Predicting the properties of salt water using neural network potentials and continuum solvent theory.

Anonymous Author(s)
Affiliation
Address
email

Abstract

Electrolyte solutions play a fundamental role in a vast range of important industrial and biological applications. Yet their thermodynamic and kinetic properties still can not be predicted from first principles. There are three central challenges that need to be overcome to achieve this. Firstly, the dynamic nature of these solutions requires long time scale simulations. Secondly, the long-range Coulomb interactions require large spatial scales. Thirdly, the short-range quantum mechanical (QM) interactions require an expensive level of QM theory. Here, we demonstrate a methodology to address these challenges. Data from a short *ab initio* molecular dynamics (AIMD) simulation of aqueous sodium chloride is used to train an equivariant graph neural network interatomic potential (NNP) that can reliably reproduce the short-range QM forces and energies at a moderate computational cost. This NNP is combined with a continuum solvent description of the long-range electrostatic interactions to enable stable long time and large spatial scale simulations. From these simulations, ion-water and ion-ion radial distribution functions (RDFs), as well as ionic diffusivities, can be determined. The ion-ion RDFs are then used in a continuum solvent approach to calculate the osmotic and activity coefficients. Good experimental agreement is demonstrated up to the solubility limit of sodium chloride in water. This result implies that classical electrostatic theory can describe electrolyte solution over a remarkably wide concentration range as long as it is combined with an accurate description of the short-range interactions. This approach should be applicable to determine the thermodynamic and kinetic properties of many important electrolyte solutions for which experimental data is insufficient.

1 Introduction

Our understanding of the liquid state lies far behind that of the solid-state or gas phase. Perhaps nothing better demonstrates this than the fact that we cannot even predict how much simple table salt can be dissolved in water.Benavides *et al.* (2016); Panagiotopoulos (2020) The prediction of important kinetic properties, such as diffusivities, is also not currently possible.Panagiotopoulos (2020)

One of the most fundamental, yet most difficult to calculate properties of electrolyte solutions is their activity coefficients. The centrality of the activity coefficients is highlighted by the fact that they can be directly related to many important properties such as the chemical equilibria, solubilities, osmotic pressure and reaction rates.Robinson and Stokes (1959); Logan (1967)

While the fundamental importance of this problem is obvious, it is also of immense direct practical concern. Activity coefficients and diffusivities are essential for modelling many important systems that involve electrolyte solutions. These systems are ubiquitous throughout chemistry, biology,

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

geology and chemical engineering. For example, batteries, mineral processing, CO_2 capture and conversion all rely centrally on electrolyte solutions.

The prediction of the properties of electrolyte solutions has been a central and fundamental problem in the field of physical chemistry for over a century. Debye-Hückel theory, a continuum solvent model developed in 1923, accurately predicts activity coefficients at very low concentrations where the known long-range electrostatic interactions dominate over the unknown short-range QM interactions, which it neglects.Robinson and Stokes (1959) Efforts to model properties at higher concentrations rely on the use of adjustable parameters. These parameters are invariably fitted to reproduce experimental measurements.Weerasinghe and Smith (2003); Fyta and Netz (2012); Dočkal *et al.* (2020) The most prominent example of this are the Pitzer equations.Pitzer (1972) This is a crude solution as there are large gaps and uncertainties in existing experimental databases.May and Rowland (2017); Vaque Aura *et al.* (2021) For example, the activity coefficients of even relatively simple electrolyte solutions such as rubidium hydroxide and lithium bicarbonate have never been measured to the best of our knowledge.

The fundamental challenge associated with predicting these quantities is that they require large scale well converged molecular simulation to capture long-range electrostatic interactions, combined with a high level of accuracy to capture the short-range interactions, only possible with sophisticated quantum chemical methods.Duignan *et al.* (2014, 2021) Hence, *ab initio* nolecular dynamics (AIMD) simulations with density functional theory (DFT) are too computationally expensive to reliably converge properties such as ion-ion radial distribution functions (RDFs) or diffusivities at low concentrations. Additionally, the density functional approximations (DFAs) used in AIMD can have non-negligible errors.Riera *et al.* (2017); Paesani *et al.* (2019); Wagle *et al.* (2021); Duignan *et al.* (2020)

Recent exciting advances in the field of deep learning,Mater and Coote (2019); Noé *et al.* (2020); White (2021) and specifically neural network potentials (NNPs) Behler (2021); Kocer *et al.* (2022) can enable a solution to this problem. NNPs are highly flexible functions that efficiently map atomic coordinates to energies/forces. They are trained to reproduce *ab initio* data and can then be used to run much longer time scale molecular dynamics (MD) simulations than is possible with AIMD itself. It has recently been demonstrated that the Neural Equivariant Interatomic Potential (NequIP)Batzner *et al.* (2022) method requires remarkably low training data requirements compared to other NNP methods. This is attributable to the use of an equivariant architecture, which encodes the known rotational symmetries of space.Batatia *et al.* (2022) A potential limitation of this approach is that it cannot currently describe long-range interactions, which is a significant concern for electrolyte solutions where long-range electrostatic interactions are essential.Yue *et al.* (2021) However, continuum solvent models are known to provide a reliable description of long-range electrostatic interactions, hence the success of Debye-Hückel theory at low concentrations.

Here, we outline a new method of combining a continuum solvent theory description of long-range electrostatic interactions with NNP MD to describe the short-range interactions, overcoming the respective limitations of each approach. This method enables the prediction of osmotic and activity coefficients and diffusivities of sodium chloride from first principles with no empirically fitted parameters using only moderate computational resources.

