

# Centroid Molecular Dynamics Can Be Greatly Accelerated through Neural Network Learned Centroid Forces Derived from Path Integral Molecular Dynamics

Timothy D. Loose, Patrick G. Sahrman, and Gregory A. Voth\*



Cite This: *J. Chem. Theory Comput.* 2022, 18, 5856–5863



Read Online

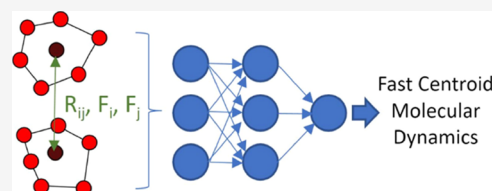
ACCESS |

Metrics & More

Article Recommendations

Supporting Information

**ABSTRACT:** For nearly the past 30 years, centroid molecular dynamics (CMD) has proven to be a viable classical-like phase space formulation for the calculation of quantum dynamical properties. However, calculation of the centroid effective force remains a significant computational cost and limits the ability of CMD to be an efficient approach to study condensed phase quantum dynamics. In this paper, we introduce a neural network-based methodology for first learning the centroid effective force from path integral molecular dynamics data, which is subsequently used as an effective force field to evolve the centroids directly with the CMD algorithm. This method, called machine-learned centroid molecular dynamics (ML-CMD), is faster and far less costly than both standard “on the fly” CMD and ring polymer molecular dynamics (RPMD). The training aspect of ML-CMD is also straightforwardly implemented utilizing the DeepMD software kit. ML-CMD is then applied to two model systems to illustrate the approach: liquid para-hydrogen and water. The results show comparable accuracy to both CMD and RPMD in the estimation of quantum dynamical properties, including the self-diffusion constant and velocity time correlation function, but with significantly reduced overall computational cost.



## 1. INTRODUCTION

The accurate simulation of quantum dynamics is limited by the computational complexity of solving the time-dependent Schrödinger equation. While classical molecular dynamics (MD) based on empirical force fields are capable of utilizing certain information from quantum calculations, the treatment of the nuclei as point particles as well as the inherent limitations of a pairwise decomposable description of the intermolecular interactions limits their accuracy.<sup>1,2</sup> The most straightforward approach to address this accuracy problem is to perform full electronic structure calculations for each timestep to calculate the force vectors on each particle, such as in the case of ab initio MD.<sup>3,4</sup> However, such methods can be prohibitively expensive for all but the smallest systems. Additionally, these calculations must be performed repeatedly over the course of a simulation which highly limits the timescales accessible to study with these methods. Many alternatives have been employed to limit the cost of representing the effects of quantum electronic structure within MD simulations, such as treating the majority of the system with classical mechanics and saving the quantum mechanical calculations for specific regions as in hybrid quantum mechanics/molecular mechanics.<sup>5</sup>

Irrespective of the underlying accuracy of the representation of the forces on the system nuclei, the challenge of quantum dynamics is compounded by the fact that many systems of interest exhibit nuclear quantum effects (NQE), even at thermal equilibrium. Path integral (PI) methods can capture

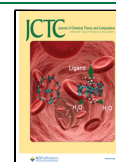
NQEs in such equilibrium circumstances. This technique relies on the imaginary time formulation of Feynman PI quantum mechanics, in which the classical principle of least action is generalized to quantum systems via functional integration over all paths that the system can take between an initial and final point.<sup>6</sup> In the case of molecular simulation, PI-based methods re-cast a quantum mechanical description of the system of interest into an isomorphic classical one where each quantum particle is represented by a set of  $P$  classical quasiparticles or “beads”, for which standard MD algorithms can be utilized.<sup>7,8</sup> In particular, it can be shown that, in the  $P \rightarrow \infty$  limit, the static equilibrium properties of a quantum particle can be described by the Boltzmann statistics generated by the following Hamiltonian:

$$H = \sum_{n=1}^P \left[ \frac{p_n^2}{2m'_n} + \frac{1}{2} m \omega_P^2 (x_n - x_{n+1})^2 + \frac{U(x_n)}{P} \right] \quad (1)$$

where  $m'_n$  are fictitious mass parameters,  $\omega_P^2 = \sqrt{P}/\beta\hbar$ , and  $P$  is the total number of replicas or beads chosen for the

Received: July 6, 2022

Published: September 14, 2022



discretized imaginary time path. Each bead represents the particle at a discrete position in imaginary time, so  $x_{p+1} = x_1$  in order to guarantee that only paths which begin and end in the same place are considered for each configuration. This Hamiltonian effectively describes a collection of classical-like particles each acting under a potential  $U(x)/P$  that are attached by harmonic oscillators to the adjacent particles in a “ring polymer” or “necklace”. (Note that the notation here is for a single quantized particle in a one-dimensional potential, but the notation is readily generalized to more than one particle in three dimensions.) MD simulations performed using the Hamiltonian in eq 1 is referred to as path integral molecular dynamics (PIMD). PIMD provides a route to calculating quantum static equilibrium properties in a computationally feasible fashion, generally achieving converged results at  $P \sim 30$  replicas for a system such as liquid water at ambient temperature.<sup>8</sup> However, the PIMD Hamiltonian cannot be used to estimate quantum dynamics and must be regarded only as a sampling tool for equilibrium statistics.<sup>7,8</sup> The PIMD approach is also still significantly slower than classical MD due to the increased complexity of the system being simulated but largely scales the same way as does classical MD in terms of computational cost (i.e., there is simply a “cost” prefactor proportional to the value of  $P$ ).

