# Development of Scalable and Generalizable Machine Learned Force Field for Polymers

Shaswat Mohanty[a], James Stevenson[b], Andrea Browning[c], Leif Jacobson[c], Karl Leswing[b], Mathew D. Halls[d], Mohammad Atif Faiz Afzal[b,*]

[a]*Department of Mechanical Engineering, Stanford University, California 94305-4040, USA*
[b]*Schrödinger, Inc., New York 10036, United States*
[c]*Schrödinger, Inc., Portland, Oregon 97204, United States*
[d]*Schrödinger, Inc., San Diego, California 92121, United States*

**Abstract**

Understanding and predicting the properties of polymers is vital to developing tailored polymer molecules for desired applications. Classical force fields may fail to capture key properties, for example, the transport properties of certain polymer systems such as polyethylene glycol. As a solution, we present an alternative potential energy surface, a charge recursive neural network (QRNN) model trained on DFT calculations made on smaller atomic clusters that generalizes well to oligomers comprising larger atomic clusters or longer chains. We demonstrate the validity of the polymer QRNN workflow by modeling the oligomers of ethylene glycol. We apply two rounds of active learning (addition of new training clusters based on current model performance) and implement a novel model training approach that uses partial charges from a semi-empirical method. Our developed QRNN model for polymers produces stable molecular dynamics (MD) simulation trajectory and captures the dynamics of polymer chains as indicated by the striking agreement with experimental values. Our model allows working on much larger systems than allowed by DFT simulations, at the same time providing a more accurate force field than classical force fields which provides a promising avenue for large-scale molecular simulations of polymeric systems.

*Keywords:* Neural Network Potential, Molecular Dynamics, Dynamics Prediction, Machine-learned Force Field, Charge Recursive Neural Network, OPLS4

## 1. Introduction

Machine learning and deep learning techniques to bridge computational shortcomings have been explored in numerous fields of computational materials science [1–4] and chemistry [5–9]. In particular, there has been a recent foray into the field of molecular dynamics (MD) to (i) lengthen the time or length scale associated with ab-initio

*Corresponding author
*Email address:* atif.afzal@schrodinger.com (Mohammad Atif Faiz Afzal)

molecular dynamics, and (ii) to improve the accuracy of classical force fields. In the current work, we will primarily focus on the properties with limited accuracy that are associated with classical force fields. To address this issue, *ab initio* molecular dynamics (AIMD) methods were developed [10, 11], primarily using density functional theory (DFT) [12]. While these methods accurately capture molecular interactions, the computational expense associated with its scalability restrict us to short time scales (around $10^1$ picoseconds) and very small systems (around $10^2$ atoms), which complicates the study of time-dependent dynamic properties of systems of large molecules such as polymers [13, 14].

There have been recent advancements in machine-learned force fields [15, 16] to capture complex behavior such as polarization and chemical reactivity [17]. Machine-learned force fields can reproduce quantum chemical calculations in both finite and extended systems to well within chemical accuracy. These force fields are based on a neural network potential energy surface (NN-PES) architecture, in which a machine learning model is trained to reproduce the total electronic energy of a reference level of theory (typically DFT) to chemical accuracy [18–20]. The model is trained by minimizing the prediction error against the reference level of theory. In the approach known as a high dimensional neural network potential (HDNNP) [21, 22], this is done by first transforming the atomic positions into local atomic descriptors which can then be mapped to atomic energy contributions. The total energy is then computed as the sum of the atomic energies. Typically HDNNPs are applied to the same chemical systems that they are trained on, enabling simulations at larger length scales and timescales which would not have been possible using traditional *ab initio* molecular dynamics techniques [23]. Recent efforts have been made to further improve HDNNP architecture and also extend to systems that are not part of the training set. Most notably, the QRNN model [24] aims to supplement and build on the success of ANI [15] by distinguishing charged species from uncharged species while keeping the generalizability that allows accuracy without task-specific training. By generalizing from clusters of small molecules up to large periodic boxes, we can use an advanced hybrid density functional like $\omega$B97X-D3BJ [25, 26] to generate training data, unlocking its greater accuracy. Moreover, since QRNN models are effective at learning partial atomic charges, these models can accurately capture dynamics for polar and ionic systems, and condensed phase systems such as electrolytes [27]. As a result, the QRNN model holds promise to model polymer dynamics more accurately than classical force fields.

In this paper, we discuss the benefits of using the QRNN approach in simulating ethylene glycol (EG) and its oligomers. EG-based solvents are used across various applications such as solvents [28], paints [29], antifreeze [30], hydraulic brake fluids [31], inks [32], plastics [33], films [34], and cosmetics [35]. More recently, EG-based solvents have been used with ionic liquids [36] and eutectic solvents [37]. Traditional biomolecular force fields, such as OPLS, have been used to model systems such as polymers [38], ethylene oxide [39], organic/inorganic nanostructures [40] and ethylene glycol [41, 42]. This force field type has been successful in capturing some thermodynamic properties such as density at a constant temperature, however, it fails to capture some dynamic

and thermodynamic properties such as self-diffusivity, viscosity, and specific heat [43]. The dynamic properties are predicted particularly poorly since the OPLS force fields result in an effective stiffening of the oligomers due to exaggerated torsional barriers, which becomes more pronounced as the chain length increases [44]. Quantum mechanics based charge-fitting can improve the performance of OPLS force field [45] for specific systems, but this is not the focus of our study. Here, we use a QRNN model to mitigate the inaccuracies from the OPLS force field as well as the computational expense posed by traditional quantum computations. In this paper, we describe the methodology for training force fields that may be applied to any class of polymers followed by a demonstration of their property prediction accuracy on ethylene glycol oligomers. Our work serves as proof of concept that the developed MLFF workflow can be used for modeling polymer materials and emerges as both a successful and economical path for exploring the mechanical and thermophysical properties, thus allowing us to design new polymer systems with higher confidence.

The paper is structured as follows. In Section 2 we discuss the workflow for the machine-learned force field training set construction followed by the MD analysis specifics. Promising improvements to the neural network potential, such as carrying out feature engineering and active learning, are also discussed in this section. Following, in Section 3 we discuss the numerical results obtained from the QRNN-MD simulations. We first show the results from the initial trained model (Section 3.2) and follow up with results from an improved model that implements feature engineering and active learning (Section 3.3). We also discuss the consequence of such model improvements on the numerical results. We conclude the paper with the findings from our study and the promise that this scalable and generalizable force field provides.

## 2. Methods

### *2.1. Classical force field Molecular Dynamics*

Given that polymers comprise repeating units of monomers, the goal is to use small oligomers ($\leq$ trimers) to train the QRNN model, which then functions as the force field for the MD simulations for larger oligomers, which serves as the test of the model's extensivity. We begin with developing a set of benchmark MD simulations, against which we will eventually compare the QRNN models performance. For the MD simulations, we use the OPLS4 force field developed by Schrödinger Inc [46]. We generate different simulation cells with increasing chain length, from pure monoethylene glycol to pure decaethlyene glycol. For creating amorphous cells, we use Disordered System Builder (DSB) tool in Schrödinger Materials Science Suite (MSS), version 2022-2 [47, 48]. The DSB tool arranges the oligomers in the simulation cell by rotating the backbone dihedrals, using a self-avoiding random walk algorithm, to ensure varying chain configurations. For each oligomer type, we create 10 replicate cells with different spatial configurations. The generated simulation cells consist of $\sim$ 1500 atoms with a density of 0.5 g/cm$^3$. The initialized configuration is then relaxed in a ten-step process, as described in Section S1.1.

