Cluster Models for Next-Generation, Machine-Learning-Based Energy Functions for Molecular Simulations

JingChun Wang, Meenu Upadhyay, Eric D. Boittier, Kham Lek Chaton, Valerii Andreichev, Mike Devereux, Shimoni Patel, Sena Aydin, Kai Töpfer, and Markus Meuwly*

Department of Chemistry, University of Basel, Klingelbergstrasse 80, CH-4056 Basel, Switzerland

E-mail: m.meuwly@unibas.ch

Abstract

Energy functions for pure and heterogenous systems are one of the backbones for molecular simulation of condensed phase systems. With the advent of machine learned potential energy surfaces (ML-PESs) a new era has started. Statistical models allow the representation of reference data from electronic structure calculations for chemical systems of almost arbitrary complexity at unprecedented detail and accuracy. Here, kernel- and neural network-based approaches for intramolecular degrees of freedom are combined with distributed charge models for long range electrostatics to describe the interaction energies of condensed phase systems. The main focus is on illustrative examples ranging from pure liquids (dichloromethane, water) to chemically and structurally heterogeneous systems (eutectic liquids, CO on amorphous solid water), reactions (Menshutkin), and spectroscopy (triatomic probes for protein dynamics). For all examples, small to medium-sized clusters are used to represent and improve the total interaction energy compared with reference quantum chemical calculations. Although remarkable accuracy can be achieved for some systems (chemical accuracy for dichloromethane and water), it is clear that more realistic models are required for van der Waals contributions and improved water models need to be used for more quantitative simulations of heterogeneous chemical and biological systems.

1 Introduction

Cluster systems consisting of a finite number of atomic and/or molecular building blocks constitute a state of matter between isolated gas phase units and the bulk. In the past, one of the main driving forces to generate, investigate, and characterize such systems was the realization that clusters of increasing sizes may provide an understanding for how condensed phase properties emerge across various length scales.[1] Furthermore, following the properties of clusters as they increase in size may provide information about the phenomenon of nucleation.[2] Historically, one of the earliest examples that was investigated are metal clusters.[3]

Clusters as finite-sized aggregates of identical atoms/molecules or mixed clusters have been investigated with great success. For example, spectroscopic and computational work on protonated water clusters of increasing size has provided fundamental insights into the relationship between structure and spectroscopy of such systems.[4,5] Likewise, the structure, energetics and thermodynamics of atomic clusters interacting through energy functions exhibiting different strengths and range was investigated.[6] From an analysis of the distribution of low-lying minima, some unusual thermodynamic properties of finite systems were determined.[7,8] Disconnectivity graphs[9] provided a compact rendering of the PES, the relationship between minima and the connectivity through transition states depending on the strength of the interatomic interactions.[8] More specifically, for pure and mixed rare gas clusters the collision induced absorption spectra were determined from classical molecular dynamics (MD) simulations and path integral MD simulations were used to characterize the structures of such clusters.[10,11] The results from such computations agreed favourably with experiments on their spectroscopy and structure. All these investigations pointed towards a pronounced dependence of the measured properties on the structure and size of the systems.

Clustering can also occur in solution. For example, the phenomenon of microheterogeneity occurs in water/alcohol mixtures and leads to local aggregation and nonuniform distribution of one type of species although the mixture on larger than molecular length scales appears more uniform.[12] Related aggregation and clustering phenomena are also relevant to and at play in separation methods such as high performance liquid chromatography (HPLC).[13–16]

Parametrization of empirical energy functions (or force fields) typically hinges on a combination of fitting to experimental reference data and information obtained from electronic structure calculations.[17–20] In addition to monomer properties, monomer–water complexes are included for parametrizing the CHARMM General Force Field(CGenFF) force field for realistically describing H-bonding interactions.[17] In the case of CGenFF the water model used was the TIP3P model.[21] Typical experimental observables are the gas phase vibrational spectrum (infrared and/or Raman), the pure liquid density, heat of vaporization, or diffusion coefficients in one-component liquids. Information that needs to be obtained from *ab initio* calculations are the partial charges on each of the atoms for which different approaches exist.[22] Estimating partial charges from experimental X-ray crystallography is in principle possible but requires highest-resolution structures.[23] In addition, *ab initio* calculations of clusters include local information on how interaction energies depend on atomic-scale details that are missing from (macroscopic) experimental data.

It is of interest to note that J. E. (Lennard-)Jones, after whom "Lennard-Jones clusters" are named and who researched the distance dependence of the intermolecular interactions between weakly interacting particles, wrote[24] already in 1924 (italics added): "Until our knowledge of the disposition and motion of the electrons in atoms and molecules is more complete, we cannot hope to make a direct calculation of the nature of the *forces* called into play during an encounter between molecules in a gas. It is true that [..] Debye [..]investigated the nature of the field in the neighbourhood of a hydrogen atom [..] and has shown how the pulsating field gives rise on the whole to a force of repulsion, as well as one of attraction on a unit negative charge. But it is difficult to see how this work can be extended to more complex systems.[..] One such method is to *assume a definite law of force*, and then by the methods of the kinetic theory to deduce the appropriate law of dependence of the viscosity of a gas on temperature." Obviously, Lennard-Jones thought of "force" instead of "energy", although the expressions that were parametrized described how the total energy changes with geometry.

In the field of machine learning potential energy surfaces encoding the chemical environment into the model plays a central role. Defining such an environment is usually done by choosing a cutoff radius which also leads to clusters of atoms surrounding a reference atom in order to generate a representation - or descriptor - suitable for fitting a machine learned PES.[25] One such descriptor is the Smooth Overlap of Atomic Positions (SOAP) that quantifies the similarity between any two neighbourhood environments, and its performance was tested in particular on small silicon clusters. Another possibility is to encode the environment through features which are trained by minimizing a suitable loss function. This is the approach followed in PhysNet.[26]

The present work considers pure and mixed clusters as a test-bed to develop accurate energy functions for condensed-phase simulations. The approach taken here uses a combination of machine learning-based techniques, combined with empirical expressions for the total energy of the system. Such an approach provides flexibility, accuracy and computational speed which are important for large(r)-scale simulations. First, the methods used are described, followed by new results on a range of paradigmatic systems, including pure substances, adsorbates on water, mixed and electrostatically dominated mixtures, reactive systems, and spectroscopic probes for condensed phase systems.

2 Methods

This chapter briefly describes the energy functions employed in the present work. More technical details are provided in each of the results subsections together with specifics of the MD simulations carried out, if applicable. All simulations were run with the CHARMM program with provisions to use machine learned energy functions.[27–30] If not otherwise mentioned, the empirical energy function used in the present work is CGenFF[17] together with the TIP3P water model[21] for consistency.

2.1 Machine-Learned Energy Functions

The machine learning-based techniques used in the present work include kernel- and neural network-based approaches.

One powerful method to construct accurate PESs uses reproducing kernel Hilbert spaces (RKHSs)[31] for which dedicated code has been made available.[28] The theory of reproducing kernel Hilbert spaces asserts that for N training values $f_i = f(\mathbf{x}_i)$ of a function $f(\mathbf{x})$ at locations $\mathbf{x}_i$, $f(\mathbf{x})$ at arbitrary position $\mathbf{x}$ can always be approximated as a linear combination of kernel products[32]

$$\widetilde{f}(\mathbf{x}) = \sum_{i=1}^{N} c_i K(\mathbf{x}, \mathbf{x}_i) \tag{1}$$

Here, the c_i are coefficients and $K(\mathbf{x}, \mathbf{x}')$ is the reproducing kernel of the RKHS. The coefficients c_i satisfy the linear relation

$$f_j = \sum_{i=1}^{N} c_i K_{ij} \tag{2}$$

where i and j are both elements of the training set and the symmetric, positive-definite kernel matrix $K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$ and c_i can therefore be calculated from the known training values f_i by solving Eq. 2 for the unknowns c_i using, e.g. Cholesky decomposition.[33] With the coefficients c_i determined, the function value at an arbitrary position $\mathbf{x}$ can be calculated using Eq. 1. Derivatives of $\widetilde{f}(\mathbf{x})$ of any order can be calculated analytically by replacing the kernel function $K(\mathbf{x}, \mathbf{x}')$ in Eq. 1 with its corresponding derivative.

The explicit form of the multi-dimensional kernel function $K(\mathbf{x}, \mathbf{x}')$ also depends on the problem to be solved. It is possible to construct D-dimensional kernels as tensor products of one-dimensional kernels $k(x, x')$

$$K(\mathbf{x}, \mathbf{x}') = \prod_{d=1}^{D} k^{(d)}(x^{(d)}, x'^{(d)}) \tag{3}$$

For the kernel functions $k(x, x')$ explicit physical knowledge can be encoded, for example the correct asymptotic decay for the long range interactions.[34] Explicit radial kernels include the reciprocal power decay kernel[35]

$$k_{n,m}(x, x') = n^2 x_>^{-(m+1)} \mathrm{B}(m+1, n)\, {}_2\mathrm{F}_1\left(-n+1, m+1; n+m+1; \frac{x_<}{x_>}\right) \tag{4}$$

where $x_>$ and $x_<$ are the larger and smaller of x and x', the integer n determines the smoothness and m controls the long-range decay (e.g. $m = 5$ for dispersion), $\mathrm{B}(a, b)$ is the beta function and ${}_2\mathrm{F}_1(a, b; c; d)$ is the Gauss hypergeometric function.

The NN-based ML-PESs were trained using the PhysNet architecture.[26] The loss function to be optimized included energies (E^{ref}), forces ($F_{i,\alpha}^{\mathrm{ref}}$), the total charge ($Q^{\mathrm{ref}}$), and dipole moments ($p_\alpha^{\mathrm{ref}}$) for N training structures

$$
\begin{aligned}
\mathcal{L} = w_E \left|E - E^{\mathrm{ref}}\right| + \frac{w_F}{3N} \sum_{i=1}^{N} \sum_{\alpha=1}^{3} \left| -\frac{\partial E}{\partial r_{i,\alpha}} - F_{i,\alpha}^{\mathrm{ref}} \right| \\
+ w_Q \left| \sum_{i=1}^{N} q_i - Q^{\mathrm{ref}} \right| + \frac{w_p}{3} \sum_{\alpha=1}^{3} \left| \sum_{i=1}^{N} q_i r_{i,\alpha} - p_\alpha^{\mathrm{ref}} \right| + \mathcal{L}_{\mathrm{nh}}.
\end{aligned} \tag{5}
$$

and was minimized using the Adam optimizer,[36,37] and α refers to the three Cartesian components of the vectorial quantities, respectively. The hyperparameters[26,29] w_i $i \in \{E, F, Q, p\}$ differentially weigh the contributions to the loss function and were $w_E = 1$, $w_F \sim 52.92$, $w_Q \sim 14.39$ and $w_p \sim 27.21$, respectively, and the term $\mathcal{L}_{\mathrm{nh}}$ is a "nonhierarchical penalty" that regularizes the loss function.[26]

PhysNet belongs to the family of message-passing NNs (MPNNs) which falls within the broader category of graph neural networks.[38] As with all MPNNs, PhysNet contains an input layer, several hidden layers ("modules") and one output layer. Each module in PhysNet consists of an interaction block and several residual blocks to facilitate training as the depth of the NN increases. In the present work, 5 hidden layers were used as in previous work.[39,40] Based on nuclear charges (the "chemistry") Z and positions $\mathbf{R}$ of all atoms of a molecule, the feature vectors describing each atom in a local chemical environment are iteratively refined, given total energies and forces for a number of reference structures.[26,29,30,41] The components of the feature vectors, which have length 128 throughout this work, were randomly initialized between $-\sqrt{3}$ and $\sqrt{3}$. As the messages propagate through the NN, the atomic feature vectors are refined to minimize the total loss function (Eq. 5).[26]

2.2 Anisotropic Electrostatics

Models for electrostatics that go beyond atom-centered point charges include atom-centered multipoles (MTP), minimally distributed charge models (MDCM),[42] and conformationally dependent MDCM where the dependence is either described by explicit parametrized functions (fMDCM)[43] or through 1-dimensional kernel functions (kMDCM) that act on internal coordinates.[44] For fMDCM models, a single internal degree of freedom, often a valence angle, parametrizes the position of a number of distributed charges in the local (molecular) frame, which is achieved through constrained least-squares fitting to the ESP for several distorted structures. The functions chosen to parameterize these 'flexible' charges are arbitrary but a polynomial expansion up to the third power is generally sufficient. This addition of an internal polarization contribution complements external polarization models such as the Drude-treatment[45–48] but specifically tuned to reproduce the electrostatic potential in the van der Waals region of the molecule. The kMDCM model, used for water, was an extension[44] which generates optimized non-equilibrium charge models using a Gaussian kernel-based representation to describe anisotropic electrostatics which adapt smoothly with molecular geometry. Further details on the fitting procedure can be found in Ref. 44.

3 Results

This section presents results for a range of systems with a focus on improvements of the total energy function. For a few cases, the sensitivity of computed observables on different models is explored to highlight changes in the performance depending on the parametrization.

3.1 Dichloromethane

Dichloromethane (DCM, CH_2Cl_2) is a widely used and extensively studied solvent.[49] Its intermolecular interactions are dominated by dispersion, short-range repulsion and electrostatic contributions from H-bonding.[50,51] Traditional empirical energy functions such as CHARMM,[17] Amber,[52] Gromos,[53] and OPLS,[54] which rely on atom-centered point charges and pairwise-additive non-bonded terms are limited to representing isotropic charge distributions. Such models inadequately describe anisotropic features such as σ-holes characteristic of halogen atoms, including chlorine.[55,56] Over time, numerous models have been developed to more realistically describe the non-bonded interactions in DCM, beginning with three- and five-site models for electrostatics[57] and five-site Lennard–Jones representations.[58] Further refinements have attempted to capture anisotropy in the bulk phase, including the use of atomic quadrupolar moments,[59] effective pair potentials,[60] polarizable pair potentials,[61] and re-parameterized van der Waals terms to better describe solvation.[62,63] Building on these developments, the present work explores the use of machine learning models as a cost-efficient alternative to explicitly complex functional terms. In particular, machine-learned dimer potentials can provide accurate short-range interaction energies, with an appropriate cutoff allowing a smooth transition to a more accurate electrostatic model or CGenFF charges at medium and long range. In this way, empirical force fields can be systematically reparametrized and extended to yield a more transferable and physically grounded description of DCM. The use of an additive dimer potential with no explicit many-body correction for bulk simulations is justified for DCM due to relatively weak H-bonding[51] and polarization contributions that limit many-body effects.[19]

To generate clusters for parametrization, a 32^3 Å^3 cubic box was generated using PACKMOL.[64] Using CHARMM[30] and the CGenFF[17] energy function, the system was relaxed through 2000 steps of Steepest Descent (SD) algorithm, followed by heating and equilibration simulations of 20 ps and 50 ps, respectively, with a timestep of $\Delta t = 1$ fs, in the NVT

and NPT ensembles. A Nosé-Hoover thermostat was used to maintain the temperature at 300 K and the total pressure was maintained at 1 atm using the Langevin Piston barostat.[65] Long range interactions were treated using particle mesh Ewald with a cutoff of 14 Å and the Lennard-Jones interactions were switched between 10 Å and 12 Å. From a 1 ns production simulations in the NPT ensemble using the Leapfrog Verlet integrator with a timestep of $\Delta t = 0.2$ fs every 100^{th} snapshot was saved. A total of 200 distinct clusters containing 20 DCM molecules were extracted by randomly choosing a DCM molecule and the 19 closest neighbors. For each cluster, the total energy of DCM_{20} and all 190 dimer energies DCM–DCM were determined at the DLPNO-MP2/cc-pVTZ level[66,67] using ORCA.[68,69]

First, the total interaction energy from the empirical energy function (CGenFF) is considered and improved by readjusting the LJ-parameters. For this, E^{inter} is defined as the total interaction energy of a given cluster from which the sum of monomer energies were subtracted. The interaction energy E_j^{inter} of cluster j with $j \in [1, 200]$ was obtained from the total energy E_j^{total} from which the sum of the 20 monomer energies $\sum_{i=1}^{20} E_{i,j}^{\mathrm{monomer}}$ were subtracted, see Eq. 6. Hence, the interaction energy

$$E_j^{\mathrm{inter}} = E_j^{\mathrm{cluster}} - \sum_{i=1}^{20} E_{i,j}^{\mathrm{monomer}} \tag{6}$$

from the electronic structure calculations can be directly compared with the external/non-bonded energy contribution for all 200 clusters, see top panels in Figure 1.