2 Methodology

Fig. 1 presents a depiction of the overall workflow used to compute the activity coefficients and diffusivities of sodium and chloride in water. In brief, a 10 ps AIMD simulation of 2.4 M NaCl in water is performed using CP2KHutter *et al.* (2014) with the strongly constrained and appropriately normed (SCAN)Sun *et al.* (2015) DFA including an additional correction term to correct for the over stabilisation of hydrogen bonds associated with this DFA.Duignan *et al.* (2020) These correction terms are parametrised using MP2/CCSD(T) level calculations performed with ORCANeese (2012) on small clusters extracted from AIMD simulations. NequIPBatzner *et al.* (2022) is then used to train a NNP using 2500 frames of coordinates and forces extracted from this short simulation to produce a mapping between the coordinates and the energies/forces. A continuum solvent model description of the long-range electrostatic interactions, i.e., long-range dielectrically screened Coulomb interactions between the ions are computed at every time step. These forces/energies are removed from the training/validation data prior to training the NNP. The long-range electrostatic Coulomb interactions

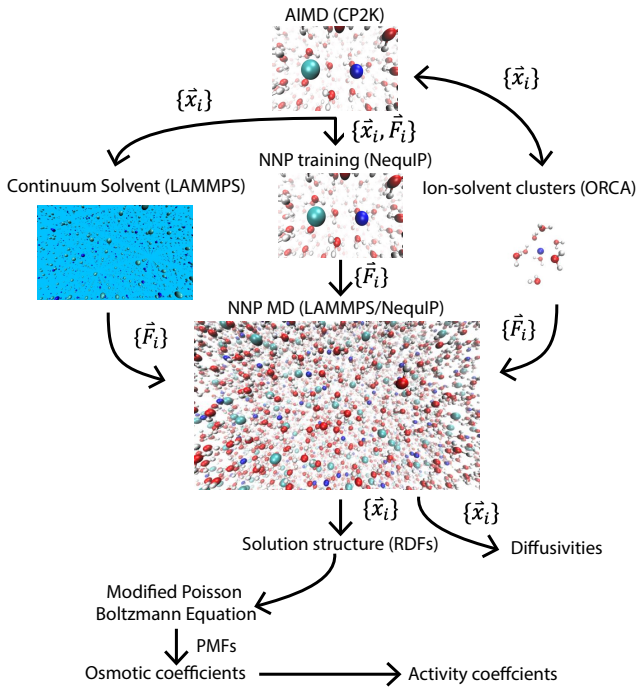


Figure 1: A depiction of the workflow used to compute electrolyte solution properties. Coordinates and forces/energies from an AIMD simulation are used to train a NNP. The NNP is used to run long time and large spatial scale MD simulations. Correction forces and long range Coulomb interactions are computed separately and added to the simulations. The coordinates from the MD simulations are used to compute important experimental properties.

are computed separately and added to the forces and energies predicted by the NNP to run much longer time scale and larger spatial scale MD simulations. This approach is a form of delta learning. These long simulations can be used to compute properties that are inaccessible with AIMD such as diffusivities and well converged ion-water and ion-ion radial distribution functions (RDFs). The validity of this method is demonstrated in the computational details section where we show that it adequately reproduces the ion-ion RDFs for a classical point charge water model.

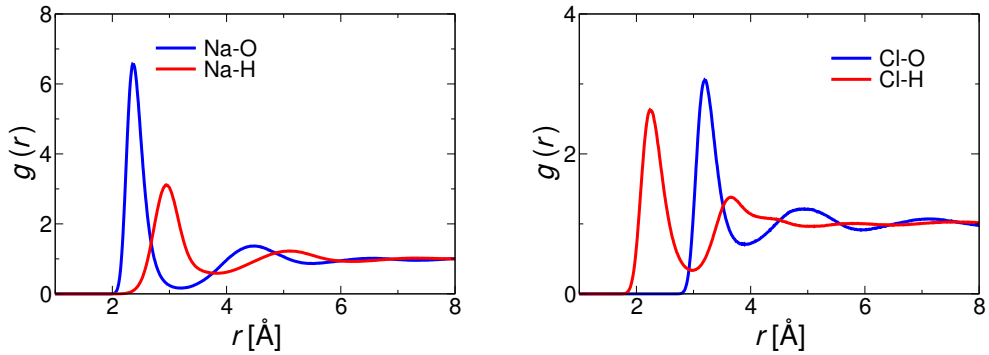
An important caveat is that although the simulation appears to be very stable for NaCl, while attempting to generalise to other electrolytes and to increase the number of parameters in the model we have observed some stability issues. This is a known issue for NNPs Stocker *et al.* (2022) where it has been observed that stability does not necessarily correlate with the mean error metrics. We believe this is an important area of future research. One potential cause of this is error/noise in the underlying AIMD data due to the poor numerical behaviour of various DFT functionals including SCAN.Lehtola and Marques (2022)

To further reduce computational demand, we use a new continuum solvent approach to determine the behaviour at many concentrations using MD simulation at a single concentration. To do this we extract the short-range non-electrostatic contribution to the ion-ion potential of mean force (PMF)Roux and Simonson (1999); Kalcher and Dzubiella (2009) using the modified Poisson-Boltzmann equation (MPBE). In spherical coordinates, this equation is given by:

$$-\epsilon_r \epsilon_o \frac{1}{r^2} \frac{d}{dr} \left(r^2 \frac{d\phi(r)}{dr} \right) = \sum_i q_i \rho_i(r) \tag{1}$$

where $\epsilon_r \epsilon_o$ gives the dielectric constant of water, r is the distance from a central ion, $\phi(r)$ is the electrostatic potential, q_i is the charge on each species in solution and $\rho_i(r)$ is the density of ions as a function of distance from a central ion. The density is approximated with the following expression:

$$\rho_i(r) = \rho_i \exp\left[-\beta \left(q_i \phi(r) + W_{\mathrm{SR}}(r)\right)\right] \tag{2}$$



(a) Sodium-water RDFs.