To obtain MD-generated values by using the OPLS4 force field, we performed an NPT simulation on the relaxed configuration for 2 ns at 300 K and 1 atm external pressure with a timestep of 2 fs. For the computation of self-diffusivity, we performed an NVT simulation for 10 ns at 300 K with a timestep of 1 fs (since any ensemble with a barostat has the potential of interfering with the system dynamics). We use the mean-squared displacement and the Einstein equation to obtain the self-diffusivity.

### *2.2. Dataset preparation for QRNN training*

From the OPLS4 trajectories, we extracted N-molecule clusters of monomers, dimers, and trimers of ethylene glycol to serve as training data. In the interest of efficient DFT labeling of the clusters, we limit our dataset to 8-molecule clusters in the case of the monomers, up to 6-molecule clusters in the case of the dimers, and up to 4-molecule clusters in the case of the trimers. We then sample several conformational clusters from the different N-molecule clusters where we find a balance between the number of clusters for representative training data and the computational expense associated with carrying out DFT calculations on the clusters. These conformational clusters contain non-equilibrium bond distances, and bond and dihedral angles which sample configurations that are representative of what the system might encounter during the course of a stable MD trajectory. The sampled conformational clusters also include density variations in addition to the bond, angle, and dihedral deformations. In addition to the sampled conformational clusters of a different number of molecules of the monomers, trimers, and dimers, we also look at a few decomposed samples of clusters. Decomposed samples are created by arbitrarily breaking/deleting bonds in our sampled configurations. These samples force the individual molecules into unfavorable configurations. The decomposed clusters prevent the QRNN model from stabilizing unrealistic molecular configurations while carrying out the MD simulation. The number of different N-molecule conformational samples from the regularly identified clusters is 210,500, whereas the total number of decomposed clusters at 106,000. The breakdown of the clusters is shown in Table S1. Out of these, a total of 217,684 clusters were used to compute the DFT energies, atomic forces, and dipoles. All the DFT data generated in this work is shared on figshare platform [49]. All DFT calculations were performed at the $\omega$B97X-D3BJ/def2-TZVPD level [25, 26, 50] with the electronic structure software package Psi4-1.3 [51]. The parameters used for the energy calculation were: 1e-10 DFT basis tolerance, 1e-10 Schwarz screening threshold, and 1e-6 linear dependency cutoff.

### *2.3. QRNN training*

The network architecture for the polymer force field predictor is trained by using a code based on the `TORCHSANI` private/internal software package. Our study extends QRNN MD simulations to systems of ethylene glycol and its oligomers from monoethylene glycol through decaethylene glycol. We explore the extent of the restriction of the training space that would yield a reliable machine-learned force field for oligomers of any arbitrary length. To this end, we train three separate models: (i) `QRNN - M`: A model trained exclusively on the conformational and decomposed clusters of ethylene

glycol, (ii) `QRNN - M,D`: A model trained exclusively on the conformational and decomposed clusters of ethylene glycol and diethylene glycol, and (iii) `QRNN - M,D,T`: A model trained exclusively on the conformational and decomposed clusters of ethylene glycol, diethylene glycol, and triethylene glycol.

The hyperparameters for training the network are inherited from previous works from the group [24, 27]. We use weight normalization with a weight decay of $10^{-4}$. The learning rate ($\alpha$) is set at $10^{-3}$. The training batch size is 256 while the validation batch size is 1024. We use 500 epochs with an epoch size of 200,000 samples during training. The training labels we use initially are total energy, atomic forces, and dipole moments. For some models (described later in the paper), we replace the dipole moments as with extended tight-binding charges as a training label.

### *2.4. QRNN aided Molecular Dynamics*

The molecular dynamics simulation from the QRNN force field was performed using the Atomic Simulation Environment (ASE) package with the QRNN being used as a force calculator. For all the QRNN MD simulations, the total time of the simulation is 1 ns with a timestep of 0.5 fs owing to the computational expense and numerical stability associated with the simulation. The Nose-Hoover thermostat was used for the NVT and NPT simulations and the Parinello-Rahman barostat was used for the NPT simulations. The thermostat relaxation time constant was fixed at 50 fs whereas the time constant for the barostat was fixed at 2.5 ps. The QRNN NPT simulations are run starting from the equilibrated structure obtained from the relaxation protocol mentioned in Section S1.1 which utilizes the OPLS4 force field. We used all the ten simulation boxes (replicates) for each of the ten oligomers of ethylene glycol that were equilibrated using the OPLS4 force field. We restrict the number of atoms in the box to $\sim$1500 to find the balance in a trade-off between computational expense and simulation accuracy. For the case of the NVT simulation, we use a snapshot from the QRNN-MD trajectory that has a density that is close to the average density of the NPT trajectory. The NPT and NVT simulations are run for a total of 1 ns. For the computation of self-diffusivity, we sample the trajectory every 0.5 ps.

### *2.5. Pair scanning*

To test the quality of the models trained by the dataset, we carry out distance scans for pairs of molecules. The process constitutes choosing one 2-molecule sample of any oligomer and rigidly translating the individual molecules from a distance of $d_{\text{min}}$ to $d_{\text{max}}$ in $n_{\text{steps}}$, where $d_{\text{min}}$ is chosen to be about 75% of the initial distance between the center-of-mass of the two molecules ($d_{\text{COM}}$) and $d_{\text{max}}$ is chosen to be 300% of $d_{\text{COM}}$. For these $n_{\text{steps}}$ configurations we compute the potential energy of the 2-molecule system using DFT, the OPLS4 force field, and the QRNN force field. For well-trained models, we would expect the energy-separation landscape from the QRNN model to be much closer to the DFT results. For our analysis, we use $n_{\text{steps}} = 500$ for ethylene glycol, diethylene glycol, and triethylene glycol. However, for tetraethylene glycol we use $n_{\text{steps}} = 300$ in the interest of computational expense.

*2.6. Active learning*

To further improve the QRNN MD accuracy, we add additional DFT samples. We do this efficiently by implementing an active learning method. Active learning constitutes a multi-step process:

- We first train an ensemble (also known as a committee) of five QRNN models by using the original dataset that we wish to improve. However, instead of using a 90-10 split for training-validation sets (as we do for training the target model), we use a 60-40 split to introduce more variation between the models.

- We follow this step by creating a large number of clusters of a different number of molecules for the monomer, dimer, and trimers of ethylene glycol. In our particular case, we create close to 800,000 clusters for the first round of active learning, details of which are shown in Table S2.

- We use all five models to compute the atomic forces across clusters and note the weight-averaged variance ($\rho^f$) of the computed forces for all the $N_a$ atoms in each cluster given by,

$$\rho^f = \sum_{i=1}^{N_a} \sum_{x,\,y,\,z} \sigma(f)/\sqrt{N_a}. \tag{1}$$

  Here, $\rho^f$ serves as a measure of difficulty in generalizing the force field to a given cluster sample.

- We select 40,000 samples with the highest $\rho^f$ and another 40,000 samples that are randomly selected from the remaining conformational clusters. To this we add the newly extracted clusters (`Cluster` column in Table S2) to form the active learning samples.

- The DFT calculations are now run on the active learning samples after which they are appended to the database.

- A new QRNN model is trained with these additional samples in the database.

We follow the same steps for the second round of active learning with a few differences. The total number of clusters extracted from our QRNN MD trajectory for the second round is $\sim 313,000$ and does not include any conformational samples. We perform the $\rho^f$ calculations directly on the extracted clusters. From this, we select 30,000 samples with the highest $\rho^f$ and another 30,000 samples that are randomly selected from the remaining extracted clusters. The details of the extracted clusters used for the second round of active learning can be found in Table S3.

