = 4$ because it is the most probable size for water clusters, making the ln $[p_X(n)]$ term in Eq. 3 go approximately to zero.

We then benchmarked calculated gas phase binding free energies of water molecules to different ions against experimental data.[95,96] Table S4 shows gas phase binding free energies for one water molecule and different ions. For all the ions (Li^+, Na^+, K^+, Cl^-, Br^-, F^-), calculations are in good agreement with the experimental data, and our errors are mostly under 5 kcal/mol. We further compared the Na^+ and K^+ binding free energies with one water molecule to other computational studies, and Table S4 shows that different calculations agree reasonably well with experimental data.[17,19] We also calculated binding free energies involving four water molecules using Schemes 1 and 2, and compared them with both experimental and other computational studies (Table S5). The reference experimental data used for comparison add water molecules one by one to the system.[95,96] Using Scheme 1, we get

very good agreement with the experimental values and our errors are under 5 kcal/mol for all ions (Li^+, Na^+, K^+, F^-, Cl^-, Br^-). The data are also in relatively close agreement with gas phase binding free energies calculated by Rempe and coworkers.[24] However, when we used Scheme 2 to calculate binding of ions to four water molecules, we obtained free energies that differ from the prior work by 10 kcal/mol. This difference is anticipated on the basis that Scheme 2 applies to solvation reactions, not gas phase association reactions.

After benchmarking our calculations, we calculated hydration free energies using Eq. 6, which relies on the cluster thermodynamic cycle as outlined in Scheme 2. We considered a data set of ions having difference sizes and charges of 2+, 1+, 1-, and 2-. Figures 1 and 2 show hydration free energies for Na^+, Mg^{2+}, Cl^-, SO_4^{2-}, and similar data are reported for all ions and shown in Figures S1-S7. Table S6 shows hydration free energies for all the solutes, and the percent error calculated by taking absolute hydration free energies from Marcus's study and adding the phase potential contribution taken from Lamoureux and Roux,[50] using Eq.1 or comparing with experimental values from Tissandier et. al.[49] We also calculated and compare hydration free energies for ion-pairs. Table S7 shows hydration free energies for ion pairs and how they agree with both Marcus et. al.[47] and Tissandier et. al.[49] For all ions we report solvation free energies computed with both BP86-D3BJ and B3LYP-D3BJ geometries to compare the relative importance of including exact exchange in these systems. In all cases, B3LYP-D3BJ geometries result with more consistent solvation energies compared to BP86-D3BJ geometries, and thus using higher levels of theory that are appropriate for each system is recommended whenever possible. In Figure 1a and 1c, for Na^+ and Mg^{2+} cations, respectively, the hydration free energies appear to get closer to the experimental data when we gradually increase the number of water molecules in the system. For Na^+ starting with 8 water molecules we get good agreement with the experimental data. Similarly, For Mg^{2+} starting with 12 water molecules we get good agreement with the experimental data. In Figures 2a and 2c, for Cl^- and SO_4^{2-} anions, respectively, the same inference does not hold. The anion hydration free energies are not particularly sensitive to

water cluster size, and hydration free energies begin to deviate more from experiment when using 16 and 20 water molecules. Thus, our initial hypothesis that adding more solvent molecules into the system should generally improve the accuracy compared to experiment was mainly true only for the cations we modeled.

We then hypothesized that different microsolvated ion clusters might have significantly different solute structures, which result in different hydration free energies, as shown in Figures 1 and 2. To test this idea, we studied geometric similarities in cluster sizes. We used the SOAP kernel to quantify the similarity between solute environments.[62,63] For the high-dimensional pair-similarity data on solute environments, we used "sketch-maps", a non-linear dimensionality reduction technique.[97,98] Sketch-maps allow us to obtain a two-dimensional map that provides a meaningful visual representation of the similarity between solvent environments and solvent arrangements.

We define the local solute environment as a sphere of radius r+2 Å, centered at the solute atom, where r is the atomic radius. For complex anions, the environment center was chosen to be the position of the central atom, and r was taken to be the average central atom-oxygen distance plus the atomic radius of the oxygen atom. This cutoff was chosen to capture the evolution of the local solvation environment around the ion as increasing numbers of water molecules were added, while disregarding solvent molecules further from the ion. The atoms within the cutoff distance (including those of the solute) contribute to a smooth representation of the atom density, which is used to define a similarity measure between structures invariant to permutation of atom indices as well as rigid translations and rotations.

