Published in final edited form as:
*J Chem Theory Comput*. 2022 April 12; 18(4): 2388–2407. doi:10.1021/acs.jctc.2c00115.

# Harnessing Deep Learning for Optimization of Lennard-Jones Parameters for the Polarizable Classical Drude Oscillator Force Field

**Payal Chatterjee** [¤], **Mert Y. Sengul** [¤], **Anmol Kumar**, **Alexander D. MacKerell Jr.**[*]
Department of Pharmaceutical Sciences, School of Pharmacy, University of Maryland Baltimore, 20 Penn Street, Baltimore, MD, 21201, USA

## Abstract

The outcomes of computational chemistry and biology research, including drug design, are significantly influenced by the underlying force field (FF) used in molecular simulations. While improved FF accuracy may be achieved via inclusion of explicit treatment of electronic polarization, such an extension must be accompanied by optimization of van der Waals (vdW) interactions, in the context of the Lennard-Jones (LJ) formalism in the present study. This is particularly challenging due to the extensive nature of chemical space combined with the correlated nature of LJ parameters. To address this challenge, a deep learning (DL)-based parametrization framework is developed allowing for sampling of wide ranges of LJ parameters targeting experimental condensed phase thermodynamic properties. The present work utilizes this framework to develop the LJ parameters for atoms associated with four distinct groups covering 10 different atom types. Final parameter selection was facilitated by quantum mechanical data on rare-gas interactions with the training set molecules. The chosen parameters were then validated through experimental hydration free energies and condensed phase thermodynamic properties of validation set molecules to confirm transferability. The ultimate outcome of utilizing this framework is a set of LJ parameters in the context of the polarizable Drude FF which

[*] Corresponding Author alex@outerbanks.umaryland.edu.
[¤] Equal First Authors

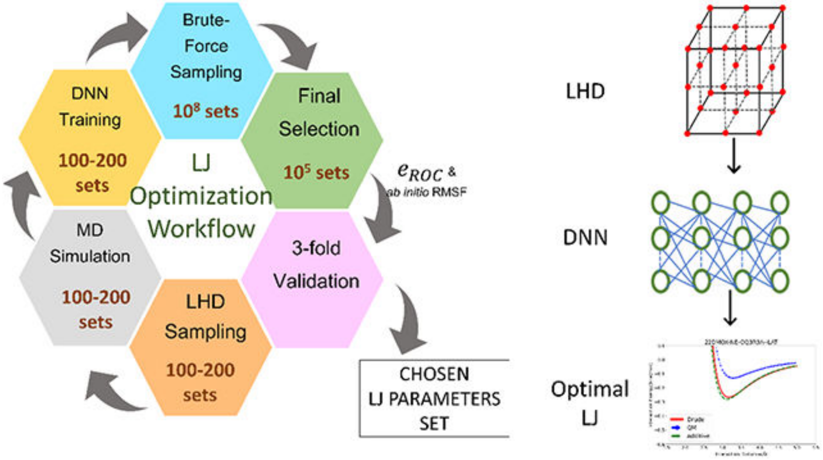
**Author Contributions:** PC contributed to conceptualization, workflow and related code development and implementation, data curation, model training, analysis of results, and writing of the article. MYS contributed to conceptualization, design of the deep learning and high throughput screening protocol, code development, and writing of the article. AK was involved in conceptualization, development and supervision of bonded and electrostatic parameter optimization and editing of the article. ADM supervised the entire research including conceptualization, analysis and writing of the article.

**Supporting Information:** The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.2c00115. Comparison of error convergence during training of alkenes DNN model (Figure SI); training and validation set molecules (Figure SII); condensed phase property correlation heatmaps (Figure SIII); convergence of the metrics mean absolute error and mean squared error (Figure SIV); correlation plots between the DL predicted and MD-calculated pure solvent condensed phase properties (Figure SV); QM vs MM potential energy scans for rare-gas interactions (Figures SVI-SVIII); temperatures used in pure solvent and crystal MD simulations (Table SI); parameter ranges used for generating parameter-based clusters (Table SII); range of LJ parameters used in the entire study (Tables SIII, SV, and SVI); comparison of the calculated vs experimental condensed phase properties of the initial and best LJ parameters for alkenes, three-membered rings, four-membered rings, and nitriles groups (Table SIV); thermodynamic properties of the pure solvents or crystals (Table SVII); additive vs Drude hydration free energies of 4-membered cyclic compounds (Table SVIII); and final optimized LJ parameters (Table SIX) (Supporting Information 1 (SI_1)); Comparison of the QM and MM parameters (Tables SI, SIII-SVI); statistical summary of the differences in the water interaction energy and distances (Table SII); QM and MM vibrational spectra (Table SVII); and topology and parameters of all molecules (Table SVIII); interaction orientations with water (Figures SI-SIV); and QM vs MM potential energy surface (Figure SV-SVIII) (Supporting Information 2 (SI_2)

**Conflict of Interest**: ADM is cofounder and CSO of SilcsBio LLC.

demonstrated improvement in the reproduction of both experimental pure solvent and crystal properties and hydration free energies of the molecules compared to the additive CHARMM General FF (CGenFF) including the ability of the Drude FF to accurately reproduce both experimental pure solvent properties and hydration free energies. The study also shows how correlations between difference in the reproduction of condensed phase data between model compounds may be used to direct the selection of new atom types and training set molecules during FF development.

## Graphical Abstract



## Keywords

van der Waals; Polarizable Force Field; CHARMM; Lennard-Jones parameters; Deep Learning; Latin Hypercube Design

## Introduction:

Molecular simulations have become indispensable in the biological and physical sciences, including their utilization in computer aided drug design (CADD). Rapid growth in computational power and increased efficiency of computational algorithms have allowed for simulations on biologically relevant timescales, extending up to milliseconds[1]. Improved computational efficiency has made it possible to address challenging problems in computational chemistry, such as accurately calculating ligand-binding affinities,[2, 3] the use of long timescale MD simulations and utilizing enhanced sampling methods to study complex conformational landscapes [4-7]. Central to the success of molecular dynamics (MD) simulations and related methods is the quality of the underlying force field (FF), dictating its ability to capture physically relevant observations *in silico*. Additive FFs are the current, widely used form of FFs, characterized by fixed point charges on each atom and other particles in the system. Examples of commonly used FFs in biomolecular systems are CHARMM[8], AMBER[9], GROMOS[10] and OPLS-AA[11]. Although additive FFs have been successfully utilized for decades, the fixed charge nature of such FFs limit their ability to respond to dynamic changes of the electronic field of the environment[12, 13]. Polarizable

FFs overcome this limitation by including the explicit treatment of electronic polarizability. Such FFs may be based on different models, including the classical Drude oscillator[14-19], fluctuating charge[20-31] and induced dipole[32-37] approaches. CHARMM's polarizable FF based on the classical Drude oscillator model has shown to be an efficient tool for capturing electrostatic interactions in a more accurate fashion[38-46]. For example, polarizable FFs like the Drude FF and AMOEBA[47] were found to improve accuracy in protein structure refinement, protein folding and simulations of intrinsically disordered proteins[42]. Polarizable FFs have also shown unique results in studies of nucleic acids, including base flipping[48], conformational sensitivity to ion type[40], ion distributions around duplexes[49] and improved modeling of RNA hairpins[50, 51]. Another recent study that compared five different force fields including Drude2017 on G-quadruplexes found Drude2017 achieved a high level of accuracy when evaluated against both quantum mechanical and experimental data[45].

The CHARMM Drude FF currently covers proteins, nucleic acids, lipids, carbohydrates, atomic ions, and a limited set of small molecules representative of those classes of molecules as well as additional species common to drug-like molecules. These include selected alkanes[52] alkenes[53], alcohols[54], ethers[55], aromatics[56], N-containing aromatic heterocyclics[57], amides[58], sulfur containing compounds[59] & halogenated aliphatic and aromatic compounds[60, 61]. However, this represents a limited range of chemical functional groups when considering broader chemical spaces, requiring significant extensions of the coverage of the FF. Examples in the context of drug-like chemical space include the full range of cyclic alkanes and heteroaromatic species, terminal and conjugated alkenes, alkynes, nitriles, amines, nitro-benzyl species, bipyrroles, biphenyl ring compounds, fused bicyclic ring compounds, thiophenes and so on.

Introduction of additional functional groups in the CHARMM-based FF approach involves consideration of the chemical connectivity of atoms and the associated atom types. The use of atom types versus typing based on, for example element and hybridization, allows for additional control of the accuracy of the force field with respect to both bonded and nonbonded parameters. Concerning the non-bonded terms, atom types differ in their Lennard-Jones (LJ) parameters, the formalism used to represent repulsion associated with Pauli's exclusion based on short-range repulsive forces between electrons with the same spin orientation and the attractive van der Waals intermolecular interactions associated London dispersion forces in the Drude FF. The LJ potential energy term as included in the Drude potential energy form is shown in equation 1.

$$U_{LJ} = \sum\nolimits_{non\,-\,bonded\ pairs} \left\{ \varepsilon_{ij} \left[ \left( \frac{R_{min,\,ij}}{r_{ij}} \right)^{12} - 2\left( \frac{R_{min,\,ij}}{r_{ij}} \right)^{6} \right] \right\} \qquad \text{Equation 1}$$

In Equation 1, $\varepsilon_{ij}$ is the LJ well depth, $r_{ij}$ is the distance between two atoms, and $R_{min,ij}$ is the distance between the two atoms i and j when the LJ potential energy surface reaches its minimum.

The LJ parameters $\varepsilon_{ij}$ and $R_{min,ij}$ are obtained from the individual parameters, $\varepsilon_i$ and $R_{min,i}$ for atom type i and $\varepsilon_j$ and $R_{min,j}$ for atom type j through combining rules, with the Lorentz-Berthelot rules used with the CHARMM FFs[62]. The individual parameters, $\varepsilon_i$ and $R_{min,i}$ are typically optimized using a set of multiple molecules sharing similar functional groups and the associated atom types[63-66]. Target data for the optimization is typically based on the experimental neat liquid or solid properties such as enthalpy of vaporization, $\Delta H_{vap}$, enthalpy of sublimation, $\Delta H_{sub,}$ molecular volume, $V_m$, dielectric constant, $\varepsilon$, isothermal compressibility, viscosity, etc. of the model compounds considered for optimization. Additional target or validation data may include experimental hydration free energies and quantum mechanical (QM) interactions between the model compounds with water, rare gases or other model compounds. Performing the optimization over multiple molecules sharing a common functional group maximizes the transferability of the LJ parameters in the context of wider chemical space occurring in more complex molecules. Tuning of the LJ parameters is the most challenging aspect of FF optimization as $\varepsilon_i$ and $R_{min,i}$ do not only include contributions from the $r^6$ and $r^{12}$ term in equation 1, but also include contributions from limitations in the electrostatic terms as well as other order terms that contribute to intermolecular interactions not directly included in the energy function. In addition, there is the problem of parameter correlation, where multiple combinations of parameters can similarly reproduce a collection of target data[63, 64]. These challenges are combined with simultaneously reproducing the experimental condensed phase properties of multiple molecules ideally requiring that the optimized LJ parameters belong to the global minimum of the LJ parameter space. Accordingly, optimization of $\varepsilon_i$ and $R_{min,i}$ for multiple atom types is a multi-variable and multi-objective problem. To address this challenge a LJ optimization approach is developed and implemented in the present study, building on an approach applied to facilitate optimization of the Reax force field (ReaxFF) [67], that harnesses the sampling capabilities of an initial design algorithm Orthogonal-maximin Latin Hypercube Design (LHD)[68], and the predictive abilities of Deep Learning (DL). Recently, a similar approach that includes LHD in conjunction with Gaussian process regression and Support Vector Machines to optimize LJ parameters for hydrofluorocarbons and ammonium perchlorate in the context of the General Amber FF[69] has been presented[70].

At first, LHD is utilized to generate LJ parameter sets $\varepsilon_i$ and $R_{min,i}$ for multiple atom types. The sampled parameter sets thus obtained are utilized in MD simulations to calculate condensed phase thermodynamic properties, $V_{m,}$ $\Delta H_{vap}$ and $\Delta H_{sub}$ of the training set molecules. These parameter sets form the input features, and the calculated properties are utilized as the output labels for building DL models to predict the condensed phase properties. The trained model is then used to comprehensively sample the LJ parameters, for example, $10^7$ combinations, allowing for prediction of the associated empirical condensed phase thermodynamic properties of the training set molecules. The resulting data is then sorted using a custom error function to select a subset of LJ parameter sets that maximize agreement with the target condensed phase thermodynamic properties. The final, optimal LJ parameters are chosen from this subset, based on their ability to reproduce the *ab initio* QM rare-gas interactions with the concerned molecule. The process flow of this newly designed approach for optimization of LJ parameters in CHARMM is depicted in Scheme 1.

In this article the developed DL-based high throughput approach for LJ parameter optimization is applied to atom types belonging to four different groups. These include the non-terminal and terminal alkenes sp2 carbons (CQ2C1A & CQ2C1B) and their corresponding hydrogens (HQ2C1A & HQ2C1B), 3- and 4-membered ring carbons (CQ3R3A & CQ3R4A) and oxygens (OQ3C3A & OQ3C4A) and nitrile carbon and nitrogen (CQ1N1 & NQ2C1) atom types. The optimized LJ parameters not only reproduce the experimental condensed phase thermodynamic properties $V_m$, $\Delta H_{vap}$ or $\Delta H_{sub}$ for both the training and validation set compounds but also their dielectric constants and hydration free energies. The total number of model compounds used for the study are 35, of which 17 belong to the training set – used for optimization of the LJ parameters, while the rest 18 belong to the validation set, meant for testing the transferability of the LJ parameters.

## Methods:

### Bonded and electrostatic parameter determination:

Prior to LJ parameter optimization, a complete set of FF parameters are required for any system. The electrostatic and bonded parameters of the training and validation molecules were obtained by following the Drude FF optimization protocol. All *ab initio* calculations were performed using the Psi4 package[71]and the molecular mechanical and condensed phase MD-based calculations were performed using CHARMM[8] and NAMD[72], with the latter used for pure-solvent systems only. The electrostatic and bonded parameters were optimized using FFParam [73], the recently developed package for FF optimization of both the additive CHARMM and Drude FFs. Although, FFParam includes a graphical user interface, an in-house alpha command line version of the package was also utilized to optimize multiple molecules together. The molecular geometries of all model compounds were optimized using MP2/6-31G(d) model chemistry. The QM optimized geometries were used to determine the Drude electrostatic parameters, including the partial atomic charges, the atomic polarizabilities (alpha), and the Thole scale factors. Alpha values represent isotropic polarizabilities of most atoms with anisotropic polarizabilities applied to selected hydrogen-bond acceptor atoms. Selected hydrogen-bond acceptor atoms also have virtual lone pair sites implemented to address the anisotropic distribution of the charges to optimize interactions with the surrounding environment[74]. Thole scale factors screen the atomic dipole-dipole interactions between 1-2 and 1-3 covalently linked atom pairs thereby by optimizing the molecular polarizability[75]. The partial atomic charges on the atoms and lone pair sites of the molecule were derived as recently described[76]. The method used an in-house adaptation of the Restrained Electrostatic Surface Potential (RESP)[77] model available in Psi4 package[71] at MP2/Sadlej model chemistry. The alpha values were obtained using a parallel implementation of the GDMA code by Stone and Misquitta [78, 79]available in Psi4 combined with the method of Heid et al [80] for charged species. Since Thole scale factors do not have a QM analog, they were determined using a Monte Carlo Simulated Annealing (MCSA) algorithm[81] to reproduce QM derived molecular dipole moments and molecular polarizability tensors scaled by a factor of 0.85. The electrostatic parameters using the above method were further assessed for their ability to reproduce the interaction of hydrogen donors and acceptors with water. For this purpose, MM interaction energies of selected atoms in the molecules with SWM4-NDP water were compared to QM water interaction energies obtained at MP2/cc-pVQZ model chemistry. The QM interaction energy was also corrected for basis set superposition error (BSSE) using the counterpoise method[82].

