










Accurate determination of solvation free energies of neutral organic compounds from first principles

Leonid Pereyaslavets^{1,8}, Ganesh Kamath^{1,8}, Oleg Butin¹, Alexey Illarionov¹, Michael Olevanov^{1,2}, Igor Kurnikov¹, Serzhan Sakipov¹, Igor Leontyev¹, Ekaterina Voronina^{1,2}, Tyler Gannon¹, Grzegorz Nawrocki¹, Mikhail Darkhovskiy¹, Ilya Ivahnenko¹, Alexander Kostikov¹, Jessica Scaranto³, Maria G. Kurnikova³, Suvo Banik^{4,5}, Henry Chan^{4,5}, Michael G. Sternberg⁴, Subramanian K. R. S. Sankaranarayanan^{4,5}, Brad Crawford⁶, Jeffrey Potoff⁶, Michael Levitt⁷, Roger D. Kornberg⁷ & Boris Fain^{1,8}

The main goal of molecular simulation is to accurately predict experimental observables of molecular systems. Another long-standing goal is to devise models for arbitrary neutral organic molecules with little or no reliance on experimental data. While separately these goals have been met to various degrees, for an arbitrary system of molecules they have not been achieved simultaneously. For biophysical ensembles that exist at room temperature and pressure, and where the entropic contributions are on par with interaction strengths, it is the free energies that are both most important and most difficult to predict. We compute the free energies of solvation for a diverse set of neutral organic compounds using a polarizable force field fitted entirely to ab initio calculations. The mean absolute errors (MAE) of hydration, cyclohexane solvation, and corresponding partition coefficients are 0.2 kcal/mol, 0.3 kcal/mol and 0.22 log units, *i.e.* within chemical accuracy. The model (ARROW FF) is multipolar, polarizable, and its accompanying simulation stack includes nuclear quantum effects (NQE). The simulation tools' computational efficiency is on a par with current state-of-the-art packages. The construction of a wide-coverage molecular modelling toolset from first principles, together with its excellent predictive ability in the liquid phase is a major advance in biomolecular simulation.

¹ InterX Inc, 805 Allston Way, Berkeley, CA 94710, USA. ² Faculty of Physics, Lomonosov Moscow State University, Moscow 119991, Russia. ³ Department of Chemistry, Carnegie Mellon University, Pittsburgh, PA 15213, USA. ⁴ Center for Nanoscale Materials, Argonne National Lab, Argonne, IL 60439, USA. ⁵ Department of Mechanical and Industrial Engineering, University of Illinois, Chicago, IL 60607, USA. ⁶ Department of Chemical Engineering and Materials Science, Wayne State University, Detroit, MI 48202, USA. ⁷ Department of Structural Biology, Stanford University School of Medicine, Stanford, CA 94304, USA. ⁸ These authors contributed equally: Leonid Pereyaslavets, Ganesh Kamath, Boris Fain. ✉email: leonid.pereyaslavets@interxinc.com; boris.fain@interxinc.com

Understanding the energetics of solvation is a fundamental part of describing biophysical processes. The liquid state properties are important in their own right, play a key role in battery design, and are a major part of more structured biological ensembles: e.g., protein shape and behavior, protein–ligand complexes and cell membranes. Because of the overwhelming complexity of ab initio calculations the underlying quantum mechanics must be represented by Newtonian models. The art and science of simulating these systems have been in development since the 1960's^{1,2} and many force fields that describe proteins and other functional groups have been created and are widely used. However, state-of-the-art wide-coverage molecular force fields^{3–9} in simulation packages that enable free energy computations derive some or all of their parameters by fitting to empirical observables. There are at least two drawbacks to this approach. First, even available experimental data (e.g., densities, heats of vaporization) are insufficient to produce models that describe existing compounds precisely; and there will always be molecules (that, for example, haven't been synthesized) that will require more precise description than is available from existing inference. Second, if an empirical model's prediction is

erroneous, it is exceedingly difficult to decide exactly which parameter(s) to remove, add, correct or adjust. A major advantage of Quantum Mechanical (QM)-parametrized physics-based molecular models (force fields)^{10,11} is that, with some caveats for molecular size, QM calculations¹² can be obtained for arbitrary molecules. Another advantage is that prediction errors can be traced to the imprecise description of the interaction energies and rectified in the model. It is therefore highly desirable to create models parameterized entirely from first-principles (ab initio) Quantum Mechanical calculations.

The value of ± 0.5 kcal/mol for the desired (“chemical”) accuracy of free energy predictions arises from several considerations. First and foremost, 0.59 kcal/mol is the thermal noise at ambient conditions (room temperature and pressure). This is the inherent fuzziness of our everyday biological world. Additionally, for example, in ligand–protein lead optimization the definition of “incremental lead improvement” is about 0.5 kcal/mol or ~ 2 – 3 -fold increase in binding affinity.