Machine learned models like PhysNet[26,29] are well-suited for describing close range intermolecular interactions but require a large amount of training data when the distance between monomers increases. Hence, combining an accurate close-range representation for dimers using a NN-PES with a more empirical long-range representation based on electrostatics and LJ-contributions is a potentially data-efficient and accurate route. For constructing the dimer-PES, a 20-monomer DCM cluster j contains 190 distinct dimer pairs. The total formation energy for cluster j can therefore be written as the sum of these 190 dimer contributions, plus a residual term accounting for many-body contributions

$$E_j^{\text{inter}} = \sum_{i=N}^{190} E_{i,j}^{\text{dimer,inter}} + E_j^{\text{residual}} \tag{7}$$

where

$$E_i^{\text{dimer,inter}} = E_i^{\text{dimer}} - \left(E_{a,i}^{\text{monomer}} + E_{b,i}^{\text{monomer}}\right) \tag{8}$$

Here, the residual energy contribution was not determined explicitly and hence $E_j^{\text{inter}} \approx \sum_{i=N}^{190} E_{i,j}^{\text{dimer,inter}}$.

For training the ML-PES for the DCM dimers, PhysNet;[26] utilizing the ML-PES fitting environment Asparagus[70] was employed. The dataset consisted of 4000 DCM-monomer structures and 38000 DCM-dimers sampled from the 200 distinct clusters resulting in a total of 42000 structures. The ML-PES was trained on reference energies, forces, dipoles and charges. All the reference calculations were carried out at the DLPNO-MP2/cc-pVTZ level of theory using ORCA.[68]

To smoothly combine these two regimes, a cutoff ($r_{cut} = 8$ Å) was employed. For $r_{cut} \leq 8$ Å, dimer energies were those from the ML-PES, whereas for $r_{\text{cut}} \geq 8$ Å the dimer interaction energy ($E_i^{\text{dimer,NB}}$) was evaluated using the CGenFF charges and LJ-parameters, see Eq. 9. The cutoff (r_{cut}) was sampled from the C-C distance in a DCM dimer at an interval of 1 Å from 3 Å to 18 Å and the choice of cutoff was based on the lowest RMSE against the reference cluster energy

$$E_i^{\text{dimer,inter}} = \begin{cases} E_i^{\text{dimer}} - \left(E_{a,i}^{\text{monomer}} + E_{b,i}^{\text{monomer}}\right), & r_i < r_{\text{cut}}, \\ E_i^{\text{dimer,NB}} & r_i \geq r_{\text{cut}} \ . \end{cases} \tag{9}$$

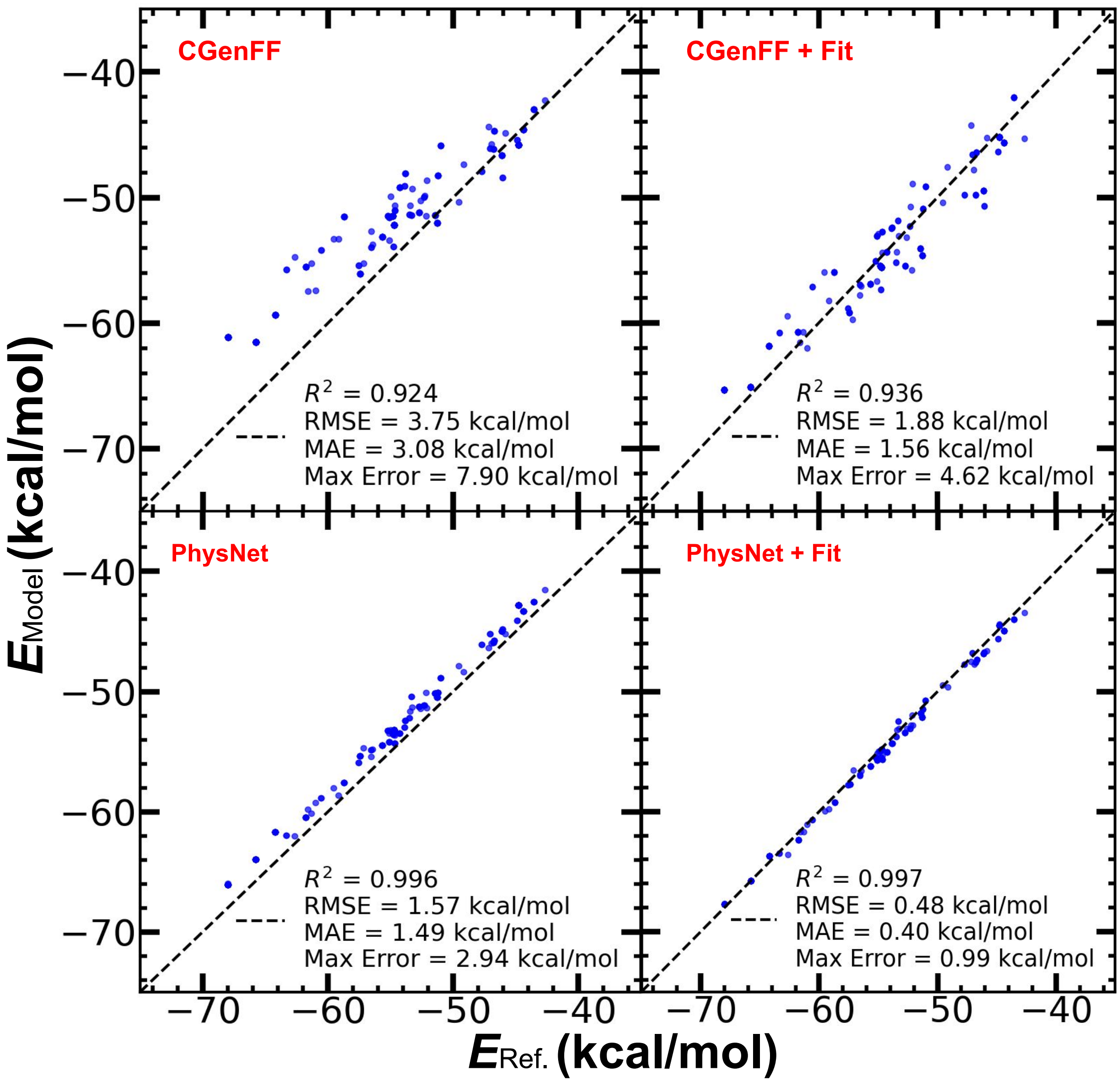

Figure 1: Correlation Plot of Model (cluster) energies vs reference energies (Cluster). The cluster ($n = 20$) geometries were extracted from an MD simulation run on CGenFF. The left panels have been marked by the respective energy functions used to get the energies of the clusters, and the Lennard-Jones potential was fitted for the corresponding right panels. The Model cluster energy is the non-bonded energy calculated by CHARMM, see panel "CGenFF". The PhysNet cluster energy is the sum of the dimer-pair energy in the corresponding cluster. The sum of dimer pairs is being correlated with the ORCA Formation Energy. The reference energies ($E_{\text{Ref.}}$) were obtained from ORCA calculations done at the DLPNO-MP2-cc-pVTZ level of theory. The ORCA Formation energy in the X-axis was obtained following Eq. 6.

The performance of the different energy functions for DCM considered here is reported in Figure 1. The conventional CGenFF energy function features RMSE, MAE and maximum error of [3.75, 3.08, 7.90] kcal/mol, respectively, with $R^2 = 0.924$ relative to reference data from DLPNO-MP2/cc-pVTZ calculations. For an empirical energy function this is rather good performance. However, improvements are possible by readjusting the LJ-parameters, see Figure 1 (CGenFF + Fit) which changes the statistical measures to [1.88, 1.56, 4.62] kcal/mol, and $R^2 = 0.936$. In other words, all errors are reduced by a factor of ~ 2. Using the ML-PES, which contains information about monomer deformation energies and 2-body intermolecular interactions between monomers, the performance measures are [1.57, 1.49, 2.94] kcal/mol, and $R^2 = 0.996$. Hence, from the perspective of errors between reference data and model, the ML-PES is already better than the readjusted empirical energy function. Most notably, the correlation coefficient is close to 1. However, the quality of the ML-PES can be further improved by readjusting the LJ-parameters, which is shown in Figure 1 (PhysNet + Fit). This decreases the errors to $[0.48, 0.40, 0.99]$ kcal/mol, and $R^2 = 0.997$, which is close to chemical accuracy.

The results for DCM demonstrate that ML-PESs and empirical energy functions can be combined in a consistent manner to arrive at high-accuracy representation of the total energy functions. Further possibilities to boost such models is to replace, for example, the CGenFF point charges by more elaborate representations of the electrostatic interactions, such as atom-centered multipoles or different flavours of distributed charge models.[43,44,71–73] Most importantly, such approaches can also be applied to larger monomers than DCM and multicomponent molecular systems.

3.2 Pure Water

Water is essential for life and involved in much of terrestrial, and interstellar, chemistry.[74–78] As a material, water in the condensed phase is also famous for its many anomalies: Despite its low molecular weight of 18 g/mol, water possesses a boiling point of 372 K and reaches a maximum density 4 degrees above its freezing temperature of 277 K, while exhibiting a high surface tension and high viscosity.[74–78] These anomalies arise in part due to the hydrogen-bonding capabilities between neighbouring water molecules in the condensed phase. For realistic computational modeling, the relevance of anisotropic intermolecular interactions (e.g. directionality of the H-bond) usually requires use of models beyond simple atom-centered point charges. Such models strive at describing higher-order multipolar interactions which can be accomplished in different ways.[79–82]

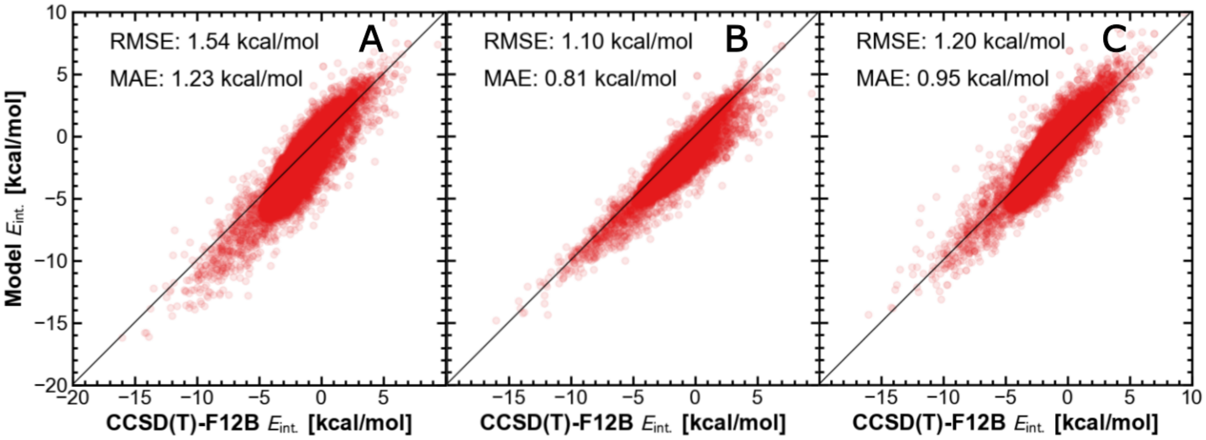

Figure 2: Performance of model energy functions versus interaction energies E_{int} calculated using the supermolecular approach at the CCSD(T)-F12B/dev-2zp level of theory for water dimers, trimers, and tetramers; using (A) TIP3P unoptimized, (B) MDCM with refit LJ parameters and (C) TIP3P with reoptimized LJ parameters.[83]

Water, as a paradigmatic "complex liquid", is an ideal system for developing and testing ML-based workflows. Accurately capturing water's phase properties is a formidable challenge. Research in water modeling for MD simulations generally follows two main directions: capturing molecular interactions with high precision by incorporating many-body interactions (usually up to four-body),[84,85] or alternatively, scalable models designed to handle large system sizes efficiently usually by fitting the model parameters to best reproduce experimentally determined quantities such as density, heat of vaporization and self-diffusion.[21] A complete many-body description offers an exceptionally accurate description of water but are often limited in their application to smaller system sizes (typically 256 monomers) due to their high computational cost. On the other hand, the mean-field, two-body empirical force field approximation, as is usually selected, usually suffer deficiencies in the PES that may become manifest in inconsistent dynamics (inaccurate rotational self-correlation life-times), and some thermodynamic properties discussed later. Within the present work, the performance of using small and medium-sized water clusters in developing a machine learning-based energy function is compared for a number of candidate solutions after optimizing the LJ-parameters, see Figure 2

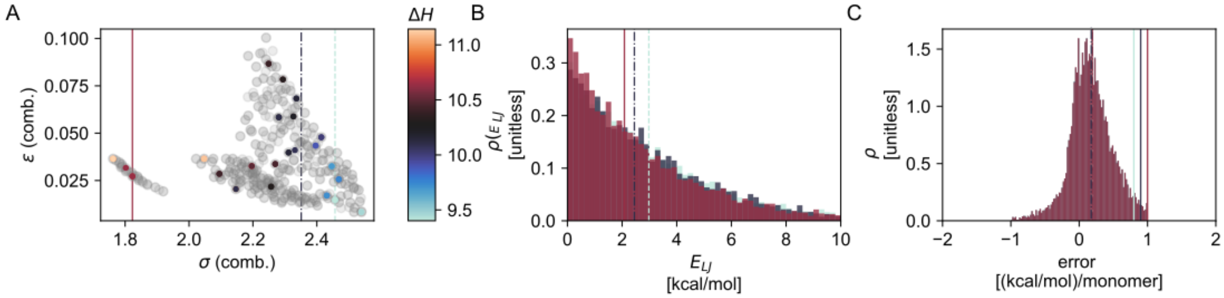

Figure 3: Panel A: The distribution of LJ parameters obtained from fitting to water cluster data (dimers, trimers and tetramers calculated at the CCSD(T)-F12B/dev-2zp level of theory). A selection of parameters were simulated in CHARMM and the resulting ΔH was obtained in units of kcal/mol. Three models were selected (red, black, and blue vertical lines) with comparable RMSEs $\sim$ 1.0 (kcal/mol)/monomer on the entire dataset. Panel B: Distribution of the LJ energy contributions for the three parameter sets. Dashed vertical lines report the mean of each distribution. Panel C: Error (per monomer) distributions for the three parameter sets. Solid vertical lines report the RMSE on the training distribution. The experimental value ΔH is 10.5 kcal/mol. As expected, the ΔH predicted by the model is largely influenced by the magnitude of the average of the LJ contribution.

Here, a generic cluster-based workflow based on a combination of machine learning-based and empirical representations of intra- and intermolecular interactions was used.[83] The total energy is decomposed into internal contributions, and electrostatic and van der Waals interactions between monomers. For the monomer potential energy surface a small neural network is combined with intermolecular interactions described by a flexible, minimally distributed charge model and van der Waals interactions. This differs from DCM fro which standard atom-centered CGenFF charges were used for the electrostatics. Remaining contributions between reference energies from electronic structure calculations and the model are fitted to standard Lennard-Jones (12-6) terms.

For water as a topical example, reference energies for the monomers are determined from CCSD(T)-F12 calculations whereas for an ensemble of cluster structures containing $[2, 60]$ and $[2, 4]$ monomers DFT and CCSD(T) energies, respectively, were used to best match the van der Waals contributions. Based on the bulk liquid density and heat of vaporization, the best-performing set of LJ(12-6) parameters was selected and a wide range of condensed phase properties were determined and compared with experiment. Figure 3A reports all fitted LJ-parameters in the $(\sigma, epsilon)-$plane and colored points provide computed heat of formation ΔH. It can, for example, be seen that larger atom radii (σ) yield lower ΔH (blue) whereas decreasing σ brings ΔH into better agreement with the measured value of 10.5 kcal/mol. Figures 3B and C report the distribution of LJ-interactions for all parameter sets shown in Figure 2 and the error distributions between reference calculations and fitted model for the three models.