The initial bonded parameters were predicted using an in-house adaptation of the CHARMM General Force Field (CGenFF) program[83, 84]. Selected bond, angle and dihedral parameters were then adjusted to optimize the agreement of MM potential energy surfaces (PES) with the respective QM (MP2/aug-cc-pVDZ) PES. Additionally, bonded parameters were optimized to reproduce QM vibrational spectra calculated at the MP2/aug-cc-pVDZ model chemistry, where the QM vibrational frequencies were scaled by a factor of 0.9590 prior to use as target data[85]. The QM vibrational frequencies were calculated in the Gaussian package[86], while the MM vibrational frequencies were empirically calculated using the MOLVIB[87] module in CHARMM.

## Pure solvent MD simulations:

Neat liquid simulations were performed by preparing a box of 216 solute molecules, such that each molecule was equally spaced 6 Å apart in each direction. The initial setup of the box was performed in CHARMM using the additive CGenFF force field, where the liquid box was heated to their experimental temperatures (Table SI of supporting information 1 (SI_1)) for 100 ps in the NVT ensemble, followed by 400 ps NPT equilibration. For both the steps, the CPT leap-frog integrator with a timestep of 1 fs was used. A smaller time step was used to maintain consistency for comparison with the Drude Polarizable Force Field (FF). The equilibrated box was then used for a 3 ns additive MD production run using NAMD, where the MD parameters were maintained from the previous step in CHARMM. The fully equilibrated additive box was then utilized as the starting configuration for the pure solvent calculations in the Drude FF. Drude particles were added to the non-hydrogen atoms of the molecules, where a mass of 0.4 amu was transferred to the Drude particles from their real atoms. In addition, lone pairs were added as required to the hydrogen-bond acceptor atoms. The Drude topologies for all molecules are included in Table SVIII of supporting information 2 (SI_2). This was followed by a steepest-decent (SD) minimization for 200 steps, where all Drude particles were allowed to relax while the real atoms were restrained using a harmonic force constant of $10^6$ (kcal/mol)/Å$^2$. This was followed by another round of minimization where all particles were allowed to relax using SD for another 500 steps.

The liquid boxes using the Drude FF were then equilibrated at the experimental temperatures (Table SI of SI_1) and 1 atm pressure with a 1 fs timestep, in the NVT ensemble for 100 ps, followed by a 400 ps equilibration in the NPT ensemble using CHARMM as the MD engine. MD simulations were performed at respective temperatures for each molecule and 1 atm pressure, using the Velocity Verlet integrator (VV2) implemented in CHARMM. The VV2 integrator approximates the self- consistent field (SCF) condition of the Drude particles through an extended Lagrangian dual thermostat formalism [14]. A separate low temperature thermostat (T=1.0 K) was used for the Drude particles to ensure that their time course approximates the self-consistent field (SCF) regimen. The equilibrated system was then further run from 600 ps to 2 ns depending on the convergence using NAMD as the MD engine. In NAMD, the extended Lagrangian dual-thermostat of CHARMM is replaced by the dual-stochastic Langevin-thermostat for the treatment of the Drude particles[88]. The systems were minimized for 1000 steps in NAMD, followed by an equilibration when the velocities were reinitialized at their experimental temperatures. The calculations of the thermodynamic properties of the molecules were then based on condensed phase and gas-phase analysis performed in CHARMM. The molecular volume was calculated as the total average box volume divided by 216 for the number of monomers, while the enthalpy of vaporization was evaluated by subtracting the average potential energy of the monomers in the liquid phase from the average potential energy of each monomer in gas phase, adding a thermal correction of RT[89]. Two of the 35 molecules used in the present study also existed in solid state, namely 2-cyanopyridine (2CYP) – a training set compound for nitriles and 3-cyanopyridine (3CYP) belonging to validation set of the same group. As the compounds are low melting crystalline solids, with a melting point at room temperature 298 – 300 K[90], both liquid and solid-state data were available for such compounds. Hence, we calculated both heats of vaporization and sublimation for these compounds. The coordinates for both crystals were obtained from the Cambridge Structural database [91] and replicated using the CRYSTAL module of CHARMM, such that there were 32 molecules for 2CYP and 3CYP crystals. The crystal configurations were then energy minimized and initiated for 3 independent simulations, using distinct seed numbers for velocity generation. The setup of the Drude systems were identical to those in liquid simulations, where each system was simulated for a total of 600 ps, where the first 100 ps was used as equilibration, while the last 500 ps was used as the production run. For the determination of their $V_m$, both the simulations were performed at 150 K, the same temperatures at which the crystallization data was collected by Kubiak et al. (2002)[90], while the $\Delta H_{sub}$ were obtained at 298.15 K , as the experimental values were measured at that temperature [92]. Crystal $V_m$ calculations were based on the total volume of the full cell used in the simulations divided by the number of molecules in that cell, 32 in the present study. The analysis and evaluation of the final properties were performed in CHARMM and identical to the pure solvent calculation illustrated above.

For all additive and Drude MD simulations the electrostatic interactions were treated using the Particle Mesh Ewald (PME) method[93, 94], where a coupling parameter of 0.34 and a sixth-order spline were used for mesh-interpolation. The non-bonded pair lists were maintained up to 14 Å, with a 10-12 Å real-space cutoff range for the electrostatic and Lennard-Jones (LJ) terms, with the LJ interactions truncated with an atom-based forced switch algorithm[95]. Long-range corrections[95] to the LJ term was implemented as previously described[62, 96]. All covalent bonds involving hydrogens were constrained using the SHAKE algorithm[97]. The Drude hardwall constraint[39] of 0.2 Å was applied only while sampling the LJ parameter space meant for the training data for DL. Once the final parameters were optimized the hardwall constraint was removed, and the systems were run for longer timescales (10-20 ns) as required for the convergence of the dielectric constant.

## Hydration free energies

The hydration free energy (HFE) calculations were performed in CHARMM, using Deng and Roux's staged implementation[98] of alchemical free energy perturbation (FEP)[99, 100]. At first, each individual molecule was solvated in a box of 250 SWM4-NDP water molecules and equilibrated for 2 ns in CHARMM, using the condensed phase MD protocol as described above. The equilibrated box was then further utilized to calculate the HFE, where the HFE denotes the change in the free energy of annihilating the solute in vacuum to that in water, with the changes in the free energy computed through the FEP method. As described in detail previously[60, 101] and applied in multiple studies [43, 102-105], the HFE is decomposed into nonpolar and electrostatic components, where the nonpolar component is further decomposed into dispersive and repulsive terms using the Weeks, Chandler & Anderson (WCA) method[106]. Thus, a coupling parameter was used for perturbing each of the three individual components: λ for electrostatic (perturbed from 0 to 1 with an increment of 0.1), staging parameter *s* for dispersion (varied as 0.0, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0) and ξ for the repulsion term (varied linearly from 0 to 1, with an increment of 0.1). While perturbing the nonpolar components *s* and ξ, all charges of the solute were set to 0, while for the electrostatic component, the states λ = 0 and λ = 1 denoted fully discharged and charged compounds, respectively. A 300 ps equilibration and 1.5 ns to 4.0 ns production was performed for each of the λ windows, such that the production phase was utilized for calculating the final value. The final reported values for the electrostatic contributions were determined using thermodynamic integration (TI)[107], while the nonpolar contribution was determined using the exponential formula with the weighted histogram analysis method (WHAM)[108]. The free energy change was thus a sum of the dispersive, repulsive and electrostatic calculation, where a long range correction[96] was included in the dispersion term by taking the difference in the LJ solvent–solute interaction energies using cutoff schemes of 12 and 50 Å. Convergence of the HFE values was confirmed by calculating one system (compound in a SWM4 or TIP3P water box) in triplicate (separate for Drude and additive, respectively) where a similar value for all three simulations indicated that the length of the simulation (1.5 to 4 ns) was enough to confirm the convergence for all molecules in the group. The reported standard deviation for the group was based on these three individual sets of simulations. All other molecules of the group were then subjected to a single set of HFE simulation using the same length of production run.

## Deep Learning Model Development:

**Data Preparation—**The training set data for the DL model included the LJ parameters $\varepsilon_i^{min}$ and $R_i^{min}$ of the targeted atom types for the molecules in each set, as features and the MD calculated pure solvent/crystal properties $V_m$, $\Delta H_{vap}$, or $\Delta H_{sub}$ as the outputs or labels. At first, the LJ parameters were generated using LHD following which the pure solvent/crystal empirical thermodynamic properties ($V_m$, $\Delta H_{vap}$, or $\Delta H_{sub}$) were calculated through MD simulations. As described above, each molecule in the training set was subjected to three distinct runs of pure solvent/crystal MD simulations, starting from three distinct fully equilibrated additive boxes where the lengths of the simulations were chosen to obtain convergence of the empirical thermodynamic properties for each molecule for the individual groups. To reproduce the experimental properties of all the training set models, several iterations of LHD parameter generation and pure solvent MD simulations were required in some cases. This involved generation of an initial set of LJ parameters and the associated thermodynamic properties, which were then compared to experimental. If the experimental properties of all training set molecules were not adequately reproduced, a new range of parameters were selected, and the process repeated until adequate agreement for all the molecules was attained. Once convergence was achieved the LJ parameters and associated thermodynamic properties from all such scans were combined to create the training data for the DL models.

**Hyperparameter Tuning and Deep Learning Model Selection:** A feed-forward Deep Neural Network (DNN) was utilized to develop the DL models based on the feature vectors from the LHD generated LJ parameters to the target empirical thermodynamic properties. Hyperparameters optimized for the models are listed in Table 1, along with the range of the parameters tested. The hyperparameter tuning was implemented using a 5-fold Cross Validation utilizing the grid search method implemented in the GridSearchCV library of Scikitlearn[109] and regression model from Keras[110], where the metric of the evaluation was “negative mean squared error”. The final hyperparameters were then utilized for training the final models, with a learning rate of 0.005, Swish as the activation function, 2 hidden layers with the same number of nodes as the number of output thermodynamic properties ($V_m$ and $\Delta H_{vap}$ or $\Delta H_{sub}$ of each molecule in the training set). The number of nodes for each layer was thus 8 for alkenes and 3- and 4-membered ring models that had 4 training set molecules each and 12 for nitriles based on 5 molecules with two different states for 2-cyanopyridine. Concerning the activation function, both Swish[111] & ReLU (Rectified Linear Unit) were tested using (Mean Absolute Error) MAE and Mean Squared Error (MSE) as the criteria. Swish is also a sigmoid activation function like ReLU, but unlike ReLU, Swish is smooth and monotonic. Swish has been previously shown to it outperform ReLU[111] and application of the two functions in the present study during development of the alkene model also showed improved convergence of the error metrics over ReLU (Figure SI of SI_1).

**Training and evaluation of the selected DL model:** Individual DL models for $V_m$ and $\Delta H_{vap}$ or $\Delta H_{sub}$ were developed for all four groups: alkenes, 3-membered rings, 4-membered rings, and nitriles. As described above, each model was a feed-forward DNN comprised of two hidden, fully connected layers utilizing Swish as the activation function to determine the non-linear relationships between the input features $\varepsilon_i$ and $R_{min,i}$ and output labels ($V_m$ and $\Delta H_{vap}$ or $\Delta H_{sub}$ – for each molecule) with a linear activation function used for the output layer. The loss function was optimized using Adaptive Moment Estimator (Adam)[112], due to its adaptive learning rate and its suitability to complex parameter space, similar to the one in our data. Early stopping method [113] was used to limit the number of epochs and to avoid overfitting of the data, using a patience value of 100 to halt the training if no significant error reduction was achieved over 100 epochs. For training, 5-fold cross validation was used with 80% of the data chosen to train the DL model and 20% of it was utilized in testing the accuracy of the model, where the correlation of the experimental values to the predicted values was evaluated to confirm the performance of the model at the end of training. Finally, we emphasize that individual DNN models had to be trained for each of the four classes of functional groups optimized in the present study. Further performance evaluation of the final models was undertaken by extracting 20-25 sets of LJ parameters from the full range of parameters subjected to Brute-Force sampling with the DNN predicted properties compared to their empirically calculated values from MD simulations.

**Error functions for LJ parameter selection:** Using the trained DL models for each functional group class, 10 million sets of stochastically selected LJ parameters were sampled. For this purpose, a “Brute-Force search algorithm”[114] also known as “*perebor* algorithm,”[115] was utilized. The top ~100,000 sets of LJ parameters were chosen from the

10 million sets generated out of this data by first comparing the DNN predicted condensed phase properties to their respective experimental values using the error function presented in the following equations.

$$e_{ROC} = \sum_{i=1}^{n} \left(\delta^{i}_{V_m} w^{i}_{V_m} + \delta^{i}_{\Delta H_{vap}} w^{i}_{\Delta H_{vap}}\right) \qquad \text{Equation 2}$$

$$\delta^{i}_{V_m} = \left| V^{i}_{m_{obs}} - V^{i}_{m_{expt.}} \right| \qquad \text{Equation 3}$$

$$\delta^{i}_{\Delta H_{vap}} = \left| \Delta H^{i}_{vap_{obs}} - \Delta H^{i}_{vap_{expt.}} \right| \qquad \text{Equation 4}$$

In equation 2, $\delta^i$ denotes the unsigned difference of the respective calculated pure solvent property from experimental values as shown in equations 3 and 4 and $w^i$ denotes the weight used for the property. In equations 3 and 4 $V^{i}_{m_{expt.}}$ and $\Delta H^{i}_{vap_{expt.}}$ are the experimental molecular volume and enthalpy of vaporization or sublimation of each molecule $i$, $V^{i}_{m_{obs}}$ and $\Delta H^{i}_{vap_{obs}}$ are the calculated molecular volume and enthalpy of vaporization or sublimation of the same molecule, where n is the total number of molecules in the training set. The weights, $w_{Vm}$ and $w_{\Delta H_{vap}}$ for the molecular volume and enthalpy of vaporization/sublimation, respectively were generated using the Rank Order Centroid (ROC) method[116] wherein weights are generated according to the number of attributes or variables associated with the decision. To obtain a good set of parameters yielding the least possible $\delta^i$ for each of the pure-solvent properties, all molecules were ranked equally, prioritizing $\Delta H_{vap}$ or $\Delta H_{sub}$ over $V_m$. Thus, weights of $w^{i}_{V_m} = 0.25$ and $w^{i}_{\Delta H_{vap}} = 0.75$ were used. This error function was used in choosing the best sets from the training data and from the Brute-Force based predictions. In equations 2 to 4 $\Delta H_{sub}$ was substituted for $\Delta H_{vap}$ when the condensed phase was represented by a crystal and its $V_m$ is equivalent to the molecular volume in solid state.