We have implemented a QM-parametrized force field in a simulation stack that covers arbitrary organic molecules and predicts solvation free energies of molecular systems to accuracy

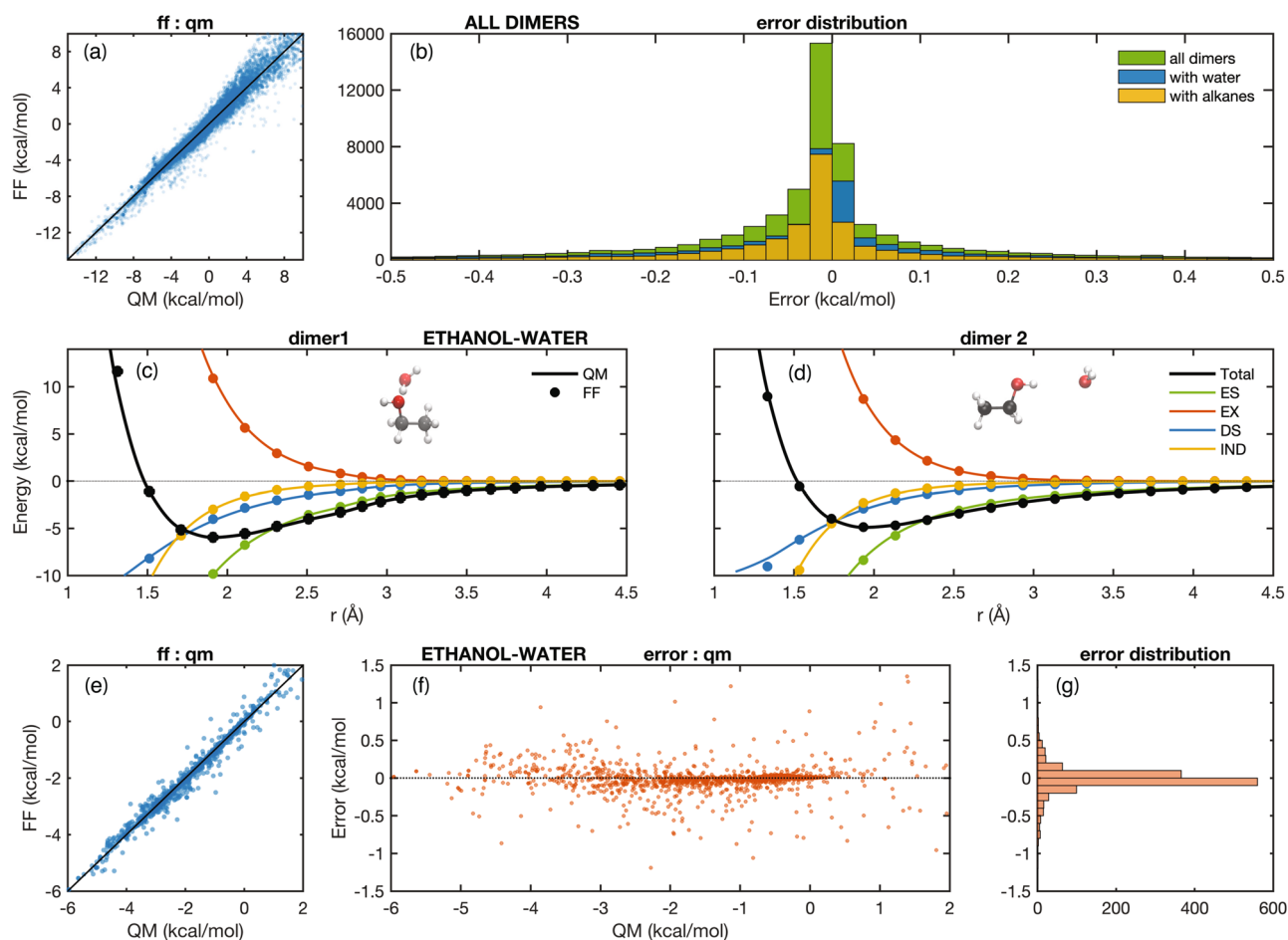


Fig. 1 QM: FF energies' correspondences and deviations. **a** FF vs. QM energies for all the dimers in our training sets. The functional form reproduces the lower energy conformations very well and is designed to permit a larger error in less important high-energy high electron overlap regions. **b** the distribution of errors for our training dimer sets. The MAE of errors are 0.17 kcal/mol for all, 0.19 kcal/mol for dimers with water (total number of dimers = 36,309), and 0.16 kcal/mol for dimers with alkanes (total number of dimers = 25,986); A specific system (ethanol-water) provides a more detailed illustration of model energies and their correspondence with QM. **c, d** dissociation curves for primary (**c**) and secondary (**d**) minima of the ethanol-water dimer. QM energies are solid lines and FF values are filled circles. The colors designate the energy components: electrostatics (ES), exchange-repulsion (EX), dispersion (DS) and induction (IND). The agreements for the total energy and for each component are excellent. **e–g** Error distributions for the ethanol-water dimer: **e** is analogous to **a**; **f** is a difference plot offering a more detailed view and is projected onto (**g**) the error distribution. The MAE for the errors in this system is 0.08 kcal/mol.

of ~ 0.3 kcal/mol for neutral species. The predictions in the liquid phase are satisfyingly accurate, and it is also satisfying that the model is created solely from ab initio computational methods without fitting to any experimental data. We demonstrate the predictive ability of the model and simulation machinery by computing solvation free energies for a wide range of chemical functional groups in water and cyclohexane.