3.3 Eutectic Mixtures

Deep eutectic mixtures (DEM) - also referred to as deep eutectic solvents (DESs) when used in practical applications - are multicomponent systems consisting of molecules acting as hydrogen bond acceptors and hydrogen bond donors at particular molar ratios.[86–88] One of the distinguishing features of DESs is that the melting point of the mixture is lower than that of the individual components, due to, for example, charge delocalization occurring through hydrogen bonding between anions and hydrogen donors.[89] Such mixtures can also contain ions which leads to pronounced crowding and strong electrostatic interactions, similar to ionic liquids,[90] and DESs are also of interest in the context of batteries and fuel cells due to the high cryostability, thermal stability, and their electrochemical stability.[91–93] The particular mixture considered here consists of water, acetamide (ACEM) and NaSCN which is present as solvated Na^+ and SCN^- (thiocyanate) ions. Acetamide forms low-temperature eutectics with a wide range of inorganic salts and the resulting non-aqueous solvents have a high ionicity. Such mixtures have also been recognized as excellent solvents and molten acetamide is known to dissolve inorganic and organic compounds.[94–97] The SCN^- anion is a suitable spectroscopic probe because the CN-stretch vibration absorbs in an otherwise empty region of the infrared spectrum. Recently, advantage has been taken of this to probe the effect of water addition to urea/choline chloride and in acetamide/water mixtures.[98,99]

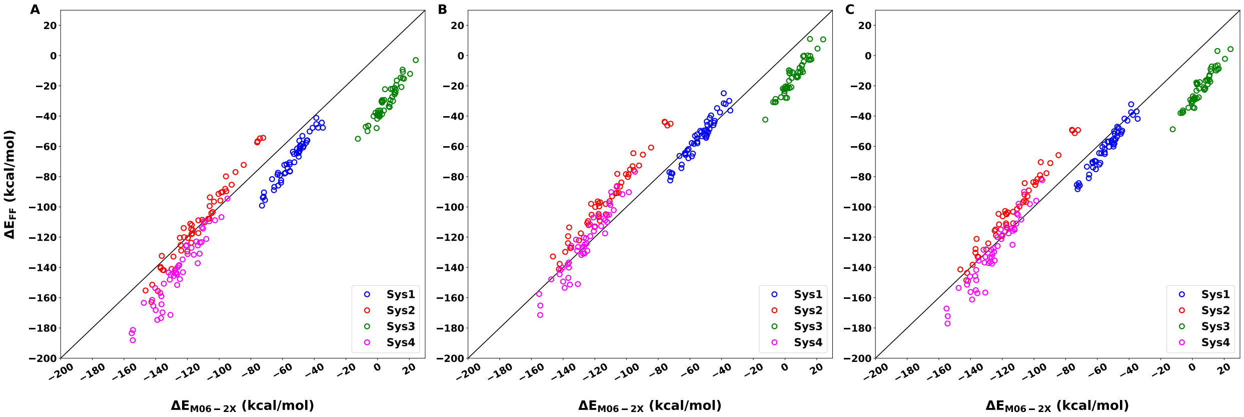

Figure 4: Correlation of interaction energies between reference DFT data and the empirical energy function for clusters extracted from simulations with [20/80] W/ACEM. Panel A: correlation before parameter optimization with the initial parameters;[99] panel B: parameters from individual optimization for the [20/80] mixture; panel C: using a transferable parameter set.

Previously, as an initial step towards the energy functions of mixed clusters, the dynamics of a deep eutectic mixture (KSCN/acetamide) was studied with different W/ACEM ratios.[99] To generate energy functions for such heterogeneous systems, a cluster-based approach which optimized the Lennard-Jones (LJ) parameters of SCN^- to the DFT calculated energetic reference was used.[100] The resulting optimal parameters enable better and more accurate predictions of viscosity and spectroscopic properties from MD simulations. In the following a cluster-based approach was applied to the deep eutectic mixture (NaSCN/acetamide) with different ratio of water contents.

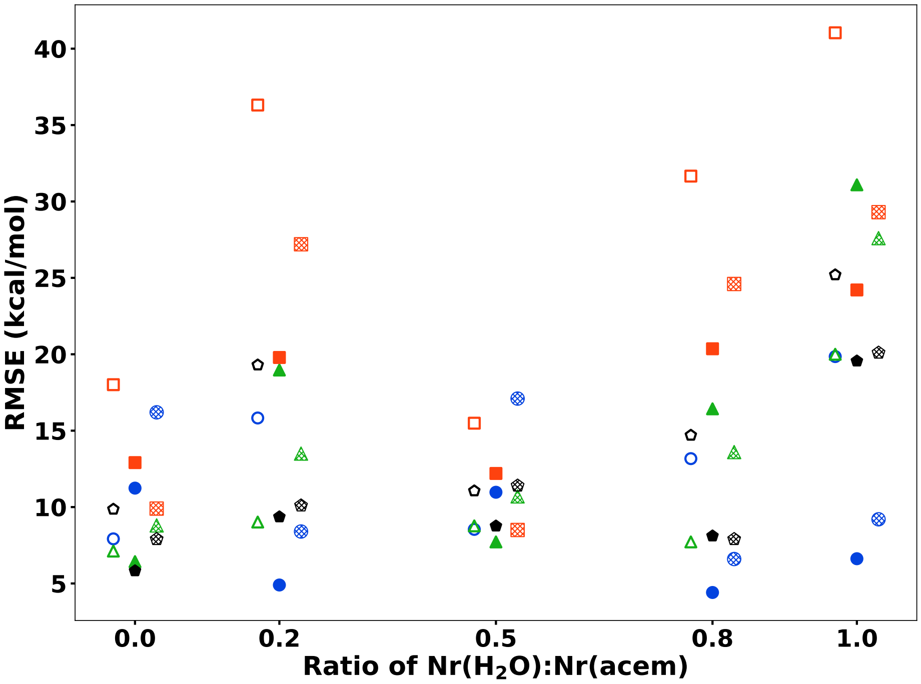

Figure 5: Summary of RMSEs between DFT calculations and final fitted outputs of different set of clusters from 5 mixtures with TIP3P water model. The hollow, filled, and hatched markers are for the initial, individually optimized, and transferable parameters. Blue, green, orange, and black symbols correspond to correspond to sys1 to sys4, respectively. Note that the labels Sys1 to Sys4 refer to specific system compositions depending on the W/ACEM ratio.

For the cluster-based optimization scheme first MD simulations of 75 Na^+ / 75 SCN^- in 5 different water / acetamide mixtures were carried out. The water / ACEM particle number ratios were 0.0, 0.2, 0.5, 0.8, and 1.0, see Table S1. In a next step, 50 clusters were extracted randomly from the MD simulations for each mixing ratio containing one SCN^-

surrounded by 4 differently organized environments (system1 to system4). The composition of these environments (clusters) for all 5 mixtures are reported in Tables S2 to S6. For each mixture, snapshots of the simulation were screened with respect to combinations of [Na$^+$, SCN$^-$, ACEM, water] with a cutoff range of 5 Å around the central SCN$^-$ and 50 cluster structures for Sys1 to Sys4 were extracted.

For each of the 200 clusters total interaction energies were calculated at the M062X/AVTZ level of theory using Gaussian16.[101] In a next step, the total interaction energy was determined from the mixed ML/MM energy function, where the ML potential describes monomer energies and the MM potential describes all nonbonded interactions, and the LJ-parameters ε and $r_{\rm min}$ of the three atoms SCN$^-$, were optimized using the truncated Newton (TNC) algorithm for the 200 clusters. The final LJ-parameters after fitting are reported in Table 1. Two different strategies were pursued. In the first, "individually optimized" LJ-parameters were determined for each of the 5 mixtures considered. This can be regarded as the maximum refinement level possible. However, such a parametrization scheme is not particularly useful if one wants to investigate mixtures of arbitrary W/ACEM combinations. For this, a more "transferable" set of parameters is more useful which yields acceptable accuracy for any amount of W and ACEM in a particular mixture. It should be further noted that a large number of possible solutions of the minimization problem exist for 6 free parameters (LJ-parameters for SCN$^-$) and 200 structures of various compositions.

It is first noted that for the individually optimized LJ-parameters the parameters for the carbon atom vary widely. This can be explained by the fact that in SCN$^-$ the central C-atom is effectively shielded from the environment due to the larger S- and N-atom. Hence, the interaction energies are not particularly sensitive to the LJ-parameters of the C-atom. Consequently, for the transferable parameter set the initial parameter values were retained. Based on the individually optimized LJ-parameters, a consensus was sought for the S- and

Table 1: The LJ parameters of SCN^- before and after fitting for the different [W/ACEM] mixtures. Two types of optimizations were considered. "Individually optimized" refers to optimization of the LJ-parameters of 200 clusters (set1 to set4) for a given [W/ACEM] mixture whereas "Transferable" refers to parameters that were initially obtained from fitting selected clusters for the [20/80], [50/50], and [80/20] mixtures with slight manual readjustments.

LJ params	Initial	Individually optimized (W/A)					Transferable
		[0/100]	[20/80]	[50/50]	[80/20]	[100/0]	
ϵ(S)	-0.364	-0.427	-0.0115	-0.568	-0.0245	-0.392	-0.250
$\tilde{r}_{\min}$(S)	2.18	2.30	2.88	2.21	2.79	2.32	2.40
ϵ(N)	-0.0741	-0.0351	-0.0149	-0.0955	-0.0306	-0.0728	-0.0100
$\tilde{r}_{\min}$(N)	2.01	2.08	2.35	2.06	2.03	2.27	2.35
ϵ(C)	-0.102	-0.200	-0.00165	-0.183	-1×10^{-4}	-1×10^{-4}	-0.102
$\tilde{r}_{\min}$(C)	1.79	1.50	2.08	1.50	1.94	1.69	1.79

N-parameters. Their performance, compared with the initial and individually optimized parameters is summarized in Table 2. Overall, the average error of the individually optimized LJ-parameters improves by 4 kcal/mol over the initial parameters whereas for the transferable LJ-parameters the improvement is 2.6 kcal/mol. This is still acceptable and all errors are heavily influenced by the quality of the TIP3P water model (see performance for [100/0]). Hence, it is expected that replacing this simple water model with an improved description will considerably boost performance, see quality for [0/100]. The performance of the fitting is a compromise over all cluster compositions and relative orientations. Hence, RMSE-values for different solvent compositions can depend on the system (Sys1 to Sys4) considered, see Figure 5.

To quantify the effect of different parameter sets on physical observables, changes in the Na^+–SCN^- pair distribution functions before and after reparametrization are reported in Figure 6. These radial distribution functions $g(r)$ were determined from simulations of the [80/20] mixture of different length, also to monitor convergence. Due to the large number of Na^+ / SCN^- pairs present, 2 ns simulations were deemed sufficient in all cases. Dashed, solid and dotted traces in Figure 6 refer to simulations using the initial, individually opti-

Table 2: Average RMSE for Table 1. Fitting individual mixtures reduces all RMSE whereas for the transferable parameters the RMSE is typically lower than for the initial parameters except for the [50/50] mixture. Note that for different [W/ACEM] mixtures sys1 to sys4 contain different numbers of W and ACEM molecules, see Tables S2 to S6.

Mixture (W/A)		Initial	Individual	Transferable
[0/100]		10.7	9.1	10.7
	sys1	7.9	11.2	16.2
	sys2	7.1	6.4	8.8
	sys3	18.0	12.9	9.9
	sys4	9.9	5.8	7.9
[20/80]		20.1	13.3	14.8
	sys1	15.8	4.9	8.4
	sys2	9.0	19.0	13.5
	sys3	36.3	19.8	27.2
	sys4	19.3	9.4	10.1
[50/50]		11.0	9.9	11.9
	sys1	8.5	11.0	17.1
	sys2	8.8	7.7	10.7
	sys3	15.5	12.2	8.5
	sys4	11.1	8.8	11.4
[80/20]		16.8	12.3	13.2
	sys1	13.2	4.4	6.6
	sys2	7.7	16.4	13.6
	sys3	31.7	20.4	24.6
	sys4	14.7	8.1	7.9
[100/0]		26.5	20.4	21.6
	sys1	19.9	6.6	9.2
	sys2	20.0	31.1	27.6
	sys3	41.0	24.2	29.3
	sys4	25.2	19.6	20.1
Overall		17.0	13.0	14.4

mized, and transferable LJ-parameters. The red, blue and green traces are for separations involving the S-, C-, and N-atoms of the SCN$^-$ anion. As Figure 6A shows, the maximum peak positions for the individually optimized parameters shorten for the Na–N$_{\rm SCN}$ and Na–C$_{\rm SCN}$ separations and increase for Na–S$_{\rm SCN}$. Using the transferable parameters, the $g(r)$ for the Na–S$_{\rm SCN}$ separation is close to that for the individually optimized set whereas $g(r)$ for Na–N$_{\rm SCN}$ differs little for the initial data set.

The peak height of $g(r)$ is a measure of the interaction strength between the atoms involved. This indicates that for the individually optimized parameters the Na–N_{SCN} is considerably stronger than for the two other parameter sets whereas for Na–C_{SCN} this differs little and for Na–S_{SCN} individually optimized and transferable parameters perform comparably whereas the initial parameter set features increased interaction strength.

For the interaction between water-oxygen atoms and each of the constituent atoms of the anion, see Figure 6B, differences between parameter sets also occur but are less pronounced overall. In general, performance of the inidividually optimized and transferable parameter sets is comparable, whereas the initial data set features shorter bond lengths and increased interaction strength. This is particularly seen for the O_W–N_{SCN} separation.

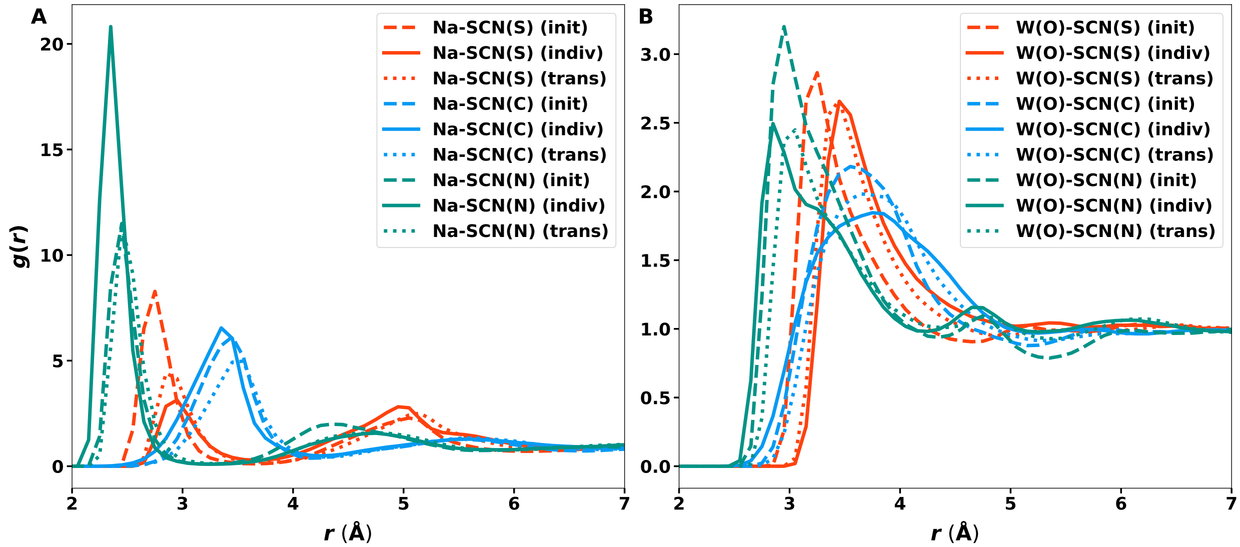

Figure 6: Comparison of the radial pair distribution function $g(r)$ between OW–X (X = S, C, N) with original (dashed lines) Lennard-Jones parameters, optimized (solid lines) Lennard-Jones parameters, and optimal (dotted) Lennard-Jones parameters from the 80/20 W/ACEM mixture. Comparison of the radial pair distribution function $g(r)$ between Na^+–X (X = S, C, N) with original (dashed lines) Lennard-Jones parameters, optimized (solid lines) Lennard-Jones parameters, and optimal (dotted) Lennard-Jones parameters from the 80/20 W/ACEM mixture.

As required, individually optimized LJ-parameters for the SCN^- probe molecule for each mixing ratio outperform the initial and transferable parametrizations. Nevertheless, the transferable data set provides a meaningful description of the intermolecular interactions. It should be noted that merely minimizing the loss function does not necessarily yield parameters within expected ranges, e.g. $r_{\mathrm{min}}/2$ for the sulfur atom increases up to 2.88 Å which is unusually large compared with standard values in CGenFF.[17] Such effects are due to both, the mathematical description of the van der Waals interactions and the type and composition of the reference data set. Further improvement of the models will require a better description of the water-water interactions, and possibly a different mathematical description of the van der Waals interactions, while transferability may be improved by adding higher order terms such as polarization that respond to changes in highly polar chemical environments such as eutectic mixtures.