Additional target data included QM rare gas (He and Ne)-model compound distance-based interaction potential energy scans (PES) focused on the concerned atom types in the molecules. The QM rare gas interaction values were obtained using the BSSE corrected MP2/cc-pVQZ model chemistry. Three interaction orientations; in-plane linear (0° from plane of target atom), in-plane lateral (90° in-plane of the target atom), and out-of-plane (90° out-of-plane of the target atom) were considered for all atom types, except CQ2C1A in internal alkenes and CQ1N1 in nitriles, where only one in-plane interaction was possible. Both QM and MM calculations were set up to perform distance-based PES, where the rare-gas-model compound distances were probed ranging from 2.5 Å to 5.0 Å. Comparison of the QM and MM interactions focused on the variation of the differences between the minimum interaction energies and distances over the different interaction orientations between the rare gases and the model compounds, not the absolute QM and MM minimum interaction energies and distances. The variance was quantified as the root mean square fluctuation (RMSF) of the differences between the MM and QM minimum interaction distances and energies, indicated by δ and ε, respectively, over all the model compounds and interaction orientations. Determination of the RMSF over all interactions orientations and model compounds first involved calculation of the absolute differences between the QM and empirical minimum interaction distances and energies as shown in equations 5a and 5b. In the equations $D_{ij,\,emp}^{K}$ and $D_{ij,QM}$ denote the empirical and QM minimum interaction distances, respectively, while $E_{ij,\,emp}^{K}$ and $E_{ij,QM}$ denote the empirical and QM minimum interaction energies, where *i* represents the model compound-rare gas interacting pairs (e.g., Ethene-helium, Ethene-neon, etc.), *j* represents the interaction orientations (in plane-linear, in plane-lateral and out-of-plane), and *K* represents each set of LJ parameters.

$$\delta_{ij}^{K} = \mid D_{ij,\,emp}^{K} - D_{ij,\,QM} \mid \qquad \text{Equation 5a}$$

$$\varepsilon_{ij}^{K} = \mid E_{ij,\,emp}^{K} - E_{ij,\,QM} \mid \qquad \text{Equation 5b}$$

The mean differences are then determined over the different interaction orientations *j* for each interacting pair *i* as denoted by equations 6a and 6b.

$$\delta_{i,\,mean}^{K} = \frac{\sum_j \delta_{ij}^{K}}{j} \qquad \text{Equation 6a}$$

$$\varepsilon_{i,\,mean}^{K} = \frac{\sum_j \varepsilon_{ij}^{K}}{j} \qquad \text{Equation 6b}$$

The RMSF about the mean differences of the interaction orientations, *j*, are then calculated for each interacting pair *i* as shown in equations 7a and 7b.

$$\delta_{i,\,rmsf}^{K} = \sqrt{\frac{\sum_j \left(\delta_{ij}^{K} - \delta_{i,\,mean}^{K}\right)^2}{j}} \qquad \text{Equation 7a}$$

$$\varepsilon_{i,\,rmsf}^{K} = \sqrt{\frac{\sum_j \left(\varepsilon_{ij}^{K} - \varepsilon_{i,\,mean}^{K}\right)^2}{j}} \qquad \text{Equation 7b}$$

Next, as shown in equations 8a and 8b, the mean of the RMSF for each LJ parameter set *K* is calculated by taking the mean of the $\delta_{i,\,rmsf}^{K}$ and $\varepsilon_{i,\,rmsf}^{K}$, respectively, over all interaction pairs per molecule *n*.

$$\delta_{n,\,mean\_rmsf}^{K} = \frac{\sum_i \left(\delta_{i,\,rmsf}^{K}\right)}{i} \qquad \text{Equation 8a}$$

$$\varepsilon^{K}_{n,\,mean\_rmsf} = \frac{\sum_i \left(\varepsilon^{K}_{i,\,rmsf}\right)}{i} \qquad \text{Equation 8b}$$

The sum of the RMSF of the distances and interaction energies are then calculated for each LJ parameter set $K$ for individual molecules $n$

$$RMSF^{K}_{n} = \delta^{K}_{n,\,mean\_rmsf} + \varepsilon^{K}_{n,\,mean\_rmsf} \qquad \text{Equation 9}$$

Finally, the overall RMSF of each LJ parameter set is calculated as the mean of $RMSF^{K}_{n}$ for all the model compounds being considered as shown in equation 10

$$RMSF^{K} = \frac{\sum_n RMSF^{K}_{n}}{n} \qquad \text{Equation 10}$$

As described previously, the $\mathrm{RMSF}^{\mathrm{K}}$ error function selects parameter sets that are balanced with respect to the relative interactions of the model compounds with the rare gases while allowing the MM energies and distances to be offset from the QM values [63-65].

**Selection of the final LJ parameters:** To choose the best set of LJ parameters from the top sets selected by the Brute-Force algorithm, rare-gas based RMSF along with a low $e_{ROC}$ was utilized. Step one involved selecting the top LJ parameter sets from the 10 million outcomes of the Brute-Force selected LJ parameters using the custom error function $e_{ROC}$. For example, for alkenes the range of $e_{ROC}$ for all 10 million sets varied from ~0.9 to ~26.0, hence only sets with $e_{ROC}$ less than ~4.0 were chosen yielding approximately 100,000 sets. Next, the chosen parameters were clustered into specific ranges of values of the LJ parameters. The ranges of the parameters were selected based on the ranges covered in the training data. Each LJ parameter was partitioned into 3 ranges as shown in Table 2 for alkenes and Table SII of SI_1 for the rest of the groups. These ranges of each of the LJ parameters were then uniquely combined into which the parameter sets were clustered. For example, for alkenes, 4 atom types HQ2C1A, HQ2C1B, CQ2C1A and CQ2C1B were being optimized, hence there were 8 different parameters ($\varepsilon_i$ and $R_{min,i}$ for each), in 3 different ranges, yielding a total of $8^3$ (512) possible clusters for the group. As an example, for alkenes, $\boldsymbol{\varepsilon_i^{HQ2C1A}}$ – Range A, $\boldsymbol{R_{min,\,i}^{HQ2C1A}}$ - Range B, $\boldsymbol{\varepsilon_i^{CQ2C1A}}$ – Range C, $\boldsymbol{R_{min,\,i}^{CQ2C1A}}$ - Range A, $\boldsymbol{\varepsilon_i^{HQ2C1B}}$ – Range C, $\boldsymbol{R_{min,\,i}^{HQ2C1B}}$ - Range **B**, $\boldsymbol{\varepsilon_i^{CQ2C1B}}$ – Range B **and** $\boldsymbol{R_i^{CQ2C1B}}$ - Range C will make one unique combination or cluster. Of the total number of possible clusters (512 in case of alkenes), only those clusters which contained at least 6 parameter sets were chosen for the RMSF analysis. Thus, the chosen subset of the data in alkenes consisted of 187 (out of 512) clusters. Next, all LJ sets, up to the top 500 based on $e_{ROC}$ ranking in each cluster, were subjected to a rare gas-based interactions in MM, which were further used to calculate the RMSF with those in QM. This type of clustering was done to ensure that all combinations of the LJ parameters present in the top ~100,000 selections were explored during the QM rare gas RMSF based analysis while avoiding the need to perform the RMSF calculation on all ~100,000 parameter sets. The final LJ parameter set was that with the lowest RMSF value along with a low $e_{ROC}$.

**Validation of the Final Parameters—**After the best LJ set was selected, the chosen set was then subjected to a threefold empirical validation process. This first involved empirically validating the predicted properties for the chosen LJ parameter set. Next, the condensed phase pure solvent/crystal properties ($V_m$ and $\Delta H_{vap}$/$\Delta H_{sub}$ & dielectric constant) were calculated for the validation set molecules. Lastly, the HFE of all the molecules in each group, as defined by the availability of the experimental value, were calculated to ensure the ability of the parameters to predict the energetics in aqueous solution.

Once the final LJ parameters were selected and validated, the electrostatic and bonded parameters were rechecked for reproducibility against the QM target data for all molecules in both training and validation sets. This included the intramolecular geometries, molecular vibrational spectra, and the PES of the selected bonds, angles, and dihedrals. In addition, water minimum interaction energies and distances, molecular dipole moments and the component vectors and molecular polarizabilities and their tensors were compared to the corresponding QM data. Figures SI-SVIII and Tables SI-SVII of SI_2 depicts all the related data for the final optimized parameters all the molecules. The FF topologies and parameters for all molecules are provided in Table SVIII of SI_2.

## Results & Discussion:

Presented is the implementation and application of a DL-based workflow for the refinement of LJ parameters. The overall DL-based LJ parameter optimization workflow is shown in Scheme 1. The procedure is composed of three main parts, namely: training, high throughput parameter selection, and validation. The training part is an iterative two step framework that is used to train a DL model that learns the relationship of the LJ parameters with the pure solvent or crystal condensed phase properties. The high-throughput selection part is used to sample the LJ parameter space using the trained DL model and find best sets of LJ parameters using the error function shown in equations 2 to 4. The final set of LJ parameters is then selected through comparison with QM rare gas-model compound interactions, thereby assuring the balance of the selected LJ parameters across the parameters themselves, atom types and molecules while still reproducing the condensed phase experimental properties. In the empirical validation part, the optimized LJ parameter set is validated through out-of-training molecules and calculation of the free energies of hydration and dielectric constants of both training and validation set molecules. The overall workflow is illustrated in Scheme 1 and detailed explanations for each part of the procedure are presented below.

### Atom type and model compound selection:

The DL-based LJ optimization approach is applied to the parameters $\varepsilon_i$ and $R_{min,i}$ belonging to atom types of four different organic functional groups not adequately optimized in the context of the Drude FF. These include the alkenes, 3- and 4-membered ring compounds, and nitriles. To initiate the optimization process initial decisions concerning the model compounds and the number of new atom types is required. For the model compounds, the training set molecules are selected based on their simplicity such that they have minimal additional atom types beyond those being targeted. In the case of the alkenes (Figure 1), these include ethene, propene, 1-butene and 2-butene. These molecules include both terminal and non-terminal sp2 carbons and the covalently linked hydrogens. Importantly, the only additional moiety on the selected model compounds are alkanes for which well optimized parameters are available[52, 117-119]. In the subsequent parameter optimization, these parameters were used for the alkyl chain atom types. Model compounds for the additional classes of molecules are shown in Figure SII of SI_1. Generally, a similar pattern in the structures of the molecules is evident within each individual functional group set. However, additional complexity in the molecules was required with the rings and nitriles to account for the lack of availability of the experimental condensed phase data for simple compounds. In addition, with the ring systems maintenance of the cyclic aliphatic carbon atom types in the presence and absence of oxygen was desired. Such considerations ultimately lead to the inclusion of compounds such as cyclic alkanes with and without ether and ketone groups in the 3- and 4-membered ring model compounds and aromatic rings in the nitrile model compounds.

The assignment of new atom types represents the second critical step in the extension of the Drude FF. The number of actual atom types required to reproduce the experimental condensed phase thermodynamic and kinetic properties has been debated [120-122] and such arguments range from suggesting individual parameters for each atom in a molecule [120, 121] or discuss the possibility of reducing the atom types to their elemental classification [122]. While too many atom types create a high level of complexity in the FF thereby limiting transferability of parameters, too few atom types limit the ability to achieve a sufficient level of accuracy. Thus, it is necessary to determine the requirement of a new atom type through a large set of molecules with a wide variety of chemical connectivity containing the same functional group. In the present study, we started with a minimal set of atom types for each functional group. Based on this minimal set, ranges for $\varepsilon_i$ and $R_{min,i}$ values were first chosen, thus constituting the LJ parameter space and LHD was used to generate the LJ parameters in a given range, which were then used to calculate the condensed phase data for all the training set molecules. In the case of alkenes, initially only three atom types were used; CQ2C1A non-terminal alkene carbon, CQ2C1B for terminal alkene carbons, while the same hydrogen type, HQ2C1A, was applied in both scenarios. The differences between the experimental and calculated condensed phase properties obtained for all the molecules in the set using all the LHD selected LJ parameters were calculated. A heat map of the correlations between experimental and calculated properties for four alkene molecules constituting the training set is shown in Figure 2. As may be seen, while reasonable correlations between some molecules were present (e.g., propene and ethene for $V_m$), in other cases the differences were anticorrelated (e.g., 1-butene and 2-butene for $V_m$ and ethene and 2-butene with $\Delta H_{vap}$). This indicates that variations in the LJ parameters for those three atom types alone would not lead to a solution that can accurately model all four model compounds. Consequently, an additional hydrogen atom type HQ2C1B, was added to allow for explicit hydrogen types for terminal C-H moieties. The resulting correlations in the differences in the reproduction of the experimental data across the molecules and LJ parameters sets significantly improved (Figure 2C and D) although no correlations were observed between ethene and the remaining three molecules. The lack of anticorrelated behavior indicated that LJ parameters that accurately modeled all four compounds could be achieved as presented below. Notably this approach could also be used to identify molecules appropriate for inclusion in the training set, as performed for the 4-membered rings (Figure SIII and associated text in SI_1), indicating its utility in facilitating decisions concerning atom types and model compounds to include during force field development.

Based on the above considerations and analysis the atom types and molecules for the DL LJ optimization were selected for all 4 groups. The atom types optimized in this work were alkene (non-terminal and terminal) carbons (CQ2C1A & CQ2C1B) and hydrogens (HQ2C1A & HQ2C1B), 3- and 4-membered ring carbons (CQ3R3A & CQ3R4A) and oxygens (OQ3C3A & OQ3C4A) and nitrile carbon and nitrogen (CQ1N1 & NQ2C1) atom types. For each group, 8-10 molecules were selected such that 4-5 of them were categorized as training set, while the rest were categorized as validation set molecules (Figure 1 and SII of SI_1). The training set molecules were used to optimize the LJ parameters and the validation-set molecules were used to test the transferability of the optimized parameters. In the remainder of the main text, the description of the application of the DL workflow will focus on the alkenes along with summary data on the remaining classes of compounds, with details included in the SI_1.

## DL model development:

DL model development is an iterative scheme composed of data generation and DL training. After the electrostatic and bonded parameters of the model compounds for a given group are parametrized, the compounds are used for the generation of training data, which involves LJ parameter sets as features and the corresponding pure solvent or crystal properties calculated using MD simulations as labels. Generation of the training data that encompasses the experimental properties for the multiple molecules in the group is challenging and requires the appropriately distributed sets of input LJ parameters. The initial LJ parameter guesses were obtained either from the LJ parameters of analogous atom types in the Drude polarizable force field or in CGenFF. Using the initial guess LJ parameters, the condensed phase properties of all the molecules in the training set were calculated. Comparison of the calculated and experimental values was then undertaken from which a range of LJ parameters for DL model development were initially selected. For the individual classes of functional groups, the range of the parameters for initial model development was proportional to the overall level of agreement between calculated and experimental values. Once an initial range of LJ parameters for all the relevant atom types was selected, LHD was utilized to generate LJ parameters that uniformly covers the selected range parameters for generation of the calculated condensed phase properties. LHD is a statistical sampling method that is used to generate evenly distributed parameter sets within a given range, thus generating non-overlapping sets. The “center maximin” type of LHD was used to generate sets of parameters uniformly sampled with a reduced pair-wise correlation and maximized “inter-site distances” between the parameters thus generated. To generate a sample size of N from x variables, LHD divides the range of each variable into N non-overlapping intervals based on an equal probability size of 1/N. The intervals within each point generated using LHD are thus uniformly distributed to represent the given LJ parameter space. As an example, N = 200 sets of LJ parameters with 8 different variables ($\varepsilon_i$ and $R_{min,i}$ of 4 different atom types) were generated for alkenes. The upper and lower limits of the LJ parameters for all four atom types are listed in Table 3 along with the total range, the sampling resolution, and the total possible number of sets present in the given LJ space interpreted by LHD for each individual LJ parameter. Thus, the 200 sets of LJ parameters generated by LHD are representative of 6.7 X $10^{17}$ possible sets of LJ parameters of alkenes.

Using the LHD selected parameters MD simulations were undertaken to calculate the associated condensed phase properties. If the calculated properties encompassed the experimental data for the training molecules, DL model training was initiated. If the calculated values did not encompass the experimental data, a new range of parameters was identified, LHD applied to select a new training set of LJ parameters and the MD calculations performed. It took different numbers of iterations to cover the experimental or near experimental properties for the four groups, which depend on the quality of the initial LJ parameters. For example, with the alkenes, the previously optimized LJ parameters for 2-butene from the Drude lipid FF[53] was used as the starting point for generating the training data, representing a high quality initial guess. In contrast, with the 4-membered rings, since there were no analogous carbon atom types in the Drude FF, the initial LJ parameters for the carbon atom were obtained from the CGenFF 4-membered ring carbon. For the 4-membered ring oxygen, the Drude FF tetrahydrofuran oxygen LJ parameters were used. The parameter ranges for alkenes along with the initial parameters are listed in Table 3, and the quality of the empirical properties of the initial and the best LJ parameters from the LHD selected training data are listed in Table 4. Data in Table 4 includes the differences and percentage differences of the calculated condensed phase properties of the four training set molecules from their respective experimental properties. With the alkenes, only a single scan was required to prepare the data used for training the DL model. This is associated with the good initial set of parameters yielding overall good agreement with experiment. Interestingly, the best set selected by LHD improved agreement with experiment in some cases (e.g., ethene ΔHvap) though not in all cases shown in Table 4. However, as LHD is designed to sample a diverse range of parameters rather than identify the ideal set, this result is expected.