Results

QM-FF agreement. We start by creating a model that represents the QM energies of the ensemble accurately enough. A description of the intermolecular functional form, the component decomposition, and the parametrization procedure is in Supplementary methods (Quantum mechanical details, force field description, force field functional form of ARROW FF, and parameter fitting), Supplementary Fig. 1 and in references^{8,13}. Though models of isolated chemical species with exquisite agreement to QM energies do exist^{14,15}, the complexity required by such precision has prevented researchers from describing arbitrary functional groups simultaneously. One of the contributions of this work is determining the degree of faithfulness that is sufficient for modeling the liquid phase of arbitrary organic molecules and mixtures while keeping the model complexity manageable.

The first step is choosing the level and accuracy of the underlying QM computations. We fit the intermolecular interactions to dimer and select multimer QM energies at the highest level of theory practical for large-scale parameterization. This “silver-like standard”¹⁶ is commonly used as a benchmark in the computational chemistry community, and is within 0.05 kcal/mol from the “gold standard”¹⁶. More details can be found in Supplementary methods (Quantum Mechanical details).

The next step is encapsulating the QM interaction energies in a physics-based analytical model^{8,13}. The required faithfulness demands a significant level of complexity from the functional form: polarizability terms enable proper transferability from dimer to bulk energies¹⁷; multipole descriptions of both the electrostatic¹⁸ and exchange-repulsion interactions permit a precise fit of the potential energy surface for all dimer orientations^{8,19}; a fairly detailed typification accounts for the difference in interaction properties of identical atoms in diverse chemical environments. The force field description including the functional form, and the parametrization workflow and pseudo-code, are discussed in detail in the Supplementary methods. The deviation (MAE) between Quantum mechanical (QM) and force field (FF) energies for all the benchmark dimers and multimers in our training set is 0.17 kcal/mol and the error distribution is centered around zero (Fig. 1a, b, e, f, g). In Fig. 1c, d, we illustrate the QM-FF agreement for a single representative system, a strongly interacting ethanol-water dimer. Additionally, the FF:QM errors for ethanol-water dimers as a function of closest distance are shown in Supplementary Fig. 2. Both the total energies as well as their individual components for this system agree to within 0.1 kcal/mol to their ab initio counterparts. To aid transferability, in addition to reproducing the total energy, we also match the individual components to their corresponding QM counterparts (Fig. 1c, d). To investigate the training-test convergence dependence of dimer space on our force field parameters we conducted this test on a subset of molecules and convergence plots are presented in Supplementary Fig. 4.

Molecular deformations (“bonded interactions”), especially torsions, are critical for correct solvation results because they determine the proper solvent accessibility. A variety of accurate models long established in the field^{3,5,6} as well as brilliant recent developments^{20,21} provide excellent reproduction of the

intramolecular energies. We take the functional form of the bonded interactions from MMFF94³, with force constants and equilibrium values fitted to QM energies at the MP2/aug-cc-pVTZ level of theory.

Solvents. We selected water and cyclohexane as our solvents for this benchmark study. Water, of course, is the most ubiquitous molecule in any biophysical model. We chose cyclohexane because it is nonpolar, it equilibrates relatively quickly, and because there is ample reliable experimental data for both cyclohexane (CHEX) solvation free energies, as well as for the cyclohexane/water (CHEX/H₂O) partition coefficients. Though the two molecules were parameterized with exactly the same procedure as every other functional group, they participate in bulk and thus warrant extra examination of their liquid-state properties.

The liquid phase must properly model not only the molecular 2-body interactions described in the previous section, but also the many-body contributions. For water, which is small, polar, and polarizable, the many-body energies are estimated to be a sizable 27% of the total²². Figure 2a shows the non-additive energies of select optimized water multimers. Additionally, we also show the non-additive behavior in the case of ethanol–water multimers, see Supplementary Fig. 3. They are in excellent agreement with their reference QM values, confirming that the energy partitioning and the induction terms of our polarizable model capture the non-additive fraction properly.