(b) Chloride-water RDFs.

where ρ_i is the bulk concentration of ions, β is the thermodynamic beta ($\frac{1}{k_{\mathrm{B}}T}$) and $W_{\mathrm{SR}}(r)$ is the short-range contribution to the potential of mean force between the ions. Note if we set $W_{\mathrm{SR}}(r)$ to be a hard sphere repulsion and linearise the exponential we arrive at classical Debye-Hückel theory. As $\frac{\rho_i(r)}{\rho_i}$ is equivalent to the radial distribution function we use the RDFs from molecular simulation at one concentration (2.4 M) to determine $W_{\mathrm{SR}}(r)$ by self consistently solving MPBE equation until the RDF predicted by MPBE matches the simulation prediction of the RDF. Once this short-range non-electrostatic PMF is determined, the MPBE equation can be used to quickly compute the RDF at many concentrations.

The osmotic coefficients over the whole solubility range of NaCl are then computed using the virial approach.Rasaiah and Friedman (1968); Kalcher and Dzubiella (2009); Vrbka *et al.* (2009) These are then converted to activity coefficients via the Gibbs-Duhem equation. More detailed computational information regarding these calculations is provided in the computational details section below.

3 Results and Discussion

The ion-water RDFs calculated from the 1 ns NNP MD are shown in Fig. 2a and 2b. The RDFs are very well converged and agree well with previous work.Dellostritto *et al.* (2020); Caruso and Paesani (2021); Duignan *et al.* (2020)

The computed diffusivities of sodium and chloride ions in water computed with the NNP MD simulation are consistent with experiment, (Fig. 3) suggesting that the simulation can reproduce important experimental properties of these ions in water without being trained to do so.

The ion-ion RDFs determined with NNP MD at 2.4 M are shown in Fig. 4. The NNP MD predicts similar peaks for the contact ion pair and solvent separated ion pair. This is consistent with other AIMD simulations of this system with several other DFAs.Timko *et al.* (2010); Duignan *et al.* (2016); Yao and Kanai (2018) While we are not aware of any AIMD calculation of the Na-Na RDF, the small shoulder has not been observed before in classical MD, this may potentially be an artifact of the larger mean error in the forces on the Na ions.

The short-range non-electrostatic contribution to the ion-ion interactions are depicted in Fig. 5. These are calculated using the self consistent solution of the MPBE method as described below. These are the solvent averaged interaction free energies of the ions in water with the long range Coulomb interaction removed. They are all close to 0 beyond 8 Å, confirming they are short range and the simulation cell is sufficiently large. The desolvation barrier associated with removing the water from around the sodium ion is clearly visible in the Na-Na and Na-Cl PMF.

Fig. 6a, 6b and 6c show the ion-ion RDFs computed with the MPBE at 0.1 M, 2.4 M and 4.9 M concentrations. They show the increasing importance of the long-range Coulomb interaction at low concentrations, which become increasingly damped at higher concentrations. Converging these low concentration RDFs with MD simulations is difficult even with classical point charge MD as it requires very large box sizes, hence the need for the MPBE approach. At higher concentrations,

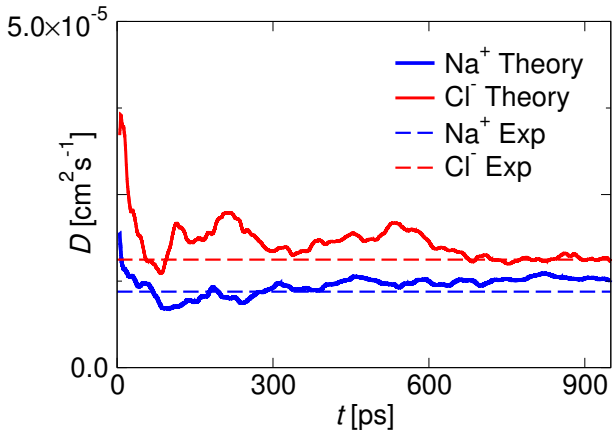


Figure 3: Sodium and chloride ion diffusivities computed from NNP MD simulation compared with experimental data. Vitagliano and Lyons (1956).

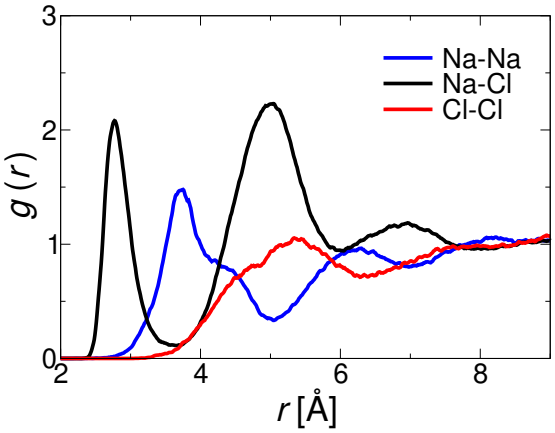


Figure 4: Sodium-sodium, sodium-chloride and chloride-chloride RDFs computed from 1 ns NNP MD simulation.

electrostatic screening significantly reduces this long-range interaction. The short-range PMF was determined self consistently to reproduce Fig. 4, hence it agrees perfectly with Fig. 6b.