## 3. Results

We compare the results obtained from the OPLS4 force field and the machine-learned force field (QRNN) against the experimental results for ethylene glycol oligomers. The OPLS4 MD simulation results carry with them higher inaccuracies in capturing certain dynamic and thermodynamic properties due to the insufficient accuracy of the force field. Part of the inaccuracy can be attributed to the size of the systems we are working with ($\sim$ 1500 atoms), which is the scale limitations imposed by the QRNN-MD model that we are comparing the OPLS4 results against. However, we demonstrate how the QRNN force field captures atomic interactions more accurately, requiring a much smaller atomic system to give comparable/better/more converged results than the OPLS4-MD obtained simulations. The experimental values for all physical and chemical properties across all oligomers are not available at 300 K, as a result of which we interpolate the values of density and self-diffusivity obtained for ethylene glycol and its oligomers at 298.15 K and 303.15 K by Hoffmann et al. [52].

### *3.1. Pair scanning*

As stated earlier, pair scanning provides a measure of the performance of a force field. In this case, the ground truth for the interaction energy is considered to be the energy computed from the DFT calculations carried out using $\omega$B97X-D3BJ [25, 26] functional and the def2-TZVPD [50] basis set with the electronic structure software package Psi4-1.3 [51], just as the protocol followed for dataset preparation. The calculations are carried out on a pair of molecules that are brought to 75% of the initial center of mass distance, $d_{\text{COM}}$ to up to 300% of $d_{\text{COM}}$. The DFT calculations are then compared against energy computed from the machine-learned force field and the OPLS4 force field. The hypothesis is that the machine-learned force field should do much better than the OPLS4 force field since it is trained on the DFT data that we are benchmarking the pair scanning results against. To illustrate our argument for the model accuracy, we test all three models: `QRNN - M`, `QRNN - M,D` and `QRNN - M,D,T` for monomers, dimers, trimers, and tetramers, as shown in Fig. 1.

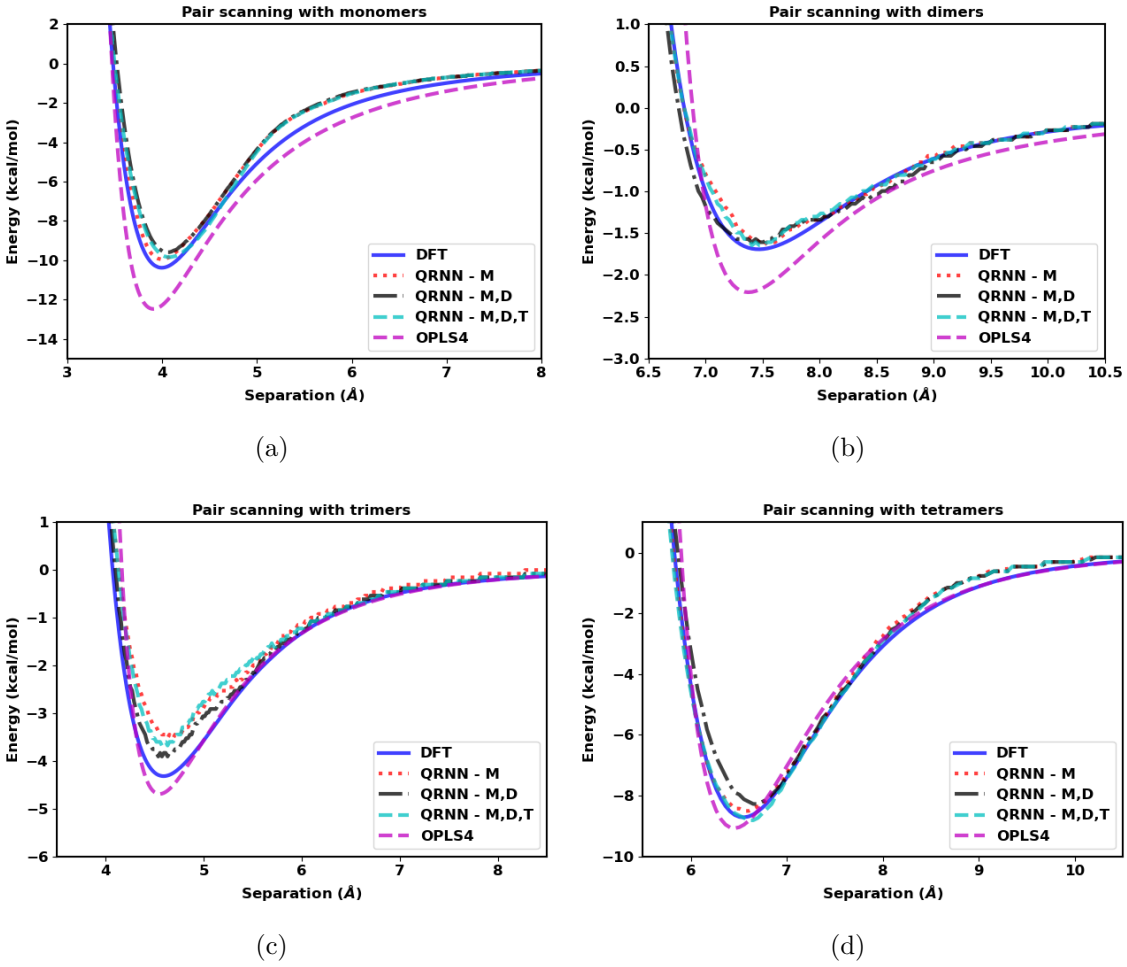


Figure 1: Separation energy between two molecules as a function of separation distance compared between the DFT calculation, OPLS4 force field, and the machine-learned force field for the (a) monomers, (b) dimers, (c) trimers, and (d) tetramers of ethylene glycol.

We see that the force field that is trained on just the monomers generalizes equally well to oligomers up to the size of tetraethylene glycol. We see that the machine-learned force fields do better than the OPLS4 force field in estimating the energy gradients at larger separations, which indicates that the QRNN models are likely to capture the dynamics better than the OPLS4 force field. Even though the largest molecules in the clusters used for training were restricted to triethylene glycol, all the models do reasonably well on predicting the interactions of tetraethylene glycol. This test corroborates the transferability and generalizability of the model to oligomers of lengths longer than the oligomers comprising the training dataset. However, a telling feature of the pair scanning is that the energy minima with the QRNN models appear to the right of the DFT calculations, which implies that the QRNN model places the clusters at a larger equilibrium distance apart. As a result, the density prediction from the QRNN models is expected to be lower. The peak position is a proxy for the density

prediction, whereas the gradients of the energy landscape (the resultant force) can be used as a proxy for the feature that captures the dynamics. Our goal would be to sample configurations such that we can shift the QRNN energy landscape to match the DFT energy landscape to get the correct density and dynamics.

### *3.2. Comparison of properties*

#### *3.2.1. Density*

The Lennard-Jones intermolecular interactions of the OPLS family of force fields are fitted to reproduce experimental properties, so it is not a surprise that the OPLS4 force field is known to capture the density of ethylene glycol oligomer systems within 1% [53, 54]. We run 10 replicates of each oligomeric system comprising 1500 atoms to see how well the OPLS4 force field does for the limited simulation size. We see that the OPLS4 force field predicted density is still within 1% and is strongly correlated with the experimentally obtained values for the oligomers in our analysis, as shown in Fig. 2. On the other hand, we see that the machine-learned force fields underpredict the density by 5%. Note, that the density from the QRNN MD is averaged over the last 800 ps of the 1 ns simulation, to avoid convergence problems.