Biological systems exist mostly at room temperature and pressure, where the shifting interplay between enthalpy and

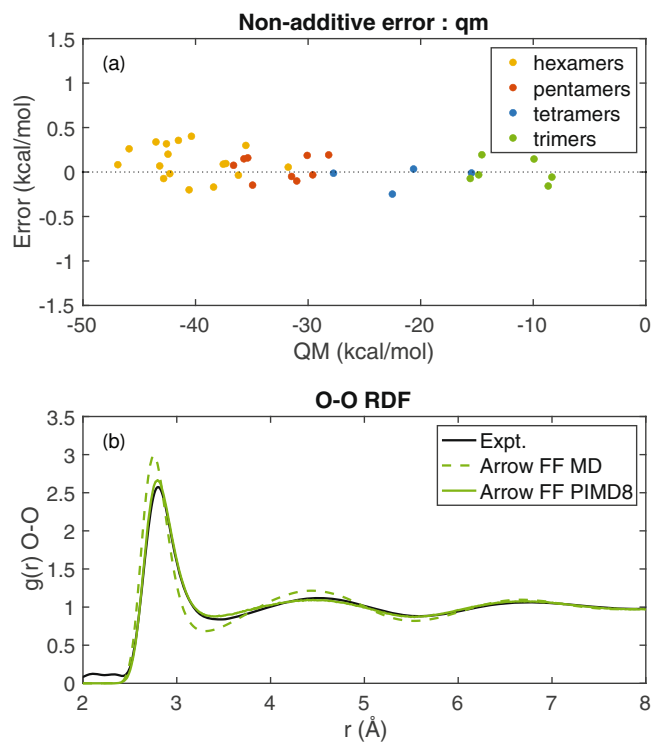


Fig. 2 Properties of the ARROW water model. **a** The non-additive many-body error for water multimers vs. their total QM intermolecular energy. All the many-body errors are below 0.5 kcal/mol or 1% of total energy, and below 3% of the many-body contributions. **b** The radial distribution function for the O–O distance in water. The MD RDF (dotted green) is overstructured compared with the experimental curve, and the presence of NQE (solid green) brings the structure of ARROW H₂O in excellent agreement with the experimental one.

entropy enables the immense variety of biological phenomena. Therefore, it is the free energies of ensembles that are both the most useful and interesting and also the most difficult to predict correctly, and what we focus on here. For solvation, in addition to capturing the enthalpy of interaction with itself and the solute, a solvent model must also reproduce the entropic effects of pushing aside and reordering molecules to create a cavity for placing the solute. This is especially important for water as it is small, highly polar, and, though called a liquid, is highly structured at room temperature and pressure. In Table 1 we list three bulk properties of our solvents: density, heat of vaporization (Hvap) and the highly informative self-solvation. The values for water agree with experimental values to within 3% or better. Additional proof that our model has captured the free energy of cavity creation in water accurately is that the hydration of anthracene, a large, non-polar molecule, is correct to within 2% (0.1 kcal/mol) (Supplementary Data 1a). The derivative of the system Hamiltonian with respect to the alchemical reaction coordinate ($\langle dH/d\lambda \rangle$) for desolvation of

anthracene in water and its accumulated statistical errors are shown in Supplementary Fig. 5. The cyclohexane predictions are slightly less accurate for two reasons: 1) it is a larger molecule so per heavy atoms the energetics are actually very good and 2) we designated its atoms to be the same atom type(s) as linear alkanes (unlike those of smaller, strained, cyclic alkanes), which introduces a slight discrepancy with QM energies. Nonetheless, the bulk energetics of cyclohexane are well within our target accuracy of 0.5 kcal/mol.

Finally, an excellent measure of how well liquid structure is captured by a model is the radial distribution function (RDF). In Fig. 2b we demonstrate that the ARROW FF reproduces the experimental water oxygen–oxygen (O–O) RDF and, therefore, describes the order of water very well. Additionally, we show that employing eight beads reaches sufficient convergence for the free energies and structural properties (see Supplementary Figs. 7 and 8 and Supplementary Table 4). The agreements for both neat properties (Table 1) and water structure (Fig. 2b) are

Table 1 Neat properties and hydration/solvation of water and cyclohexane.

H ₂ O	Density (g/cc)	Hvap (kcal/mol)	Hydration (kcal/mol)	Self-solvation (kcal/mol)
expt	0.997	10.51	−6.30	−6.30
ARROW FF (MD)	1.027	11.98	−6.81	−6.81
ARROW FF (PIMD8)	1.027	10.63	−6.13	−6.13
CHEX	Density (g/cc)	Hvap (kcal/mol)	Hydration (kcal/mol)	Self-solvation (kcal/mol)
expt	0.790	7.91	1.20	−4.42
ARROW FF (MD)	0.803	8.04	−0.10	−4.23
ARROW FF (PIMD)	0.786	7.83	1.08	−4.02

Predictions were performed by classical simulations and with inclusion of NQE. All numbers are in good agreement with the experimental values, with PIMD simulations being significantly closer than the classical MD ones. The self-solvation of water is a succinct measure of model accuracy and we recommend its determination for all water models. For water the self-solvation and hydration are obviously identical.