The osmotic coefficients are then calculated from the PMFs and RDFs using the MPBE equation and the virial approach as described below. Fig. 7a shows the resulting good experimental agreement.Robinson and Stokes (1959); Pitzer and Mayorga (1973) These can be converted to activities using the Gibbs-Duhem relation as shown in Fig. 7b. Remarkably, experimental agreement extends up to the solubility limit of NaCl (5.4 M.) We do not rely on a concentration-dependent dielectric constant to achieve this agreement. This means no experimental parameters are required for this work except for the dielectric constant of pure water. It should be noted that these properties are notoriously sensitive to small variations in the interaction potentials.Kalcher and Dzubiella (2009) They are therefore an excellent demonstration of the accuracy of this approach.

The consistency of our computational results with experiment throughout this section indicates that this method is accurately reproducing the structure of aqueous sodium chloride solutions. The fact that we achieve agreement using the MPBE equation implies that the simple classical electrostatic theory is remarkably reliable as long as it is combined with an accurate description of short-range interactions.

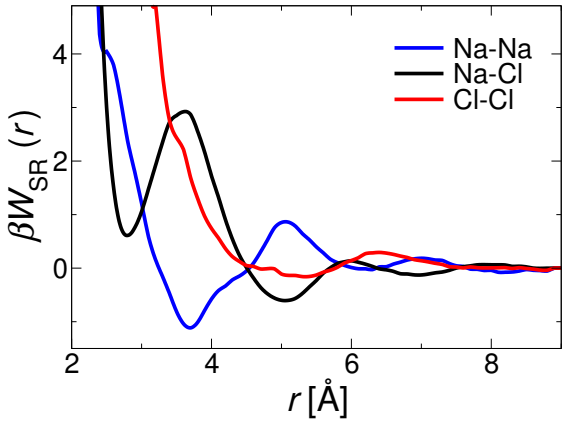
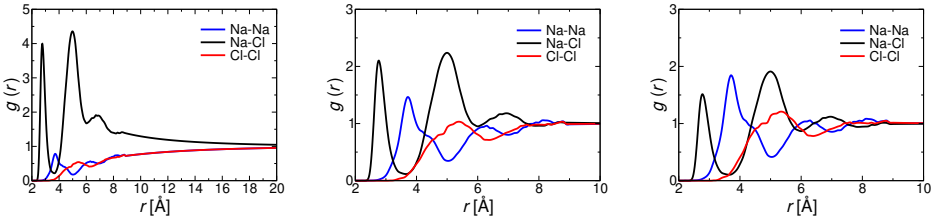
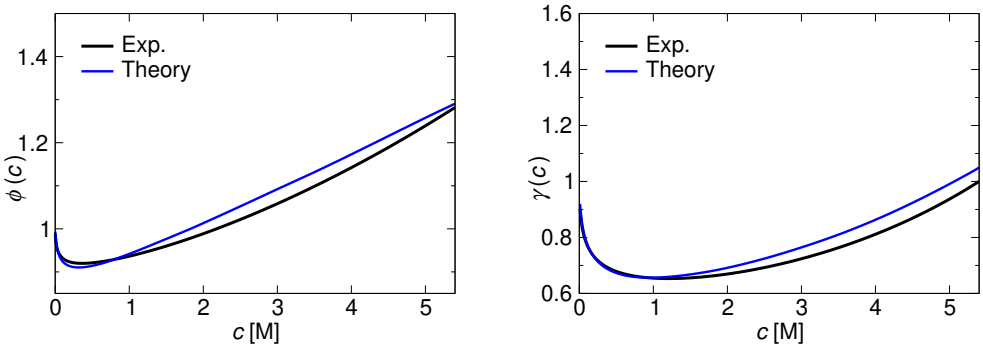


Figure 5: Sodium-sodium, sodium-chloride and chloride-chloride short-range PMFs computed with MPBE.



(a) 0.1 M (b) 2.4 M (c) 4.9 M

Figure 6: Sodium-sodium, sodium-chloride and chloride-chloride RDFs computed with MPBE at three concentrations.



(a) Osmotic coefficients. (b) Activity coefficients.

Figure 7: Thermodynamic properties of NaCl calculated with MPBE and the virial route compared with experiment. Robinson and Stokes (1959); Pitzer and Mayorga (1973)

Future work should focus on extending this methodology to other ions, particularly important ions that have never had their activities measured such as lithium bicarbonate. Further effort to improve the accuracy of this method also needs to be undertaken. For instance the error in the force on the sodium ion is substantially larger than the other species in the solution as shown below. It is likely the numerical stability of the underlying DFT functional is a significant source of this error.Lehtola and Marques (2022) Additionally, it should be feasible to use a higher level of theory such as density corrected SCANDasgupta *et al.* (2021) or the random phase approximationYao and Kanai (2021) to generate the training data.

4 Conclusion

We have demonstrated a methodology for computing the solvation structure, the activity and osmotic coefficients as well as the diffusivities of aqueous sodium chloride from first principles using equivariant graph neural network interatomic potentials to describe the short-range interactions combined with continuum solvent theory to describe the long-range interactions. While, some challenges remain, primarily associated with the accuracy of the DFAs and the stability of the NNPs, this work outlines a pathway for long standing problem in physical chemistry. It demonstrates that a classical electrostatic description of electrolyte solutions is remarkably reliable as long as it is combined with an accurate description of short-range interactions. Critically, this approach requires only moderate computational resources. This methodology should in principle be applicable to a much wider class of solutes and solvents allowing the prediction of many important thermodynamic and kinetic properties of electrolyte solutions from first principles. This approach should enable a transition away from the current reliance on limited and unreliable experimental databases for the determination of the properties of electrolyte solutions. It also demonstrates the potential of equivariant graph NNPs to accelerate accurate molecular simulations and provide answers to many long-standing important scientific questions.