In our previous study using SOAP, we applied the pairwise similarity between configurations to comment on structural analogies.[99] To provide an intuitive representation of the relationships between all pairs within a group of structures, we used a sketch-map based on the SOAP metric. Each point on the map represents a solute environment. Data points in close proximity indicate systems with high similarity in local solvent environments. The

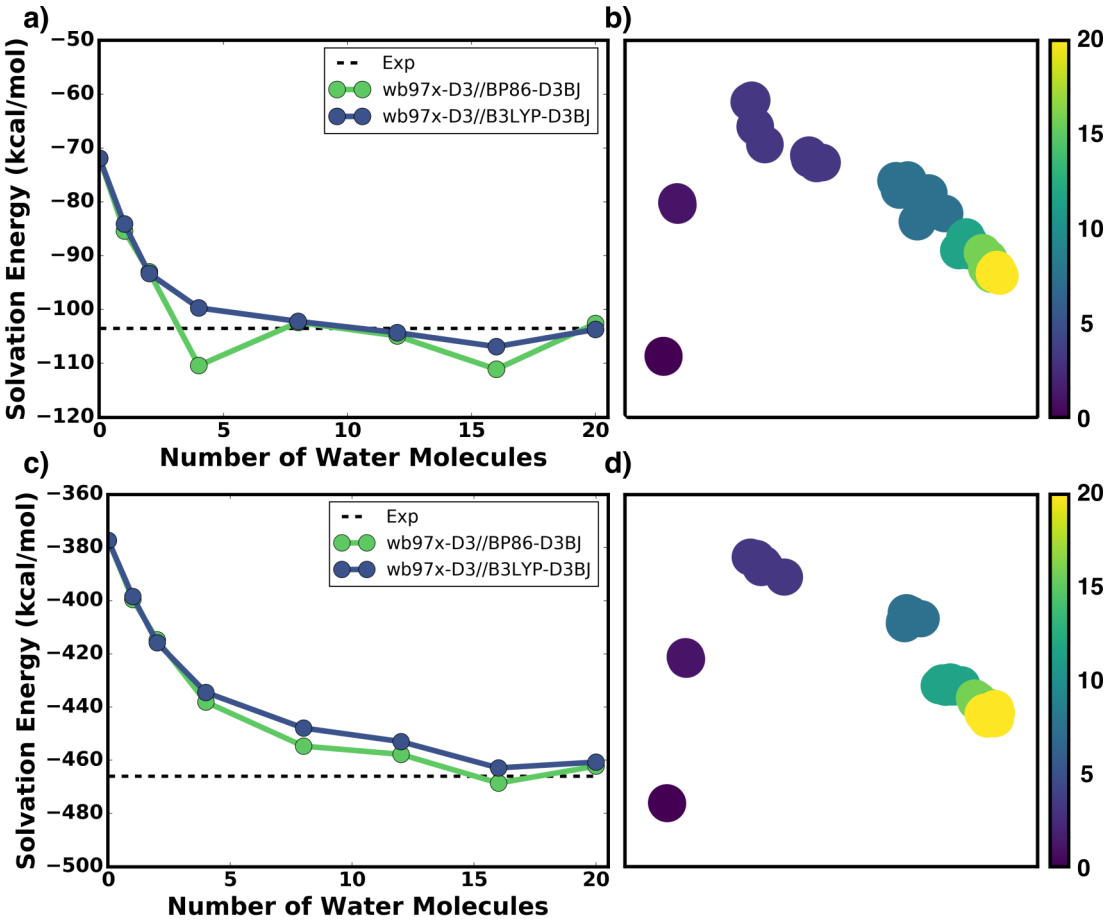


Figure 1: Hydration free energy plots and sketch-maps for Na^+ and Mg^{2+}. Plots show (a) hydration free energies calculated with Eq.6 for Na^+, (b) SOAP/sketch-map analysis for Na^+, (c) hydration free energies calculated with Eq.6 for Mg^{2+}, (d) SOAP/sketch-map analysis for Mg^{2+}. Data are from ωB97X-D3/def2-TZVP calculations on BP86-D3BJ/def2-SVP or B3LYP-D3BJ/def2-SVP geometries. Color bar shows the number of water molecules in the system.

sketch-map algorithm follows a non-linear optimization procedure where the discrepancy between pairwise Euclidean distances in low dimension and the kernel-induced metric is minimized. A sigmoid function is applied to focus the optimization on the most relevant range of distances, e.g. disregarding thermal fluctuations. The parameters of this filter are in the format, sigma-a-b. In all cases, we used a=3 and b=8, while sigma values were adapted to different systems following the heuristics described in Ref.[100] In Figures 1 and 2, we used a coloring scheme in which the color gets lighter when the size of the cluster increases. Yellow dots represent the clusters that have 20 water molecules, and dark purple dots represent the