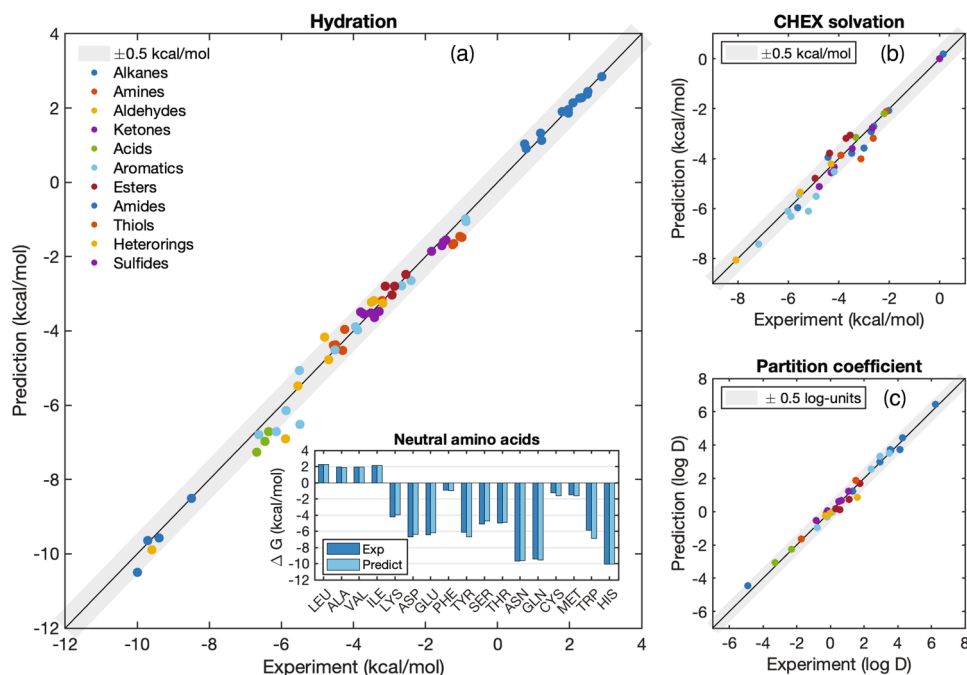


Fig. 3 ARROW force field solvation predictions. **a** Predicted vs. experimental free energy of hydration for a diverse set of compounds. The straight line is a line of perfect agreement between experimental and theoretical values, and the gray bar is the range of chemical accuracy. The predicted and experimental free energies of hydration for neutral amino acid analogs are inset. **b** Predicted vs. experimental free energy of solvation in cyclohexane. The molecules here are a subset of those in **a** because only those with experimental values for CHEX solvation can be included. **c** H₂O/CHEX partition coefficient for the same set as **b**. The free energy predictions are well within chemical accuracy.

significantly improved by including NQE^{13,15,23}. Satisfyingly, the small errors in initial model parameterization are not amplified through the chain of model construction and simulation machinery.

Solutes and solvation predictions. We chose representative solutes containing all of the common neutral chemical functional groups: carboxylic acids, alkanes, alkenes, aromatics, aldehydes, ketones, alcohols, amides, esters, thiols, sulfides, disulfides, and heterocycles²⁴. The simulations were performed independently by four groups using their own respective computational resources and architectures, and then averaged. The graphical summaries of the solvation and hydration free energies predictions' are in Fig. 3a, b, and we list the results for each molecule in Supplementary Data 1a. We also provide the free energy results as reproduced by our collaborators in Supplementary Data 1d. Because aqueous protein and protein–ligand systems are of special importance, and because accurate prediction of solvation and desolvation of amino acids is critical for modeling of these systems²⁵, we highlight the results for neutral amino-acid analogs separately (Fig. 3a inset), see Supplementary Data 1b for raw data. The partition coefficient is a valuable measure of the model's simultaneous compatibility with both polar (e.g., aqueous) and non-polar (e.g., membranes and proteins) environments which is crucial for describing bio-molecular systems, and we show it in Fig. 3c.

The proper art of simulation^{26,27} is also essential for obtaining accurate predictions. Accurate treatment of long range electrostatic (e.g., Particle Mesh Ewald^{28,29}) and dispersion²⁷ interactions, proper thermodynamic modeling (thermostats and barostats)^{30,31}, enhanced sampling techniques and the Path Integral formulation of nuclear motion^{32–34} all help to translate the FF-QM agreements to correct free-energy values. Further details are provided in Supplementary methods (Simulation details and protocols). We also provide computational performance of the ARROW FF stack for CPU and CPU+GPU implementations for both classical and path-integral simulations in Supplementary Table 2.

The error (MAE) for the free energies of hydration is 0.2 kcal/mol and for the neutral amino-acid subset is 0.23 kcal/mol. The largest hydration errors seen for o-cresol and 3-methyl-indole are only ~1 kcal/mol. For solvation in cyclohexane the MAE is 0.3 kcal/mol, and for the partition coefficient it is 0.22 log units. These predictions are very good: most are within experimental and simulation uncertainty, and are uniformly correct across a diverse range of chemical groups of varying sizes and interaction strengths.