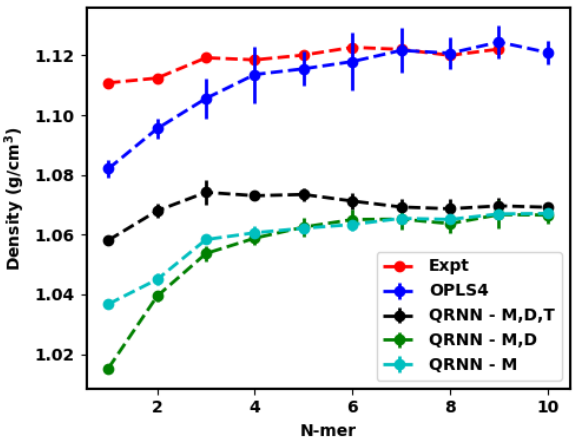


Figure 2: Predicted density from the OPLS4 force field and the machine-learned force field compared against the experimental density.

Experimentally, we see the density values converging to 1.12 g/cm$^3$ [52] as the oligomer length of ethylene glycol increases. This trend is accurately captured by the OPLS4 force field where the density prediction becomes closer to the experimental values with the increase in the oligomer length. The density prediction from the QRNN models is still successful in capturing the convergent density prediction, although it happens to be lower than the expected density. This highlights that the QRNN model can be trained on smaller constituent molecules of a polymer to run simulations on larger polymer chains due to the generalizability of the HDNNP.

*3.2.2. Self-diffusivity*

The OPLS4 force field has its drawbacks in simulating the dynamics of polymeric networks. Typically, longer oligomers are artificially stiff using the OPLS4 force field, as shown in Fig. 3. For the monomer ethylene glycol, we see that the measured diffusivity is in the same order of magnitude as the experimentally observed self-diffusivity. However, for oligomers starting as small as the dimer, we see a sharp drop-off in the self-diffusivity. This rigid locking of the oligomers might result in them getting trapped in solid-like configurations.

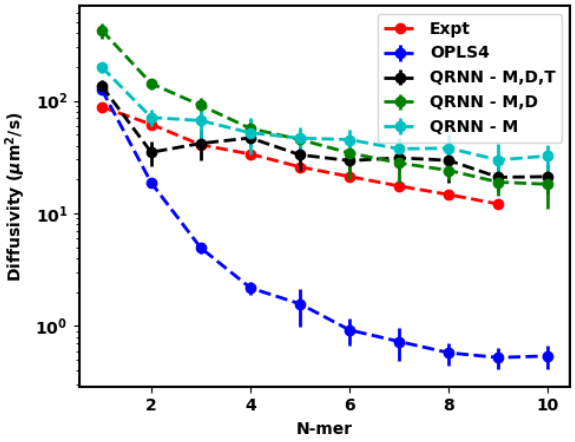


Figure 3: Predicted diffusivity from the OPLS4 force field and the machine-learned force field compared against the experimental diffusivity.

On the other hand, we see that the QRNN force field does not reduce the larger oligomers to such configurational locking allowing for a more diffusive trajectory during the simulation. However, the self-diffusivity from the QRNN models overpredicts the experimental self-diffusivity. This observation can be attributed to the lower density configurations that are generated by the QRNN MD simulation. The lower density affords the oligomers an ease in migration within the simulation box resulting in the prediction of a higher self-diffusivity.

*3.2.3. Specific Heat Capacity*

The specific heat capacity for molecular simulations is computed by using the fluctuation-dissipation theorem, where the specific heat at constant pressure, $c_p$, is given by,

$$c_p = \frac{\langle E^2 \rangle_{\mathrm{NPT}} - \langle E \rangle^2_{\mathrm{NPT}}}{k_B T^2}, \tag{2}$$

where $E$ is the total energy, $k_B$ is the Boltzmann constant and $T$ is the temperature held by a thermostat in the NPT simulation. For ethylene glycol and its oligomers, we have observed a consistent overprediction of the $c_p$, which is commonly observed for the OPLS force field without quantum corrections [43], as confirmed by our simulations

(see Fig. 4). For our QRNN model trained on *ab initio* calculations of the force field, we expect the quantum effects to be captured more thoroughly than the traditional OPLS4 force field. Across the three models that we have implemented, we see that the $c_p$ predictions are much closer to the experimentally observed values, even though they are still overestimated.

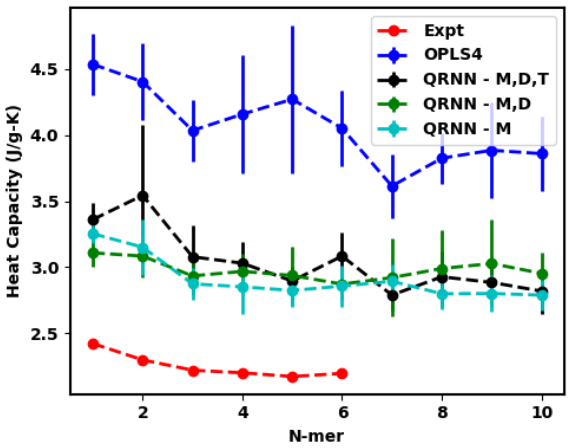


Figure 4: Predicted specific heat capacity from the OPLS4 force field and the machine-learned force field compared against the experimental specific heat capacity.

### *3.3. Model revisions*

The machine-learned force field showed promising results. The MD trajectory was stable across all replicates, however, the density predictions are not as close to the experimental values as the OPLS4 force field predictions are. The heat capacity and diffusivity predictions are better than the OPLS4 force field, however, they are still not in agreement with the experimentally obtained values. The recurring inference to all our results was that the mismatch in the density predictions could all be stemming from the low-density configurations that the QRNN MD simulation was running into. To address this density mismatch we follow a two-pronged approach to improve our force field: (i) add samples to the dataset by doing active learning, and (ii) train the model on extended tight binding (GFN2-xTB, hereby referred to as xTB) charges instead of the dipole moment.

For active learning, as described in the Section 2.6, we train five models on the database for `QRNN - M,D,T` database by using a 60-40 train test split to introduce variability within the trained models. We then generate $\sim 800,000$ conformational samples from various monomer, dimer, and trimer clusters as described in Table S2 (see Section S1.3). The weighted variance of the predicted forces on each atom given by,

$$\rho^f = \sum_{i=1}^{N_a} \sum_{x,y,z} \sigma(f)/\sqrt{N_a}. \tag{3}$$

We choose 40,000 of the worst-performing conformational clusters and we randomly choose 40,000 clusters from the remaining samples. These were selected from the $800,000$ conformational clusters created by altering the conformations of clusters that are extracted ($\sim 12,000$) from the NPT trajectory of the QRNN run MD simulation. We use the extracted clusters and the active learning sampled clusters (a total of $\sim 92,000$ clusters for the first round of active learning) to run DFT calculations and add them to the training database. In the second round of active learning, we use $313,000$ extracted clusters and identify 30,000 worst $\rho_f$ clusters and 30,000 randomly sampled clusters. The $\rho^f$ is expected to be directly correlated to the energy prediction error, $\epsilon = |E_{\text{DFT}} - E_{\text{QRNN}}|$.