With the remaining groups additional scans were required as the initial guess of the LJ parameters was not as good as with the alkenes. For the 4-membered rings it took 5 scans to cover the experimental properties for most of the molecules in the set (Table SIII of SI_1). However, this process included using the correlation analysis of the differences in the condensed phase properties to identify that 2-oxetanone was inappropriate as model compound (Figure SIII and associated text in SI_1). Once 2-oxetanone was identified as problematic an additional scan was performed with both 2- and 3-oxetanone followed by a single, final scan with 2-oxetonane omitted. For the 3-membered rings and nitriles the initial LJ from CGenFF produced near experimental properties in two scans. The number of sets selected from LHD used for training each model varied from 97-200. The details of the LJ parameter ranges for the rest of the three groups, along with the quality of the initial set to the best set from the LHD-selected training data is presented in Tables SIII, SIV and SV of the SI_1. Thus, the presented DL approach appears to require one to two scans to identify the appropriate region of LJ parameters space in cases where the suitable model compounds are identified. When multiple scans were required, all the LJ parameter sets for all the scans on which MD simulations were performed and empirical condensed phase data obtained were used for training of the final DL models for each group.

Although LHD samples the selected LJ parameter space uniformly it does so at a low resolution such that identifying the regions of parameter space that most accurately represent the experimental pure solvent properties for all chosen molecules is not achieved. DNNs have the ability to extract complex information from data, even when the training data is composed of much simpler information [123]. Our models use the LJ parameters $\varepsilon_i$ and $R_{min,i}$ of the given atom types as features to predict the pure solvent properties $V_m$ and $\Delta H_{vap}/\Delta H_{sub}$ as the outputs or labels for each molecule in the set. The convergence of the error metrics (MAE and MSE) during the training is shown in Figure SIV of SI_1. In all cases the models are largely converged after 100 epochs though training continued until the exit criteria discussed above was met. The resulting DL models were able to predict the MD simulation-based target data based on condensed phase data from simulations of a subset of the LHD selected parameter sets. For the final models the correlations between the MD-based true and DL-based predicted values of all target properties in the test split from the 5-fold cross validation for each group are shown in Figure 3 for all four groups. The average $R^2$ of all the models was 0.96 ± 0.03, which depicts their high predictive ability.

With the 3- and 4-membered ring groups, a subset of the condensed phase simulations based on the LHD selected parameters were not stable. This is due, for example, to certain combinations of LJ parameters having $R_{min,i}$ values that are too large or $\varepsilon_i$ values that are not favorable enough over the different atom types in the group such that the interactions between the monomers in the simulations were not favorable enough to maintain a condensed phase. In these cases, the liquids expanded into gases during the NPT simulations. Such a behavior, associated with what is termed infeasible vs. feasible parameter sets, was anticipated as LHD explores the full range of the multivariate LJ parameter space, thus potentially resulting in infeasible combinations of $R_{min,i}$ and $\varepsilon_i$ as occurred with the cyclic compounds. Notably, the number of infeasible sets was limited to 53 out of 220 for 3-membered rings, 18 out of 225 for 4-membered rings (without optimizing the oxygen) and 10 out of 120 (with optimizing the ring oxygen) (Tables SIII and SIV of SI_1). Thus, the final models for such groups were trained on the remaining, feasible LJ parameter sets, while there were no infeasible sets in alkenes and nitriles groups. The number of sets used for DL training for each group are included in Table SIII of SI_1.

## High-throughput LJ Parameter Selection:

The trained DL models were applied to sample from a broad range of LJ parameters space at high resolution to identify top ranking LJ parameter sets for the four groups. This involved using the DL models to predict the pure-solvent or crystal properties of 10 million input parameter sets ($\varepsilon_i$ and $R_{min,i}$), where the input parameters were generated by stochastically sampling throughout the entire LJ parameter space of up to $10^{18}$ possible parameter sets using the Brute-Force algorithm. The Brute-Force algorithm is a straight-forward problem-solving method where all possible solutions to a problem are tested individually, retaining only those that are close to the actual solution. Some well-known examples of its applications include the implementation of chess in Artificial Intelligence[114, 124] and cryptography[115, 125]. Such an algorithm was recently used in a DL-based force field parametrization framework for ReaxFF [67]. Following this strategy, $V_m$ and $\Delta H_{vap}/\Delta H_{sub}$ were predicted for 10 million LJ parameter sets generated stochastically within the specified range of parameters for the training-set molecules. Such a large number of sets were sampled to allow the DL models to interpolate and predict pure solvent empirical properties for LJ parameters that were not covered by the LHD selected parameters.

Once the 10 million LJ parameter sets were sampled using Brute-Force, the next step involved determining the sets of parameters which yield empirical condensed-phase properties closest to their experimental values. However, determining a single set of parameters that yields objectives closest to the target represents a significant challenge associated with the present multi-variable, multi-objective problem. Thus, a custom error function, $e_{ROC}$, was applied to choose a collection of best sets for each group. $e_{ROC}$ is based on the weighted unsigned differences between predicted and experimental values of the observables (Equations 2 to 4). Presented in Figure 4 are the distributions of $e_{ROC}$ for the four functional groups. Evident are the broad distributions, with the distributions biased towards low $e_{ROC}$ values for all four groups, with the widest range of errors for 4-membered groups and least for the nitriles. However, in all cases it is evident that a large number of parameter sets have low scoring $e_{ROC}$ values. Accordingly, additional target data was required to select the final parameter sets, as described in the next section.

## Selection of the final LJ parameters:

To select the final sets of parameters *ab initio* QM interactions of rare gas elements (He and Ne) with the training set molecules were used. The use of rare gas-model compound *ab initio* data has previously been used in CHARMM and Drude FF LJ parameter optimization [63-65]. This approach focuses on balancing the interaction energies and distances over the different molecules and orientations while the magnitude of both terms may be systematically offset from the QM values, as required to allow for accurate reproduction of experimental condensed phase data. This approach was designed to address the parameter correlation problem where more extreme values of LJ parameters can yield good agreement with experimental data as, for example, an unphysically large $\varepsilon_i$ with one atom type may compensate for an unphysically small $\varepsilon_i$ on a second atom type during optimization.

The final parameter set selection was initiated by selecting approximately the top 100,000 LJ parameter sets from the 10 million sets subjected to Brute-Force analysis with the DNNs. This was performed by identifying an $e_{ROC}$ cutoff that yielded approximately the top 100,000 sets. For alkenes, the top ~100,000 sets encompassed an $e_{ROC}$ less than 4.0. The chosen subset of the data was divided into unique clusters of LJ parameters, with a minimum of 6 LJ sets in each cluster. These clusters were then subjected to the rare gas-model compounds RMSF calculations for up to the top 500 sets in a cluster ranked based on the $e_{ROC}$ values. The clustering ensured a uniform sampling of the LJ parameters in the top 100,000 sets while focusing on lower $e_{ROC}$ values as well as avoiding the need to perform the RMSF calculation on all 100,000 sets. The LJ parameter set corresponding to the lowest RMSF value along with a low $e_{ROC}$ was chosen as the final set.

### Importance of inclusion of both condensed phase and ab initio QM target data in LJ parameter optimization.

Shown in Figure 5 are rare gas-model compounds PES for 1-propene, 1-butene and 2-butene, targeting the terminal and non-terminal carbons in the double bond for different interaction orientations. The figure represents PES from the *ab initio* QM calculations and for the final LJ parameter set selected based on the lowest RMSF along with those from CGenFF. As is evident there are significant differences between the QM and MM PES, with the MM PES for the Drude force field being systematically more favorable and with minima at shorter distances; a similar trend occurs with CGenFF. This emphasizes the need for the use of the DNN to facilitate the selection of LJ parameters that reproduce the experimental data while the RMSF of the differences between the MM and QM minimum interaction energies and distances over the rare gas-model compound interaction orientations selects LJ parameters that balance the interactions as a function of orientation. The similarity of the Drude and additive CGenFF PES indicates that the difference between QM and MM PES largely reflect limitations in the use of dimers alone to model dispersion interactions that also yield appropriate condensed phase properties with an MM model. Additional limitations in the treatment of long-range dispersion contributions in the QM model[126-129] will also contribute to the differences in Figure 5 and Figure SVI to SVIII in SI_1 for the remaining groups.

Additional analysis was undertaken to better quantify the use of the DL model for LJ parameter optimization and the impact of the use of the RMSF metric for final parameter set. Shown in Table 5 are average difference and percent difference in condensed phase properties for the training sets for the four groups for LJ parameters obtained at different steps in the parametrization workflow. In addition, the $e_{ROC}$ and RMSF from the final two steps of with workflow are included. As is evident going from the initial guess LJ parameters to the best of the LHD selected parameters to the parameters from the DL Brute-Force sampling based on the $e_{ROC}$ metric alone generally leads to improvement in the overall agreement with the experimental condensed phase properties. Inclusion of the RMSF metric in addition to $e_{ROC}$ when selecting parameters leads to poorer agreement with experiment in the majority of cases associated with an increase in the $e_{ROC}$ metric while the RMSF value decreased as expected. With the nitriles, the lowest $e_{ROC}$ LJ parameter set also corresponded to the lowest RMSF associated set. Thus, as expected the inclusion of the RMSF metric leads to suboptimal $e_{ROC}$ values and a degradation in the agreement with the average condensed phase properties, though the differences are not statistically significant in the majority of cases. However, the inclusion of the RMSF metric yields LJ parameters with an improved balance in the interactions between the rare gases and the model compounds as function of orientation and, importantly yields the overall good agreement with the experimental condensed phase properties for both training and validation set molecules as presented below.

### Validation of the final parameters:

The final LJ parameters chosen from the high throughput selection process were validated through pure solvent calculations on the separate validation set molecules and on the calculation of dielectric constants and HFEs of both the training and validation set compounds. Table 6 presents the final pure solvent and crystal properties including $V_m$, $\Delta H_{vap}$, $\Delta H_{sub}$ and dielectric constants along with the HFE of the training and validation set molecules for alkene group. Table 7 lists the average unsigned differences and the percent differences of the $V_m$ and $\Delta H_{vap}/\Delta H_{sub}$ of the final calculated properties to their experimental values for all four groups with the HFE values for the four groups shown in Table 8. For the alkenes, the differences in the energetic terms are typically less than 0.5 kcal/mol and the percent difference less than 10 % indicating general agreement within chemical accuracy [130, 131] of experiment for the studied properties. Analysis of Tables 7 and 8 indicate that the Drude model generally shows improvement over the additive FF when taking all the molecules into account. An exception occurred with the alkenes, where the average differences were smaller with the additive model, though the Drude performs better with the validation set molecules. Specifically, for 4-membered ring compounds, on an average the Drude FF reproduced $V_m$ by 4.09±0.36 $Å^3$ for the training set compounds, 1.04±0.22 $Å^3$ for validation set compounds, while the quality of the $\Delta H_{vap}$ were similar with 0.11±0.03 kcal/mol for the training set and −0.18±0.15 kcal/mol for the validation set. For alkenes, the overall quality of the optimized LJ parameters was similar or better than CGenFF, where Drude FF was better than CGenFF by 0.02±0.12 $Å^3$ for $V_m$ and 0.12±0.08 kcal/mol for $\Delta H_{vap}$, while similar to CGenFF for dielectric constant (−0.02±0.01) and better by 0.24±0.03 kcal/mol for HFEs. The highest difference in HFE was in the case of alkenes (−0.6 kcal/mol) with cyclohexene. Overall, the condensed phase properties $V_m$ and $\Delta H_{vap}$ or $\Delta H_{sub}$ for the validation set molecules and the dielectric constants and HFEs for all compounds were all close to their experimental values using the optimized set of LJ parameters.

## Conclusion:

Optimization of LJ parameter is a complex multi-variable, multi-objective problem that requires extensive numbers of condensed phase simulations during the optimization process. The LJ parameters, which are limited to one or two atom types specific for a functional group, must be able to reproduce the experimental thermodynamic properties of multiple molecules that contain that functional group. In addition, there is the parameter correlation problem where the LJ parameters on different atom types can compensate for unphysical parameter values in the individual atom types. This issue will be addressed in more detail in a forthcoming manuscript. Finally, there is the broad range of chemical space that needs to be covered by a given force field. In combination these represent a significant challenge. To address such a challenge, we have re-designed the workflow for the LJ parameter optimization in the Drude FF by taking advantage of the sampling power of Latin Hypercube Design (LHD) and the predictive power of Deep Learning to allow for the extensive sampling of LJ parameter space while being able to include condensed phase data into the optimization process. In addition, QM data is used to overcome the parameter correlation problem. Using this approach, we obtained high quality parameters for four groups of molecules representing different functional groups including alkenes, 3- and 4-membered ring compounds and nitriles.

Our method at first utilizes LHD to generate 97 to 200 parameter sets uniformly sampled from the multidimensional LJ parameter space. These parameter sets are used to calculate pure solvent/crystal properties of the training set compounds for each group. The generated data is then used for training the DL model. When selecting such wide ranges of LJ parameters for MD simulations to obtain the condensed phase data for training the DL models, certain combinations of parameters led to unstable systems associated with infeasible parameter sets for the 3 and 4-membered ring compounds, and, therefore, were eliminated from the training data.

The trained DL model is then used in a Brute-Force search algorithm to predict the properties from 10 million sets of $\varepsilon_i$ and $R_{min,i}$ over the atom types being optimized for the training set molecules. From this data top parameter sets are selected based on a weighted error function that includes experimental $V_m$ and $\Delta H_{vap}$ or $\Delta H_{sub}$ condensed phase data and then clustered based on their LJ parameters. The final parameter set out of the 10 million sets is chosen based on good agreement with the experimental data as indicated by the $e_{ROC}$ metric and on the lowest RMSF between the MM and QM minimum interaction energies and distances of the rare gas elements He and Ne with the training set compounds. The final chosen set is then validated by determining the experimental values of the training set molecules to confirm their quality and tested for transferability by testing them on an out-of-training validation set molecules. In addition, the dielectric constant and the HFE of the molecules are determined. The final LJ parameters optimized using the current workflow yielded parameters which reproduced the experimental properties of the training and validation set compounds with an average unsigned error of 3.32±0.94 Å$^3$ for $V_m$, 0.68±0.22 kcal/mol for $\Delta H_{vap}$, 1.28±0.59 for dielectric constant and 0.42±0.18 kcal/mol for HFE, where the uncertainties represent standard error.

The quality of the final parameters and the resulting empirical pure solvent/crystal properties indicated the overall strength of the workflow. The overall agreement of the pure solvent properties of the compounds in Drude FF was improved over the additive CGenFF. In addition, the Drude model can also reproduce the HFE values well as the pure solvent or crystal properties, as seen previously[132], thus emphasizing the importance of the explicit inclusion of polarization in a FF.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements:

Appreciation to Ron Kasl for systems support. Financial support from NIH GM131710 and computational support from the University of Maryland Computer-Aided Drug Design Center and Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. XSEDE access was through the Texas Advanced Computer Center and the San Diego Computer Center and allocation TG-MCA98N017.

## References:

1. Shaw DE; Maragakis P; Lindorff-Larsen K; Piana S; Dror RO; Eastwood MP; Bank JA; Jumper JM; Salmon JK; Shan Y; Wriggers W, Atomic-level characterization of the structural dynamics of proteins. Science 2010, 330 (6002), 341–6. [PubMed: 20947758]

2. Kuzmanic A; Bowman GR; Juarez-Jimenez J; Michel J; Gervasio FL, Investigating cryptic binding sites by molecular dynamics simulations. Acc. Chem. Res 2020, 53 (3), 654–661. [PubMed: 32134250]

3. Pan AC; Xu H; Palpant T; Shaw DE, Quantitative characterization of the binding and unbinding of millimolar drug fragments with molecular dynamics simulations. J. Chem. Theory Comput. 2017, 13 (7), 3372–3377. [PubMed: 28582625]

4. Harpole TJ; Delemotte L, Conformational landscapes of membrane proteins delineated by enhanced sampling molecular dynamics simulations. BBA-Biomembranes 2018, 1860 (4), 909–926. [PubMed: 29113819]

5. Kharche SA; Sengupta D, Dynamic protein interfaces and conformational landscapes of membrane protein complexes. Curr. Opin. Struct. Biol 2020, 61, 191–197. [PubMed: 32036279]

6. Heo L; Feig M, Experimental accuracy in protein structure refinement via molecular dynamics simulations. Proc. Natl. Acad. Sci. U.S.A 2018, 115 (52), 13276–13281. [PubMed: 30530696]

7. Campbell EC; Correy GJ; Mabbitt PD; Buckle AM; Tokuriki N; Jackson CJ, Laboratory evolution of protein conformational dynamics. Curr. Opin. Struct. Biol 2018, 50, 49–57. [PubMed: 29120734]

8. Brooks BR; Bruccoleri RE; Olafson BD; States DJ; Swaminathan S. a.; Karplus M, CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J. Comput. Chem 1983, 4 (2), 187–217.