3.4 CO on Amorphous Solid Water

Surface processes are primary for the genesis of molecules in the interstellar medium. In cold molecular clouds, dust behaves as a suitable substrate for the deposition and chemical synthesis of molecules in a bottom-up manner. The dominant form of interstellar ice is amorphous solid water (ASW),[102,103] whose intrinsically disordered hydrogen-bond network gives rise to a broad variety of adsorption sites and binding environments.[104] This structural complexity makes ASW a challenging substrate to model, yet it also places it at the center of astrochemistry, as it governs how molecules adhere, diffuse, and react on grain surfaces.[105] Among the species of astrochemical interest, carbon monoxide (CO) is of particular interest, serving both as the main carbon reservoir[106] and as the second most abundant molecule in molecular clouds.

To capture these microscopic processes, simulations must describe the interaction between adsorbed molecules and the ice surface with very high accuracy. This requirement is especially stringent at the low temperatures of molecular clouds ($10 - 50$ K), where even small errors in adsorption energies can lead to large deviations in desorption rates, diffusion barriers, and ultimately, reaction kinetics.[105] Moreover, proper modeling must also account for interaction within the ice H_2O molecules: during molecular formation, excess reaction energy couples back into the water network for energy dissipation, altering its local structure and influencing subsequent reactivity. Thus, an accurate description must encompass both intermolecular interactions between adsorbates and ASW, as well as intramolecular interactions within the ice matrix.[107]

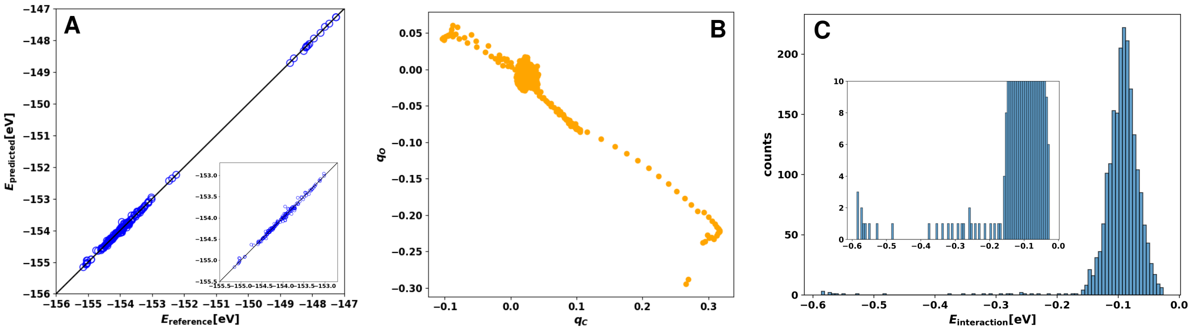

Figure 7: Panel A: Correlation between predicted and reference energies for a test set of 195 CO/water clusters, with an RMSE of 0.0399 eV and an MAE of 0.0279 eV. The low-energy structures are from MD simulations at 50 K using the KKY water model,[108,109] while the high-energy structures were generated using xTB[110] to add higher-energy configurations. Panel B: Atomic charges of the C and O atoms in CO across all 2007 clusters on ASW calculated using NN model. Panel C: Distribution of interaction energies for all 2007 clusters at M062X/aug-cc-pVTZ + D3 level.

Traditionally, MD simulations have employed hybrid approaches combining QM (or ML) with MM with either mechanical or electrostatic embedding.[108,111–113] While these hybrid approaches reduce computational cost, they rely on empirical potentials that sacrifice accuracy. This presents challenges in accurately capturing the diverse, heterogeneous adsorption environments on the ASW and the energy redistribution effects (i.e., coupling between internal degrees of freedom of the adsorbate and the adsorbent) essential for understanding surface chemistry. In contrast, fully data-driven pure ML potentials trained directly on *ab* *initio* data offer a more accurate and internally consistent framework which is done for CO on ASW. In other words, the total interaction energy of the clusters in question is represented using a NN-PES.

To determine the appropriate cluster size, the interaction energy (E_{int}) between CO and water clusters containing 9 to 18 water molecules was computed at the M06-2X/aug-cc-pVTZ + D3[114,115] level using ORCA.[68,69] The results in Figure S8 show that E_{int} varies noticeably with cluster size due to the influence of cavity shape and local water arrangement. Balancing accuracy and computational cost, 14 water molecules were selected for dataset generation. Cluster structures were extracted from long MD simulations of CO diffusion on the water surface (from earlier work[108]) to capture the diverse arrangements and cavities present on ASW. Energies, forces, and dipole moments were calculated for 2007 clusters, and the dataset was trained using Asparagus[70] with an 80/10/10 split for training, validation, and test sets. The correlation between the predicted and reference energies for the test set is shown in Figure 7A. Panel 7B shows the atomic charges of the C and O atoms in CO, and panel 7C the interaction energy range with amorphous solid water, both evaluated across 2007 clusters. Strong interactions occur mainly in cavities, which explains the low probability of highly negative interaction energies in panel 7C.

3.5 Menshutkin Reaction

The Menshutkin reaction is a key S_N2 reaction in organic and bioorganic chemistry and has been widely studied since the original paper back in 1890.[116,117] In this reaction, neutral reactants form ionic products carrying opposite charges. It is known that the reaction behaves differently in solution compared to the gas phase; it proceeds much faster in polar solvents than in less polar ones.[118] While it is clear that solvent effects strongly influence the reaction energetics, much less is known about the molecular details of the process, particularly about the solvent dynamics and reorganization along the reaction path.[117,119,120]

In this part, the reaction between NH_3 and CH_3Cl was investigated in the presence of $n = 0, 1, 2,$ and 5 explicit water molecules using a machine-learning-based potential energy surface (ML-PES). Similar to CO on ASW the total interaction energy was represented using PhysNet. To build the training dataset, three types of structures were generated: (i) normal-mode samples, (ii) metadynamics snapshots, and (iii) local minima obtained via the GOAT algorithm.[121]

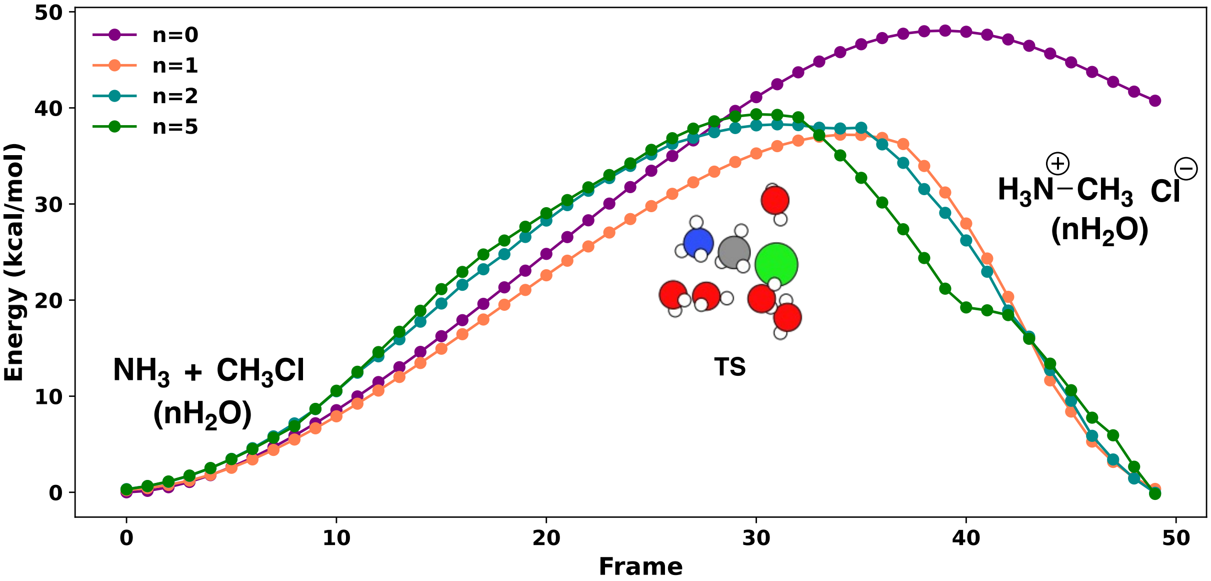

Figure 8: One-dimensional energy profile along the N-C and C-Cl bond for Menshutkin reaction with different number of water molecules. Predominantly, the first half of the reaction (frames 1 to 25) involves C–Cl bond elongation whereas C–N bond formation occurs during the second half. However, the two coordinates are coupled throughout the chemical transformation as is known for S_N2 reactions. The energy function is the ML-PES represented using PhysNet.

The reaction paths for systems with varying numbers of water molecules were first determined using the NEB-TS algorithm, as implemented in the ORCA software package, at the B3LYP/ma-def2-TZVP level of theory.[68,122,123] All generated structures were then used for normal-mode sampling at 100, 300, 500, and 1000 K at the PBE/def2-SVP level of theory using the Asparagus software.[70] In addition, metadynamics simulations were carried out with GFN2-xTB across the same temperature range.[124]

Because solvent organization plays a key role in understanding reaction mechanisms and the role of solvation, the GOAT algorithm was further applied to identify local minima corresponding to different cluster organizations. Altogether, this procedure yielded 33,300 structures, which were subsequently used to determine energies, forces and dipole moments at the RI-MP2/cc-pVTZ+cc-pVTZ/C level of theory using ORCA.[125] The resulting dataset was then employed to train a machine-learning model within the Asparagus framework.

These steps yielded a stable ML model capable of simulating systems with varying numbers of surrounding water molecules. The values of MAE are 0.22 kcal/mol for energies and 0.35 kcal/mol·Å for forces, while corresponding RMSE values are 0.74 kcal/mol and 1.23 kcal/mol·Å, respectively, indicating the high quality of the trained model. Using this model, a one-dimensional scan was carried out along the reaction coordinate defined by the N–C and C–Cl separations. The scan started at an N—C distance of 3.2 Å and a C—Cl distance of 1.8 Å, and the trajectory was divided into 50 frames up to final distances of 1.48 Å (N–C) and 3.0 Å (C–Cl). During this process, the N, C, and Cl atoms were fixed, while the remainder of the system was allowed to relax. All optimizations were performed using the ML-PES with the ASE framework.[126]

The paths including one or several water molecules provide an impression for the effect of solvation for a gas-phase–like reaction, see Figure 8. The results clearly show that in the absence of coordinating solvent water the charged product is significantly destabilized. However, addition of a single water molecule already increases the overall stability of the product by ~ 10 kcal/mol. Notably it is found that the barrier height does not strongly depend on the number of water molecules included. It should be noted, however, that a 1-dimensional

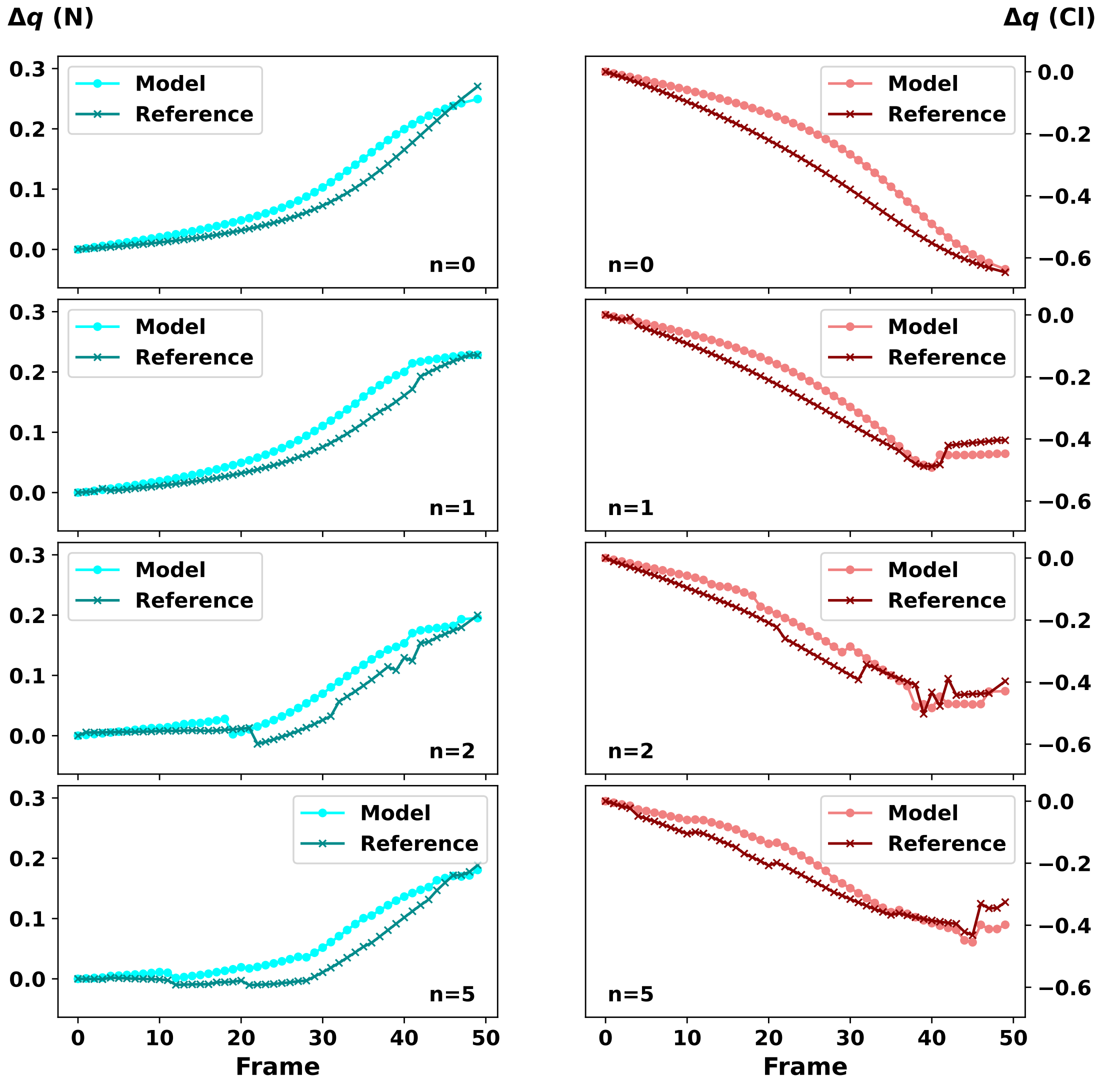

Figure 9: Comparison between model predicted and Hirshfeld charge evolution on the N atom (left panel) and on the Cl atom (right panel) along one-dimensional scan for the Menshutkin reaction with different number of water molecules.

picture is does not provide a complete description of this reaction.

Because PhysNet provides (fluctuating) atomic charges it is also of interest to monitor changes on the N- and Cl-atoms along the 1-dimensional reaction, providing insight into charge redistribution along this reaction path, see Figure 9. Setting the reactant state as the reference, close agreement between the charge evolution along the reaction pathway, as predicted by the ML model, and the Hirshfeld charges[127] calculated for each snapshot at the RI-MP2/cc-pVTZ+cc-pVTZ/C level of theory is found. The results also show that fluctuating charges are required to realistically describe such a reaction both, in the gas phase and in solution. The change in the partial charges on the N- and Cl-atoms depends on the degree of hydration. Again, the first water molecule has the largest influence. It reduces the charge on the N- and Cl-atoms between reactant and product state by 30 %. Addition of one or 4 water molecules does not change the charge in the product state but influences the curvature of the curve on the reactant side. The observed decrease in the N- and Cl-charges can be attributed to the presence of water molecules in the outer coordination sphere of the reactants and products. As illustrated in the figure, this effect becomes more apparent with an increasing number of water molecules. These molecules act as a compensating network, thus facilitating charge redistribution by providing additional contributors to the overall charge distribution. For the nitrogen atom, the charge remains nearly constant up to the transition state. A noticeable increase in charge is observed only after the transition state, indicating the role of the solvent in facilitating charge compensation. This is consistent with earlier work on the Menshutkin reaction.[117]

3.6 Spectroscopic Probes

Spectroscopic probes are small molecules that can be used to label proteins or ligands for characterization of the energetics and dynamics in condensed phase environments. Specific examples include cyanide (–CN), azide (–N_3), or nitric oxide (–NO). Importantly, these small molecules exhibit infrared signatures that set them clearly apart from the vibrational modes of the guest molecule. As an example, protein vibrational spectra extend up to $\sim$ 1800 cm^{-1}, followed by a largely "empty" region up to 2800 cm^{-1} above which the X-H stretch vibrations are located (X = C, N, O). Hence, by modifying residues such as alanine using such a label, its spectroscopy and vibrational dynamics can be followed with great precision. For example, IR spectroscopy can distinguish stretching frequencies associated with CN^- or N_3^-,[128–130] which reveal differences in bond order and electron delocalization. By employing these spectroscopic probes, the characteristic properties of these species can be unraveled to understand their roles in various fields of chemistry.