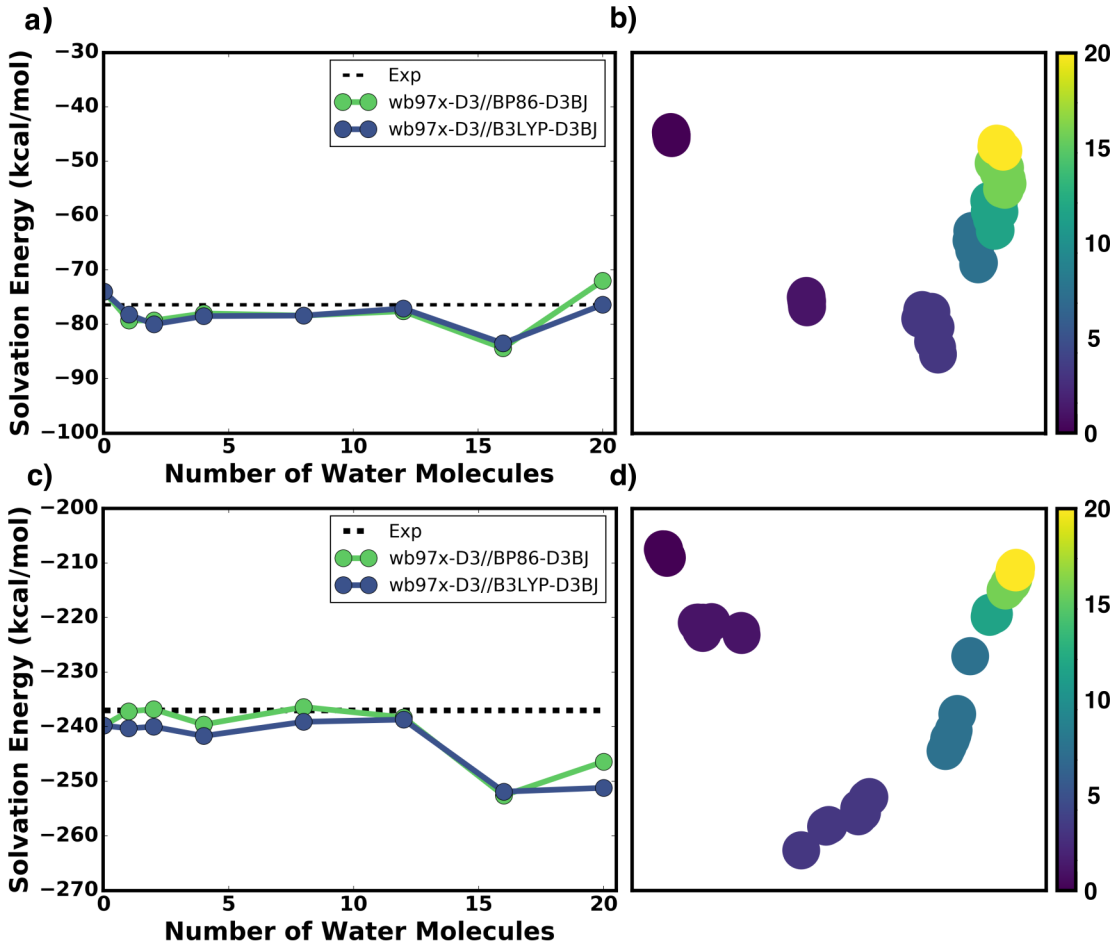


Figure 2: Hydration free energy plots and sketch-maps for Cl^- and SO_4^{2-}. Plots show (a) hydration free energies calculated Eq.6 for Cl^-, (b) SOAP/sketch-map analysis for Cl^-, (c) hydration free energies calculated with Eq.6 for SO_4^{2-}, (d) SOAP/sketch-map anaylsis for SO_4^{2-}. Data are from ωB97X-D3/def2-TZVP calculations on BP86-D3BJ/def2-SVP or B3LYP-D3BJ/def2-SVP geometries. Color bar shows the number of water molecules in the system.

clusters with only one water molecule.

Figure 1b demonstrates that models for Na^+ starting with eight or more water molecules give good agreement with the experimental values and all structures exhibit a similar solvent environment around the ion. Similarly for Mg^{2+}, Figure 1d shows a similar set of results except that models should have 12 or more water molecules. Given the short-range cutoff for the SOAP descriptors, these results indicate that the relatively accurate hydration free energies for the cations appear to be correlated with a similar local solvent environment.

However, this correspondence between structures and energetics is not observed for the anions (Cl^-, SO_4^{2-}). The sketch-map for Cl^- in Figure 2b shows that local solvation environments start becoming progressively similar with $n = 8$, but there is a 10 kcal/mol deviation in energy with clusters that involve 16 and 20 water molecules. A similar trend is seen for the sketch-map for SO_4^{2-} in Figure 2d. As with the cations, we see the local solvation structures start to become more similar with about $n = 12$, but the solvation energies for the 16 and 20 molecule water clusters are inaccurate compared to experiment by 15 kcal/mol. This discrepancy suggests that the errors shown in Figure 2 with the 16 and 20 molecule water clusters likely arise from an imbalance in anharmonic effects in Scheme 2 because the ion-water clusters have more anharmonicity than the water-water clusters. Nevertheless we can still use sketch-maps to identify the number of solvent molecules needed to see a analogous solvent arrangements. For example with Cl^-, the local solvation environment for progressively larger microsolvation environments starts becoming similar with $n = 8$ and when we calculate the hydration free energy with eight water molecules we get a relatively good agreement with the experimental data. Similarly with SO_4^{2-}, we identify 12, 16, and 20 water clusters have similar solvent arrangements. When we calculate the hydration free energy with 12 waters, we also get a very good agreement with the experimental values. Table S6 summarizes all of our calculations and compares calculated hydration free energies with experimental data. Thus, the sketch-map analysis appears to be useful for identifying how many solvent molecules are needed to calculate an accurate hydration free energy.