References

A. L. Benavides, J. L. Aragones and C. Vega, *J. Chem. Phys.*, 2016, **144**, 124504.

A. Z. Panagiotopoulos, *J. Chem. Phys.*, 2020, **153**, 010903.

R. A. Robinson and R. H. Stokes, *Electrolyte solutions*, Butterworth & Co., Devon, 1959.

S. R. Logan, *Trans. Faraday Soc.*, 1967, **63**, 3004–3008.

S. Weerasinghe and P. E. Smith, *J. Chem. Phys.*, 2003, **119**, 11342–11349.

M. Fyta and R. R. Netz, *J. Chem. Phys.*, 2012, **136**, 124103.

J. Dočkal, F. Moučka and M. Lísal, *J. Chem. Theory Comput.*, 2020, **16**, 3677–3688.

K. S. Pitzer, *J. Phys. Chem.*, 1972, **77**, 268–277.

P. M. May and D. Rowland, *J. Chem. Eng. Data*, 2017, **62**, 2481–2495.

S. Vaque Aura, J.-S. Roa Pinto, N. Ferrando, J.-C. de Hemptinne, A. ten Kate, S. Kuitunen, N. Diamantonis, T. Gerlach, M. Heilig, G. Becker and M. Brehelin, *J. Chem. Eng. Data*, 2021, **66**, 2976–2990.

T. T. Duignan, D. F. Parsons and B. W. Ninham, *Chem. Phys. Lett.*, 2014, **608**, 55–59.

T. T. Duignan, S. M. Kathmann, G. K. Schenter and C. J. Mundy, *Acc. Chem. Res.*, 2021, **54**, 2833–2843.

M. Riera, N. Mardirossian, P. Bajaj, A. W. Götz and F. Paesani, *J. Chem. Phys.*, 2017, **147**, 161715.

F. Paesani, P. Bajaj and M. Riera, *Adv. Phys. X*, 2019, **4**, 1631212.

K. Wagle, B. Santra, P. Bhattarai, C. Shahi, M. R. Pederson, K. A. Jackson and J. P. Perdew, *J. Chem. Phys.*, 2021, **154**, 094302.

T. T. Duignan, G. Schenter, C. J. Mundy and X. S. Zhao, *J. Chem. Theory Comput.*, 2020, **16**, 5401–5409.

A. C. Mater and M. L. Coote, *J. Chem. Inf. Model.*, 2019, **59**, 2525–2559.

F. Noé, A. Tkatchenko, K.-R. Müller and C. Clementi, *Annu. Rev. Phys. Chem.*, 2020, **71**, 361–390.

A. D. White, *Deep Learning for Molecules and Materials*, 2021, `https://dmol.pub`.

J. Behler, *Chem. Rev.*, 2021, **121**, 10037–10072.

E. Kocer, T. W. Ko and J. Behler, *Annu. Rev. Phys. Chem.*, 2022, **73**, 163–186.

S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt and B. Kozinsky, *Nat. Commun.*, 2022, **13**, 2453.

I. Batatia, S. Batzner, D. P. Kovács, A. Musaelian, N. C. Gregor Simm, R. Drautz, C. Ortner, B. Kozinsky and G. Csányi, *arXiv*, 2022, arXiv:2205.06643.

S. Yue, M. C. Muniz, M. F. Andrade, L. Zhang, R. Car and A. Z. Panagiotopoulos, *J. Chem. Phys.*, 2021, **154**, 034111.

J. Hutter, M. Iannuzzi, F. Schiffmann and J. VandeVondele, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2014, **4**, 15–25.

J. Sun, A. Ruzsinszky and J. P. Perdew, *Phys. Rev. Lett.*, 2015, **115**, 036402.

F. Neese, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2012, **2**, 73–78.

S. Stocker, J. Gasteiger, F. Becker, S. Günnemann and J. T. Margraf, *ChemRxiv*, 2022, 10.26434/chemrxiv–2022–mc4gb.

S. Lehtola and M. A. L. Marques, *ArXiV*, 2022, arXiv:2206.14062v3.

B. Roux and T. Simonson, *Biophys. Chem.*, 1999, **78**, 1–20.

I. Kalcher and J. Dzubiella, *J. Chem. Phys.*, 2009, **130**, 134507.

J. C. Rasaiah and H. L. Friedman, *J. Chem. Phys.*, 1968, **48**, 2742–2752.

L. Vrbka, M. Lund, I. Kalcher, J. Dzubiella, R. R. Netz and W. Kunz, *J. Chem. Phys.*, 2009, **131**, 154109.

V. Vitagliano and P. A. Lyons, *J. Am. Chem. Soc.*, 1956, **78**, 1549–1552.

K. S. Pitzer and G. Mayorga, *J. Phys. Chem.*, 1973, **77**, 2300–2308.

M. Dellostritto, J. Xu, X. Wu and M. L. Klein, *Phys. Chem. Chem. Phys.*, 2020, **22**, 10666–10675.

A. Caruso and F. Paesani, *J. Chem. Phys.*, 2021, **155**, 064502.

J. Timko, D. Bucher and S. Kuyucak, *J. Chem. Phys.*, 2010, **132**, 114510.

T. T. Duignan, M. D. Baer and C. J. Mundy, *Curr. Opin. Colloid Interface Sci.*, 2016, **23**, 58–65.

Y. Yao and Y. Kanai, *J. Chem. Theory Comput.*, 2018, **14**, 884–893.

S. Dasgupta, E. Lambros, J. P. Perdew and F. Paesani, *Nat. comm*, 2021, **12**, 6359.

Y. Yao and Y. Kanai, *J. Phys. Chem. Lett.*, 2021, **12**, 6354–6362.

J. VandeVondele, M. Krack, F. Mohamed, M. Parrinello, T. Chassaing and J. Hutter, *Comput. Phys. Commun.*, 2005, **167**, 103–128.