In this last example the behavior of -SCN, -SNO, and -N_3 probes in different water cluster environments was investigated. To establish a consistent framework, methyl-substituted reference systems (Me-X) were constructed, with CH_3 group Lennard-Jones (LJ) parameters obtained from the CHARMM[30] force field via CGenFF,[17] while the probe atoms were explicitly parapmetrized. For each system, a dedicated fMDCM charge model[43] was generated. The total charge of each system was set to zero. The resulting electrostatics were employed to refine the Lennard-Jones (LJ) parameters, thereby ensuring a balanced description of nonbonded interactions.

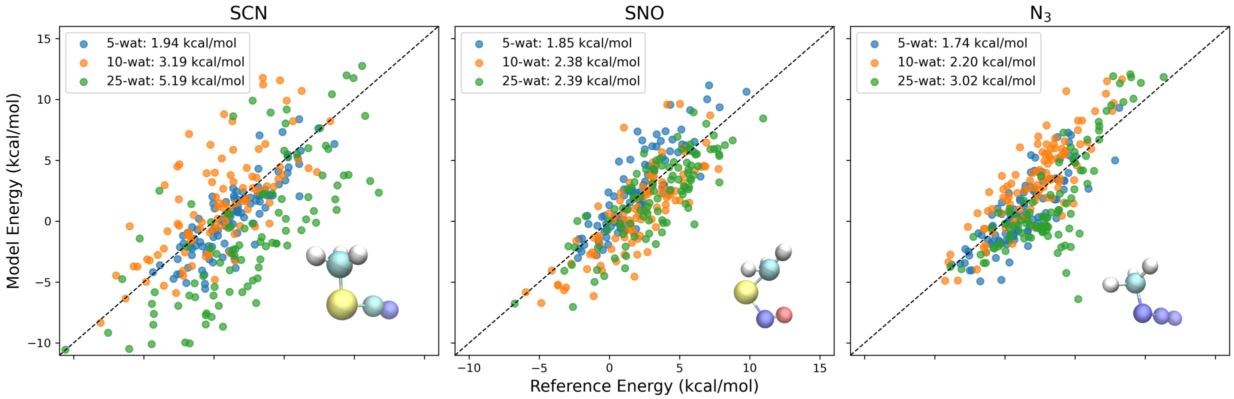

Figure 10: Correlation between reference interaction energies from B3LYP/aug-cc-pVDZ calculations with the model interaction energies for CH_3-X clusters using the fMDCM electrostatic model. Blue, orange, and green colors represent the clusters that contain 5, 10, and 25 waters, respectively.

To account for solvation effects, probe-water clusters of varying sizes, containing 5, 10, and 25 water molecules, were examined. From molecular dynamics simulations, 100 configurations were randomly sampled for each cluster, with the exception of the CH_3-SCN 10- and 25-water cluster (99 configurations), the CH_3-SNO with 5- and 10-water clusters (98 and 97 configurations, respectively), and the CH_3-N_3 25-water cluster (96 configurations). Total interaction energies for all configurations were subsequently computed using Gaussian[101] at the B3LYP/aug-cc-pVDZ level of theory. This was the reference data for fitting the LJ parameters within the CHARMM framework in conjunction with the fMDCM model for each spectroscopic probe. For the fitting, curve fit in SciPy[131] was used, which minimizes the sum of squared residuals between the model and the reference data using the non-linear least-squares solver. CHARMM nonbonded interaction energies were compared to quantum chemical cluster energies with monomer energies removed, similar to DCM and water nonbonded energies described above.

The fitted LJ parameters, ϵ and $r_{min}/2$ for each water-cluster-specific probe are summarized in Tables S7 to S9. The tables also report the atomic charges obtained from the fMDCM model (up to 4 charges per atom). The correlation between the reference interaction energies and those from the fitted models are shown in Figure 10. In all cases the RMSE decreases considerably compared with the initial parameters. For -SNO, -N_3, and -SCN surrounded by 5 water molecules the RMSE ranges from 4.3 to 5.5 kcal/mol which decreases to 1.8 to 1.9 kcal/mol after readjusting the LJ-parameters. Increasing the size of the solvent shell to 10 and 25 water molecules, the errors using the initial LJ-parameters increase to $\sim$ 8 kcal/mol and $\sim$ 13 kcal/mol, see Table S10. Hence, the error scales with the number of water molecules which is indicative of a considerable gain that can be expected from using improved water models in the future. After fitting the LJ-parameters, the RMSE-values range from 2.2 to 3.2 kcal/mol (10 water molecules) and 2.4 to 5.2 kcal/mol (25 water molecules), respectively. Hence, the LJ-fitted models improve by a factor of up to $\sim$ 5 in terms of reproducing the reference calculations.

4 Conclusion

This manuscript uses pure and mixed molecular clusters for improving partial or full ML-based energy functions. For this, finite-sized clusters are extracted from condensed-phase simulations. In a next step, reference data at the highest affordable levels of quantum chemistry are determined from which interaction energies are obtained. This data set constitutes the reference data to optimize in particular Lennard-Jones parameters. This is a meaningful approach because for internal degrees of freedom (bonds, valence angles, dihedrals - in the language of empirical energy functions) highly accurate ML-PESs can be obtained from either kernel- or NN-based approaches. Both have been used in the present work. For the electrostatics on the other hand, a range of methods to best describe the electrostatic potential are available. Here, the minimal distributed charge models either without or with conformational adjustments are employed. From the perspective of a non-polarizable empirical energy function the only remaining contribution is then the van der Waals term.

Within the broader perspective to generate next-generation energy functions for condensed phase simulations, the role of experiments also needs to be discussed. The present work clearly shows that representing reference data from electronic structure calculations and (re)adjusting certain key contributions to the total energy (here the LJ-parameters were improved) can provide qualitatively and quantitatively improved energy functions. On the other hand, given the large amount of available data that can be generated from *ab initio* calculations invariably leads to overdetermined fitting problems with a multitude of competitive solutions. Constraining this target space of equally likely solutions can be accomplished through calculation of *experimental observables* and comparing with measurements. This was, for example, done recently for water.[83] Earlier efforts, based on more empirical expressions for the total energy, yielded high-quality models for water.[132] Similarly, for the infrared spectroscopy of trialanine in water, a Bayesian reweighting approach based on the measured IR spectrum yielded an improved conformational ensemble in dihedral angle space

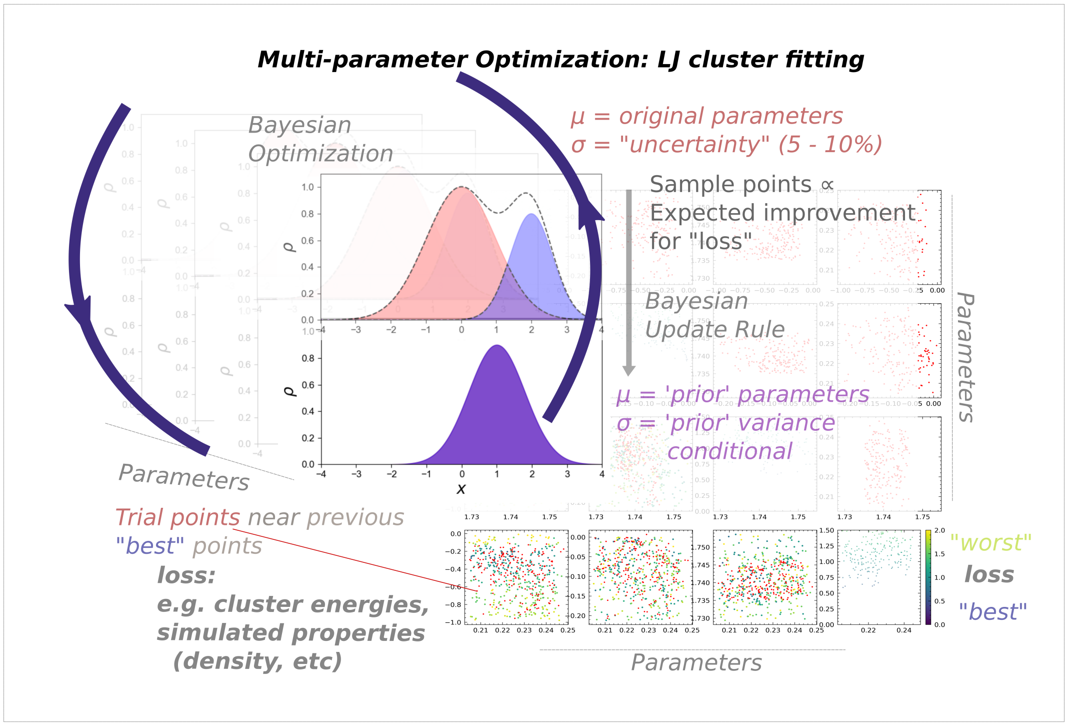
Multi-parameter Optimization: LJ cluster fitting
Bayesian Optimization
μ = original parameters
σ = "uncertainty" (5 - 10%)
Sample points ∝ Expected improvement for "loss"
Bayesian Update Rule
μ = 'prior' parameters
σ = 'prior' variance conditional
Parameters
Parameters
Parameters
Trial points near previous "best" points
loss:
e.g. cluster energies, simulated properties (density, etc)
"worst"
loss
"best"
ρ
x

Figure 11: An uncertainty-aware, multi-parametrization optimization strategy for obtaining optimized LJ-parameters using cluster formation energies and simulated properties. Bayesian optimization marginalizes previous knowledge or 'priors' with new information, for example from MD simulations, to make informed decisions on new areas of parameter space to explore. Parameter searches can prioritize exploration (sample points based on variance) or exploitation (sampling based on expected) depending on the updating rule for selecting trial points.

(Ramachandran plot).[133] Interesingly, this ensemble was consistent with subsequent simulations using improved empirical energy functions which also correctly described the IR spectrum.[134] Yet an alternative approach is to *morph* entire PESs to improve agreement between computed and measured observables which has, however, only been done for gas phase systems so far.[135,136] There have also been efforts to go beyond Bayesian reweighting to improve empirical energy functions for specific systems.[137]

As has been demonstrated here, there are clear limitations due to the functional form of the PES, e.g. parametric dependence of van der Waals interactions, that impacts the performance of total energy functions derived from cluster data. other challenges such as overfitting are equally important from a technical stand-point.[138,139] Statistical approaches such as bootstrapping are well known in the community. Related Bayesian interpretations of convergence, e.g. of the loss function, are often extremely helpful in this high dimensional multi-objective optimizations; in fact any least squares optimization can be restated as imposing a prior distribution of parameter probabilities (i.e. Bayesian) (Figure 11). The prior widths, e.g. restraining LJs parameters to be within a certain percentage of literature values, reflect the expected variations of the parameters during the optimization, and, in an empirical Bayes approach, can be obtained through resampling (and updating the prior beliefs), implemented in such efforts as Force Balance.[138,139] A Bayesian approach helps combine cluster based energy models with different modalities of training data (such as some desired simulated properties).

As the examples discussed in the present work indicate, a cluster-based approach yields models for improved total energies which eventually can be used in molecular simulation (here demonstrated for eutectic liquids). The work highlights that improvements in either parametrized expressions for describing van der Waals interaction or resorting to ML-based approaches for this contribution will further boost models for intermolecular interactions in “simple” and “multi-component” molecular systems, in particular when viewed in the context and together with experimentally measured properties amenable to molecular simulation.

Supplementary Material

Data Availability

The codes and data for the present study are available from `https://github.com/MMunibas/cluster` upon publication.

Acknowledgment

The authors gratefully acknowledge financial support from the Swiss National Science Foundation through grants 200020_219779 (MM), 200021_215088 (MM), and the University of Basel (MM). This article is also based upon work within COST Action COSY CA21101, supported by COST (European Cooperation in Science and Technology) (to MM).

References

(1) Jr., A. C.; Keesee, R. Clusters: Bridging the Gas and Condensed Phases. *Acc. Chem. Res.* **1986**, *19*, 413–419.

(2) Jr., A. C.; Keesee, R. Gas-Phase Clusters: Spanning the States of Matter. *Science* **1988**, *241*, 36–42.

(3) Cotton, F. A.; Haas, T. E. A Molecular Orbital Treatment of the Bonding in Certain Metal Atom Clusters. *Inorg. Chem.* **1964**, *3*, 10–17.

(4) Shin, J.-W.; Hammer, N. I.; Diken, E. G.; Johnson, M. A.; Walters, R. S.; Jaeger, T. D.; Duncan, M. A.; Christie, R. A.; Jordan, K. D. Infrared Signature of Structures Associated with the $H^+(H_2O)_n$ (n = 6 to 27) Clusters. *Science* **2004**, *304*, 1137–1140.

(5) Headrick, J. M.; Diken, E. G.; Walters, R. S.; Hammer, N. I.; Christie, R. A.; Cui, J.; Myshakin, E. M.; Duncan, M. A.; Johnson, M. A.; Jordan, K. D. Spectral Signatures of Hydrated Proton Vibrations in Water Clusters. *Science* **2005**, *308*, 1765–1769.

(6) Wales, D. J. Structure, Dynamics, and Thermodynamics of Clusters: Tales from Topographic Potential Surfaces. *Science* **1996**, *271*, 925–929.

(7) Hill, T. L. *Thermodynamics of Small Systems, Part I*; Frontiers in Chemistry; W. A. Benjamin, Inc., 1963; Reprinted by Dover Publications in 2014 as Parts I & II.

(8) Wales, D. J. Exploring Energy Landscapes. *Annu. Rev. Phys. Chem.* **2018**, *69*, 401–425.

(9) Becker, O. M.; Karplus, M. The Topology of Multidimensional Potential Energy Surfaces: Theory and Application to Peptides. *J. Chem. Phys.* **1997**, *106*, 1495–1517.

(10) Meuwly, M.; Doll, J. D. Dynamical studies of mixed rare-gas clusters: Collision-induced absorption in $(\mathrm{Ne})_n$–$(\mathrm{Ar})_m$ ($n + m \leq 100$). *Phys. Rev. A* **2002**, *66*, 023202.

(11) Meuwly, M.; Doll, J. D. Finite-temperature quantum simulations of mixed rare gas clusters. *J. Chem. Phys.* **2010**, *132*, 044305.

(12) Požar, M.; Buršíc, L.; Smith, A.-S.; Vranesic, H.; Primorac, D.; Zlatic, D. Micro-heterogeneity versus clustering in binary mixtures of ethanol with water or alkanes. *Phys. Chem. Chem. Phys.* **2016**, *18*, 30148–30158.

(13) Dorsey, J. G.; Dill, K. A. The molecular mechanism of retention in reversed-phase liquid chromatography. *Chem. Rev.* **1989**, *89*, 331–346.

(14) Braun, J.; Fouqueau, A.; Bemish, R. J.; Meuwly, M. Solvent structures of mixed water/acetonitrile mixtures at chromatographic interfaces from computer simulations. *Phys. Chem. Chem. Phys.* **2008**, *10*, 4765.

(15) Hage, K. E.; Gupta, P. K.; Bemish, R. J.; Meuwly, M. Molecular Mechanisms Underlying Solute Retention at Heterogeneous Interfaces. *J. Phys. Chem. Lett.* **2017**, *8*, 4600–4607.

(16) Hage, K. E.; Bemish, R. J.; Meuwly, M. From in silica to in silico: retention thermodynamics at solid–liquid interfaces. *Phys. Chem. Chem. Phys.* **2018**, *20*, 18610–18622.

(17) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I., et al. 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.

(18) Polêto, M. D.; Lemkul, J. A. Integration of Experimental Data and Use of Automated Fitting Methods in Developing Protein Force Fields. *Commun. Chem.* **2022**, *5*, 38.