To obtain a more detailed understanding of where the errors come from, we performed MD simulations using the AMEOBA force field[92] with the TINKER[93] software package to account for both chemical and thermal energy scales. We performed simulations for Na^+ and Cl^- and used a cubic box starting with 500 solvent molecules. We picked 100 frames from the trajectory and carved out clusters having eight water molecules. Figure 3 shows SOAP/sketch-map analysis for these 100 structures and our DFT optimized structures. For Na^+, DFT optimization resulted in similar structures as those found from the MD simulation

regardless of the functional used in the optimization. However for Cl^-, DFT optimization resulted in different structures than those found from the MD simulation. We believe this is a result of enhanced anharmonic effects in ion-water clusters with anions. The lack of error cancellation in Scheme 2 with anions suggests that one needs to perform BOMD or LD simulations within Scheme 1 to capture the correct geometry for outer-shell contributions and account for anharmonicity in the inner-shell contributions to hydration free energy, as demonstrated recently.[27,28,30]

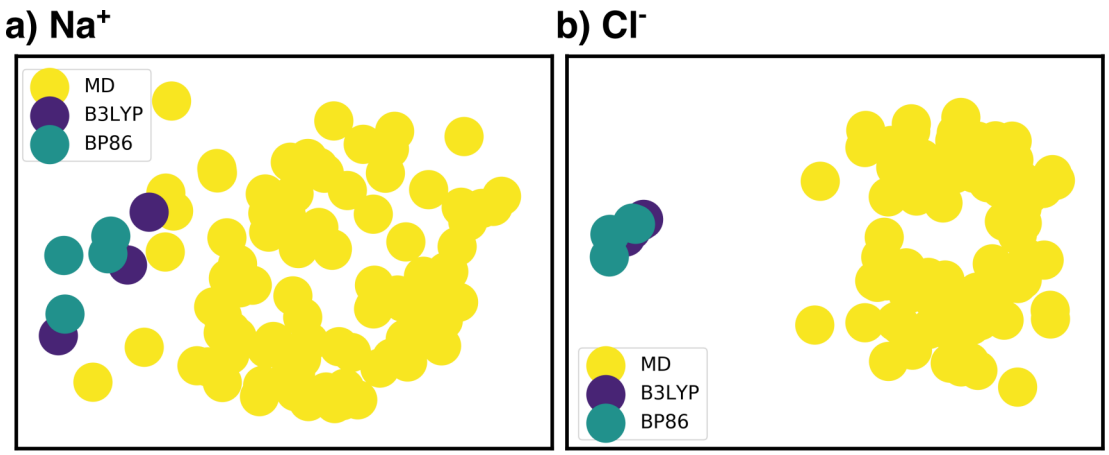


Figure 3: SOAP/sketch-map analysis for Na^+ and Cl^- with eight water clusters. Yellow data points are obtained from an MD trajectory, purple and green data points are obtained from a full QM optimization in this work.

Therefore, in the absence of well-parameterized force fields or computationally intensive BOMD simulations, we propose the following practical treatment (that exploits error cancellations) for automated calculations of real solvation free energies: 1) Calculate solvation free energies using the thermodynamic cycle in Scheme 2 for various values of n. 2) Use a SOAP sketch-map analysis to identify a relatively similar local solvent environment. 3) Use the smallest possible cluster to calculate solvation free energy contributions to minimize unbalanced errors that appear to arise from a CSM analysis of outer-shell contributions combined with harmonic analysis of inner-shell contributions to solvation free energy.

5 Conclusion

We have demonstrated an automatable cluster-continuum (i.e., mixed implicit/explicit) modeling approach to calculating solvation free energies of ions and small molecules that should be applicable for other molecules not considered in this study. We elucidated Scheme 1 in practical applications limits analysis to small n and CSM models to account for outer-shell contributions, and the results agree with absolute solvation free energies. In contrast, Scheme 2 analyzes large n and large cluster sizes, where CSM models drop out and phase potentials enter the calculation. We also showed how adding explicit solvent molecules improves calculated solvation free energies by creating a more physical local solvation environment, but adding too many solvent molecules leads to significant errors in the CSM combined with harmonic analysis of inner-shell. Overall, we show an approach to systematically investigate atomic scale microsolvation environments along with corresponding solvation free energies. The SOAP/sketch-map analysis can be combined with global optimization techniques such as ABCluster to minimize the required prior knowledge needed to compute an accurate solvation free energy using quantum chemistry. We expect this approach will be applicable to other ions in water as well as ions in different solvents.

Acknowledgement

We acknowledge support from the R. K. Mellon Foundation and the National Science Foundation (CBET-1653392 and CBET-1705592). We thank the University of Pittsburgh Center for Research Computing for computing time and technical support. We also acknowledge support from the Center for Integrated Nanotechnologies, an Office of Science User Facility operated for the U.S. Department of Energy (DOE) Office of Science, and Sandia National Laboratories (SNL) LDRD program. SNL is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. DOEs National Nuclear Security

Administration under contract DE-NA-0003525. The views expressed in the article do not necessarily represent the views of the U.S. DOE or the United States Government.