J. VandeVondele and J. Hutter, *J. Chem. Phys.*, 2007, **127**, 114105.

S. Goedecker, M. Teter and J. Hutter, *Phys. Rev. B*, 1996, **54**, 1703–1710.

J. Hutter, *https://github.com/juerghutter/GTH/*, 2021.

G. Miceli, J. Hutter and A. Pasquarello, *J. Chem. Theory Comput.*, 2016, **12**, 3456–3462.

G. J. Martyna, M. L. Klein and M. Tuckerman, *J. Chem. Phys.*, 1992, **97**, 2635–2643.

T. T. Duignan, *chemRxiv*, 2021, 10.26434/chemrxiv–2021–7jxpq–v2.

P. B. Calio, G. M. Hocky and G. A. Voth, *J. Chem. Theory Comput.*, 2020, **16**, 5675–5684.

T. H. Dunning, Jr., *J. Chem. Phys.*, 1989, **90**, 1007–1023.

D. E. Woon and T. H. Dunning, Jr., *J. Chem. Phys.*, 1994, **100**, 2975–2988.

R. W. Hockney, *Computer simulation using particles*, A. Hilger, 1988, p. 540.

S. Plimpton, *J. Comput. Phys.*, 1995, **117**, 1–19.

W. Humphrey, A. Dalke and K. Schulten, *Journal of Molecular Graphics*, 1996, **14**, 33–38.

Wolfram Research Inc., *Mathematica*, 2019.

T. M. Herrington, A. D. Pethybrldge and M. G. Roffey, *J. Chem. Eng. Data*, 1986, **31**, 31–34.

P. M. Sipos, G. Hefter and P. M. May, *J. Chem. Eng. Data*, 2000, **45**, 613–617.

CRC Handbook of Chemistry and Physics, Internet Version, ed. D. R. Lide, Taylor and Francis, Boca Raton, FL, eighty sev edn., 2007.

H. L. Friedman, *J. Solution Chem.*, 1972, **1**, 413–417.

J.-P. Simonin, *J. Chem. Soc. Faraday Trans.*, 1996, **92**, 3519–3523.

I.-C. Yeh and G. Hummer, *J. Phys. Chem. B*, 2004, **108**, 15873–15879.

5 Appendix: Computational Details

5.1 AIMD (CP2K)

We used Born-Oppenheimer *ab initio* molecular dynamics simulations within a constant volume NVT (300 K) ensemble with periodic boundary conditions. The `CP2K` simulation suite (http:www.cp2k.org) containing the *QuickStep* module for the DFT calculations VandeVondele *et al.* (2005) was used with a 0.5 fs time step. We used a double ζ basis set that has been optimized for the condensed phaseVandeVondele and Hutter (2007) in conjunction with GTH pseudopotentials Goedecker *et al.* (1996) optimised for SCANSun *et al.* (2015); Hutter (2021) and a 1200 Ry cutoff.Miceli *et al.* (2016); Yao and Kanai (2018) A Nosé-Hoover thermostat was attached to every degree of freedom to ensure equilibration. Martyna *et al.* (1992) The energies were accumulated for $\approx$ 10 ps after 2 ps of equilibration. The simulation consisted of 4 sodium ions 4 chloride ions and 80 water molecules in a box of fixed dimensions of 13.92^3 Å^3 giving an electrolyte concentration of 2.4 M.

5.2 Hydrogen bond correction (ORCA)

An exponential repulsive correction to the hydrogen bond interaction between the oxygen atoms and the neighbouring hydrogen atoms is added as described in a previous publicationDuignan (2021) as a minimal biasCalio *et al.* (2020) to reduce the error in the water structure. This correction was included in the CP2K simulation using the multiple force evaluation option with the FIST method. This correction was fitted to remove the error in the water dimer interaction computed with MP2 level as outlined in Ref. 49 and is given by:

$$A \exp^{-br} \tag{3}$$

With A = 0.06 and b = 1.3 in atomic units.

It is also known that SCAN has a similar issue with describing the interaction of anions with water.Wagle *et al.* (2021) We therefore also add an exponential repulsion between the chloride anion

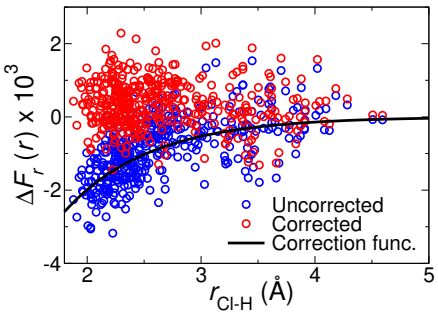


Figure 8: Error of SCAN compared to MP2 in the radial component of the force on water molecules as a function of distance between the central chloride ion and the closest hydrogen atom with and without the correction potential. The negative of the force corresponding to the correction potential is also overlaid in black.

and neighbouring hydrogen atoms. In order to determine the parameters for this correction we follow Ref. 16 and optimise them to minimise the error in the radial component of the force on eight water molecules surrounding a central chloride anion. The parameters are given by $A = 0.04$ and $b = 0.7$ in atomic units. The radial error is computed from the projection of the error of the total force on the water molecule in the r_{ClH} direction and is shown in Fig. 8 along with the correction force used to remove this error. The torque on the water molecules was also computed to confirm that this correction potential also removes a significant error there too.