9. Cornell WD; Cieplak P; Bayly CI; Gould IR; Merz KM; Ferguson DM; Spellmeyer DC; Fox T; Caldwell JW; Kollman PA, A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J. Am. Chem. Soc 1995, 117 (19), 5179–5197.

10. Oostenbrink C; Villa A; Mark AE; Van Gunsteren WF, A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. J. Comput. Chem 2004, 25 (13), 1656–1676. [PubMed: 15264259]

11. Jorgensen WL; Maxwell DS; Tirado-Rives J, Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J. Am. Chem. Soc 1996, 118 (45), 11225–11236.

12. Vanommeslaeghe K; MacKerell AD Jr., CHARMM additive and polarizable force fields for biophysics and computer-aided drug design. Biochim Biophys Acta Gen Subj 2015, 1850 (5), 861–871.

13. Lin F-Y; MacKerell AD Jr., Force Fields for Small Molecules. Methods Mol Biol 2019, 2022, 21–54. [PubMed: 31396898]

14. Lamoureux G; Roux B, Modeling induced polarization with classical Drude oscillators: Theory and molecular dynamics simulation algorithm. J. Chem. Phys 2003, 119 (6), 3025–3039.

15. Lemkul JA; Huang J; Roux B; MacKerell AD Jr., An empirical polarizable force field based on the classical drude oscillator model: development history and recent applications. Chem. Rev 2016, 116 (9), 4983–5013. [PubMed: 26815602]

16. Kunz AP; van Gunsteren WF, Development of a nonlinear classical polarization model for liquid water and aqueous solutions: COS/D. J. Phys. Chem. A 2009, 113 (43), 11570–11579. [PubMed: 19663490]

17. Rick SW, A polarizable, charge transfer model of water using the drude oscillator. J Comput Chem 2016, 37 (22), 2060–6. [PubMed: 27296874]

18. Lamoureux G; MacKerell AD Jr.; Roux B, A simple polarizable model of water based on classical Drude oscillators. J. Chem. Phys 2003, 119 (10), 5185–5197.

19. Lamoureux G; Harder E; Vorobyov IV; Roux B; MacKerell AD Jr., A polarizable model of water for molecular dynamics simulations of biomolecules. Chem. Phys. Lett 2006, 418 (1-3), 245–249.

20. Bauer BA; Patel S, Recent applications and developments of charge equilibration force fields for modeling dynamical charges in classical molecular dynamics simulations. Theor. Chem. Acc 2012, 131 (3), 1–15.

21. Patel S; Davis JE; Bauer BA, Exploring ion permeation energetics in gramicidin A using polarizable charge equilibration force fields. J. Am. Chem. Soc 2009, 131 (39), 13890–13891. [PubMed: 19788320]

22. Patel S; MacKerell AD Jr.; Brooks CL,III, CHARMM fluctuating charge force field for proteins: II protein/solvent properties from molecular dynamics simulations using a nonadditive electrostatic model. J. Comput. Chem 2004, 25 (12), 1504–1514. [PubMed: 15224394]

23. Zhong Y; Patel S, Binding structures of tri-N-acetyl-β-glucosamine in hen egg white lysozyme using molecular dynamics with a polarizable force field. J. Comput. Chem 2013, 34 (3), 163–174. [PubMed: 23109228]

24. Soniat M; Hartman L; Rick SW, Charge transfer models of zinc and magnesium in water. J. Chem. Theory Comput. 2015, 11 (4), 1658–1667. [PubMed: 26574375]

25. Oostenbrink C; van Gunsteren WF, Free energies of ligand binding for structurally diverse compounds. Proc. Natl. Acad. Sci. U.S.A 2005, 102 (19), 6750–6754. [PubMed: 15767587]

26. Rick SW; Stuart SJ; Bader JS; Berne B, Fluctuating charge force fields for aqueous solutions. J. Mol. Liq 1995, 65, 31–40.

27. Rick SW; Berne B, Dynamical fluctuating charge force fields: the aqueous solvation of amides. J. Am. Chem. Soc 1996, 118 (3), 672–679.

28. Bryce RA; Vincent MA; Malcolm NO; Hillier IH; Burton NA, Cooperative effects in the structuring of fluoride water clusters: ab initio hybrid quantum mechanical/molecular mechanical model incorporating polarizable fluctuating charge solvent. J. Chem. Phys 1998, 109 (8), 3077–3085.

29. Yoshii N; Miyauchi R; Miura S; Okazaki S, A molecular-dynamics study of the equation of state of water using a fluctuating-charge model. Chem. Phys. Lett 2000, 317 (3-5), 414–420.

30. Chen B; Xing J; Siepmann JI, Development of polarizable water force fields for phase equilibrium calculations. J. Phys. Chem. A 2000, 104 (10), 2391–2401.

31. Llanta E; Ando K; Rey R, Fluctuating charge study of polarization effects in chlorinated organic liquids. J. Phys. Chem. B 2001, 105 (32), 7783–7791.

32. Wang ZX; Zhang W; Wu C; Lei H; Cieplak P; Duan Y, Strike a balance: optimization of backbone torsion parameters of AMBER polarizable force field for simulations of proteins and peptides. J. Comput. Chem. 2006, 27 (6), 781–790. [PubMed: 16526038]

33. Kaminski GA; Ponomarev SY; Liu AB, Polarizable Simulations with Second-Order Interaction Model Force Field and Software for Fast Polarizable Calculations: Parameters for Small Model Systems and Free Energy Calculations. J. Chem. Theory Comput. 2009, 5 (11), 2935–2943. [PubMed: 20209038]

34. Harder E; Kim B; Friesner RA; Berne BJ, Efficient simulation method for polarizable protein force fields: Application to the simulation of BPTI in liquid water. J. Chem. Theory Comput. 2005, 1 (1), 169–180. [PubMed: 26641127]

35. Shi Y; Xia Z; Zhang J; Best R; Wu C; Ponder JW; Ren P, Polarizable atomic multipole-based AMOEBA force field for proteins. J. Chem. Theory Comput. 2013, 9 (9), 4046–4063. [PubMed: 24163642]

36. Zhang C; Lu C; Jing Z; Wu C; Piquemal J-P; Ponder JW; Ren P, AMOEBA polarizable atomic multipole force field for nucleic acids. J. Chem. Theory Comput. 2018, 14 (4), 2084–2108. [PubMed: 29438622]

37. Elking D; Darden T; Woods RJ, Gaussian induced dipole polarization model. J. Comput. Chem 2007, 28 (7), 1261–1274. [PubMed: 17299773]

38. Harder E; MacKerell AD Jr.; Roux B, Many-body polarization effects and the membrane dipole potential. J. Am. Chem. Soc 2009, 131 (8), 2760–2761. [PubMed: 19199514]

39. Chowdhary J; Harder E; Lopes PE; Huang L; MacKerell AD Jr.; Roux B, A polarizable force field of dipalmitoylphosphatidylcholine based on the classical drude model for molecular dynamics simulations of lipids. J. Phys. Chem. B 2013, 117 (31), 9142–9160. [PubMed: 23841725]

40. Yu H; Whitfield TW; Harder E; Lamoureux G; Vorobyov I; Anisimov VM; MacKerell AD Jr; Roux B, Simulating monovalent and divalent ions in aqueous solution using a Drude polarizable force field. J. Chem. Theory Comput. 2010, 6 (3), 774–786. [PubMed: 20300554]

41. Bedrov D; Piquemal J-P; Borodin O; MacKerell AD Jr.; Roux B; Schröder C, Molecular Dynamics Simulations of Ionic Liquids and Electrolytes Using Polarizable Force Fields. Chemical Reviews 2019, 119 (13), 7940–7995. [PubMed: 31141351]

42. Wang A; Zhang Z; Li G, Higher Accuracy Achieved in the Simulations of Protein Structure Refinement, Protein Folding, and Intrinsically Disordered Proteins Using Polarizable Force Fields. J. Phys. Chem. Lett 2018, 9 (24), 7110–7116. [PubMed: 30514082]

43. Li H; Ngo V; Da Silva MC; Salahub DR; Callahan K; Roux B; Noskov SY, Representation of ion–protein interactions using the drude polarizable force-field. J. Phys. Chem. B 2015, 119 (29), 9401–9416. [PubMed: 25578354]

44. Kamenik AS; Handle PH; Hofer F; Kahler U; Kraml J; Liedl KR, Polarizable and non-polarizable force fields: Protein folding, unfolding, and misfolding. J. Chem. Phys 2020, 153 (18), 185102. [PubMed: 33187403]

45. Li N; Zhu T, Extensive Evaluation of Force Fields for G-Quadruplexes. 2021.

46. Hazel AJ; Walters ET; Rowley CN; Gumbart JC, Folding free energy landscapes of β-sheets with non-polarizable and polarizable CHARMM force fields. J. Chem. Phys 2018, 149 (7), 072317. [PubMed: 30134731]

47. Ren P; Ponder JW, Consistent treatment of inter-and intramolecular polarization in molecular mechanics calculations. J. Comput. Chem 2002, 23 (16), 1497–1506. [PubMed: 12395419]

48. Lemkul JA; Savelyev A; MacKerell AD Jr., Induced polarization influences the fundamental forces in DNA base flipping. J. Phys. Chem. Lett 2014, 5 (12), 2077–2083. [PubMed: 24976900]

49. Lemkul JA; MacKerell AD Jr., Polarizable force field for DNA based on the classical Drude oscillator: II. Microsecond molecular dynamics simulations of duplex DNA. J. Chem. Theory Comput. 2017, 13 (5), 2072–2085. [PubMed: 28398748]

50. Sengul MY; MacKerell AD Jr., Accurate modeling of RNA hairpins through the explicit treatment of electronic polarizability with the classical Drude oscillator force field. J. Comput. Biophys. Chem 2021.

51. Yuan Y; Fu S; Huo D; Su W; Zhang R; Wei J, Multipolar electrostatics for hairpin and pseudoknots in RNA: Improving the accuracy of force field potential energy function. J. Comput. Chem 2021, 42 (11), 771–786. [PubMed: 33586809]

52. Vorobyov IV; Anisimov VM; MacKerell AD Jr., Polarizable empirical force field for alkanes based on the classical drude oscillator model. J. Phys. Chem. B 2005, 109 (40), 18988–18999. [PubMed: 16853445]

53. Li H; Chowdhary J; Huang L; He X; MacKerell AD Jr.; Roux B, Drude polarizable force field for molecular dynamics simulations of saturated and unsaturated zwitterionic lipids. J. Chem. Theory Comput. 2017, 13 (9), 4535–4552. [PubMed: 28731702]

54. Anisimov VM; Vorobyov IV; Roux B; MacKerell AD Jr., Polarizable Empirical Force Field for the Primary and Secondary Alcohol Series Based on the Classical Drude Model. J. Chem. Theory Comput. 2007, 3 (6), 1927–1946. [PubMed: 18802495]

55. Vorobyov I; Anisimov VM; Greene S; Venable RM; Moser A; Pastor RW; MacKerell AD Jr., Additive and classical drude polarizable force fields for linear and cyclic ethers. J. Chem. Theory Comput. 2007, 3 (3), 1120–1133. [PubMed: 26627431]

56. Lopes PE; Lamoureux G; Roux B; MacKerell AD Jr., Polarizable empirical force field for aromatic compounds based on the classical drude oscillator. J. Phys. Chem. B 2007, 111 (11), 2873–2885. [PubMed: 17388420]

57. Lopes PE; Lamoureux G; MacKerell AD Jr., Polarizable empirical force field for nitrogen-containing heteroaromatic compounds based on the classical Drude oscillator. J. Comput. Chem 2009, 30 (12), 1821–1838. [PubMed: 19090564]

58. Harder E; Anisimov VM; Whitfield T; MacKerell AD Jr.; Roux B, Understanding the dielectric properties of liquid amides from a polarizable force field. J. Phys. Chem. B 2008, 112 (11), 3509–3521. [PubMed: 18302362]

59. Zhu X; MacKerell AD Jr, Polarizable empirical force field for sulfur-containing compounds based on the classical Drude oscillator model. J. Comput. Chem 2010, 31 (12), 2330–2341. [PubMed: 20575015]

60. Lin F-Y; MacKerell AD Jr, Polarizable empirical force field for halogen-containing compounds based on the classical Drude oscillator. J. Chem. Theory Comput. 2018, 14 (2), 1083–1098. [PubMed: 29357257]

61. Lin F-Y; MacKerell AD Jr, Improved modeling of halogenated ligand–protein interactions using the drude polarizable and CHARMM additive empirical force fields. J Chem Inf Model 2018, 59 (1), 215–228. [PubMed: 30418023]

62. Allen M; Tildesley D, Computer simulation of liquids: Oxford university press. 1989.

63. Yin D; MacKerell AD Jr., Combined ab initio/empirical approach for optimization of Lennard–Jones parameters. J. Comput. Chem 1998, 19 (3), 334–348.

64. Chen IJ; Yin D; MacKerell AD Jr., Combined ab initio/empirical approach for optimization of Lennard-Jones parameters for polar-neutral compounds. J. Comput. Chem 2002, 23 (2), 199–213. [PubMed: 11924734]

65. Yin D; MacKerell AD Jr., Ab initio calculations on the use of helium and neon as probes of the van der Waals surfaces of molecules. J. Phys. Chem 1996, 100 (7), 2588–2596.

66. Boulanger E; Huang L; Rupakheti C; MacKerell AD Jr.; Roux B, Optimized Lennard-Jones parameters for druglike small molecules. J. Chem. Theory Comput. 2018, 14 (6), 3121–3131. [PubMed: 29694035]

67. Sengul MY; Song Y; Nayir N; Gao Y; Hung Y; Dasgupta T; van Duin AC, INDEEDopt: a deep learning-based ReaxFF parameterization framework. Npj Comput. Mater 2021, 7 (1), 1–9.

68. Joseph VR; Hung Y, Orthogonal-maximin Latin hypercube designs. Stat. Sin 2008, 171–186.

69. Wang J; Wolf RM; Caldwell JW; Kollman PA; Case DA, Development and testing of a general amber force field. J. Comput. Chem 2004, 25 (9), 1157–1174. [PubMed: 15116359]

70. Befort BJ; DeFever RS; Tow GM; Dowling AW; Maginn EJ, Machine Learning Directed Optimization of Classical Molecular Modeling Force Fields. J. Chem. Inf. Model 2021, 61 (9), 4400–4414. [PubMed: 34402301]

71. Turney JM; Simmonett AC; Parrish RM; Hohenstein EG; Evangelista FA; Fermann JT; Mintz BJ; Burns LA; Wilke JJ; Abrams ML; Russ NJ; Leininger ML; Janssen CL; Seidl ET; Allen WD; Schaefer HF; King RA; Valeev EF; Sherrill CD; Crawford TD, Psi4: an open-source ab initio electronic structure program. Wiley Interdiscip. Rev. Comput. Mol. Sci 2012, 2 (4), 556–565.

72. Phillips JC; Braun R; Wang W; Gumbart J; Tajkhorshid E; Villa E; Chipot C; Skeel RD; Kalé L; Schulten K, Scalable molecular dynamics with NAMD. J. Comput. Chem 2005, 26 (16), 1781–1802. [PubMed: 16222654]

73. Kumar A; Yoluk O; MacKerell AD Jr., FFParam: Standalone package for CHARMM additive and Drude polarizable force field parametrization of small molecules. J. Comput. Chem 2020, 41 (9), 958–970. [PubMed: 31886576]

74. Harder E; Anisimov VM; Vorobyov IV; Lopes PEM; Noskov SY; MacKerell AD Jr.; Roux B, Atomic Level Anisotropy in the Electrostatic Modeling of Lone Pairs for a Polarizable Force Field Based on the Classical Drude Oscillator. J. Chem. Theory Comput. 2006, 2 (6), 1587–1597. [PubMed: 26627029]

75. Baker CM; MacKerell AD Jr., Polarizability rescaling and atom-based Thole scaling in the CHARMM Drude polarizable force field for ethers. J Mol Model 2010, 16 (3), 567–76. [PubMed: 19705172]

76. Kumar A; Pandey P; Chatterjee P; MacKerell AD Jr, Deep Neural Network Model to Predict the Electrostatic Parameters in the Polarizable Classical Drude Oscillator Force Field. J. Chem. Theory Comput 2022.