Supporting Information Available

- Table S1: Comparison of different solvation scales in kcal/mol.
- Table S2: Comparison of calculated electronic energies of water clusters in kcal/mol.
- Scheme S1: Thermodynamic cycle for the formation of water clusters.
- Table S3: Energy contributions (in kcal/mol) used in solvation free energy calculations of $(H_2O)_n$ clusters at $T = 298.15$ K in kcal/mol.
- Table S4: Comparison of gas phase binding free energies for $[X(H_2O)]$ in kcal/mol.
- Table S5: Gas phase binding free energies of $[X(H_2O)_4]$ in kcal/mol.
- Table S6: Comparison of experimental and calculated solvation free energies in kcal/mol.
- Table S7: Comparison of experimental and calculated solvation free energies for ion pairs in kcal/mol.
- Table S8: CGENFF parameters used with ABCluster.
- Figure S1-S7: Hydration free energy plots and sketch-maps for the remaining solutes.

References

(1) Balbuena, P. B.; Johnston, K. P.; Rossky, P. J. Molecular dynamics simulation of electrolyte solutions in ambient and supercritical water. 1. Ion solvation. *J. Phys. Chem.* **1996**, *100*, 2706–2715.

(2) Pickard IV, F. C.; König, G.; Simmonett, A. C.; Shao, Y.; Brooks, B. R. An efficient protocol for obtaining accurate hydration free energies using quantum chemistry and reweighting from molecular dynamics simulations. *Bioorg. Med. Chem.* **2016**, *24*, 4988–4997.

(3) Eggimann, B. L.; Siepmann, J. I. Size Effects on the Solvation of Anions at the Aqueous Liquid- Vapor Interface. *J. Phys. Chem. C* **2008**, *112*, 210–218.

(4) Tomasi, J.; Persico, M. Molecular interactions in solution: an overview of methods based on continuous distributions of the solvent. *Chem. Rev.* **1994**, *94*, 2027–2094.

(5) Miertuš, S.; Scrocco, E.; Tomasi, J. Electrostatic interaction of a solute with a continuum. A direct utilizaion of AB initio molecular potentials for the prevision of solvent effects. *Chem. Phys.* **1981**, *55*, 117–129.

(6) Cammi, R.; Tomasi, J. Remarks on the use of the apparent surface charges (ASC) methods in solvation problems: Iterative versus matrix-inversion procedures and the renormalization of the apparent charges. *J. Comput. Chem.* **1995**, *16*, 1449–1458.

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

(8) Barone, V.; Cossi, M. Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model. *J. Phys. Chem. A* **1998**, *102*, 1995–2001.

(9) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *J. Phys. Chem. B* **2009**, *113*, 6378–6396.

(10) Plata, R. E.; Singleton, D. A. A case study of the mechanism of alcohol-mediated Morita Baylis–Hillman reactions. The importance of experimental observations. *J. Am. Chem. Soc.* **2015**, *137*, 3811–3826.

(11) Pratt, L. R.; Rempe, S. B. Quasi-chemical theory and implicit solvent models for simulations. *AIP Conference Proceedings* **1999**, *492*, 172–201.

(12) Asthagiri, D.; Dixit, P.; Merchant, S.; Paulaitis, M.; Pratt, L.; Rempe, S.; Varma, S. Ion selectivity from local configurations of ligands in solutions and ion channels. *Chem. Phys. Lett.* **2010**, *485*, 1 – 7.

(13) Rogers, D. M.; Jiao, D.; Pratt, L. R.; Rempe, S. B. *Annual Reports in Computational Chemistry*; Elsevier, 2012; Vol. 8; pp 71–127.

(14) Rempe, S. B.; Pratt, L. R.; Hummer, G.; Kress, J. D.; Martin, R. L.; Redondo, A. The hydration number of Li^+ in liquid water. *J. Am. Chem. Soc.* **2000**, *122*, 966–967.

(15) Rempe, S. B.; Pratt, L. R. The hydration number of Na^+ in liquid water. *Fluid Phase Equilib.* **2001**, *183*, 121–132.

(16) Rempe, S. B.; Asthagiri, D.; Pratt, L. R. Inner shell definition and absolute hydration free energy of K^+(aq) on the basis of quasi-chemical theory and ab initio molecular dynamics. *Phys. Chem. Chem. Phys.* **2004**, *6*, 1966–1969.

(17) Varma, S.; Rempe, S. B. Structural transitions in ion coordination driven by changes in competition for ligand binding. *J. Am. Chem. Soc.* **2008**, *130*, 15405–15419.

(18) Sabo, D.; Jiao, D.; Varma, S.; Pratt, L.; Rempe, S. Case study of Rb^+(aq), quasi-chemical theory of ion hydration, and the no split occupancies rule. *Annual Reports Section C (Physical Chemistry)* **2013**, *109*, 266–278.

(19) Soniat, M.; Rogers, D. M.; Rempe, S. B. Dispersion-and exchange-corrected density

functional theory for sodium ion hydration. *J. Chem. Theory Comput.* **2015**, *11*, 2958–2967.

(20) Chaudhari, M. I.; Soniat, M.; Rempe, S. B. Octa-coordination and the aqueous Ba^{2+} ion. *J. Phys. Chem. B* **2015**, *119*, 8746–8753.

(21) Chaudhari, M. I.; Rempe, S. B. Strontium and barium in aqueous solution and a potassium channel binding site. *J. Chem. Phys.* **2018**, *148*, 222831.

(22) Chaudhari, M. I.; Pratt, L. R.; Rempe, S. B. Utility of chemical computations in predicting solution free energies of metal ions. *Mol. Simulat.* **2018**, *44*, 110–116.