ORCANeese (2012) was used to calculate the cluster forces at the MP2 level of theory. 50 clusters of 8 water molecules surrounding a chloride anion were used in the cluster correction calculation using the same trajectories as in Ref. 16. The aug-cc-pVDZ basis set was used for the oxygen, hydrogen and chloride atoms.Dunning, Jr. (1989) Similarly, the cc-pCVDZ basis set was used for the sodium ion Woon and Dunning, Jr. (1994). Frozen cores were used for the MP2 calculations. For the SCAN cluster energy calculations, `CP2K` was used with the periodicity none option and a larger cell size to remove any box size dependence. Otherwise, the same parameters, basis sets and pseudo-potentials as in the simulation described above were used.

5.3 NNP fitting (NequIP)

Forces and energies from 2000 frames extracted from CP2K were used to train the NNP with NequIP. 500 frames were held out as a validation set. Coulomb interactions screened by the dielectric constant of water (78.3) between all the ions were subtracted before training. These were calculated with LAMMPS by placing appropriately screened charges on the ions to reproduce dielectric screening of 78.3 and with the particle-particle particle-mesh method.Hockney (1988) An equal weighting on forces and energies was used in the default loss function.Batzner *et al.* (2022) We decrease the initial learning rate of 0.01 by a decay factor of 0.5 whenever the validation RMSE in the forces has not seen an improvement for five epochs. A radial cutoff distance of 5 Å was used. Two interaction blocks were used with the maximum l set to one each with 8 features. Only even parity was used. All the other parameters were set to the defaults. Fig. 9 shows the learning curve.

RMSE on the validation set was 128 meV/Å for the forces and 0.341 meV/atom for the energies. Note it was possible to achieve a significantly lower mean error by increasing the number of parameters in the model. However, we observed counter-intuitively, this would usually decrease the stability of the simulation. This is consistent with other research showing that lower mean error does not necessarily correlate with better stability. Stocker *et al.* (2022) Fig. 10 shows the correlation between the forces calculated using the ab inito method with CP2K from the origianl 10 ps trajecory compared with the predictions with NequIP. It is clear the biggest source of error by far is the force on the sodium ion. Understanding and reducing the magnitude of this error is an important future research goal.

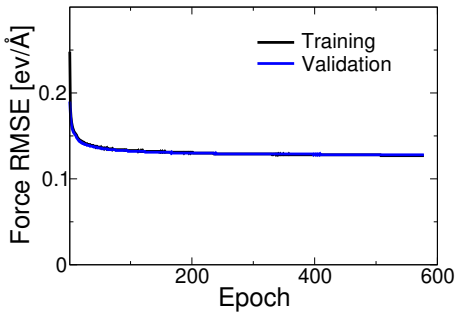


Figure 9: Learning curve for the root mean square error on the forces predicted by NequIP.

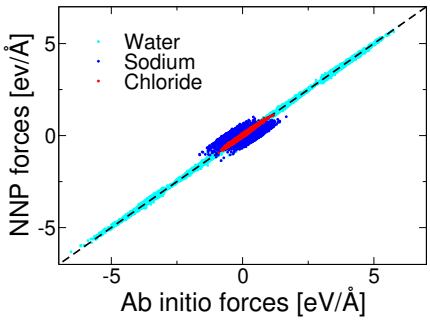


Figure 10: Correlation between forces calculated from the *ab initio* MD trajectory compared with those predicted by the NNP.

5.4 NNP MD (NequIP/LAMMPS)

The NequIP plugin for LAMMPSPlimpton (1995) was used to perform NVT simulations at 300 K for 1 ns. A Nosé-Hoover thermostat was attached to every degree of freedom to ensure equilibration Martyna *et al.* (1992). The chloride hydrogen bond exponential correction term was added to the NNP MD simulation using tabulated data as this correction was not included in the original CP2K SCAN simulations. The long-range Coulomb interactions were added to the simulation using LAMMPS hybrid overlay method. Without this correction, we observed strong over stabilisation of cation-anion pairs. The reliability of this method was verified by training on forces and energies from a classical MD run and demonstrating that the NNP MD reproduced the ion-ion RDFs with reasonable accuracy as shown in Fig. 11. No initial data was discarded as the initial frame was taken from the end of the AIMD simulation. A dielectric constant of 70.7 is used to better match the dielectric constant of SPC/E water, although this has a minimal effect.

We use constant volume simulations with a simple approximation using the ion size to estimate the box size to avoid relying on experimental information. However, we have repeated these simulations at the exact experimental density of sodium chloride to confirm that the RDFs remain unaltered. A cell size of 20.24^3 Å^3 was used which was larger than the the cell size used for the AIMD simulation data (13.92^3 Å^3). To demonstrate that this was reliable we performed simulations in the smaller cell size, which shows that within the cell the agreement is good, (Fig. 12) however the larger cell size is needed as there is still significant oscillation beyond 7 Å.

VMDHumphrey *et al.* (1996) was used to create the RDFs and images in Fig. 1.

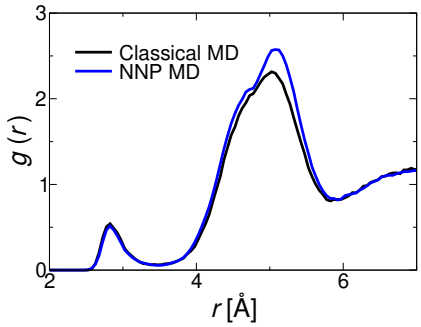


Figure 11: Comparison of a Na-Cl RDF obtained from classical point charge MD simulation with a flexible SPC/E water model and an NNP simulation using the protocol outlined.