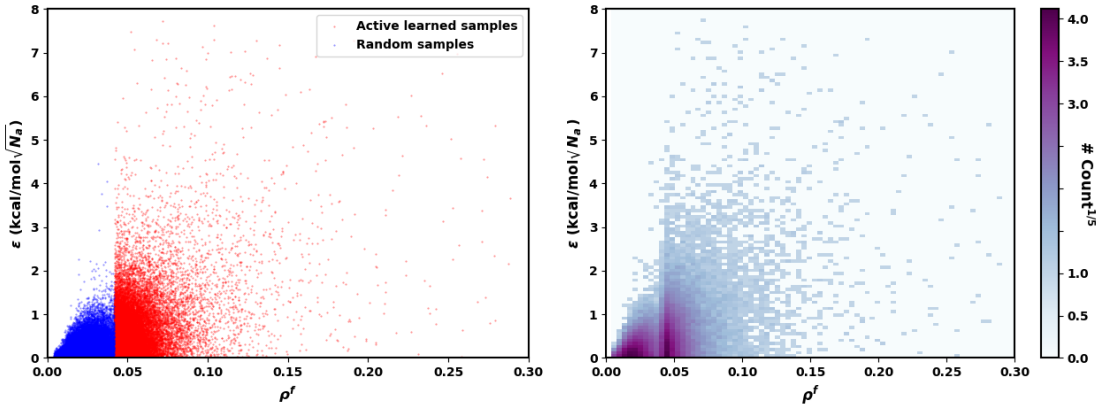


Figure 5: (a) Location of the active learned samples and random samples on the $\epsilon$ vs $\rho^f$ landscape. (b) Heat map (by $\#\text{counts}^{1/5}$) of the $\epsilon$ vs $\rho^f$ landscape for the conformational samples selected for retraining.

We see that the low $\rho^f$ occurs more frequently, explaining why the low frequency – high variance clusters are not what the model has generalized to. Adding these samples to the training, as a result, should improve the model significantly and can help bridge the density prediction mismatch.

The default workflow is designed to train the network (one network per atom) to learn the total energy, the forces per degree of freedom (from the backpropagation gradient of the energy landscape), and the dipole moment. As an alternative feature engineering approach, we can replace the dipole moment as a feature with the xTB charges. This can be a helpful modification since ethylene glycol and its oligomers are symmetric as a result of which they have insignificant dipole moments. This makes the weights corresponding to the dipole moment label difficult to train. xTB charges as a label provide a solution since the computed partial charges will not be diminishingly small.

For brevity and to demonstrate the incremental improvement shown by active learning and the replacement of dipole moment as a training label with xTB charges, we

present two models from the two rounds of active learning and compare against `QRNN - M,D,T`. The first model (`QRNN - AL-TL`) is developed by carrying out transfer learning on `QRNN - M,D,T` by using the composite database at the end of the first active learning cycle (comprising the original database appended with the samples from the first round of active learning). The second model (`QRNN - AL2.0-TL-xTB`) is using the composite database at the end of the second active learning cycle and it transfer learns from a model that is trained on xTB charge labels (`QRNN - AL-TL-xTB`, a first active learning cycle model that is briefly described in Section S1.3). With both the newly trained models, we carry out the same analysis as before: (i) We use the NPT trajectory to compute the density and the specific heat capacity; (ii) we use the NPT snapshot closest to the average density over the NPT run as the starting configuration for the NVT simulation – the resulting trajectory is used to compute diffusivity and finally (iii) we carry out pair scanning on the monomer through tetramer of ethylene glycol to check the model performance.

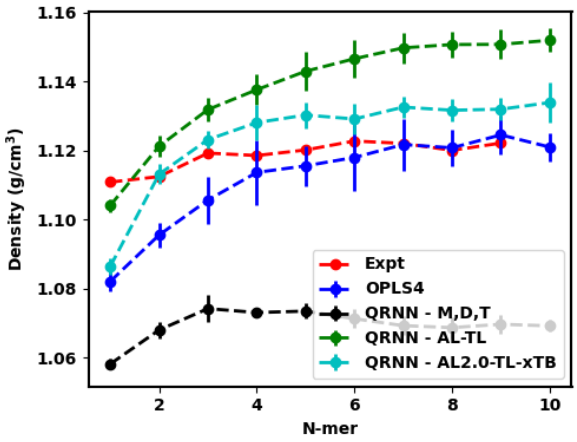


Figure 6: Predicted density from the OPLS4 force field and the machine-learned force field compared against the experimental density.

The model trained on the initial clusters resulted in under-predicting the experimental density. On carrying out transfer learning on the initial model with active learning samples appended to the original dataset, we see that the model, `QRNN - AL-TL`, now shows an overprediction of the density as seen in Fig. 6. We see that the active learned model underpredicts the density for ethylene glycol before transitioning to over-predicting the density for longer oligomers. The density from this new model eventually converges to 1.15 g/cm$^3$, which is an overprediction by 3%. However, we see that `QRNN - AL2.0-TL-xTB`, model from the second round of active learning, circumvents this shortcoming with an overprediction of the density by under 1%. This shows that multiple rounds of active learning could help the model eventually converge to the experimental density, which appears to be the mean of the previous model and the newer model with the active learned samples. We also note that the model trained

on the xTB charges instead of the dipole moments provides a more convergent density and shows a very good correlation ($R^2 = 0.81$ – see Fig. S8) with the experimentally observed density.

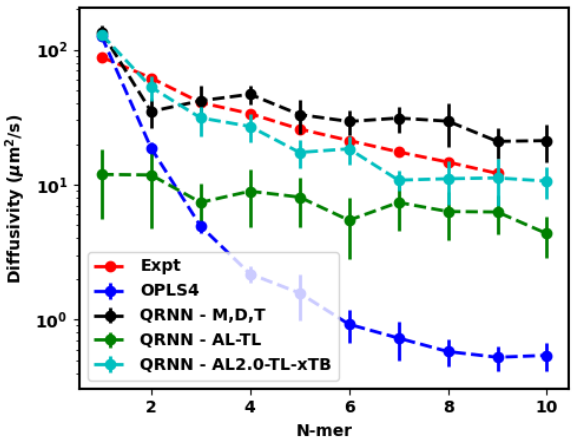


Figure 7: Predicted diffusivity from the OPLS4 force field and the machine-learned force field compared against the experimental diffusivity.

Dynamics prediction appears to be the bottleneck associated with the OPLS4 force field as well as our initial set of models. Even though our model circumvents the artificial locking/freezing of longer oligomers that the OPLS4 force field induces we still observe overprediction by $40 - 80\%$. With higher densities being predicted by the two new test models, we expect to see a lower self-diffusivity, which is corroborated by our results shown in Fig. 7. We see that `QRNN - AL-TL` shows an underprediction of the self-diffusivity from the experimental values by about $40 - 60$. The model trained on the xTB charges, `QRNN - AL2.0-TL-xTB`, at the end of the second active learning cycle performs much better in the prediction of the diffusivity and it most accurately captures the drop in diffusivity with the oligomer length (within 30%), showing the best correlation with the experimental self-diffusivity ($R^2 = 0.91$ – see Fig. S9).

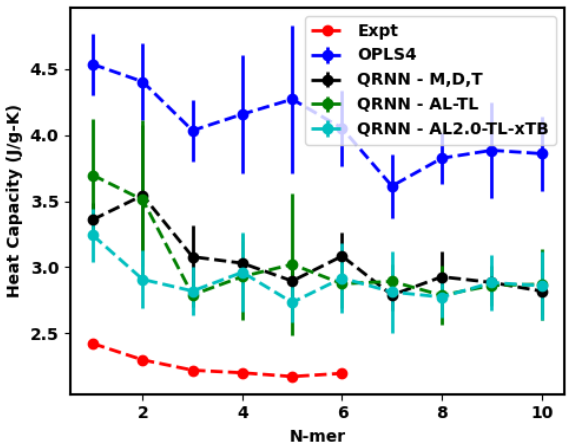


Figure 8: Predicted specific heat capacity from the OPLS4 force field and the machine-learned force field compared against the experimental specific heat capacity.