We recently highlighted the importance of including NQE when modeling alkanes^{13,35}. The results presented in this manuscript suggest that NQE must be taken into account for precision calculations for all molecular systems. We illustrate this in Fig. 4a where we plot the hydration predictions of classical simulations alongside those performed with PIMD. Proper accounting of the quantum nature of nuclear motion systematically shifts the predictions towards the experimental values and improves the prediction error from MAE of 0.78 to 0.2 kcal/mol.

Comparison with other force fields. The main advance reported in this paper is three-fold: our model is a wide-coverage force field and simulation stack parameterized exclusively from QM data which produces accurate predictions. It is of interest to gauge the relative performance of ARROW FF to existing wide-coverage state-of-the-art models for prediction accuracy. Most of the QM-parameterized FF's^{10,36} are not currently enabled in a simulation stack which produces free energy predictions, so we

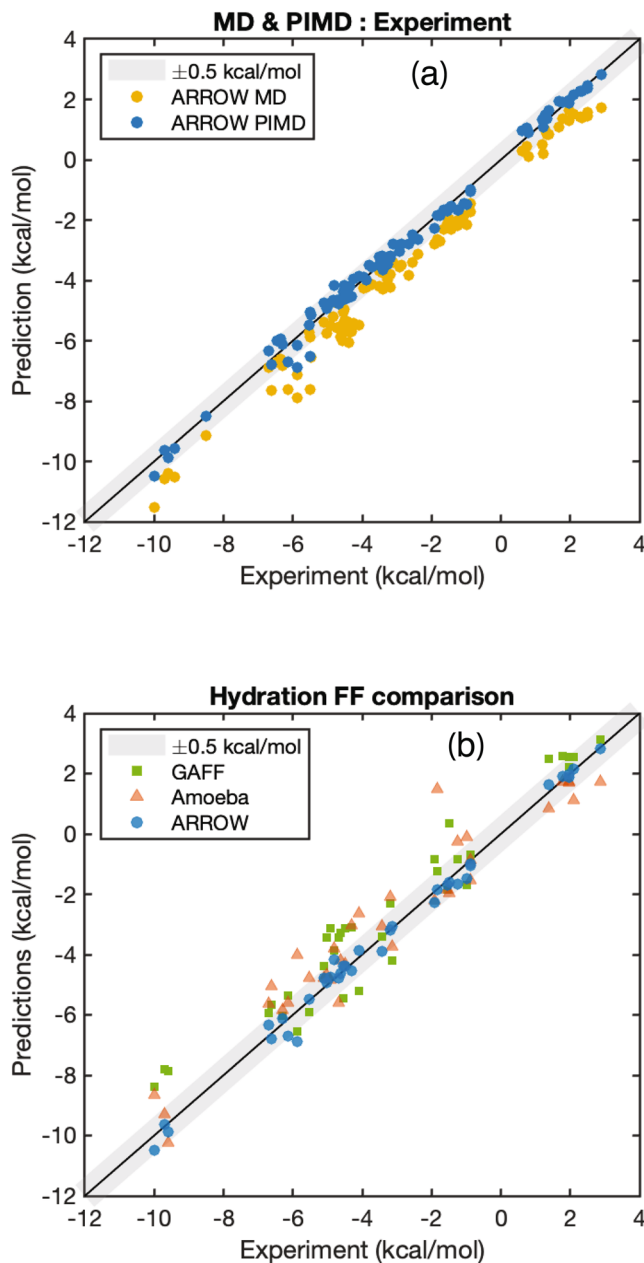


Fig. 4 NQE effect and comparison of hydration predictions. **a** A visual comparison of the hydration predictions vs. experimental values for PIMD8 vs. classical MD values. The inclusion of NQE systematically improves the predictions and decreases the overall error (MAE) from 0.78 to 0.2 kcal/mol. **b** A comparison of the hydration free energies to state-of-the-art wide coverage Force Fields. The molecules shown include the major functional groups that have been parametrized by all three models and are therefore a subset of those shown in Fig. 3a. The errors (MAE) for GAFF, AMOEBA, and ARROW FF are 0.88, 0.76, and 0.22 kcal/mol, respectively.

selected two widely-used empirical models to compare with. One is GAFF⁶, a representative of the many fixed-charge models, and the other is a polarizable model AMOEBA^{9,18}. To avoid reproduction discrepancies the comparison is made on the available published subset of functional groups and is plotted in Fig. 4b. The MAE's for this subset are, respectively, 0.88 (GAFF AM1-BCC)³⁷, 0.76 (AMOEBA)⁹ and 0.22 (ARROW) kcal/mol. A list of molecules and their predicted hydration values for each model is in Supplementary Data 1c. Additionally, in Supplementary Data 1e, Supplementary Table 3, Supplementary Fig. 6, and

Supplementary methods (Comparison to Implicit solvent models and Machine learning models) we summarize and discuss the comparative performance of several excellent tools from a variety of methodologies that focus specifically on prediction of solvation energies^{38,39}. In Supplementary Data 1f we also show the QM-FF agreement of ARROW FF on the S22 and S66 datasets as well as a comparison with the same for geometry, frequency, non-covalent force field (GFN-FF)^{11,39}, the MAE's for such datasets can be found in Supplementary Table 1.