(19) Devereux, M.; Boittier, E. D.; Meuwly, M. Systematic improvement of empirical energy functions in the era of machine learning. *J. Comput. Chem* **2024**, *45*, 1899–1913.

(20) Chaton, K. L.; Meuwly, M. Enhancing empirical energy functions using physics- and machine learning-based extensions. *J. Comput. Chem.* **2025**,

(21) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.

(22) Poater, J.; Illas, F.; Solà, M. The Atomic Partial Charges Arboretum: Trying to See the Forest for the Trees. *Chem. Phys. Chem.* **2020**, *21*, 688–696.

(23) Jelsch, C.; Teeter, M. M.; Lamzin, V.; Pichon-Lesme, V.; Blessing, B.; Lecomte, C. Accurate protein crystallography at ultra-high resolution: valence electron distribution in crambin. *Proc. Nat. Acad. Sci. USA* **2000**, *97*, 3171–3176.

(24) Jones, J. E. On the Determination of Molecular Fields. I. From the Variation of the Viscosity of a Gas with Temperature. *Proc. R. Soc. Lond. A* **1924**, *106*, 441–462.

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

(26) Unke, O. T.; Meuwly, M. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *J. Chem. Theor. Comp.* **2019**, *15*, 3678–3693.

(27) Brooks, B. R. et al. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem* **2009**, *30*, 1545–1614.

(28) Unke, O. T.; Meuwly, M. Toolkit for the construction of reproducing kernel-based representations of data: Application to multidimensional potential energy surfaces. *J. Chem. Inf. Model.* **2017**, *57*, 1923–1931.

(29) Song, K.; Käser, S.; Töpfer, K.; Vazquez-Salazar, L. I.; Meuwly, M. PhysNet meets CHARMM: A framework for routine machine learning/molecular mechanics simulations. *J. Chem. Phys.* **2023**, *159*.

(30) Hwang, W.; Austin, S. L.; Blondel, A.; Boittier, E. D.; Boresch, S.; Buck, M.; Buckner, J.; Caflisch, A.; Chang, H.-T.; Cheng, X., et al. CHARMM at 45: Enhancements in Accessibility, Functionality, and Speed. *J. Phys. Chem. B* **2024**, *128*, 9976–10042.

(31) Hollebeek, T.; Ho, T.-S.; Rabitz, H. Constructing multidimensional molecular potential energy surfaces from ab initio data. *Annu. Rev. Phys. Chem.* **1999**, *50*, 537–570.

(32) Schölkopf, B.; Herbrich, R.; Smola, A. J. A Generalized Representer Theorem. International Conference on Computational Learning Theory. 2001; pp 416–426.

(33) Golub, G. H.; Van Loan, C. F. *Matrix Computations*; JHU Press Baltimore, 2012; Vol. 3.

(34) Soldán, P.; Hutson, J. M. On the long-range and short-range behavior of potentials from reproducing kernel Hilbert space interpolation. *J. Chem. Phys.* **2000**, *112*, 4415–4416.

(35) Ho, T.-S.; Rabitz, H. A general method for constructing multidimensional molecular potential energy surfaces from ab initio calcul ations. *J. Chem. Phys.* **1996**, *104*, 2584–2597.

(36) Kingma, D. P. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**,

(37) Reddi, S. J.; Kale, S.; Kumar, S. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237* **2019**,

(38) Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; Monfardini, G. The Graph Neural Network Model. *IEEE Trans. Neural Netw.* **2009**, *20*, 61–80.

(39) Käser, E. D.; Boittier, M.; Upadhyay, M.; Meuwly, M. Transfer Learning to CCSD(T): Accurate Anharmonic Frequencies from Machine Learning Models. *J. Chem. Theor. Comp.* **2021**, *17*, 3687–3699.

(40) Upadhyay, M.; Meuwly, M. Thermal and vibrationally activated decomposition of the syn-ch3choo criegee intermediate. *ACS Earth Space Chem.* **2021**, *5*, 3396–3406.

(41) Käser, S.; Vazquez-Salazar, L. I.; Meuwly, M.; Töpfer, K. Neural network potentials for chemistry: concepts, applications and prospects. *Dig. Disc.* **2023**, *2*, 28–58.

(42) Unke, O. T.; Devereux, M.; Meuwly, M. Minimal distributed charges: Multipolar quality at the cost of point charge electrostatics. *J. Chem. Phys.* **2017**, *147*, 161712.

(43) Boittier, E. D.; Devereux, M.; Meuwly, M. Molecular dynamics with conformationally dependent, distributed charges. *J. Chem. Theor. Comp.* **2022**, *18*, 7544–7554.

(44) Boittier, E.; Töpfer, K.; Devereux, M.; Meuwly, M. Kernel-Based Minimal Distributed Charges: A Conformationally Dependent ESP-Model for Molecular Simulations. *J. Chem. Theor. Comp.* **2024**, *20*, 8088–8099.

(45) Lamoureux, G.; MacKerell, J., Alexander D.; Roux, B. A simple polarizable model of water based on classical Drude oscillators. *J. Chem. Phys.* **2003**, *119*, 5185–5197.

(46) Lemkul, J. A.; Huang, J.; Roux, B.; MacKerell, A. D. J. An Empirical Polarizable Force Field Based on the Classical Drude Oscillator Model: Development History and Recent Applications. *Chem. Rev.* **2016**, *116*, 4983–5013.

(47) Huang, J.; Simmonett, A. C.; Pickard, F. C.; MacKerell, A. D.; Brooks, B. R. Mapping the Drude polarizable force field onto a multipole and induced dipole model. *J. Chem. Phys.* **2017**, *147*, 161702.

(48) Teng, X.; Yu, W.; MacKerell, A. D. J. Computationally Efficient Polarizable MD Simulations: A Simple Water Model for the Classical Drude Oscillator Polarizable Force Field. *J. Phys. Chem. Lett.* **2025**, *16*, 1016–1023.

(49) Bálint, S.; Bakó, I.; Grósz, T.; Megyes, T. Structure of liquid methylene chloride:

Molecular dynamics simulation compared to diffraction experiments. *J. Mol. Liq.* **2007**, *136*, 257–266.

(50) Almásy, L.; Bende, A. Intermolecular Interaction in Methylene Halide (CH_2F_2, CH_2Cl_2, CH_2Br_2 and CH_2I_2) Dimers. *Molecules* **2019**, *24*, 1810.

(51) Allen, F. H.; Wood, P. A.; Galek, P. T. Role of chloroform and dichloromethane solvent molecules in crystal packing: an interaction propensity study. *Struct. Sci.* **2013**, *69*, 379–388.

(52) Weiner, S. J.; Kollman, P. A.; Case, D. A.; a nd C. Ghio, U. S.; Alagona, G.; Profeta Jr, S.; Weiner, P. A new force-field for molecular mechanical simulation of nucleic-acids and proteins. *J. Am. Chem. Soc.* **1984**, *106*, 765–784.

(53) Hermans, J.; Berendsen, H. J. C.; van Gunsteren, W. F.; Postma, J. P. M. A consistent empirical potential for water-protein interactions. *Biopol.* **1984**, *23*, 1513–1518.

(54) Jorgensen, W. L.; Tirado-Rives, J. The OPLS potential functions for proteins - energy minimization s for crystals of cyclic-peptides and crambin. *J. Am. Chem. Soc.* **1988**, *110*, 1657–1666.

(55) Clark, T.; Hennemann, M.; Murray, J. S.; Politzer, P. Halogen bonding: The σ-hole: Proceedings of “Modeling interactions in biomolecules II”, Prague, September 5th–9th, 2005. *J. Mol. Model.* **2007**, *13*, 291–296.

(56) El Hage, K.; Bereau, T.; Jakobsen, S.; Meuwly, M. Impact of quadrupolar electrostatics on atoms adjacent to the sigma-hole in condensed-phase simulations. *J. Chem. Theor. Comp.* **2016**, *12*, 3008–3019.

(57) Bálint, S.; Bakó, I.; Grósz, T.; Megyes, T. Structure of liquid methylene chloride: Molecular dynamics simulation compared to diffraction experiments. *J. Mol. Liq.* **2007**, *136*, 257–266, EMLG/JMLG 2006.

(58) Ferrario, M.; Evans, M. W. Computer simulation of dichloromethane. II. Molecular dynamics. *Chem. Phys.* **1982**, *72*, 147–154.

(59) Torii, H. Atomic quadrupolar effect in intermolecular electrostatic interactions of chloroalkanes: the cases of chloroform and dichloromethane. *J. Mol. Liq.* **2005**, *119*, 31–39.

(60) Böhm, H. J.; Ahlrichs, R. Molecular dynamics simulation of liquid CH_2Cl_2 and $CHCl_3$ with new pair potentials. *Mol. Phys.* **1985**, *54*, 1261–1274.

(61) Dang, L. X. Intermolecular interactions of liquid dichloromethane and equilibrium properties of liquid–vapor and liquid–liquid interfaces: A molecular dynamics study. *J. Chem. Phys.* **1999**, *110*, 10113–10122.

(62) Pollice, R.; Bot, M.; Kobylianskii, I. J.; Shenderovich, I.; Chen, P. Attenuation of London Dispersion in Dichloromethane Solutions. *J. Am. Chem. Soc.* **2017**, *139*, 13126–13140.

(63) Savoy, A.; Paenurk, E.; Pollice, R.; H unenberger, P. H.; Chen, P. Solvation Free Energies of Ion Dissociations in Dichloromethane: En Route to Accurate Computations. *J. Phys. Chem. B* **2025**,

(64) Martínez, L.; Andrade, R.; Birgin, E. G.; Martínez, J. M. PACKMOL: A package for building initial configurations for molecular dynamics simulations. *J. Comput. Chem* **2009**, *30*, 2157–2164.

(65) Constant pressure molecular dynamics simulation: The Langevin piston method. *J. Chem. Phys.* **1995**, *103*, 4613–4621.

(66) Pinski, P.; Riplinger, C.; Valeev, E. F.; Neese, F. Sparse maps—A systematic infrastructure for reduced-scaling electronic structure methods. I. An efficient and simple

linear scaling local MP2 method that uses an intermediate basis of pair natural orbitals. *J. Chem. Phys.* **2015**, *143*.

(67) Neugebauer, H.; Pinski, P.; Grimme, S.; Neese, F.; Bursch, M. Assessment of DLPNO-MP2 approximations in double-hybrid DFT. *J. Chem. Theor. Comp.* **2023**, *19*, 7695–7703.

(68) Neese, F. The ORCA program system. *WIRES Comput. Molec. Sci.* **2012**, *2*, 73–78.

(69) Neese, F. Software update: the ORCA program system, version 5.0. *WIRES Comput. Molec. Sci.* **2022**, *12*, e1606.

(70) Töpfer, K.; Vazquez-Salazar, L. I.; Meuwly, M. Asparagus: A toolkit for autonomous, user-guided construction of machine-learned potential energy surfaces. *Comput. Phys. Commun.* **2025**, *308*, 109446.

(71) Bereau, T.; Kramer, C.; Meuwly, M. Leveraging Symmetries of Static Atomic Multipole Electrostatics in Molecular Dynamics Simulations. *J. Chem. Theor. Comp.* **2013**, *9*, 5450–5459.

(72) Hedin, F.; Hage, K. E.; Meuwly, M. A Toolkit to Fit Nonbonded Parameters from and for Condensed Phase Simulations. *J. Chem. Theor. Comp.* **2016**, *12*, 1479–1489.

(73) Devereux, M.; Raghunathan, S.; Fedorov, D. G.; Meuwly, M. A Novel, Computationally Efficient Multipolar Model Employing Distributed Charges for Molecular Dynamics Simulations. *J. Chem. Theor. Comp.* **2014**, *10*, 4229–4241.

(74) Wolfenden, R. Benchmark Reaction Rates, the Stability of Biological Molecules in Water, and the Evolution of Catalytic Power in Enzymes. *Ann. Rev. Biochem.* **2011**, *80*, 645–667.

(75) Lynch, C. I.; Rao, S.; Sansom, M. S. P. Water in Nanopores and Biological Channels: A Molecular Simulation Perspective. *Chem. Rev.* *120*, 10298–10335.

(76) Gallo, P. et al. Water: A Tale of Two Liquids. *Chem. Rev.* **2016**, *116*, 7463–7500.

(77) Clark, G. N.; Cappa, C. D.; Smith, J. D.; Saykally, R. J.; Head-Gordon, T. The Structure of Ambient Water. *Mol. Phys. 108*, 1415–1433.

(78) Cisneros, G. A.; Wikfeldt, K. T.; Ojamäe, L.; Lu, J.; Xu, Y.; Torabifard, H.; Bartók, A. P.; Csányi, G.; Molinero, V.; Paesani, F. Modeling Molecular Interactions in Water: From Pairwise to Many-Body Potential Energy Functions. *Chem. Rev. 116*, 7501–7528.

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

(80) Tröster, P.; Lorenzen, K.; Tavan, P. Polarizable Water Models from Mixed Computational and Empirical Optimization. *J. Phys. Chem. B* **2013**, *117*, 9486–9500.

(81) Qi, R.; Wang, Q.; Ren, P.; Wang, L.-P.; Pande, V. S. United Polarizable Multipole Water Model for Molecular Mechanics Simulation. *J. Chem. Phys.* **2015**, *142*.

(82) Sidler, D.; Meuwly, M.; Hamm, P. An efficient water force field calibrated against intermolecular THz and Raman spectra. *J. Chem. Phys.* **2018**, *148*, 244504.

(83) Boittier, E. D.; Käser, S.; Meuwly, M. Towards Large-Scale Condensed Phase Simulations using Machine Learned Energy Functions. *arXiv preprint arXiv:2506.23272* **2025**,

(84) Yu, Q.; Qu, C.; Houston, P. L.; Conte, R.; Nandi, A.; Bowman, J. M. q-AQUA: A many-body CCSD(T) water potential, including four-body interactions, demonstrates the quantum natu re of water from clusters to the liquid phase. *J. Phys. Chem. Lett.* **2022**, *13*, 5068–5074.

(85) Zhu, X.; Riera, M.; Bull-Vulpe, E. F.; Paesani, F. MB-pol(2023): Sub-chemical Accuracy for Water Simulations from the Gas to the Liquid Phase. *J. Chem. Theor. Comp.* **2023**, *19*, 3551–3566.

(86) Abbott, A. P.; Capper, G.; Davies, D. L.; Rasheed, R. K.; Tambyrajah, V. Novel solvent properties of choline chloride/urea mixtures. *Chem. Commun.* **2003**, 70–71.

(87) Marcus, Y. *Deep Eutectic Solvents*; Springer, 2019; pp 185–191.

(88) Martins, M. A.; Pinho, S. P.; Coutinho, J. A. Insights into the nature of eutectic and deep eutectic mixtures. *J. Solut. Chem.* **2019**, *48*, 962–982.

(89) Smith, E. L.; Abbott, A. P.; Ryder, K. S. Deep eutectic solvents (DESs) and their applications. *Chem. Rev.* **2014**, *114*, 11060–11082.

(90) Arriaga, S.; Aizpuru, A. In *Advances and Applications of Partitioning Bioreactors*; Huerta-Ochoa, S., Castillo-Araiza, C. O., Quijano, G., Eds.; Advances in Chemical Engineering; Academic Press, 2019; Vol. 54; pp 299–348.

(91) De Sloovere, D.; Vanpoucke, D. E. P.; Paulus, A.; Joos, B.; Calvi, L.; Vranken, T.; Reekmans, G.; Adriaensens, P.; Eshraghi, N.; Mahmoud, A.; Boschini, F.; Safari, M.; Van Bael, M. K.; Hardy, A. Deep Eutectic Solvents as Nonflammable Electrolytes for Durable Sodium-Ion Batteries. *Adv. Energy Sustain. Res.* **2022**, *3*, 2100159.

(92) Hariyanto, Y.; Ng, Y. K.; Siew, Z. Z.; Soon, C. Y.; Fisher, A. C.; Kloyer, L.; Wong, C. W.; Chan, E. W. C. Deep Eutectic Solvents for Batteries and Fuel Cells: Biosubstitution, Advantages, Challenges, and Future Directions. *Energy & Fuels* **2023**, *37*, 18395–18407.

(93) Zhou, K.; Dai, X.; Li, P.; Zhang, L.; Zhang, X.; Wang, C.; Wen, J.; Huang, G.; Xu, S. Recent advances in deep eutectic solvents for next-generation lithium batteries: Safer and greener. *Prog. Mater. Sci.* **2024**, *146*, 101338.