The increase in density that is enforced by the modified force field results in the slowing of the dynamics. We can expect the variance in the total energy to be limited since a major contribution to the energy fluctuations is brought about by the kinetic energy contribution. Our hypothesis is corroborated by the specific heat capacity calculations in Fig. 8. Additionally, we see that the `QRNN - AL-TL` and `QRNN - AL2.0-TL-xTB` models predict the $c_p$, that is much more correlated to the experimental values than the OPLS4 force field or the initial `QRNN - M,D,T` model. Furthermore, the $c_p$ prediction made by the `QRNN - AL2.0-TL-xTB` is better for smaller oligomers before converging to the long chain $c_p$ predicted by the other trained models.

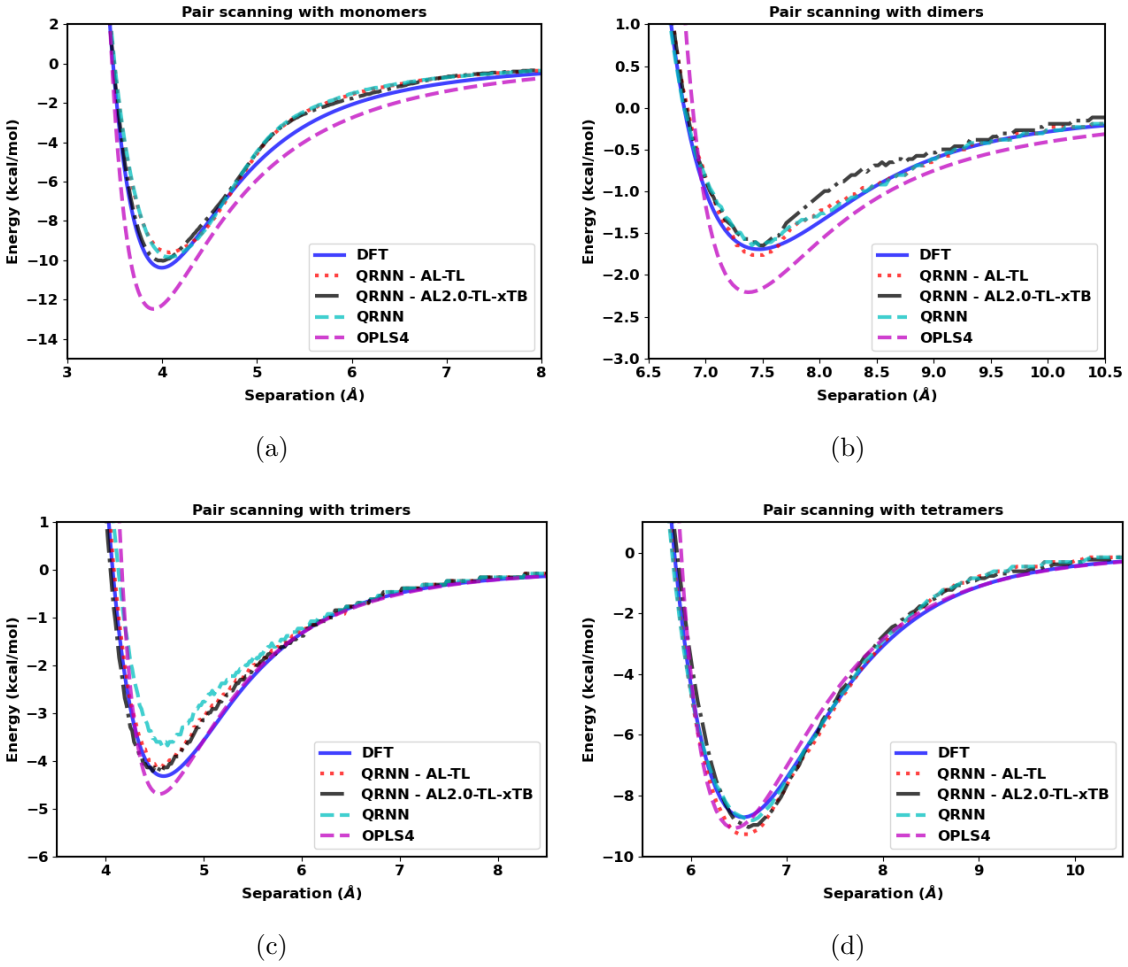


Figure 9: Separation energy between two molecules as a function of separation distance compared between the DFT calculation, OPLS4 force field, and the machine-learned force field for the (a) monomers, (b) dimers, (c) trimers and (d) tetramers of ethylene glycol.

As a final test for the quality of the force field that is newly trained, we carry out the pair scanning test. The QRNN models are qualitatively similar in their performance across oligomers of all sizes, as shown in Fig. 9. Much like the previous analysis, the QRNN models perform better than the OPLS4 force field in terms of the force prediction. In addition, we also see a left shift in the energy well obtained from the QRNN models after the proposed changes to the training protocol (energy well from the red dotted lines and the black dash-dot lines is to the left of that of the blue dashed line which represents the initial model), implying higher density. The eventual goal for any material system would be to train a model through iterative active learning such that the energy well from the DFT calculations and the QRNN model overlap for the most accurate density and dynamics prediction.

### *3.4. Model summary and extensibility*

From these results, we conclude that both our strategies, (i) active learning and (ii) active learning with xTB labels, were successful in achieving our desired behavior of increasing the density predictions. The first learning cycle resulted in the density under-prediction transitioning to density over-prediction. This trend is then improved by the second active learning cycle, which demonstrates the success of active learning in nudging the model toward ideal behavior. As a result, we can draw insights on how to smartly sample configurations initially to overcome the issue with the under or overprediction of densities and the subsequent dynamics that arise from it. Carrying out multiple active learning rounds or smartly managing the distribution of the samples during active learning can help arrive at a convergent force field in fewer active learning iterations. Since the diffusivity and the specific heat capacity are tied to the dynamics governed by the simulation box of a specific density, we can hypothesize that these properties will be predicted better once we arrive at a more convergent machine-learned force field.

To further test the stability of the developed QRNN model, we applied it to a longer polyethylene glycol chain containing 25 monomers. The resulting simulation was stable and demonstrates chain entanglement (see Fig. S11a). Visual inspection from different snapshots of the chains shows more realistic configurations of long chains (see Fig. S11b). Furthermore, a single chain in the QRNN MD shows a more dynamic nature when compared to OPLS4 dynamics.

## 4. Conclusion

In this paper, we present the development and performance of a QRNN force field for ethylene glycol and its oligomers by comparing it against results from experiments and the OPLS4 force field. The QRNN model is trained and validated against a reference level DFT theory to address the shortcomings of the OPLS4 force field in capturing the dynamic and thermodynamic properties of polymeric systems, such as self-diffusivity and heat capacity. The QRNN force field performs marginally poorer than the OPLS4 force field in estimating the density, however, it performs much better in capturing the specific heat capacity and the self-diffusivity. We see promising results from the QRNN force field in which the model trained on small oligomers generalizes well to longer oligomers.

Our QRNN model is trained only to small ethylene glycol oligomers, but generalizes well to longer oligomers - the QRNN model does relatively well in capturing the density, self-diffusivity, and specific heat capacity of the oligomer up to decaethylene glycol. Due to the superior quasi-linear scaling of the QRNN model [55], compared to the cubic scaling of *ab initio* calculations, we are able to run simulations of systems of 1500 atoms where the QRNN model calculations are $10^5$ times faster than AIMD simulations. The enhanced accuracy of the QRNN force field will allow us to capture processes and dynamics that a larger or longer OPLS MD simulation might not even be able to capture due to the shortcomings of the traditional force field.