We have shown that a QM-parametrized, physics-based force field embedded in a simulation and analysis stack predicts the free energies of solvation of arbitrary organic molecules to an accuracy better than thermal noise at room temperature (± 0.5 kcal/mol). The correspondence from quantum mechanics to ensemble predictions is established via several important links. First, the benchmark QM calculations need to be of sufficient accuracy. Second, the model should provide a faithful description of the QM potential energy surface (PES), which imposes a significant yet computationally manageable level of complexity on the functional form. Third, the established art of molecular ensemble averaging must be performed with care. Finally, the dynamics of sampling the system should account for nuclear quantum effects. The ARROW FF is likely at the limit of complexity feasible for a wide-coverage analytical force field, and so it is satisfying that this model results in excellent prediction of properties in the liquid phase.

Data availability

The scripts, tools, and data used in this work are available from the corresponding authors and InterX Inc. upon request. The full results' data has been included in Supplementary Information Tables and further data is also available upon request.

Code availability

The codes, tools and data needed to reproduce the data presented in this article is available on github https://github.com/freecurve/interx_solvation_suite.

Received: 18 June 2021; Accepted: 3 January 2022;

Published online: 20 January 2022

References

- Lifson, S. & Warshel, A. Consistent force field for calculations of conformations, vibrational spectra, and enthalpies of cycloalkane and n-alkane molecules. *J. Chem. Phys.* **49**, 5116–5129 (1968).
- Levitt, M. & Lifson, S. Refinement of protein conformations using a macromolecular energy minimization procedure. *J. Mol. Biol.* **46**, 269–279 (1969).
- Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **17**, 490–519 (1996).
- Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **118**, 11225–11236 (1996).
- MacKerell, A. D. et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**, 3586–3616 (1998).
- Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **25**, 1157–1174 (2004).
- Mackerell, A. D. Jr Empirical force fields for biological macromolecules: overview and issues. *J. Comput. Chem.* **25**, 1584–1604 (2004).
- Donchev, A. G. et al. Assessment of performance of the general purpose polarizable force field QMPFF3 in condensed phase. *J. Comput. Chem.* **29**, 1242–1249 (2008).
- Ponder, J. W. et al. Current status of the AMOEBA polarizable force field. *J. Phys. Chem. B* **114**, 2549–2564 (2010).
- Xu, P., Guidez, E. B., Bertoni, C. & Gordon, M. S. Perspective: ab initio force field methods derived from quantum mechanics. *J. Chem. Phys.* **148**, 090901 (2018).
- Spicher, S. & Grimme, S. Robust atomistic modeling of materials, organometallic, and biochemical systems. *Angew. Chem. Int. Ed. Engl.* **59**, 15665–15673 (2020).
- Jensen, F. *Introduction to Computational Chemistry* (Wiley, 2017).
- Pereyaslavets, L. et al. On the importance of accounting for nuclear quantum effects in ab initio calibrated force fields in biological simulations. *Proc. Natl. Acad. Sci. USA* **115**, 8878–8882 (2018).
- Babin, V., Leforestier, C. & Paesani, F. Development of a 'First Principles' water potential with flexible monomers: dimer potential energy surface, VRT spectrum, and second virial coefficient. *J. Chem. Theory Comput.* **9**, 5395–5403 (2013).
- Medders, G. R., Babin, V. & Paesani, F. Development of a 'First-Principles' water potential with flexible monomers. III. Liquid phase properties. *J. Chem. Theory Comput.* **10**, 2906–2910 (2014).
- Burns, L. A., Marshall, M. S. & Sherrill, C. D. Appointing silver and bronze standards for noncovalent interactions: a comparison of spin-component-scaled (SCS), explicitly correlated (F12), and specialized wavefunction approaches. *J. Chem. Phys.* **141**, 234111 (2014).
- Cieplak, P., Dupradeau, F.-Y., Duan, Y. & Wang, J. Polarization effects in molecular mechanical force fields. *J. Phys. Condens. Matter* **21**, 333102 (2009).
- Ren, P. & Ponder, J. W. Polarizable atomic multipole water model for molecular mechanics simulation. *J. Phys. Chem. B* **107**, 5933–5947 (2003).
- Van Vleet, M. J., Misquitta, A. J., Stone, A. J. & Schmidt, J. R. Beyond Born–Mayer: improved models for short-range repulsion in ab initio force fields. *J. Chem. Theory Comput.* **12**, 3851–3870 (2016).