77. Bayly CI; Cieplak P; Cornell W; Kollman PA, A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. J. Phys. Chem. A 1993, 97 (40), 10269–10280.

78. Stone AJ, Distributed multipole analysis: Stability for large basis sets. J. Chem. Theory Comput. 2005, 1 (6), 1128–1132. [PubMed: 26631656]

79. Misquitta AJ; Stone AJ, Distributed polarizabilities obtained using a constrained density-fitting algorithm. J Chem Phys 2006, 124 (2), 024111. [PubMed: 16422575]

80. Heid E; Fleck M; Chatterjee P; Schröder C; MacKerell AD Jr, Toward prediction of electrostatic parameters for force fields that explicitly treat electronic polarization. J. Chem. Theory Comput. 2019, 15 (4), 2460–2469. [PubMed: 30811193]

81. Vanderbilt D; Louie SG, A Monte Carlo simulated annealing approach to optimization over continuous variables. J. Comput. Phys 1984, 56 (2), 259–271.

82. Boys SF; Bernardi F, The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors. Mol. Phys 1970, 19 (4), 553–566.

83. Vanommeslaeghe K; MacKerell AD Jr., Automation of the CHARMM General Force Field (CGenFF) I: bond perception and atom typing. J. Chem. Inf. Model 2012, 52 (12), 3144–3154. [PubMed: 23146088]

84. Vanommeslaeghe K; Raman EP; MacKerell AD Jr., Automation of the CHARMM General Force Field (CGenFF) II: assignment of bonded parameters and partial atomic charges. J. Chem. Inf. Model 2012, 52 (12), 3155–3168. [PubMed: 23145473]

85. Johnson III RD, NIST 101. Computational chemistry comparison and benchmark database. 1999.

86. Frisch M; Trucks G; Schlegel H; Scuseria G; Robb M; Cheeseman J; Montgomery J Jr; Vreven T; Kudin K; Burant J Gaussian 03, Revision C. 02. Wallingford, CT: Gaussian, 2004.

87. Kuczera K; Wiorkiewicz J; Karplus M, MOLVIB: Program for the Analysis of Molecular Vibrations. CHARMM, Harvard University 1993.

88. Jiang W; Hardy DJ; Phillips JC; MacKerell AD Jr.; Schulten K; Roux B, High-performance scalable molecular dynamics simulations of a polarizable force field based on classical Drude oscillators in NAMD. J. Phys. Chem. Lett 2011, 2 (2), 87–92. [PubMed: 21572567]

89. Dill KA; Bromberg S; Stigter D, Molecular driving forces: statistical thermodynamics in biology, chemistry, physics, and nanoscience. Garland Science: 2010.

90. Kubiak R; Janczak J; Śledź M, Crystal structures of 2-and 3-cyanopyridine. J. Mol. Struct 2002, 610 (1-3), 59–64.

91. Allen FH, The Cambridge Structural Database: a quarter of a million crystal structures and rising. Acta Crystallogr., Sect. B: Struct. Sci 2002, 58 (3), 380–388.

92. Bickerton J; Pilcher G; Al-Takhin G, Enthalpies of combustion of the three aminopyridines and the three cyanopyridines. J. Chem. Thermodyn 1984, 16 (4), 373–378.

93. Darden T; York D; Pedersen L, Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems. J. Chem. Phys 1993, 98 (12), 10089–10092.

94. Essmann U; Perera L; Berkowitz ML; Darden T; Lee H; Pedersen LG, A smooth particle mesh Ewald method. J. Chem. Phys 1995, 103 (19), 8577–8593.

95. Steinbach PJ; Brooks BR, New spherical-cutoff methods for long-range forces in macromolecular simulation. J. Comput. Chem 1994, 15 (7), 667–683.

96. Lagüe P; Pastor RW; Brooks BR, Pressure-based long-range correction for Lennard-Jones interactions in molecular dynamics simulations: application to alkanes and interfaces. J. Phys. Chem. B 2004, 108 (1), 363–368.

97. Ryckaert J-P; Ciccotti G; Berendsen HJ, Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J. Comput. Phys 1977, 23 (3), 327–341.

98. Deng Y; Roux B, Hydration of amino acid side chains: Nonpolar and electrostatic contributions calculated from staged molecular dynamics free energy simulations with explicit water molecules. J. Phys. Chem. B 2004, 108 (42), 16567–16576.

99. Becker OM; MacKerell AD Jr.; Roux B; Watanabe M, Computational Biochemistry and Biophysics. Crc Press: 2001.

100. Kollman P, Free energy calculations: applications to chemical and biochemical phenomena. Chem. Rev 1993, 93 (7), 2395–2417.

101. Kognole AA; Aytenfisu AH; MacKerell AD Jr., Balanced polarizable Drude force field parameters for molecular anions: phosphates, sulfates, sulfamates, and oxides. J. Mol. Model 2020, 26, 1–11.

102. Lemkul JA; MacKerell AD Jr, Balancing the interactions of Mg2+ in aqueous solution and with nucleic acid moieties for a polarizable force field based on the classical Drude oscillator model. J. Phys. Chem. B 2016, 120 (44), 11436–11448. [PubMed: 27759379]

103. Savelyev A; MacKerell AD Jr., Balancing the interactions of ions, water, and DNA in the Drude polarizable force field. J. Phys. Chem. B 2014, 118 (24), 6742–6757. [PubMed: 24874104]

104. Riahi S; Rowley CN, Solvation of hydrogen sulfide in liquid water and at the water–vapor interface using a polarizable force field. J. Phys. Chem. B 2014, 118 (5), 1373–1380. [PubMed: 24498909]

105. Baker CM; Best RB, Matching of additive and polarizable force fields for multiscale condensed phase simulations. J. Chem. Theory Comput. 2013, 9 (6), 2826–2837. [PubMed: 23997691]

106. Weeks JD; Chandler D; Andersen HC, Perturbation theory of the thermodynamic properties of simple liquids. J. Chem. Phys 1971, 55 (11), 5422–5423.

107. Kirkwood JG, Statistical mechanics of fluid mixtures. J. Chem. Phys 1935, 3 (5), 300–313.

108. Kumar S; Rosenberg JM; Bouzida D; Swendsen RH; Kollman PA, The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. J. Comput. Chem 1992, 13 (8), 1011–1021.

109. Feurer M; Klein A; Eggensperger K; Springenberg JT; Blum M; Hutter F, Auto-sklearn: efficient and robust automated machine learning. In Automated Machine Learning, Springer, Cham: 2019; pp 113–134.

110. Gulli A; Pal S, Deep learning with Keras. Packt Publishing Ltd: 2017.

111. Ramachandran P; Zoph B; Le QV, Searching for Activation Functions. ArXiv 2018, abs/1710.05941.

112. Kingma DP; Ba J, Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 2014.

113. Prechelt L, Early stopping-but when? In Neural Networks: Tricks of the trade, Springer: 1998; pp 55–69.

114. Schaeffer J; Lu P; Szafron D; Lake R In A re-examination of brute-force search, Proceedings of the AAAI Fall Symposium on Games: Planning and Learning, 1993; pp 51–58.

115. Trakhtenbrot BA, A survey of Russian approaches to perebor (brute-force searches) algorithms. IEEE Ann. Hist. Comput 1984, 6 (4), 384–400.

116. Barron FH; Barrett BE, Decision quality using ranked attribute weights. Manage. Sci 1996, 42 (11), 1515–1523.

117. Klauda JB; Brooks BR; MacKerell AD Jr.; Venable RM; Pastor RW, An ab initio study on the torsional surface of alkanes and its effect on molecular simulations of alkanes and a DPPC bilayer. J. Phys. Chem. B 2005, 109 (11), 5300–5311. [PubMed: 16863197]

118. Klauda JB; Kučerka N; Brooks BR; Pastor RW; Nagle JF, Simulation-based methods for interpreting x-ray data from lipid bilayers. Biophys. J 2006, 90 (8), 2796–2807. [PubMed: 16443652]

119. Klauda JB; Brooks BR; Pastor RW, Dynamical motions of lipids and a finite size effect in simulations of bilayers. J. Chem. Phys 2006, 125 (14), 144710. [PubMed: 17042634]

120. Kantonen SM; Muddana HS; Schauperl M; Henriksen NM; Wang LP; Gilson MK, Data-Driven Mapping of Gas-Phase Quantum Calculations to General Force Field Lennard-Jones Parameters. J Chem Theory Comput 2020, 16 (2), 1115–1127. [PubMed: 31917572]

121. Mobley DL; Bannan CC; Rizzi A; Bayly CI; Chodera JD; Lim VT; Lim NM; Beauchamp KA; Slochower DR; Shirts MR; Gilson MK; Eastman PK, Escaping Atom Types in Force Fields Using Direct Chemical Perception. J Chem Theory Comput 2018, 14 (11), 6076–6092. [PubMed: 30351006]

122. Schauperl M; Kantonen S; Wang LP; Gilson MK, Data-driven analysis of the number of Lennard-Jones types needed in a force field. Commun Chem 2020, 3 (1).

123. LeCun Y; Bengio Y; Hinton G, Deep learning. Nature 2015, 521 (7553), 436–444. [PubMed: 26017442]

124. Gillogly JJ, The technology chess program. Artif. Intell 1972, 3, 145–163.

125. Kiktenko EO; Kudinov MA; Fedorov AK In Detecting brute-force attacks on cryptocurrency wallets, International Conference on Business Information Systems, Springer: 2019; pp 232–242.

126. Dobson JF, Beyond pairwise additivity in London dispersion interactions. Int. J. Quantum Chem. 2014, 114 (18), 1157–1161.

127. Kronik L; Tkatchenko A, Understanding molecular crystals with dispersion-inclusive density functional theory: pairwise corrections and beyond. Acc. Chem. Res 2014, 47 (11), 3208–3216. [PubMed: 24901508]

128. Hapka M; Krzemińska A; Pernal K, How Much Dispersion Energy Is Included in the Multiconfigurational Interaction Energy? J. Chem. Theory Comput. 2020, 16 (10), 6280–6293. [PubMed: 32877179]

129. Reilly AM; Tkatchenko A, van der Waals dispersion interactions in molecular materials: beyond pairwise additivity. Chem. Sci 2015, 6 (6), 3289–3301. [PubMed: 28757994]

130. Rufa DA; Bruce Macdonald HE; Fass J; Wieder M; Grinaway PB; Roitberg AE; Isayev O; Chodera JD, Towards chemical accuracy for alchemical free energy calculations with hybrid physics-based machine learning / molecular mechanics potentials. In bioRxiv, 2020; p 2020.07.29.227959.

131. Mobley DL; Graves AP; Chodera JD; McReynolds AC; Shoichet BK; Dill KA, Predicting absolute ligand binding free energies to a simple model site. J Mol Biol 2007, 371 (4), 1118–1134. [PubMed: 17599350]

132. Rupakheti CR; MacKerell AD Jr.; Roux B, Global Optimization of the Lennard-Jones Parameters for the Drude Polarizable Force Field. J Chem Theory Comput 2021, 17 (11), 7085–7095. [PubMed: 34609863]

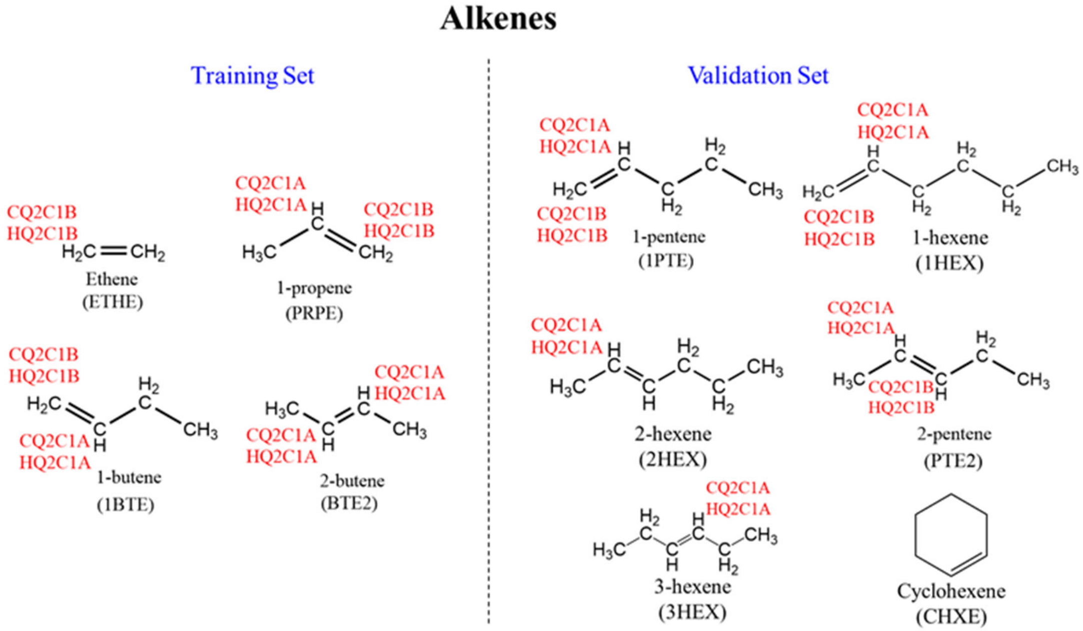


**Figure 1:**

Training and validation set compounds for alkenes along with the atom types included in the optimization represented by each compound. Figure SII of SI_1 presents the structures of molecules in rest of the three groups.

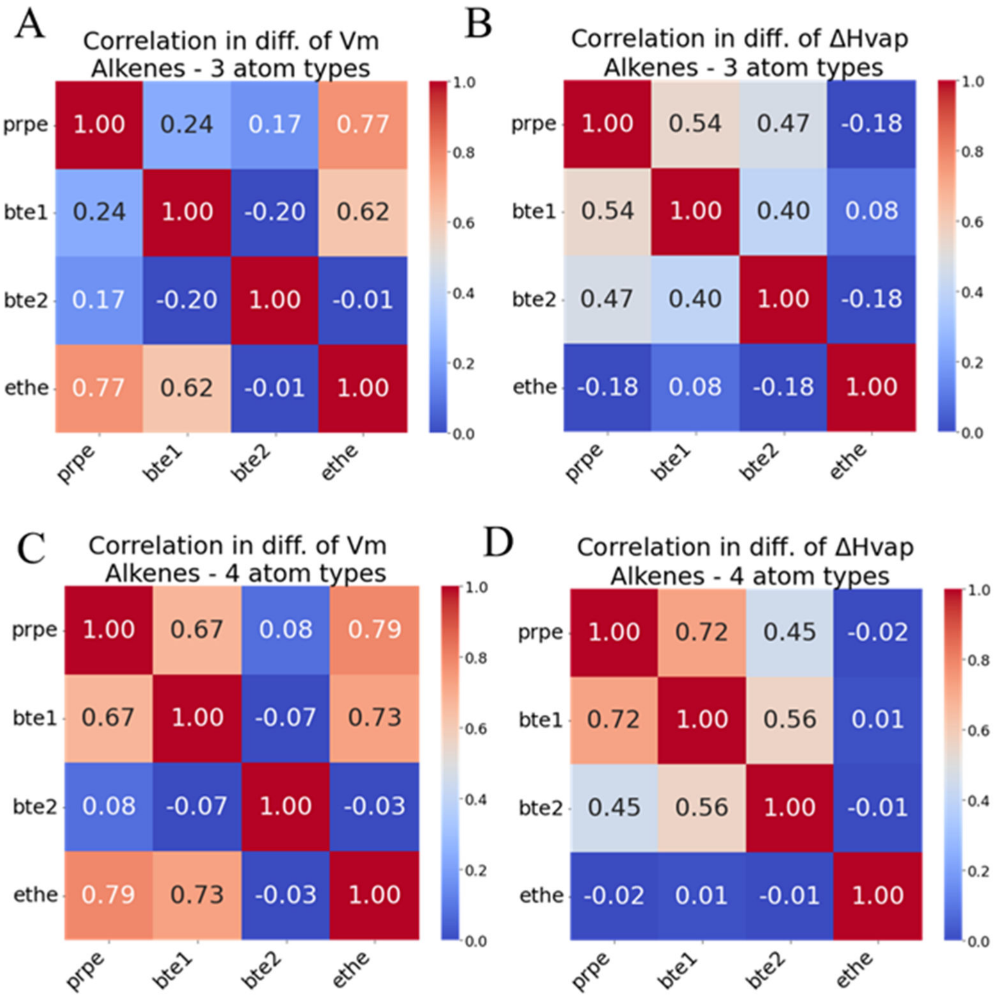


**Figure 2:**

Correlation heatmaps of the differences between the calculated and experimental values of $V_m$ (A and C) and $\Delta H_{vap}$ (B and D) of the alkene training set molecules based on the set of LJ parameters selected by LHD during DL model development. Results are shown for 3 atom types (A and B) and for 4 atom types (C and D). (Molecule abbreviations: ethe - ethene, prpe - propene, bte1 – 1-butene and bte2 – 2-butene)