To further improve the force field in the area of dynamics and the system density, we carry out two approaches: active learning and training on xTB charges instead of the dipole moment. We see that these approaches have the desired effect of incrementally improving the density and dynamics prediction of the machine-learned force fields. A smarter implementation of active learning by carefully including sampled configurations helps in improving the QRNN model, as demonstrated by this paper. Our work shows that training generalizable models for polymer systems is possible even with smaller building blocks such as monomers, dimers, and trimers. We also show that with a smarter training protocol and conformational sampling to cover the configuration space, we can develop a force field that is more adept than the classical force field at capturing the dynamics of the target system.

This approach paves the way for developing QRNN force fields for polymeric systems with limited computational expense (by training on smaller oligomers), which can generalize well to polymers of arbitrary length. As a result, we can now assess rare events and dynamical phenomena in such systems that were earlier inaccessible by traditional force fields (due to force-field inaccuracy) and ab-initio methods (due to length-scale limitations) – two problems that are well addressed by the QRNN force field.

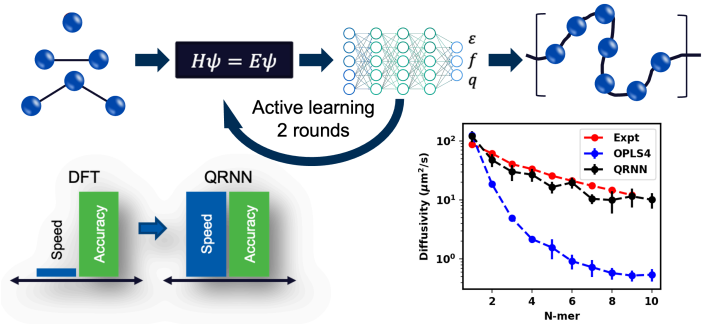


Figure 10: Graphical abstract.

## Conflict of Interest

The authors declare no competing interests.

## Data Access

All the DFT data generated in this work is available on the figshare platform [49]. We provide additional data on cluster sampling and modeling methods in the Supplementary appendices.

## References

[1] Jonathan Schmidt, Mário RG Marques, Silvana Botti, and Miguel AL Marques. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials*, 5(1):1–36, 2019.

[2] Jing Wei, Xuan Chu, Xiang-Yu Sun, Kun Xu, Hui-Xiong Deng, Jigen Chen, Zhongming Wei, and Ming Lei. Machine learning in materials science. *InfoMat*, 1(3): 338–358, 2019.

[3] Janine George. Automation in dft-based computational materials science. *Trends in Chemistry*, 3(9):697–699, 2021.

[4] Kamal Choudhary, Brian DeCost, Chi Chen, Anubhav Jain, Francesca Tavazza, Ryan Cohn, Cheol Woo Park, Alok Choudhary, Ankit Agrawal, Simon JL Billinge, et al. Recent advances and applications of deep learning methods in materials science. *npj Computational Materials*, 8(1):1–26, 2022.

[5] Garrett B Goh, Nathan O Hodas, and Abhinav Vishnu. Deep learning for computational chemistry. *Journal of computational chemistry*, 38(16):1291–1307, 2017.

[6] Maria Korshunova, Boris Ginsburg, Alexander Tropsha, and Olexandr Isayev. Openchem: A deep learning toolkit for computational chemistry and drug design. *Journal of Chemical Information and Modeling*, 61(1):7–13, 2021.

[7] Adam C Mater and Michelle L Coote. Deep learning in chemistry. *Journal of chemical information and modeling*, 59(6):2545–2559, 2019.

[8] John A Keith, Valentin Vassilev-Galindo, Bingqing Cheng, Stefan Chmiela, Michael Gastegger, Klaus-Robert Müller, and Alexandre Tkatchenko. Combining machine learning and computational chemistry for predictive insights into chemical systems. *Chemical reviews*, 121(16):9816–9872, 2021.

[9] Brian B Goldman and W Patrick Walters. Machine learning in computational chemistry. *Annual Reports in Computational Chemistry*, 2:127–140, 2006.

[10] Dominik Marx and Jurg Hutter. Ab initio molecular dynamics: Theory and implementation. *Modern methods and algorithms of quantum chemistry*, 1(301-449): 141, 2000.

[11] Mark E Tuckerman. Ab initio molecular dynamics: basic concepts, current trends and novel applications. *Journal of Physics: Condensed Matter*, 14(50):R1297, 2002.

[12] Robert G Parr. Density functional theory. In *Electron Distributions and the Chemical Bond*, pages 95–100. Springer, 1982.

[13] Aron J Cohen, Paula Mori-Sánchez, and Weitao Yang. Challenges for density functional theory. *Chemical reviews*, 112(1):289–320, 2012.

[14] Pragya Verma and Donald G Truhlar. Status and challenges of density functional theory. *Trends in Chemistry*, 2(4):302–318, 2020.

[15] Justin S Smith, Olexandr Isayev, and Adrian E Roitberg. Ani-1: an extensible neural network potential with dft accuracy at force field computational cost. *Chemical science*, 8(4):3192–3203, 2017.

[16] Oliver T Unke and Markus Meuwly. Physnet: A neural network for predicting energies, forces, dipole moments, and partial charges. *Journal of chemical theory and computation*, 15(6):3678–3693, 2019.

[17] Connor W Coley, Wengong Jin, Luke Rogers, Timothy F Jamison, Tommi S Jaakkola, William H Green, Regina Barzilay, and Klavs F Jensen. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical science*, 10(2):370–377, 2019.

[18] Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018.

[19] Linfeng Zhang, Jiequn Han, Han Wang, Roberto Car, and EJPRL Weinan. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Physical review letters*, 120(14):143001, 2018.

[20] Tsz Wai Ko, Jonas A Finkler, Stefan Goedecker, and Jörg Behler. A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer. *Nature communications*, 12(1):1–11, 2021.

[21] Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical review letters*, 98(14):146401, 2007.

[22] Tsz Wai Ko, Jonas A Finkler, Stefan Goedecker, and Jörg Behler. General-purpose machine learning potentials capturing nonlocal charge transfer. *Accounts of Chemical Research*, 54(4):808–817, 2021.

[23] Georg Kresse and Jürgen Hafner. Ab initio molecular dynamics for liquid metals. *Physical review B*, 47(1):558, 1993.

[24] Leif D Jacobson, James M Stevenson, Farhad Ramezanghorbani, Delaram Ghoreishi, Karl Leswing, Edward D Harder, and Robert Abel. Transferable neural network potential energy surfaces for closed-shell organic molecules: Extension to ions. *Journal of Chemical Theory and Computation*, 18(4):2354–2366, 2022.

[25] Stefan Grimme, Stephan Ehrlich, and Lars Goerigk. Effect of the damping function in dispersion corrected density functional theory. *Journal of computational chemistry*, 32(7):1456–1465, 2011.

[26] Narbe Mardirossian and Martin Head-Gordon. ωb97x-v: A 10-parameter, range-separated hybrid, generalized gradient approximation density functional with non-local correlation, designed by a survival-of-the-fittest strategy. *Physical Chemistry Chemical Physics*, 16(21):9904–9924, 2014.

[27] Steven Dajnowicz, Garvit Agarwal, James M Stevenson, Leif D Jacobson, Farhad Ramezanghorbani, Karl Leswing, Richard A Friesner, Mathew D Halls, and Robert Abel. High-dimensional neural network potential for liquid electrolyte simulations. *The Journal of Physical Chemistry B*, 2022.

[28] R Nagarajan and Chien-Chung Wang. Theory of surfactant aggregation in water/ethylene glycol mixed solvents. *Langmuir*, 16(12):5242–5251, 2000.