(94) Guchhait, B.; Al Rasid Gazi, H.; Kashyap, H. K.; Biswas, R. Fluorescence Spectroscopic Studies of (Acetamide + Sodium/Potassium Thiocyanates) Molten Mixtures: Composition and Temperature Dependence. *J. Phys. Chem. B* **2010**, *114*, 5066–5081.

(95) Kalita, G.; Rohman, N.; Mahiuddin, S. Viscosity and Molar Volume of Potassium Thiocyanate + Sodium Thiocyanate + Acetamide Melt Systems. *J. Chem. Eng. Data* **1998**, *43*, 148–151.

(96) Hu, Y.; Li, H.; Huang, X.; Chen, L. Novel room temperature molten salt electrolyte based on LiTFSI and acetamide for lithium batteries. *Electrochem. Commun.* **2004**, *6*, 28–32.

(97) Wallace, R. A. Solubility of potassium halides in fused acetamide. *Inorg. Chem.* **1972**, *11*, 414–415.

(98) Sakpal, S. S.; Deshmukh, S. H.; Chatterjee, S.; Ghosh, D.; Bagchi, S. Transition of a deep eutectic solution to aqueous solution: A dynamical perspective of the dissolved solute. *J. Phys. Chem. Lett.* **2021**, *12*, 8784–8789.

(99) Töpfer, K.; Pasti, A.; Das, A.; Salehi, S. M.; Vazquez-Salazar, L. I.; Rohrbach, D.; Feurer, T.; Hamm, P.; Meuwly, M. Structure, Organization, and Heterogeneity of Water-Containing Deep Eutectic Solvents. *J. Am. Chem. Soc.* **2022**, *144*, 14170–14180.

(100) Töpfer, K.; Wang, J.; Patel, S.; Meuwly, M. In *Recent Developments of Molecular Electronic Structure Theory*; Hoggan, P. E., Coletti, C., Eds.; Adv. Quantum Chem.; Academic Press, 2025; Vol. 91; pp 317–339.

(101) Frisch, M. J. et al. Gaussian~16 Revision C.01. 2016; Gaussian Inc. Wallingford CT.

(102) Wakelam, V.; Bron, E.; Cazaux, S.; Dulieu, F.; Gry, C.; Guillard, P.; Habart, E.; Hornekær, L.; Morisset, S.; Nyman, G., et al. H_2 formation on interstellar dust grains:

The viewpoints of theory, experiments, models and observations. *Mol. Astrophys.* **2017**, *9*, 1–36.

(103) Hagen, W.; Tielens, A.; Greenberg, J. The infrared spectra of amorphous solid water and ice Ic between 10 and 140 K. *Chem. Phys.* **1981**, *56*, 367–379.

(104) Bovolenta, G. M.; Molpeceres, G.; Furuya, K.; Kästner, J.; Vogt-Geisse, S. CO Adsorption Sites on Interstellar Water Ices Explored with Machine Learning Potentials. Binding energy distributions and snowline. *arXiv preprint arXiv:2508.14219* **2025**,

(105) Cuppen, H.; Walsh, C.; Lamberts, T.; Semenov, D.; Garrod, R.; Penteado, E. M.; Ioppolo, S. Grain surface models and data for astrochemistry. *Space Sci. Rev.* **2017**, *212*, 1–58.

(106) Tielens, A. G. G. M. The molecular universe. *Rev. Mod. Phys.* **2013**, *85*, 1021.

(107) Hama, T.; Watanabe, N. Surface processes on interstellar amorphous solid water: Adsorption, diffusion, tunneling reactions, and nuclear-spin conversion. *Chem. Rev.* **2013**, *113*, 8783–8839.

(108) Upadhyay, M.; Meuwly, M. CO_2 and NO_2 formation on amorphous solid water. *Astron. Astrophys.* **2024**, *689*, A319.

(109) Kumagai, N.; Kawamura, K.; Yokokawa, T. An Interatomic Potential Model for H_2O: Applications to Water and Ice Polymorphs. *Mol. Spect.* **1994**, *12*, 177–186.

(110) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.

(111) Gao, J. Methods and applications of combined quantum mechanical and molecular mechanical potentials. *Rev. Comput. Chem.* **1996**, 119–185.

(112) Senn, H. M.; Thiel, W. QM/MM methods for biomolecular systems. *Angew. Chem. Int. Ed.* **2009**, *48*, 1198–1229.

(113) Pezzella, M.; Meuwly, M. O_2 formation in cold environments. *Phys. Chem. Chem. Phys.* **2019**, *21*, 6247–6255.

(114) Zhao, Y.; Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **2008**, *120*, 215–241.

(115) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132*.

(116) Menschutkin, N. Über die Affinitätskoeffizienten der Alkylhaloide und der Amine: Zweiter Teil. Über den Einfluss des chemisch indifferenten flüssigen Mediums auf die Geschwindigkeit der Verbindung des Triäthylamins mit den Alkyljodiden. *Z. Physikal. Chem.* **1890**, *6*, 41–57.

(117) Turan, H. T.; Brickel, S.; Meuwly, M. Solvent Effects on the Menshutkin Reaction. *J. Phys. Chem. B* **2022**, *126*, 1951–1961.

(118) Su, P.; Ying, F.; Wu, W.; Hiberty, P. C.; Shaik, S. The Menshutkin reaction in the gas phase and in aqueous solution: a valence bond study. *ChemPhysChem* **2007**, *8*, 2603–2614.

(119) Dutta Dubey, K.; Stuyver, T.; Kalita, S.; Shaik, S. Solvent organization and rate regulation of a menshutkin reaction by oriented external electric fields are revealed by combined MD and QM/MM calculations. *J. Am. Chem. Soc.* **2020**, *142*, 9955–9965.

(120) Aziz, M.; Prindle, C. R.; Lee, W.; Zhang, B.; Schaack, C.; Steigerwald, M. L.; Zandkarimi, F.; Nuckolls, C.; Venkataraman, L. Evaluating the ability of external electric fields to accelerate reactions in solution. *J. Phys. Chem. B* **2024**, *128*, 9553–9560.

(121) de Souza, B. GOAT: A Global Optimization Algorithm for Molecules and Atomic Clusters. *Angew. Chem. Int. Ed.* **2025**, *64*, e202500393.

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

(123) Rappoport, D.; Furche, F. Property-optimized Gaussian basis sets for molecular response calculations. *J. Chem. Phys.* **2010**, *133*.

(124) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *J. Chem. Theor. Comp.* **2019**, *15*, 1652–1671.

(125) Weigend, F.; Häser, M.; Patzelt, H.; Ahlrichs, R. RI-MP2: optimized auxiliary basis sets and demonstration of efficiency. *Chem. Phys. Lett.* **1998**, *294*, 143–152.

(126) Larsen, A. H. et al. The atomic simulation environment—a Python library for working with atoms. *J. Phys. Condens. Matter.* **2017**, *29*, 273002.

(127) Spackman, M. A.; Jayatilaka, D. Hirshfeld surface analysis. *CrystEngComm* **2009**, *11*, 19–32.

(128) Fafarman, A. T.; Webb, L. J.; Chuang, J. I.; Boxer, S. G. Site-Specific Conversion of Cysteine Thiols into Thiocyanate Creates an IR Probe for Electric Fields in Proteins. *J. Am. Chem. Soc.* **2006**, *128*, 13356–13357, PMID: 17031938.

(129) Aydin, S.; Salehi, S. M.; Töpfer, K.; Meuwly, M. SCN as a local probe of protein structural dynamics. *J. Chem. Phys.* **2024**, *161*, 055101.

(130) Salehi, S. M.; Koner, D.; Meuwly, M. Vibrational Spectroscopy of N3– in the Gas and Condensed Phase. *J. Phys. Chem. B* **2019**, *123*, 3282–3290, PMID: 30830786.

(131) Virtanen, P. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **2020**, *17*, 261–272.

(132) Wang, L.-P.; Head-Gordon, T.; Ponder, J. W.; Ren, P.; Chodera, J. D.; Eastman, P. K.; Martinez, T. J.; Pande, V. S. Systematic Improvement of a Classical Molecular Model of Water. *J. Phys. Chem. B* **2013**, *117*, 9956–9972.

(133) Feng, C.-J.; Dhayalan, B.; Tokmakoff, A. Refinement of Peptide Conformational Ensembles by 2D IR Spectroscopy: Application to Ala–Ala–Ala. *Biophysical Journal* **2018**, *114*, 2820–2832.

(134) Mondal, P.; Cazade, P.-A.; Das, A. K.; Bereau, T.; Meuwly, M. Multipolar Force Fields for Amide-I Spectroscopy from Conformational Dynamics of the Alanine Trimer. *Journal of Physical Chemistry B* **2021**, *125*, 10928–10938.

(135) Meuwly, M.; Hutson, J. M. Morphing ab initio potentials: A systematic study of Ne–HF. *J. Chem. Phys.* **1999**, *110*, 8338–8347.

(136) Horn, K. P.; Vázquez-Salazar, L. I.; Koch, C. P.; Meuwly, M. Improving Potential Energy Surfaces Using Measured Feshbach Resonance States. *Sci. Adv.* **2024**, *10*, eadi6462.

(137) Köfinger, J.; Hummer, G. Empirical Optimization of Molecular Simulation Force Fields by Bayesian Inference. *Eur. Phys. J. B* **2021**, *94*, 1–12.

(138) Wang, L.-P.; Martinez, T. J.; Pande, V. S. Building Force Fields: An Automatic, Systematic, and Reproducible Approach. *J. Phys. Chem. Lett. 5*, 1885–1891.

(139) Wang, Y. et al. On the Design Space between Molecular Mechanics and Machine Learning Force Fields. *Appl. Phys. Rev. 12*, 021304.

SUPPORTING INFORMATION: Cluster Models for Next-Generation Energy Functions for Molecular Simulations

S1 Additional Figures for Water

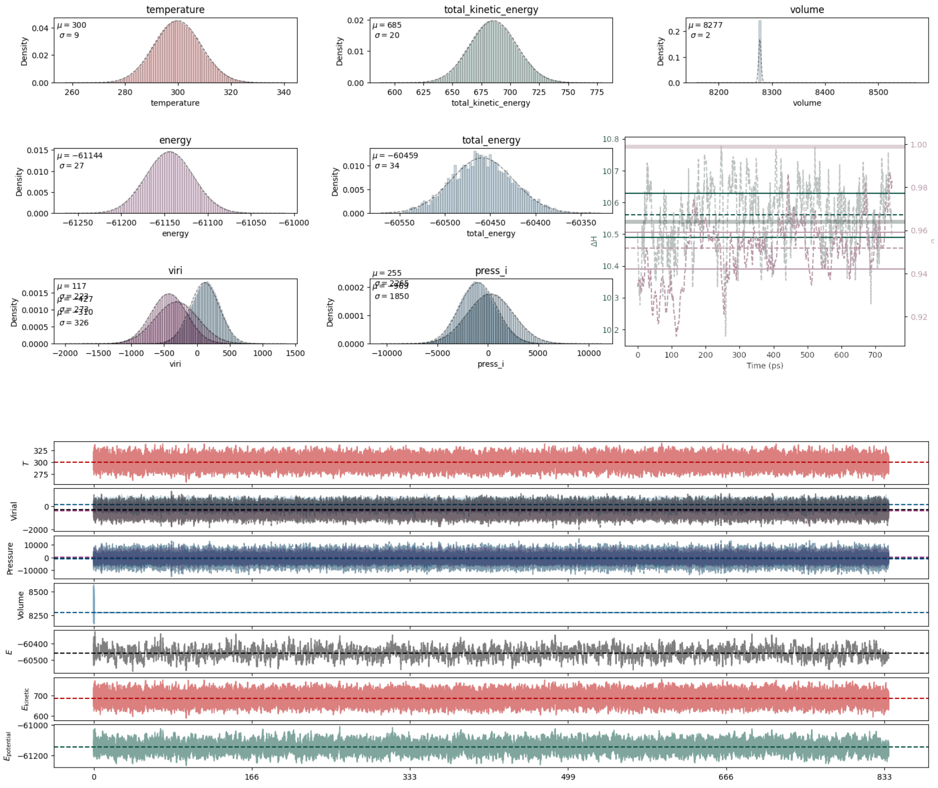

Figure S1: Refinement of LJ parameters may require a large number of trial simulations. For liquid water at ambient conditions, the predicted enthalpy of vaporization and density fluctuate with a few percent of the experimental values after a relatively short (1 ns) heating and equilibration producing stable NVT behavior (time series and distributions), arriving at canonical distributions within an additional 1 ns of simulation time. The predictive uncertainty associated with fluctuations before and after equilibration can be used to inform future trial parameters.

S1.1 Specification of eutectic mixtures

The different mixtures studied here are specified in Table S1.

Table S1: Specification of mixtures studied in the simulation

Index	Cation	Water percentage	Acetamide percentage
sod0	Na^+	0%	100%
sod20	Na^+	20%	80%
sod50	Na^+	50%	50%
sod70	Na^+	70%	30%
sod80	Na^+	80%	20%
sod90	Na^+	90%	10%
sod100	Na^+	100%	0%
hyb0	$45Na^+$, $30K^+$	0%	100%
hyb1	$45Na^+$, $30K^+$	100%	0%

S2 LJ parameter fitting for NaSCN in TIP3P/acetamide

Table S2: Composition of clusters with 0% water.

	SCN^-	Acetamide	TIP3P	Na^+
sys1	1	6	0	2
sys2	1	5	0	1
sys3	2	5	0	2
sys4	2	4	0	3

Table S3: Composition of clusters with 20% water + 20% ACEM.

	SCN^-	Acetamide	TIP3P	Na^+
sys1	1	5	1	0
sys2	1	4	2	1
sys3	2	5	1	0
sys4	2	4	2	2

Table S4: Composition of clusters with 50% water + 50% ACEM.

	SCN^-	Acetamide	TIP3P	Na^+
sys1	1	6	1	2
sys2	1	5	2	1
sys3	2	4	1	2
sys4	2	3	2	3

Table S5: Composition of clusters with 80% water + 20% ACEM.

	SCN^-	Acetamide	TIP3P	Na^+
sys1	1	2	4	0
sys2	1	1	5	1
sys3	2	2	3	0
sys4	2	1	4	2

Table S6: Composition of clusters with 100% water.

	SCN^-	Acetamide	TIP3P	Na^+
sys1	1	0	16	0
sys2	1	0	14	1
sys3	2	0	14	0
sys4	2	0	12	1

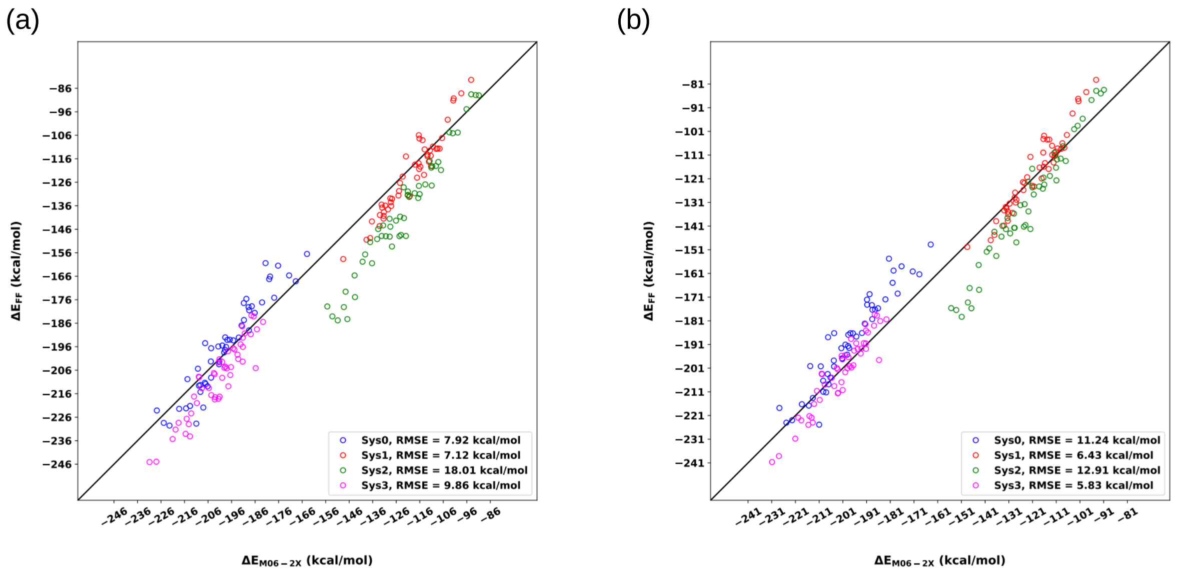

Figure S2: Correlation of interaction energies for clusters in 100% acetamide solutions between calculations with (a) initial parameters; (b) fitted parameters and DFT.