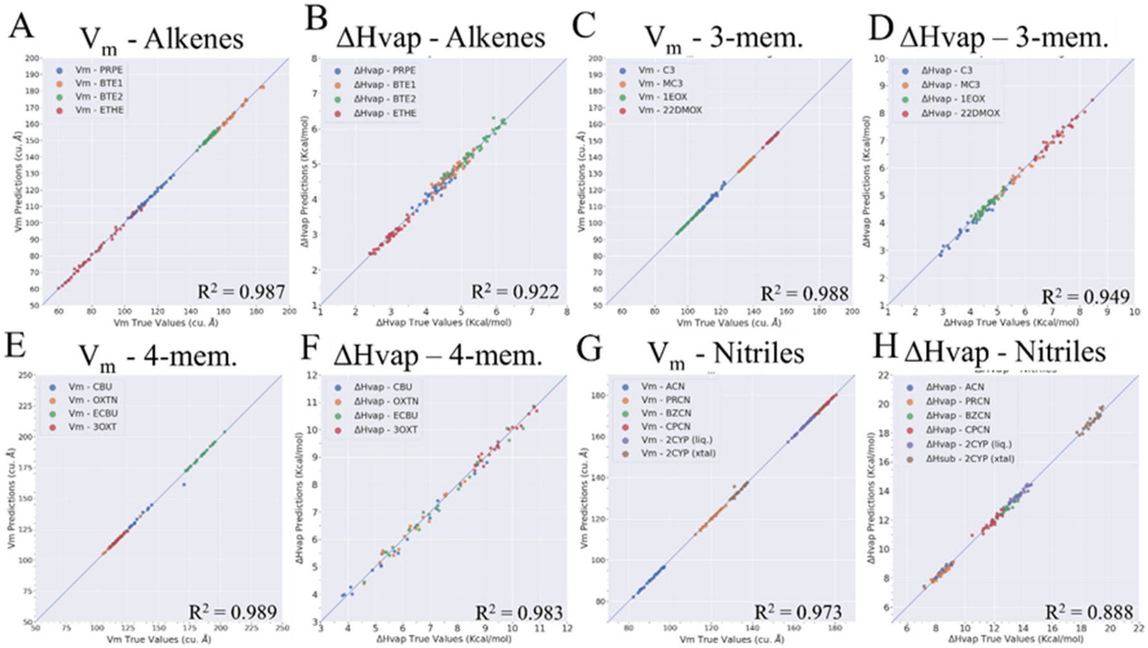


**Figure 3:**

Correlation plots for MD-based true vs. DL-based predicted properties for of all target properties in the test split from the 5-fold cross validation applied during DL training. A & B) alkenes model; C & D) 3-membered ring compound model; E & F) 4-membered ring compound model; G & H – Nitriles model, where the correlations of $V_m$ and $\Delta H_{vap}/\Delta H_{sub}$ are depicted individually.

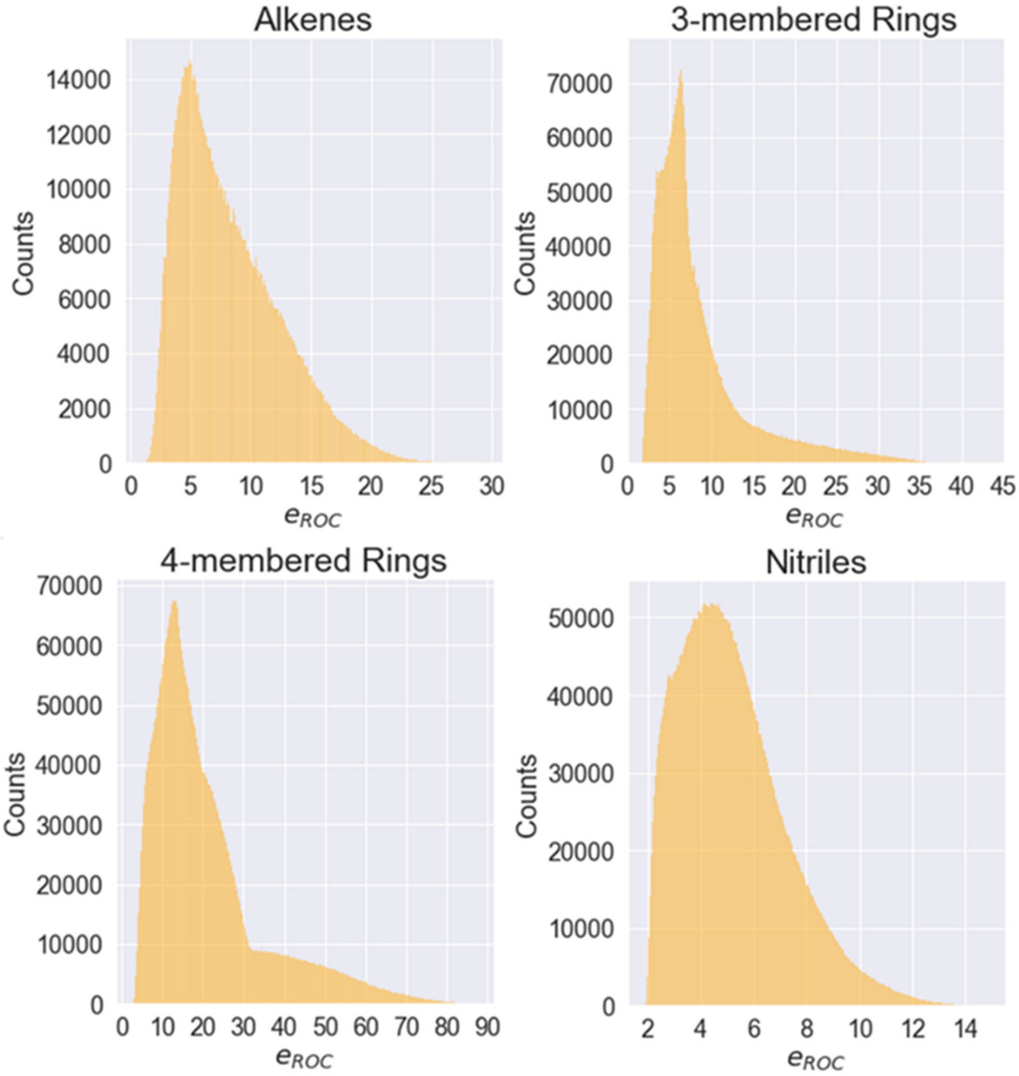
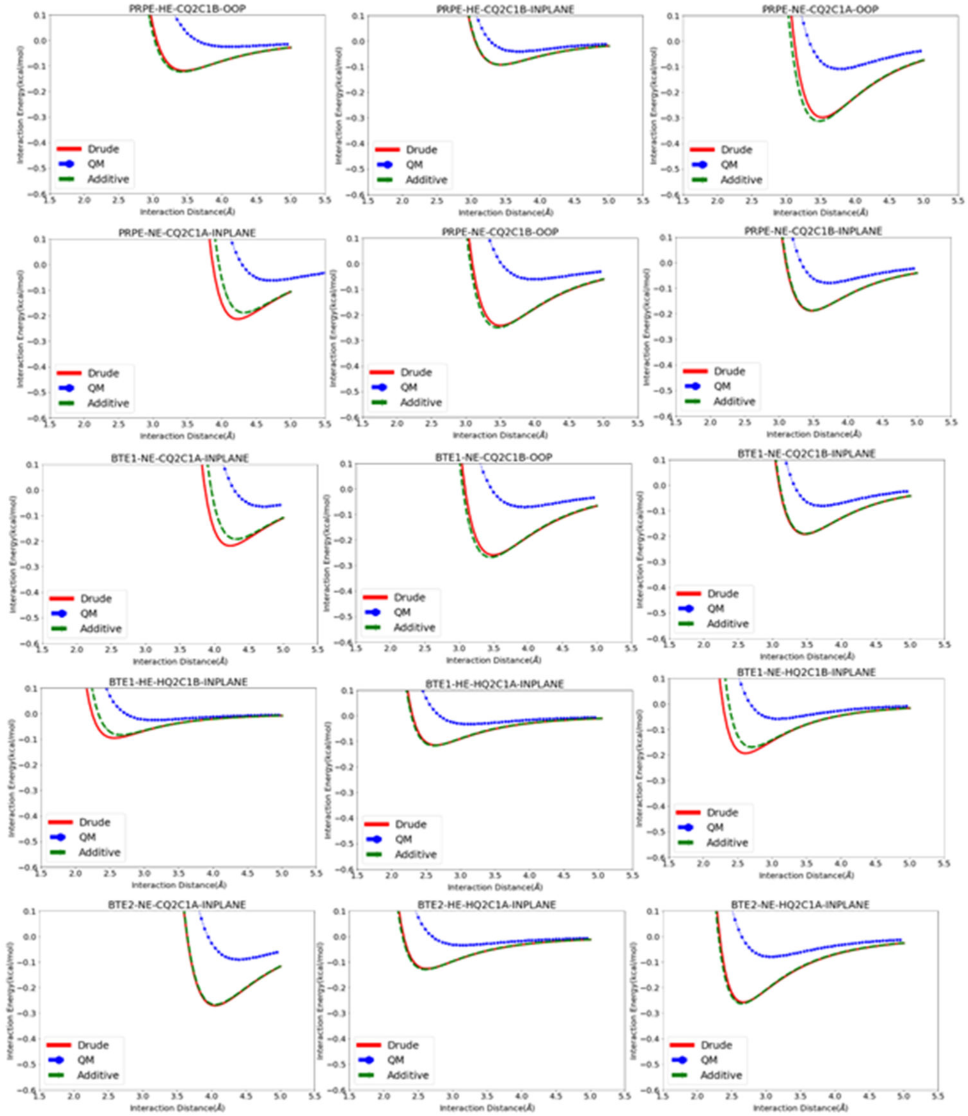
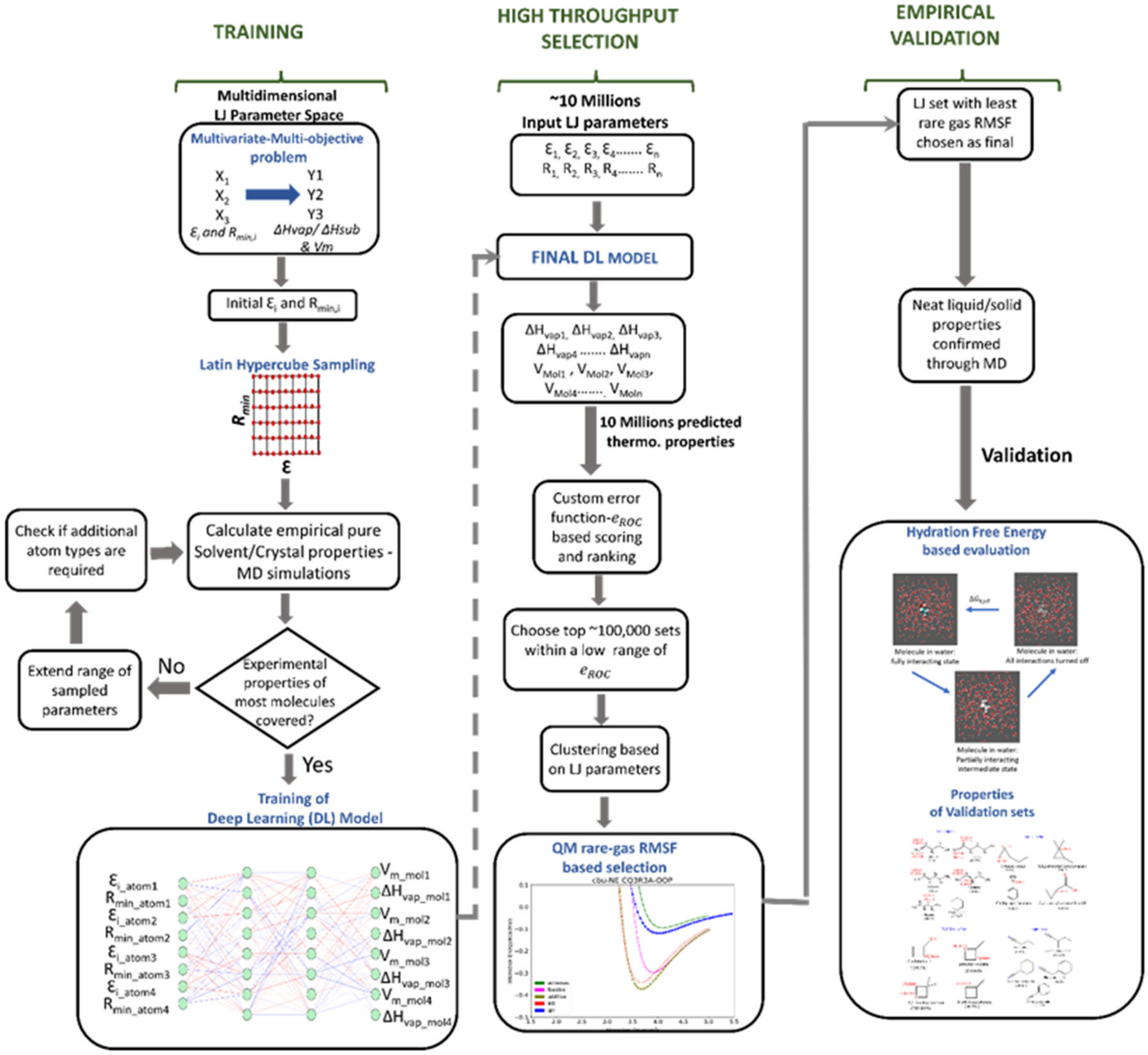


**Figure 4:**

Histograms showing probability distributions of the error $e_{ROC}$ of the brute-force scan data for each functional group.



Potential energy surfaces of the rare gas-alkene model compound interactions for 1-propene (prpe), 1-butene (1-bte) and 2-butene (bte2), at different angles of interaction: in-plane and out-of-plane (OOP) with Helium (He) and Neon (Ne) for the QM, final Drude FF and additive CGenFF model chemistries.


TRAINING
HIGH THROUGHPUT SELECTION
EMPIRICAL VALIDATION
Multidimensional LJ Parameter Space
Multivariate-Multi-objective problem
X1 X2 X3 ε_i and R_min,i
Y1 Y2 Y3 ΔHvap/ ΔHsub & Vm
Initial ε_i and R_min,i
Latin Hypercube Sampling
R_min
ε
Check if additional atom types are required
Calculate empirical pure Solvent/Crystal properties - MD simulations
Extend range of sampled parameters
No
Experimental properties of most molecules covered?
Yes
Training of Deep Learning (DL) Model
ε_i_atom1 R_min_atom1 ε_i_atom2 R_min_atom2 ε_i_atom3 R_min_atom3 ε_i_atom4 R_min_atom4
V_m_mol1 ΔH_vap_mol1 V_m_mol2 ΔH_vap_mol2 V_m_mol3 ΔH_vap_mol3 V_m_mol4 ΔH_vap_mol4
~10 Millions Input LJ parameters
ε1, ε2, ε3, ε4....... εn
R1, R2, R3, R4....... Rn
FINAL DL MODEL
ΔHvap1, ΔHvap2, ΔHvap3, ΔHvap4 ....... ΔHvapn
VMol1, VMol2, VMol3, VMol4........ VMoln
10 Millions predicted thermo. properties
Custom error function-e_ROC based scoring and ranking
Choose top ~100,000 sets within a low range of e_ROC
Clustering based on LJ parameters
QM rare-gas RMSF based selection
cbu-NE CG3R3A-OOP
LJ set with least rare gas RMSF chosen as final
Neat liquid/solid properties confirmed through MD
Validation
Hydration Free Energy based evaluation
ΔG_hyd
Molecule in water: fully interacting state
Molecule in water: All interactions turned off
Molecule in water: Partially interacting intermediate state
Properties of Validation sets

Schematic of the new process of optimization of LJ parameters in Drude Polarizable Force Field. Dotted arrow represents use of the same model after training, solid arrows represent continuity to the next steps.