(23) Biomolecular Hydration Mimicry in Ion Permeation through Membrane Channels. *Acc. Chem. Res.* **2019**,

(24) Asthagiri, D.; Pratt, L. R.; Paulaitis, M. E.; Rempe, S. B. Hydration structure and free energy of biomolecularly specific aqueous dications, including Zn^{2+} and first transition row metals. *J. Am. Chem. Soc.* **2004**, *126*, 1285–1289.

(25) Jiao, D.; Leung, K.; Rempe, S. B.; Nenoff, T. M. First principles calculations of atomic nickel redox potentials and dimerization free energies: A study of metal nanoparticle growth. *J. Chem. Theory Comput.* **2010**, *7*, 485–495.

(26) Chaudhari, M. I.; Rempe, S. B.; Pratt, L. R. Quasi-chemical theory of F^-(aq): The "no split occupancies rule" revisited. *J. Chem. Phys.* **2017**, *147*, 161728.

(27) Muralidharan, A.; Pratt, L. R.; Chaudhari, M. I.; Rempe, S. B. Quasi-chemical theory with cluster sampling from ab initio molecular dynamics: Fluoride (F^-) anion hydration. *J. Phys. Chem. A* **2018**, *122*, 9806–9812.

(28) Muralidharan, A.; Pratt, L. R.; Chaudhari, M. I.; Rempe, S. B. Quasi-Chemical Theory for Anion Hydration and Specific Ion Effects: Cl^-(aq) *vs.* F^-(aq). *Chem. Phys. Lett.* **2019**,

(29) Ashbaugh, H. S.; Asthagiri, D.; Pratt, L. R.; Rempe, S. B. Hydration of krypton and consideration of clathrate models of hydrophobic effects from the perspective of quasi-chemical theory. *Biophys. Chem.* **2003**, *105*, 323–338.

(30) Sabo, D.; Varma, S.; Martin, M. G.; Rempe, S. B. Studies of the thermodynamic properties of hydrogen gas in bulk water. *The Journal of Physical Chemistry B* **2008**, *112*, 867–876.

(31) Clawson, J. S.; Cygan, R. T.; Alam, T. M.; Leung, K.; Rempe, S. B. Ab initio study of hydrogen storage in water clathrates. *J. Comput. Theor. Nanosci.* **2010**, *7*, 2602–2606.

(32) Jiao, D.; Rempe, S. B. CO_2 solvation free energy using quasi-chemical theory. *J. Chem. Phys.* **2011**, *134*, 224506.

(33) Chaudhari, M. I.; Nair, J. R.; Pratt, L. R.; Soto, F. A.; Balbuena, P. B.; Rempe, S. B. Scaling atomic partial charges of carbonate solvents for lithium ion solvation and diffusion. *J. Chem. Theory Comput.* **2016**, *12*, 5709–5718.

(34) Varma, S.; Rempe, S. B. Tuning ion coordination architectures to enable selective partitioning. *Biophys. J.* **2007**, *93*, 1093 – 1099.

(35) Varma, S.; Sabo, D.; Rempe, S. B. K^+/Na^+ Selectivity in K Channels and Valinomycin: Over-coordination Versus Cavity-size constraints. *J. Mol. Biol.* **2008**, *376*, 13 – 22.

(36) Varma, S.; Rogers, D. M.; Pratt, L. R.; Rempe, S. B. Design principles for K^+ selectivity in membrane transport. *J. Gen. Physiol.* **2011**, *137*, 479–488.

(37) Rogers, D. M.; Rempe, S. B. Probing the thermodynamics of competitive ion binding using minimum energy structures. *J. Phys. Chem. B* **2011**, *115*, 9116–9129.

(38) Jiao, D.; Rempe, S. B. Combined Density Functional Theory (DFT) and Continuum Calculations of pK_a in Carbonic Anhydrase. *Biochem.* **2012**, *51*, 5979–5989.

(39) Rossi, M.; Tkatchenko, A.; Rempe, S. B.; Varma, S. Role of methyl-induced polarization in ion binding. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 12978–12983.

(40) Stevens, M. J.; Rempe, S. L. Ion-specific effects in carboxylate binding sites. *J. Phys. Chem. B* **2016**, *120*, 12519–12530.

(41) Mennucci, B.; Martínez, J. M. How to model solvation of peptides? Insights from a quantum-mechanical and molecular dynamics study of N-methylacetamide. 1. Geometries, infrared, and ultraviolet spectra in water. *J. Phys. Chem. B* **2005**, *109*, 9818–9829.

(42) Fennell, C. J.; Dill, K. A. Physical modeling of aqueous solvation. *J. Stat. Phys.* **2011**, *145*, 209–226.

(43) Pratt, L. R. Contact potentials of solution interfaces: phase equilibrium and interfacial electric fields. *J. Phys. Chem.* **1992**, *96*, 25–33.

(44) Leung, K.; Rempe, S. B.; von Lilienfeld, O. A. Ab initio molecular dynamics calculations of ion hydration free energies. *J. Chem. Phys.* **2009**, *130*, 204507.

(45) Lyklema, J. Interfacial Potentials: Measuring the Immeasurable? *Substantia* **2017**, *1*.