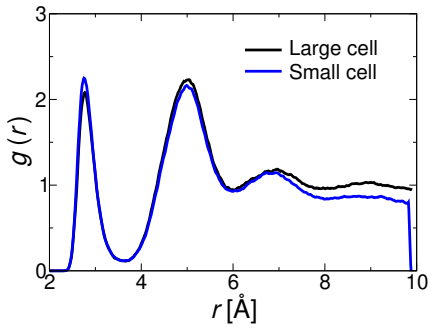


Figure 12: Comparison of Na-Cl RDFs obtained from a 20.24^3 Å^3 and a 13.92^3 Å^3 NNP MD simulation.

5.5 Modified Poisson-Boltzmann Equation (MPBE)

To compute $W_{\mathrm{SR}}(r)$ we first assume it is 0 and then solve the MPBE equation to determine the electrostatic potential and then determine a new estimate for $W_{\mathrm{SR}}(r)$ using this rearrangement of the expression for the density:

$$\beta W_{\mathrm{SR}}(r) = -\beta q_i \phi(r) - \ln \frac{\rho_i(r)}{\rho_i} \tag{4}$$

where $\frac{\rho_i(r)}{\rho_i}$ is the RDF taken from the NNP MD simulations. This procedure is repeated iteratively until the RDFs predicted with the MPBE (Eq. 1) and Eq. 2 agree well with the ones obtained from the NNP MD simulations. The boundary conditions used to solve the MPBE are the electric field of a point charge at 2 Å and the electric potential of 0 at large separations (60 Å). These settings ensure reliable numerical solutions are found using the shooting method implemented in Mathematica.Wolfram Research Inc. (2019) Smoothing of the nearest 5 points (0.05 Å) was applied to improve the convergence of the numerical differential equation solution. The ionic density is integrated to confirm a net counter charge of +/- 1 as required to satisfy the electroneutrality condition. The RDF is fitted up to 8.8 Å which ensures a smooth transition to the long-range only region of the RDF. A fixed dielectric constant of water of 78.3 is used throughout.

Note that there are two ways to solve the MPBE one with the cation as the central ion and one with the anion. Fig. 13 shows the cation-anion short-range PMF extracted using this approach for both cases indicating minimal difference based on this assumption.

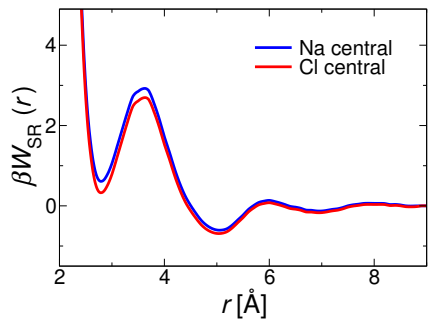


Figure 13: Short range contribution to the sodium chloride interaction PMF calculated using either the sodium cation as the central ion or the chloride anion as the central ion demonstrating minimal difference.

5.6 Activities and osmotic coefficients calculation

Once the short-range contributions to the PMF ($W_{\mathrm{SR}}(r)$) have been determined we can use the MPBE to calculate the RDF at many concentrations including down to infinite dilution with minimal computational demands. $W_{\mathrm{SR}}(r)$ can then be input into the virial expression for the osmotic coefficients to estimate them at many concentration.Rasaiah and Friedman (1968); Kalcher and Dzubiella (2009); Vrbka *et al.* (2009)

$$\phi(\rho) = 1 - \frac{\pi}{3}\rho \sum_{i,j} \int_0^\infty g_{ij}(r) \frac{dW_{ij}(r)}{dr} r^3 dr \tag{5}$$

where $W_{ij}(r)$ is the ion-ion infinite dilution PMF, i.e., $W_{\mathrm{SR}}(r)$ plus the Coulomb interaction. We solve the MPBE and compute osmotic coefficients with both chloride and sodium as the central ion taking the average of the two approaches for the final prediction.

This gives reasonable agreement with osmotic coefficients as shown in the main text. A more accurate approximation can be obtained with the Hyper-Netted Chain (HNC) closure of the Ornstein–Zernike (OZ) equation,Kalcher and Dzubiella (2009); Vrbka *et al.* (2009) which incorporates ion correlation effects. However, the relative consistency of the PMF methods using the cation or the anion as the central molecule (Fig. 13) indicates that this effect can plausibly be neglected. And extracting the correct short-range PMF to use in the OZ equation is very difficult without the MPBE method used here.

We do not use a concentration-dependent dielectric constant as this would require the method to be empirically parameterised and this effect will be incorporated into the PMF which is fitted to reproduce the RDF.

The Gibbs-Duhem relationship between osmotic and activity coefficients is used to convert osmotic coefficients into activities:

$$\ln\gamma = \phi - 1 + \int_0^m \frac{\phi(m) - 1}{m} dm \tag{6}$$

Experimental densitiesHerrington *et al.* (1986); Sipos *et al.* (2000); Lide (2007) were used to convert theoretical osmotic coefficients from the McMillan-Mayer to the Lewis-Randal level to enable the correct thermodynamic comparison.Friedman (1972); Simonin (1996) Experimental values are obtained from the Pitzer equations and converted to molarity again using experimental densities.Pitzer and Mayorga (1973)

5.7 Diffusion coefficients calculation

Diffusion coefficients were computed from the mean squared displacements (MSD) of sodium and chloride ions in our NNP MD trajectories. This conversion was carried out using the diffusion coefficient-MSD relationship described below:

$$D = \frac{MSD}{6t} \tag{7}$$

The results were finally adjusted by finite size corrections Yeh and Hummer (2004). Here, we have used the experimental value for the viscosity of pure water when determining the finite size correction. Experimental values Vitagliano and Lyons (1956) for the value of sodium and chloride ion diffusivities in a 2.4 M NaCl solution were used to validate the results.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [No]
   (b) Did you include complete proofs of all theoretical results? [No]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [No]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]