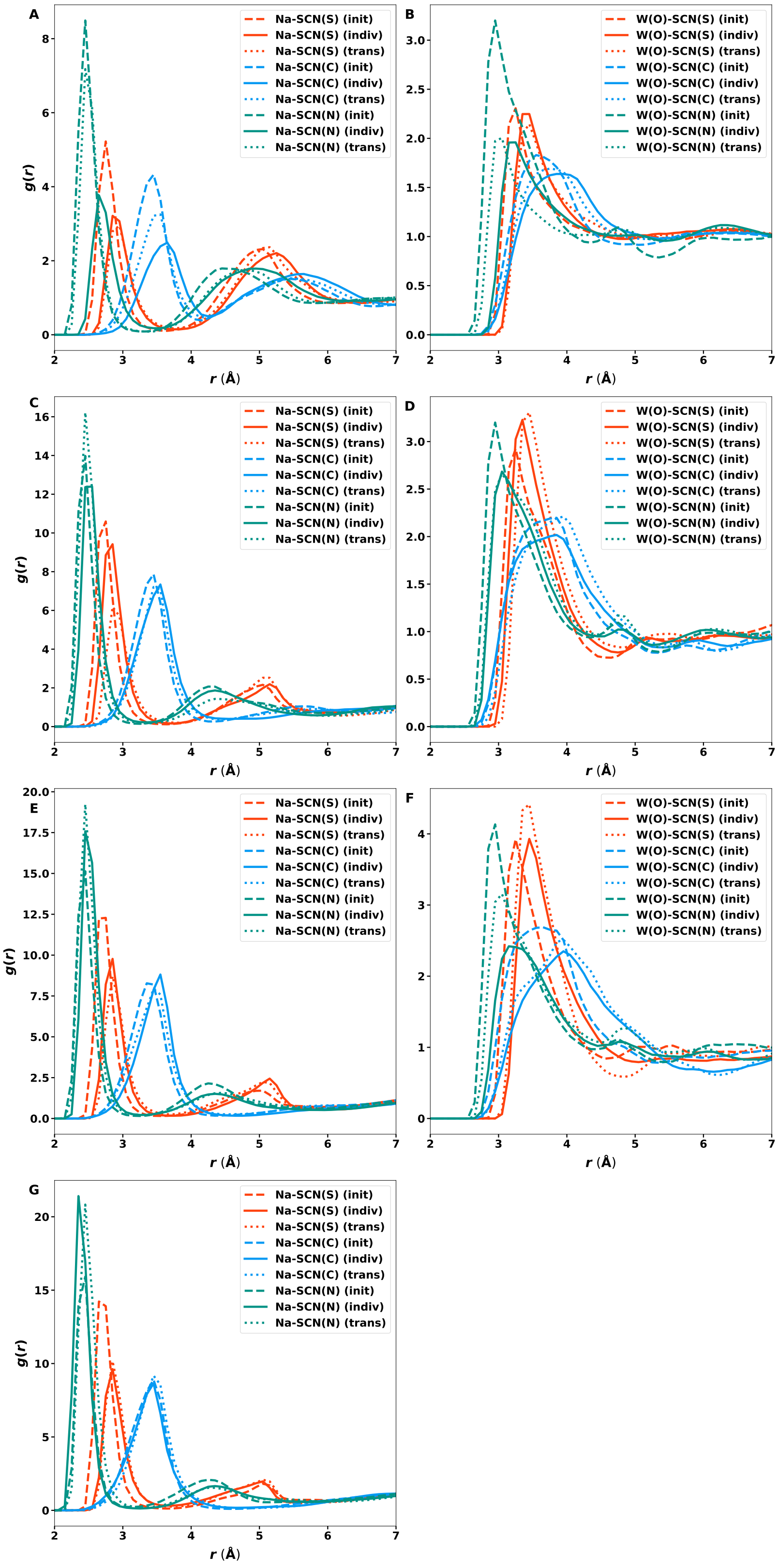

Figure S3: Comparison of the radial pair distribution function $g(r)$ between OW–X (X = S, C, N) with original (dashed lines) Lennard-Jones parameters, optimized (solid lines) Lennard-Jones parameters, and optimal (dotted) Lennard-Jones parameters from the 100/0 (A and B), 50/50 (C and D), 20/80 (E and F), and 0/100 (G, no water) W/ACEM mixtures.

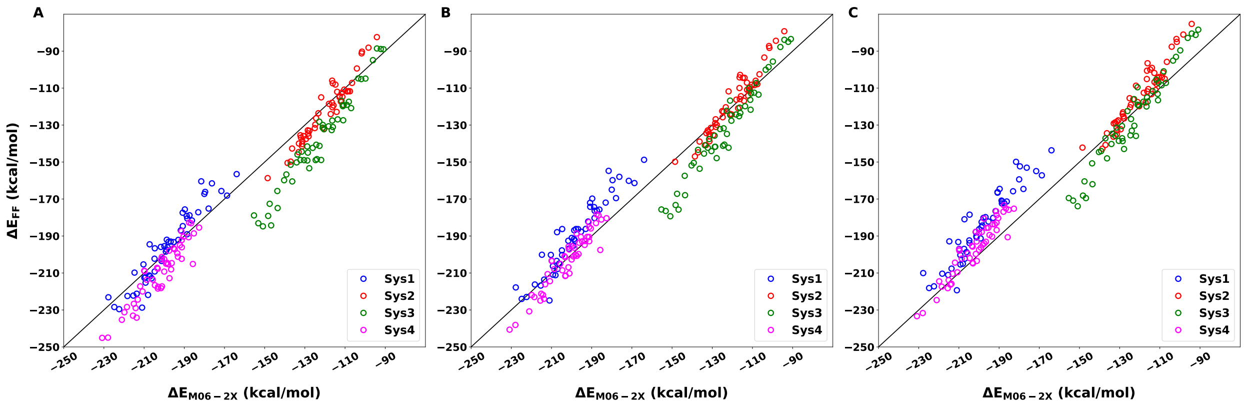

Figure S4: Correlation of interaction energies between reference DFT data and the empirical energy function for clusters extracted from simulations with [0/100] W/ACEM. Panel A: correlation before parameter optimization with the initial parameters;[99] panel B: parameters from individual optimization; panel C: transferable set of parameters.

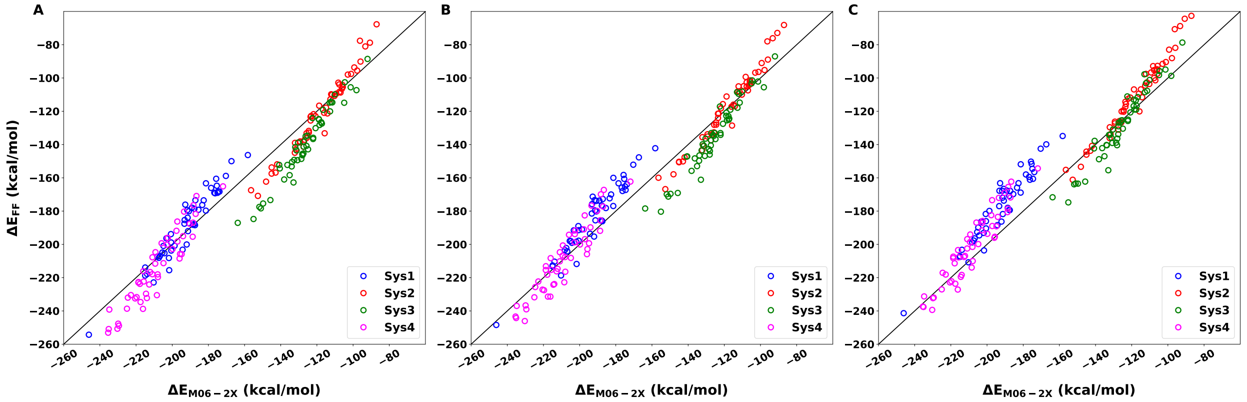

Figure S5: Correlation of interaction energies between reference DFT data and the empirical energy function for clusters extracted from simulations with [50/50] W/ACEM. Panel A: correlation before parameter optimization with the initial parameters; panel B: parameters from individual optimization; panel C: transferable set of parameters.

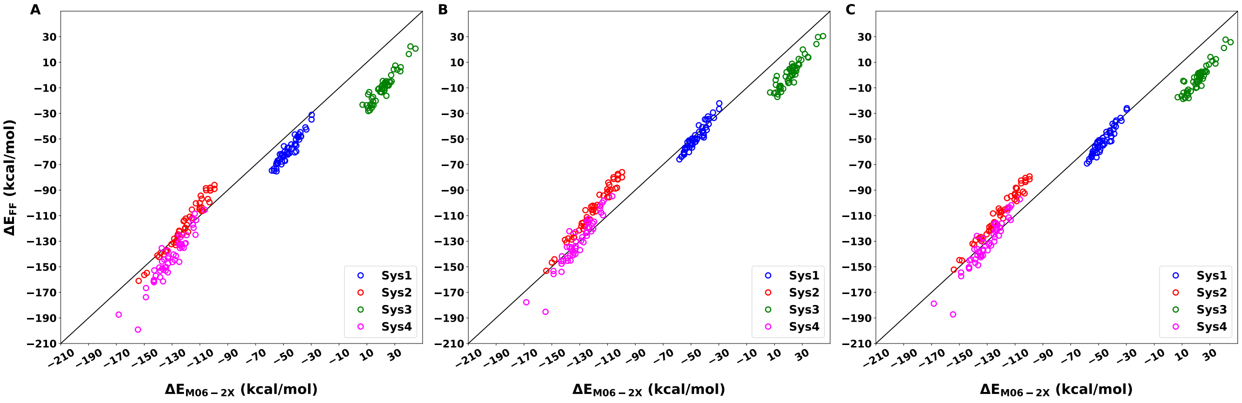

Figure S6: Correlation of interaction energies between reference DFT data and the empirical energy function for clusters extracted from simulations with [80/20] W/ACEM. Panel A: correlation before parameter optimization with the initial parameters; panel B: parameters from individual optimization; panel C: transferable set of parameters.

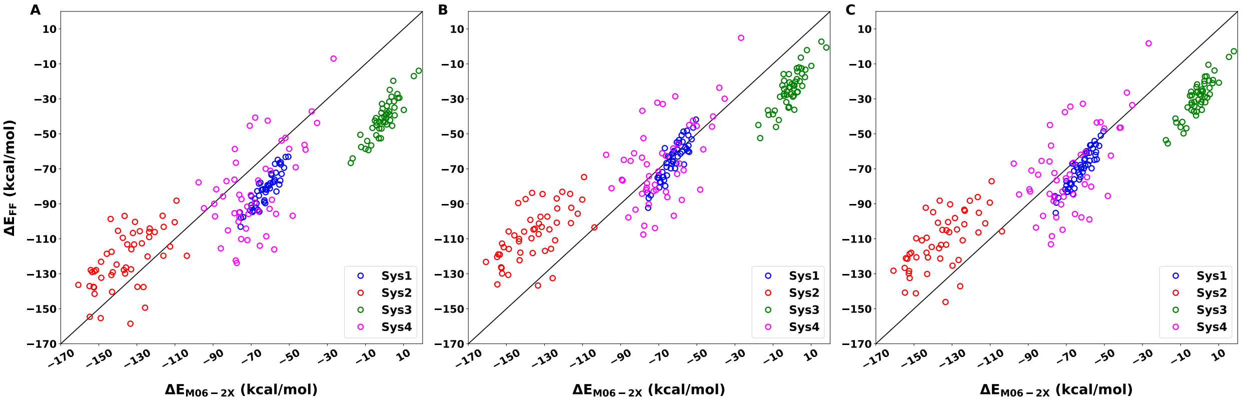

Figure S7: Correlation of interaction energies between reference DFT data and the empirical energy function for clusters extracted from simulations with [100/0] W/ACEM. Panel A: correlation before parameter optimization with the initial parameters; panel B: parameters from individual optimization; panel C: transferable set of parameters.

S3 Spectroscopic Probes

Table S7: Atomic charges from the fMDCM fit for CH_3SCN, along with Lennard-Jones (LJ) parameters obtained using the model. For the methyl group, LJ parameters were taken from the CGenFF force field.

Atom type	Charges				LJ parameters		
	Q1	Q2	Q3	Q4	5-wat	10-wat	25-wat
S	0.997	-0.890	-0.890	0.488	-0.20072 1.82607	-0.6 2.4	-0.6 1.5
C	-0.305	0.844			-0.18 2.17363	-0.008 1.6	-0.008 1.6
N	-0.439	-0.864	0.864		-0.4 1.79027	-0.01 2.24215	-0.08421 2.4
C2	0.010	-0.010					
H1-H2-H3	0.064						

Table S8: Atomic charges from the fMDCM fit for CH_3SNO, along with Lennard-Jones (LJ) parameters obtained using the model. For the methyl group, LJ parameters were taken from the CGenFF force field.

Atom type	Charges				LJ parameters		
	Q1	Q2	Q3	Q4	5-wat	10-wat	25-wat
S	1.000	-0.636	-0.636	0.070	-0.6 1.92449	-0.6 2.2	-0.15 2.2
N	0.463	0.963	-0.236		-0.23760 2.2	-0.01 1.0	-0.08006 2.2
O	-0.991	-0.072	-0.320		-0.05214 2.0	-0.17867 2.0	-0.01991 2.0
C2	-0.761	0.692					
H1-H2-H3	0.059						

Table S9: Atomic charges from the fMDCM fit for CH_3N_3, along with Lennard-Jones (LJ) parameters obtained using the model. For the methyl group, LJ parameters were taken from the CGenFF force field.

Atom type	Charges				LJ parameters		
	Q1	Q2	Q3	Q4	5-wat	10-wat	25-wat
N1	-0.931	0.448	-1.000		-0.05 2.26148	-0.05 2.21190	-0.04352 2.4
N2	0.992	0.796			-0.01 2.32741	-0.01 2.32741	-0.03102 2.4
N3	-0.887	-0.658	0.910		-0.01 2.21731	-0.01 2.19236	-0.01 2.19236
C2	0.505	-0.301					
H1-H2-H3	0.042						

Table S10: RMSE values (kcal/mol) of the fMDCM electrostatic models with unfitted LJ-parameters from CGenFF, relative to reference QM interaction energies. Columns show different molecules, rows are different cluster sizes.

	CH_3SCN	CH_3SNO	CH_3N_3
5-wat	4.32	5.12	5.49
10-wat	7.73	8.53	8.10
25-wat	13.08	14.13	11.59

S4 CO on Amorphous Solid Water

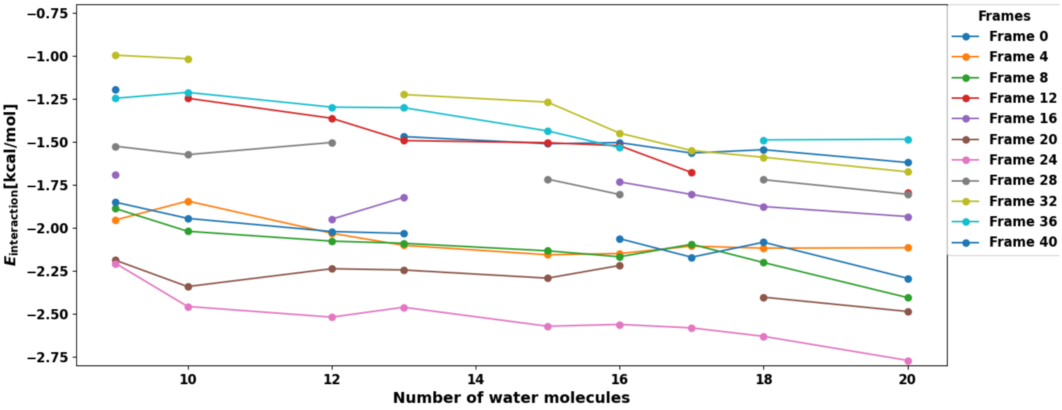

Figure S8: E_{int} v/s number of water molecules at M06-2X/aug-cc-pVTZ + D3 level for 11 different frames (orientations from MD simulations).