[29] FG Hoogland and JJ Boon. Development of maldi-ms and nano-esi-ms methodology for the full identification of poly (ethylene glycol) additives in artists' acrylic paints. *International Journal of Mass Spectrometry*, 284(1-3):66–71, 2009.

[30] Jan M Hollis, Frank J Lovas, Philip R Jewell, and LH Coudert. Interstellar antifreeze: ethylene glycol. *The Astrophysical Journal*, 571(1):L59, 2002.

[31] Rodney J Boatman and James B Knaak. Ethers of ethylene glycol and derivatives. *Patty's Toxicology*, 2001.

[32] Jung-Tang Wu, Steve Lien-Chung Hsu, Ming-Hsiu Tsai, and Weng-Sing Hwang. Conductive silver patterns via ethylene glycol vapor reduction of ink-jet printed silver nitrate tracks on a polyimide substrate. *Thin Solid Films*, 20(517):5913–5917, 2009.

[33] Guo-Qiang Chen and Martin K Patel. Plastics derived from biological sources: present and future: a technical and environmental review. *Chemical reviews*, 112 (4):2082–2099, 2012.

[34] Sadhana Sharma, Robert W Johnson, and Tejal A Desai. Evaluation of the stability of nonfouling ultrathin poly (ethylene glycol) films for silicon-based microdevices. *Langmuir*, 20(2):348–356, 2004.

[35] Yu-Wei Cheng, Yen-Ting Lin, Kun-Ho Liu, Jung-San Chen, Shih-Hsuan Wang, and Ting-Yu Liu. In situ and initiator-free atmospheric plasma-induced functionalization of poly (ethylene glycol) methacrylate on nonwoven cosmetic masks for the evaluation of the bacteria inhibitory effect. *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, 642:128719, 2022.

[36] Mohammed Taghi Zafarani-Moattar, Hemayat Shekaari, and Elnaz Mazaher Haji Agha. Phase equilibrium study in aqueous solutions containing ionic liquid 1-butyl-3-methyl imidazolium chloride and poly (propylene glycol) 400 or poly (ethylene glycol) dimethyl ether 250 via a vapor–liquid equilibrium study at t= 298.15 k. *Journal of Chemical & Engineering Data*, 64(10):4298–4305, 2019.

[37] Sushma P Ijardar. Deep eutectic solvents composed of tetrabutylammonium bromide and peg: Density, speed of sound and viscosity as a function of temperature. *The Journal of Chemical Thermodynamics*, 140:105897, 2020.

[38] Kateri H DuBay, Michelle Lynn Hall, Thomas F Hughes, Chuanjie Wu, David R Reichman, and Richard A Friesner. Accurate force field development for modeling conjugated polymers. *Journal of chemical theory and computation*, 8(11):4556–4569, 2012.

[39] Chan-En Fang, Yi-Chen Tsai, Christoph Scheurer, and Chi-Cheng Chiu. Revised atomic charges for opls force field model of poly (ethylene oxide): Benchmarks and applications in polymer electrolyte. *Polymers*, 13(7):1131, 2021.

[40] Hendrik Heinz, Tzu-Jen Lin, Ratan Kishore Mishra, and Fateme S Emami. Thermodynamically consistent force fields for the assembly of inorganic, organic, and biological nanostructures: the interface force field. *Langmuir*, 29(6):1754–1765, 2013.

[41] Borys Szefczyk and M Natália DS Cordeiro. Physical properties at the base for the development of an all-atom force field for ethylene glycol. *The Journal of Physical Chemistry B*, 115(12):3013–3019, 2011.

[42] L Saiz, JA Padro, and E Guardia. Structure of liquid ethylene glycol: A molecular dynamics simulation study with different force fields. *The Journal of Chemical Physics*, 114(7):3187–3199, 2001.

[43] Carl Caleman, Paul J Van Maaren, Minyan Hong, Jochen S Hub, Luciano T Costa, and David Van Der Spoel. Force field benchmark of organic liquids: density, enthalpy of vaporization, heat capacities, surface tension, isothermal compressibility, volumetric expansion coefficient, and dielectric constant. *Journal of chemical theory and computation*, 8(1):61–74, 2012.

[44] William M Haynes, David R Lide, and Thomas J Bruno. *CRC handbook of chemistry and physics*. CRC press, 2016.

[45] Shirley WI Siu, Kristyna Pluhackova, and Rainer A Böckmann. Optimization of the opls-aa force field for long hydrocarbons. *Journal of Chemical theory and Computation*, 8(4):1459–1470, 2012.

[46] Chao Lu, Chuanjie Wu, Delaram Ghoreishi, Wei Chen, Lingle Wang, Wolfgang Damm, Gregory A Ross, Markus K Dahlgren, Ellery Russell, Christopher D Von Bargen, et al. Opls4: Improving force field accuracy on challenging regimes of chemical space. *Journal of chemical theory and computation*, 17(7):4291–4300, 2021.

[47] Materials Science Suite, verion 2022-2. Schrödinger, LLC: New York. URL https://www.schrodinger.com/platform/materials-science.

[48] Mohammad Atif Faiz Afzal, Jefferey M. Sanders, Alexander Goldberg, Andrea R. Browning, and Mathew D. Halls. Using molecular simulation with high-temperature composites resins. 2019.

[49] Shaswat Mohanty, James Stevenson, Andrea R. Browning, Leif D. Jacobson, Karl Leswing, Mathew D. Halls, and Mohammad Atif Faiz Afzal. Development of Scalable and Generalizable Machine Learned Force Field for Polymers. 12 2022. doi: 10.6084/m9.figshare.21720881.v1. URL https://figshare.com/articles/dataset/Development_of_Scalable_and_Generalizable_Machine_Learned_Force_Field_for_Polymers/21720881.

[50] Florian Weigend. Accurate coulomb-fitting basis sets for h to rn. *Physical chemistry chemical physics*, 8(9):1057–1065, 2006.

[51] Justin M Turney, Andrew C Simmonett, Robert M Parrish, Edward G Hohenstein, Francesco A Evangelista, Justin T Fermann, Benjamin J Mintz, Lori A Burns, Jeremiah J Wilke, Micah L Abrams, et al. Psi4: an open-source ab initio electronic structure program. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2(4):556–565, 2012.

[52] Markus M Hoffmann, Rachel H Horowitz, Torsten Gutmann, and Gerd Buntkowsky. Densities, viscosities, and self-diffusion coefficients of ethylene glycol oligomers. *Journal of Chemical & Engineering Data*, 66(6):2480–2500, 2021.

[53] Mohammad Atif Faiz Afzal, Andrea R Browning, Alexander Goldberg, Mathew D Halls, Jacob L Gavartin, Tsuguo Morisato, Thomas F Hughes, David J Giesen, and Joseph E Goose. High-throughput molecular dynamics simulations and validation of thermophysical properties of polymers for various applications. *ACS Applied Polymer Materials*, 3(2):620–630, 2020.

[54] Mohammad Atif Faiz Afzal, Aditya Sonpal, Mojtaba Haghighatlari, Andrew J Schultz, and Johannes Hachmann. A deep neural network model for packing density predictions and its application in the study of 1.5 million organic molecules. *Chemical science*, 10(36):8374–8383, 2019.

[55] James M Stevenson, Leif D Jacobson, Yutong Zhao, Chuanjie Wu, Jon Maple, Karl Leswing, Edward Harder, and Robert Abel. Schrödinger-ANI: An eight-element neural network interaction potential with greatly expanded coverage of druglike chemical space. *arXiv preprint arXiv:1912.05079*, 2019.