(46) Doyle, C. C.; Shi, Y.; Beck, T. L. The Importance of the Water Molecular Quadrupole for Estimating Interfacial Potential Shifts Acting on Ions Near the Liquid-Vapor Interface. *J. Phys. Chem. B* **2019**,

(47) Marcus, Y. Thermodynamics of solvation of ions. Part 5.Gibbs free energy of hydration at 298.15 K. *J. Chem. Soc., Faraday Trans.* **1991**, *87*, 2995–2999.

(48) Marcus, Y. The thermodynamics of solvation of ions. Part 4.Application of the tetraphenylarsonium tetraphenylborate (TATB) extrathermodynamic assumption to

the hydration of ions and to properties of hydrated ions. *J. Chem. Soc., Faraday Trans. 1* **1987**, *83*, 2985–2992.

(49) Tissandier, M. D.; Cowen, K. A.; Feng, W. Y.; Gundlach, E.; Cohen, M. H.; Earhart, A. D.; Coe, J. V.; Tuttle, T. R. The proton's absolute aqueous enthalpy and Gibbs free energy of solvation from cluster-ion solvation data. *J. Phys. Chem. A* **1998**, *102*, 7787–7794.

(50) Lamoureux, G.; Roux, B. Absolute hydration free energy scale for alkali and halide ions established from simulations with a polarizable force field. *J. Phys. Chem. B* **2006**, *110*, 3308–3322.

(51) Bryantsev, V. S.; Diallo, M. S.; Goddard III, W. A. Computational study of copper (II) complexation and hydrolysis in aqueous solutions using mixed cluster/continuum models. *J. Phys. Chem. A* **2009**, *113*, 9559–9567.

(52) Wu, W.; Kieffer, J. New Hybrid Method for the Calculation of the Solvation Free Energy of Small Molecules in Aqueous Solutions. *J. Chem. Theory Comput.* **2018**, *15*, 371–381.

(53) Kemp, D. D.; Gordon, M. S. Theoretical study of the solvation of fluorine and chlorine anions by water. *The Journal of Physical Chemistry A* **2005**, *109*, 7688–7699.

(54) da Silva, E. F.; Svendsen, H. F.; Merz, K. M. Explicitly representing the solvation shell in continuum solvent calculations. *J. Phys. Chem. A* **2009**, *113*, 6404–6409.

(55) Zhang, J.; Dolg, M. ABCluster: the artificial bee colony algorithm for cluster global optimization. *Phys. Chem. Chem. Phys.* **2015**, *17*, 24173–24181.

(56) Zhang, J.; Dolg, M. Global optimization of clusters of rigid molecules using the artificial bee colony algorithm. *Phys. Chem. Chem. Phys.* **2016**, *18*, 3003–3010.

(57) Bryantsev, V. S.; Diallo, M. S.; Goddard Iii, W. A. Calculation of solvation free energies of charged solutes using mixed cluster/continuum models. *J. Phys. Chem. B* **2008**, *112*, 9709–9719.

(58) Hutchinson, S. T.; Kobayashi, R. Solvent-Specific Featurization for Predicting Free Energies of Solvation through Machine Learning. *Journal of chemical information and modeling* **2019**, *59*, 1338–1346.

(59) Jaquis, B. J.; Li, A.; Monnier, N. D.; Sisk, R. G.; Acree, W. E.; Lang, A. S. Using Machine Learning to Predict Enthalpy of Solvation. *Journal of Solution Chemistry* **2019**, *48*, 564–573.

(60) Moorthy, N. H. N.; Martins, S. A.; Sousa, S. F.; Ramos, M. J.; Fernandes, P. A. Classification study of solvation free energies of organic molecules using machine learning techniques. *RSC Advances* **2014**, *4*, 61624–61630.

(61) Lim, H.; Jung, Y. Delfos: Deep Learning Model for Prediction of Solvation Free Energies in Generic Organic Solvents. *Chemical Science* **2019**,

(62) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, 184115.

(63) De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754–13769.

(64) Pliego, J. R.; Riveros, J. M. The cluster- continuum model for the calculation of the solvation free energy of ionic species. *J. Phys. Chem. A* **2001**, *105*, 7241–7247.

(65) Zhan, C.-G.; Dixon, D. A. Absolute hydration free energy of the proton from first-principles electronic structure calculations. *J. Phys. Chem. A* **2001**, *105*, 11534–11540.

(66) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. Aqueous solvation free energies of ions and

ion- water clusters based on an accurate value for the absolute aqueous solvation free energy of the proton. *J. Phys. Chem. B* **2006**, *110*, 16066–16081.

(67) Riccardi, D.; Guo, H.-B.; Parks, J. M.; Gu, B.; Liang, L.; Smith, J. C. Cluster-continuum calculations of hydration free energies of anions and group 12 divalent cations. *J. Chem. Theory Comput.* **2012**, *9*, 555–569.

(68) Pliego, J. R. Shells theory of solvation and the long-range Born correction. *Theoretical Chemistry Accounts* **2011**, *128*, 275–283.

(69) Pliego Jr, J. R. Cluster expansion of the solvation free energy difference: Systematic improvements in the solvation of single ions. *The Journal of chemical physics* **2017**, *147*, 034104.