- Smith, J. S. et al. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Commun.* **10**, 2903 (2019).
- von Lilienfeld, O. A., Müller, K.-R. & Tkatchenko, A. Exploring chemical compound space with quantum-based machine learning. *Nat. Rev. Chem.* **4**, 347–358 (2020).
- Stone, A. *The Theory of Intermolecular Forces*. (OUP Oxford, 2013).
- Cerioti, M. et al. Nuclear quantum effects in water and aqueous systems: experiment, theory, and current challenges. *Chem. Rev.* **116**, 7529–7550 (2016).
- Horta, B. A. C. et al. A GROMOS-compatible force field for small organic molecules in the condensed phase: the 2016H66 parameter set. *J. Chem. Theory Comput.* **12**, 3825–3850 (2016).
- Bash, P. A., Singh, U. C., Langridge, R. & Kollman, P. A. Free energy calculations by computer simulation. *Science* **236**, 564–568 (1987).
- Levitt, M. A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* **104**, 59–107 (1976).
- Allen, M. P. & Tildesley, D. J. *Computer Simulation of Liquids* (Oxford University Press, 2017).
- Ewald, P. P. Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Ann. Phys.* **369**, 253–287 (1921).
- Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: an N-log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).
- Martyna, G. J., Klein, M. L. & Tuckerman, M. Nosé–Hoover chains: the canonical ensemble via continuous dynamics. *J. Chem. Phys.* **97**, 2635–2643 (1992).
- Martyna, G. J., Tuckerman, M. E., Tobias, D. J. & Klein, M. L. Explicit reversible integrators for extended systems dynamics. *Mol. Phys.* **87**, 1117–1157 (1996).
- Tuckerman, M. E., Berne, B. J., Martyna, G. J. & Klein, M. L. Efficient molecular dynamics and hybrid Monte Carlo algorithms for path integrals. *J. Chem. Phys.* **99**, 2796–2808 (1993).
- Feynman, R. P., Hibbs, A. R. & Styer, D. F. *Quantum Mechanics and Path Integrals* (Courier Corporation, 2010).
- Martyna, G. J., Hughes, A. & Tuckerman, M. E. Molecular dynamics algorithms for path integrals at constant pressure. *J. Chem. Phys.* **110**, 3275–3290 (1999).
- Balog, E., Hughes, A. L. & Martyna, G. J. Constant pressure path integral molecular dynamics studies of quantum effects in the liquid state properties of n-alkanes. *J. Chem. Phys.* **112**, 870–880 (2000).
- Grimme, S. A general quantum mechanically derived force field (QMDF) for molecules and condensed phase simulations. *J. Chem. Theory Comput.* **10**, 4497–4514 (2014).
- Mobley, D. L., Bayly, C. I., Cooper, M. D., Shirts, M. R. & Dill, K. A. Small molecule hydration free energies in explicit solvent: an extensive test of fixed-charge atomistic simulations. *J. Chem. Theory Comput.* **5**, 350–358 (2009).
- Weinreich, J., Browning, N. J. & von Lilienfeld, O. A. Machine learning of free energies in chemical compound space using ensemble representations: reaching experimental uncertainty for solvation. *J. Chem. Phys.* **154**, 134113 (2021).
- Ehlerst, S., Stahn, M., Spicher, S. & Grimme, S. Robust and efficient implicit solvation model for fast semiempirical methods. *J. Chem. Theory Comput.* **17**, 4250–4261 (2021).

Acknowledgements

The authors thank InterX Inc. for their generous support. The authors would also thank Alexander Donchev, Oleg Khoruzhii, Alston Misquitta, and participants of the Telluride “Many-Body Interactions: From Quantum Mechanics to Force Fields” workshop for useful and stimulating discussions. We also thank Sean Greenslade, Christopher Lock, Hulda Chen, Meredith Roberts, Erik Ven, Micheal Feese, and David Bushnell. The authors also acknowledge the use of the Center for Nanoscale Materials and Office of Science user facilities, which was supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. DE-AC02-06CH11357. S.S. acknowledges support by the U.S. Department of Energy through BES award DE-SC0021201.

Author contributions

L.P., B.F., and G.K. designed research; L.P., G.K., I.K., Serz.S., E.V., G.N., M.D., O.B., A.I., I.L., A.K., T. G., I.L., M.O., and B.F. wrote tools and performed research; L.P., G.K., and B.F. analyzed data; J. S., M.G.K., M.S., S.B., H.C., M.G.S., S.S., B.C., and J.P. validated the models and the results; and M.L., R.D.K., and B.F. wrote the manuscript which was revised and reviewed by all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-022-28041-0>.

Correspondence and requests for materials should be addressed to Leonid Pereyaslavets or Boris Fain.

Peer review information *Nature Communications* thanks Bernd Hartke and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022