**Table 1:**

Hyperparameters used for model optimization

| Hyperparameter | Range of values | Final Hyperparameter |
|---|---|---|
| Number of hidden layers | 2,3,4,6,8,10 | 2 |
| Number of nodes in each layer | 2,4,6,8,10,12,14 | Number of output thermodynamic properties |
| Learning Rate | 0.1,0.05,0.005,0.001 | 0.005 |
| Batch Size | 2,4,6,8,10 | 6 |
| Activation function | ReLU, Swish | Swish |

**Table 2:**

Parameter ranges used for generating parameter-based clusters in alkenes.

| **Range** | $\varepsilon_i^{HQ2C1A}$ | $R_{min,i}^{HQ2C1A}$ | $\varepsilon_i^{HQ2C1B}$ | $R_{min,i}^{HQ2C1B}$ | $\varepsilon_i^{CQ2C1A}$ | $R_{min,i}^{CQ2C1A}$ | $\varepsilon_i^{CQ2C1B}$ | $R_{min,i}^{CQ2CB}$ |
|---|---|---|---|---|---|---|---|---|
| Range A | < −0.0350 | < 1.1000 | < −0.0350 | < 1.1000 | < −0.0675 | < 1.8000 | < −0.0675 | < 1.8000 |
| Range B | −0.0350 to −0.0290 | 1.10 to 1.35 | −0.0351 to −0.0290 | 1.10 to 1.35 | −0.0674 to −0.0555 | 1.8 to 2.1 | −0.0675 to −0.0555 | 1.8 to 2.1 |
| Range C | > −0.0290 | > 1.3500 | > −0.0290 | > 1.3500 | > −0.0555 | > 2.1 | > −0.0555 | > 2.1 |

**Table 3:**

The initial and the range of parameters for all atom types of alkenes, used in training the DL LJ model. $\varepsilon_i$ in Kcal/mol and $R_{min,i}$ in Å.

| **Group** | **Alkenes** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Atom Types** | **CQ2C1A** | | **CQ2C1B** | | **HQ2C1A** | | **HQ2C1B** | |
| **LJ Parameters** | $\varepsilon_i$ | $R_{min,i}$ | $\varepsilon_i$ | $R_{min,i}$ | $\varepsilon_i$ | $R_{min,i}$ | $\varepsilon_i$ | $R_{min,i}$ |
| **Initial** | −0.066 | 2.07 | −0.066 | 2.07 | −0.034 | 1.2099 | −0.034 | 1.2099 |
| **Lower Limit** | −0.0759 | 1.7595 | −0.0759 | 1.7595 | −0.0391 | 1.0284 | −0.0391 | 1.0284 |
| **Upper Limit** | −0.0561 | 2.3805 | −0.0561 | 2.3805 | −0.0289 | 1.3914 | −0.0289 | 1.3914 |
| **Range** | 0.0198 | 0.621 | 0.0198 | 0.621 | 0.0102 | 0.363 | 0.0102 | 0.363 |
| **Sampling Resolution** | 0.0001 | 0.0031 | 0.0001 | 0.0031 | 0.0001 | 0.0018 | 0.0001 | 0.0018 |
| **No. of points** | 198 | 200 | 198 | 200 | 102 | 202 | 102 | 202 |
| **Total number of possible sets** | | 6.66E+17 | | | | | | |
| **Number of LHD generated sets** | | 200 | | | | | | |

## Table 4:

Pure solvent properties of alkenes training set molecules, using initial LJ and the best LJ from the training data selected by Latin Hypercube Design.

| Alkenes Training Set | | | | | |
|---|---|---|---|---|---|
| **Molecule** | **Description** | **Molecular Volume (cu. Å)** | | **ΔHvap (Kcal/mol)** | |
| | | **Diff** | **%Diff** | **Diff** | **%Diff** |
| Ethene | Initial | −3.79 | −4.84% | 1.36 | 28.87% |
| | Best of training | 2.61 | 3.08 | −0.08 | −2.45 |
| Propene | Initial | −1.96 | −1.74% | 0.1 | 2.22% |
| | Best of training | −0.42 | −0.37% | 0.03 | 4.74% |
| 1-butene | Initial | 2.00 | 1.25% | −0.11 | −2.33% |
| | Best of training | 4.74 | 2.91% | 0.05 | 1.02% |
| 2-butene | Initial | −1.98 | −1.29% | −0.20 | −3.82% |
| | Best of training | −4.35 | −2.88% | −0.28 | −5.43% |

**Table 5:**

Comparison of the eroc and RMSF metrics along with the average differences and percent difference between the calculated and experimental pure solvent condensed phase properties for the LJ parameters from the initial guess, the best LJ parameters selected by Latin Hypercube Design and from the DL Brute-Force sampling based on the $e_{ROC}$ metric alone and based on both the $e_{ROC}$ and RMSF metrics. Averages are over the training set compounds in each group.

| **Group** | **Description** | **$e_{ROC}$** | **RMSF** | **Vm Diff.** | **Vm** | **ΔHvap Diff.** | **ΔHvap** |
|---|---|---|---|---|---|---|---|
| | | | | **(cu. Å)** | **% Diff.** | **(kcal/mol)** | **% Diff.** |
| **Alkenes** | **Initial** | NA | NA | 2.43±0.45 | 2.28±0.86 | 0.44±0.31 | 9.31±6.53 |
| | **Best of LHD** | NA | NA | 3.03±0.99 | 2.31±0.65 | 0.11±0.06 | 3.41±1.02 |
| | **$e_{ROC}$ selected** | 0.8500 | 0.0818 | 1.01±0.56 | 0.67±0.34 | 0.10±0.05 | 2.24±1.21 |
| | **$e_{ROC}$ /RMSF selected (final)** | 2.5100 | 0.0752 | 1.31±0.23 | 1.20±0.43 | 0.18±0.08 | 5.07±2.99 |
| **3 mem. Cyclic** | **Initial** | NA | NA | 13.69±5.77 | 10.19±3.90 | 1.19±0.37 | 29.06±16.60 |
| | **Best of LHD** | NA | NA | 0.38±0.27 | 0.36±0.28 | 0.51±0.22 | 8.72±4.32 |
| | **$e_{ROC}$ selected** | 1.6725 | 0.1937 | 0.28±0.10 | 0.24±0.09 | 0.57±0.27 | 10.48±5.88 |
| | **$e_{ROC}$ /RMSF selected (final)** | 2.1750 | 0.1804 | 0.46±0.13 | 0.40±0.14 | 0.58±0.26 | 6.17±0.48 |
| **4 mem. Cyclic** | **Initial** | NA | NA | 12.12±2.86 | 9.22±2.67 | 2.44±1.30 | 26.54±6.14 |
| | **Best of LHD** | NA | NA | 3.82±0.80 | 2.79±0.21 | 0.50±0.15 | 6.95±1.64 |
| | **$e_{ROC}$ selected** | 3.8225 | 0.1478 | 1.26±0.85 | 0.76±0.44 | 0.89±0.17 | 15.23±5.05 |
| | **$e_{ROC}$ /RMSF selected (final)** | 3.8900 | 0.1447 | 1.53±0.47 | 1.11±0.28 | 0.95±0.19 | 16.46±5.15 |
| **Nitriles** | **Initial** | NA | NA | 1.37±0.35 | 1.10±0.33 | 0.96±0.28 | 0.42±0.56 |
| | **Best of LHD** | NA | NA | 1.38±0.31 | 0.86±0.19 | 0.61±0.15 | 4.79±0.91 |
| | **$e_{ROC}$ selected** | 1.8887 | 0.0844 | 1.13±0.32 | 0.82±0.21 | 0.53±0.31 | 3.79±1.65 |
| | **$e_{ROC}$ /RMSF selected (final)** | 1.8887 | 0.0844 | 1.13±0.32 | 0.82±0.21 | 0.53±0.31 | 3.79±1.65 |

## Table 6:

Thermodynamic properties (Vm, ΔHvap, dielectric constant) and Hydration Free Energies of the Alkenes. NA indicates that experimental data is not available. Differences in molecular volumes - Vm in Å$^3$ and enthalpies of vaporization and sublimation in Kcal/mol.

| Training Set | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Molecule** | **Force Field** | **Vm** | | **ΔHvap** | | **Dielectric Constant** | | **Hydration Free Energy** | |
| | | **Diff** | **%Diff** | **Diff** | **%Diff** | **Diff** | **%Diff** | **Diff** | **%Diff** |
| Ethene | Additive | 1.14 | 1.37% | −0.11 | −3.40% | NA | NA | −0.30 | −30.00% |
| | Drude Final | −1.97 | −2.46% | −0.41 | −13.95% | NA | NA | 0.04 | 2.83% |
| Propene | Additive | 1.19 | 1.03% | −0.04 | −0.92% | −0.06 | −2.80% | −0.15 | −13.04% |
| | Drude Final | −1.05 | −0.92% | 0.09 | 2.00% | 0.27 | 11.20% | −0.21 | −19.01% |
| 1-butene | Additive | 4.78 | 2.93% | −0.12 | −2.55% | −0.28 | −14.89% | −0.26 | −22.81% |
| | Drude Final | 1.21 | 0.76% | −0.15 | −3.21% | −0.02 | −0.93% | 0.35 | 20.13% |
| 2-butene | Additive | −5.65 | −3.77% | 0.05 | 0.91% | −0.19 | −10.80% | NA | NA |
| | Drude Final | −0.99 | −0.64% | −0.06 | −1.12% | −0.20 | −11.43% | NA | NA |
| **Validation Set** | | | | | | | | | |
| 1-pentene | Additive | 4.02 | 2.16% | 0.01 | 0.16% | −0.13 | −7.20% | −0.68 | −67.22% |
| | Drude Final | 2.88 | 3.20% | 0.22 | 3.49% | 0.26 | 11.37% | −0.1 | −6.38% |
| 1-hexene | Additive | 2.68 | 1.27% | −0.14 | −1.95% | −0.14 | −7.54% | 0.00 | 0.00% |
| | Drude Final | 1.14 | 2.04% | −0.11 | −1.53% | −0.04 | −1.93% | −0.02 | −1.19% |
| 2-pentene | Additive | −0.36 | −0.20% | 0.46 | 6.73% | NA | NA | NA | NA |
| | Drude Final | 4.48 | 2.00% | 0.16 | 2.45% | NA | NA | NA | NA |
| 2-hexene | Additive | −1.98 | −0.96% | 0.29 | 3.70% | −0.23 | −12.97% | NA | NA |
| | Drude Final | 3.77 | 1.50% | 0.12 | −0.91% | 0.10 | 3.48% | NA | NA |
| 3-hexene | Additive | −1.01 | −0.50% | 0.73 | 8.85% | −0.19 | −10.80% | NA | NA |
| | Drude Final | 5.31 | 2.52% | 1.25 | 14.29% | −0.19 | −10.80% | NA | NA |
| Cyclohexene | Additive | 4.52 | 2.62% | 0.38 | 4.59% | NA | NA | −0.07 | −24.20% |
| | Drude Final | 8.60 | 4.87% | −0.85 | −11.90% | NA | NA | −0.63 | 244.68% |

## Table 7:

Pure solvent and crystal properties averaged over the four groups. Molecular volumes, Vm in $\text{Å}^3$ and enthalpies of vaporization and sublimation in Kcal/mol. Differences and percent differences are unsigned, reported uncertainties represent standard error values.

| Group | Force Field | Training Set | | | | Validation Set | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Vm Diff. | Vm % Diff. | ΔHvap Diff. | ΔHvap % Diff. | Vm Diff. | Vm % Diff. | ΔHvap Diff. | ΔHvap % Diff |
| **Alkenes** | **Additive** | 3.19±1.18 | 2.27±0.65 | 0.08±0.02 | 1.94±0.62 | 2.43±0.91 | 1.29±0.47 | 0.34±0.13 | 4.31±1.59 |
| | **Drude** | 1.31±0.23 | 1.20±0.43 | 0.18±0.08 | 5.57±2.86 | 4.36±1.07 | 2.69±0.61 | 0.45±0.28 | 5.87±3.24 |
| **3 mem. Cyclic** | **Additive** | 2.80±1.73 | 2.73±1.93 | 0.59±0.13 | 10.48±2.88 | 6.76±0.30 | 3.91±0.32 | 0.97±0.29 | 9.79±2.20 |
| | **Drude** | 0.46±0.13 | 0.40±0.14 | 0.58±0.26 | 6.17±0.48 | 9.57±2.61 | 5.55±1.40 | 0.70±0.29 | 7.29±3.77 |
| **4 mem. Cyclic** | **Additive** | 5.62±1.19 | 4.19±0.84 | 0.84±0.14 | 10.90±1.53 | 6.91±1.63 | 4.64±0.97 | 1.23±0.36 | 13.87±1.53 |
| | **Drude** | 1.53±0.47 | 1.11±0.28 | 0.95±0.19 | 16.46±5.15 | 5.87±1.20 | 4.29±1.25 | 1.05±0.07 | 12.69±2.10 |
| **Nitriles** | **Additive** | 2.99±1.94 | 2.02±1.34 | 0.82±0.34 | 5.59±1.75 | 2.83±2.00 | 1.96±1.42 | 1.80±0.95 | 12.35±5.36 |
| | **Drude** | 1.13±0.32 | 0.82±0.21 | 0.53±0.31 | 3.79±1.65 | 2.29±1.33 | 1.25±0.63 | 1.01±0.32 | 7.68±1.93 |

## Table 8:

Hydration Free Energy (HFE) and dielectric constants, averaged over the four groups. HFE in Kcal/mol, where uncertainties represent standard error values.

| Group | Force Field | Training Set | | | | Validation Set | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | HFE Diff. | HFE % Diff. | Dielectric Constant Diff. | Dielectric Constant % Diff. | HFE Diff. | HFE % Diff. | Dielectric Constant Diff. | Dielectric Constant % Diff. |
| Alkenes | Additive | 0.24±0.04 | 21.95±4.26 | 0.18±0.06 | 8.85±2.68 | 0.25±0.19 | 30.51±17.0 | 0.17±0.02 | 10.51±1.20 |
| | Drude | 0.20±0.09 | 14.08±5.55 | 0.16±0.06 | 8.32±3.22 | 0.25±0.19 | 83.19±79.5 | 0.15±0.05 | 6.56±2.28 |
| 3 mem. rings | Additive | 1.11±0.36 | 49.18±2.26 | 17.00±0.0[*] | 57.78±0.0[*] | NA | NA | NA | NA |
| | Drude | 0.29±0.17 | 22.95±14.37 | 0.72±0.0[*] | 5.48±0.0[*] | NA | NA | NA | NA |
| 4 mem. rings | Additive | NA | NA | 0.22±0.0[*] | 12.29±0.0[*] | NA | NA | NA | NA |
| | Drude | NA | NA | 0.19±0.0[*] | 10.39±0.0[*] | NA | NA | NA | NA |
| Nitriles | Additive | 0.32±0.08 | 8.95±2.79 | 16.00±0.27 | 151.93±32.31 | 1.20±0.41 | 19.64±1.21 | 12.36±0 | 99.12±0 |
| | Drude | 0.15±0.05 | 3.98±1.20 | 4.04±2.29 | 23.75±12.81 | 0.44±0.21 | 7.90±3.03 | 8.67±0 | 53.65±0 |

[*] Indicates data from only one molecule (due to non-availability of experimental data for others).NA indicates non-availability of experimental data.