(70) de Lima, G. F.; Duarte, H. A.; Pliego Jr, J. R. Dynamical discrete/continuum linear response shells theory of solvation: convergence test for NH4+ and OH- ions in water solution using DFT and DFTB methods. *The Journal of Physical Chemistry B* **2010**, *114*, 15941–15947.

(71) Beck, T. L.; Paulaitis, M. E.; Pratt, L. R. *The Potential Distribution Theorem and Models of Molecular Solutions*; Cambridge University Press, 2006.

(72) Roux, B.; Yu, H. Assessing the accuracy of approximate treatments of ion hydration based on primitive quasichemical theory. *The Journal of chemical physics* **2010**, *132*, 06B606.

(73) Gutowski, K. E.; Dixon, D. A. Predicting the energy of the water exchange reaction and free energy of solvation for the uranyl ion in aqueous solution. *The Journal of Physical Chemistry A* **2006**, *110*, 8840–8856.

(74) Pollard, T.; Beck, T. L. Quasichemical analysis of the cluster-pair approximation for the thermodynamics of proton hydration. *J. Chem. Phys.* **2014**, *140*, 224507.

(75) Pollard, T. P.; Beck, T. L. The thermodynamics of proton hydration and the electrochemical surface potential of water. *J. Chem. Phys.* **2014**, *141*, 18C512.

(76) Pollard, T. P.; Beck, T. L. Re-examining the tetraphenyl-arsonium/tetraphenyl-borate (TATB) hypothesis for single-ion solvation free energies. *J. Chem. Phys.* **2018**, *148*, 222830.

(77) Shi, Y.; Beck, T. L. Length scales and interfacial potentials in ion hydration. *J. Chem. Phys.* **2013**, *139*, 044504.

(78) Beck, T. L. The influence of water interfacial potentials on ion hydration in bulk water and near interfaces. *Chem. Phys. Lett* **2013**, *561*, 1–13.

(79) Engel, E. A.; Anelli, A.; Ceriotti, M.; Pickard, C. J.; Needs, R. J. Mapping uncharted territory in ice from zeolite networks to ice structures. *Nature communications* **2018**, *9*, 2173.

(80) Bartók, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R.; Csányi, G.; Ceriotti, M. Machine learning unifies the modeling of materials and molecules. *Science advances* **2017**, *3*, e1701816.

(81) Musil, F.; De, S.; Yang, J.; Campbell, J. E.; Day, G. M.; Ceriotti, M. Machine learning for the structure–energy–property landscapes of molecular crystals. *Chemical science* **2018**, *9*, 1289–1300.

(82) Ceriotti, M. Unsupervised machine learning in atomistic simulations, between predictions and understanding. *The Journal of chemical physics* **2019**, *150*, 150901.

(83) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **2010**, *31*, 671–690.

(84) Perdew, J. P. Density-functional approximation for the correlation energy of the inhomogeneous electron gas. *Phys. Rev. B* **1986**, *33*, 8822.

(85) Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **1988**, *38*, 3098.

(86) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **2011**, *32*, 1456–1465.

(87) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

(88) Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(89) Neese, F. The ORCA program system. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2*, 73–78.

(90) Chai, J.-D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620.

(91) Abascal, J. L.; Vega, C. A general purpose model for the condensed phases of water: TIP4P/2005. *J. Chem. Phys.* **2005**, *123*, 234505.

(92) Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio Jr, R. A.; Head-Gordon, M. Current status of the AMOEBA polarizable force field. *J. Phys. Chem. B* **2010**, *114*, 2549–2564.

(93) Ponder, J. W. TINKER: Software tools for molecular design. **2004**,

(94) Wales, D. J.; Hodges, M. P. Global minima of water clusters $(H_2O)_n$, $n \leq 21$, described by an empirical potential. *Chem. Phys. Lett* **1998**, *286*, 65–72.

(95) Arshadi, M.; Yamdagni, R.; Kebarle, P. Hydration of the halide negative ions in the gas phase. II. Comparison of hydration energies for the alkali positive and halide negative ions. *J. Phys. Chem.* **1970**, *74*, 1475–1482.

(96) Dzidic, I.; Kebarle, P. Hydration of the alkali ions in the gas phase. Enthalpies and entropies of reactions $M^+(H_2O)_{n-1} + H_2O = M^+ (H_2O)_n$. *J. Phys. Chem.* **1970**, *74*, 1466–1474.

(97) Ceriotti, M.; Tribello, G. A.; Parrinello, M. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 13023–13028.

(98) Ceriotti, M.; Tribello, G. A.; Parrinello, M. Demonstrating the transferability and the descriptive power of sketch-map. *J. Chem. Theory Comput.* **2013**, *9*, 1521–1532.

(99) Basdogan, Y.; Keith, J. A. A paramedic treatment for modeling explicitly solvated chemical reaction mechanisms. *Chem. Sci.* **2018**, *9*, 5341–5346.

(100) De, S.; Musil, F.; Ingram, T.; Baldauf, C.; Ceriotti, M. Mapping and classifying molecules from a high-throughput structural database. *J. Cheminform.* **2017**, *9*, 6.

Graphical TOC Entry