Centroid molecular dynamics (CMD), introduced by Cao and Voth nearly thirty 30 ago, provides a means to estimate certain quantum dynamical information from the discretized imaginary time PI via the dynamics of the imaginary time path centroid moving in a classical-like fashion under the mean centroid force.<sup>9,10</sup> This method is motivated by Feynman’s observation that the imaginary time path centroid is the most classical-like variable of a quantum system.<sup>6,11</sup> The original papers introducing the CMD concept relied largely on ad hoc arguments to justify the method, but two subsequent papers<sup>12,13</sup> in 1999 provided an exact formulation of centroid quantum dynamics and also a route to deriving CMD as an approximation to those exact dynamics. These latter two papers are sometimes not cited by authors when discussing CMD so the primary content of that work is briefly reviewed here for completeness.

Formally, a quasi-density operator (QDO) for the centroid density can be defined which can be used to formulate the exact dynamics of the imaginary time path centroids.<sup>12</sup> In one dimension, this QDO is equal to

$$\hat{\phi}(x_c, p_c) = \frac{\hbar}{2\pi} \int_{-\infty}^{\infty} d\zeta \int_{-\infty}^{\infty} d\eta e^{i\zeta(\hat{x}-x_c) + i\eta(\hat{p}-p_c) - \beta\hat{H}}, \quad (2)$$

where  $x_c$  and  $p_c$  are centroid positions and momenta,  $\hat{H}$  is the system’s Hamiltonian and  $\beta = 1/k_B T$ . The centroid distribution function can be obtained by tracing this operator.

$$\rho(x_c, p_c) = \text{Tr}[\hat{\phi}(x_c, p_c)]. \quad (3)$$

Evaluation of this trace gives a classical-like form for the centroid distribution function which separates the position and momentum components when a Cartesian coordinate system is used, such that

$$\rho(x_c, p_c) = e^{-\beta p_c^2/2m} \rho(x_c) = e^{-\beta p_c^2/2m} e^{-\beta V_c(x_c)} \quad (4)$$

where  $V_c(x_c)$  is the effective centroid quantum potential of mean force. Integrating over the centroid position and momentum variables thus yields the standard quantum partition function

$$Z = \iint \frac{dx_c dp_c}{2\pi\hbar} \rho(x_c, p_c). \quad (5)$$

This partition function can be used to calculate the average of a physical observable corresponding to an operator

$$\hat{A}$$

as

$$\langle \hat{A} \rangle = \frac{1}{Z} \iint \frac{dx_c dp_c}{2\pi\hbar} \rho(x_c, p_c) \hat{A}(x_c, p_c; t), \quad (6)$$

where the time-dependent centroid variable  $A_c(x_c, p_c; t)$  is defined as

$$\hat{A}(x_c, p_c; t) = \text{Tr}[\hat{\phi}(x_c, p_c) e^{i\hat{H}t/\hbar} \hat{A} e^{-i\hat{H}t/\hbar}] / \rho(x_c, p_c). \quad (7)$$

For the exact dynamics of the centroid variables, a normalized time-dependent QDO can also be defined as

$$\hat{\delta}_c(t; x_c, p_c) = e^{-i\hat{H}'t/\hbar} \hat{\delta}_c(x_c, p_c) e^{i\hat{H}'t/\hbar} \quad (8)$$

$$\hat{\delta}_c(x_c, p_c) = \hat{\phi}(x_c, p_c) / \rho(x_c, p_c). \quad (9)$$

where  $\hat{H}'$  is a time-independent Hamiltonian upon which the system evolves ( $\hat{H}' = \hat{H}$  in the usual equilibrium case). It is important to note that the QDO in this derivation is not a typical density operator: while it is Hermitian with non-negative diagonal elements, it is not positive-definite. This operator can be used to describe the exact dynamics of the path centroids; however, this is not a useful approach for nontrivial systems for which the quantum Liouville equation cannot be solved.<sup>12</sup> Instead, various approximations to the QDO can be made which result in various forms of PI-based methods including linearized quantum dynamics, centroid Hamiltonian dynamics, and CMD.<sup>13</sup> In the case of CMD, the centroid phase space variables are propagated quasi-classically by virtue of the following *ansatz*:

$$\hat{\delta}_c(t; x_c, p_c) \approx \hat{\delta}_c(x_c(t), p_c(t)) \quad (10)$$

$$m\dot{x}_c(t) = p_c(t); p_c(t) \approx F_c(x_c(t)) = F_{\text{CMD}}(t) \quad (11)$$

The approximation made here assumes that the QDO is the same at  $t = 0$  as at later times except for the placement of the centroids. This mean field-like assumption is reasonable for cases in which linear response theory approximates the dynamics of the system well and for systems that have strong regression to equilibrium behavior. This perspective reveals that CMD is likely to be most accurate for systems at equilibrium and in which coherent (purely quantum) aspects of the dynamics are not likely to have enough time to influence the system significantly before the correlations die out (decohere).

The key to CMD is thus to numerically calculate the effective equilibrium forces felt by the path centroids in one way or another. Sampling of the noncentroid imaginary time PI modes at each centroid position determines an effective potential on which to propagate the centroids in a classical-like fashion. However, fully sampling these modes at each timestep in a simulation is usually very computationally expensive and so it is arguably this feature of CMD that has precluded its application to certain problems over the years. Numerical implementations of CMD instead attempt to adiabatically separate the centroid and noncentroid imaginary time PI

(Matsubara) modes, of which the zero frequency mode is the centroid. Adiabatic separation is achieved by setting the fictitious masses of the noncentroid modes to be much lower than that of the centroid, which is set to the physical mass of the particle. Then, one attaches thermostats to the noncentroid modes to help more rapidly sample them. While the adiabatic approximation enables “on-the-fly” calculation of the centroid effective force, generating centroid trajectories in CMD still involves significant computational overhead.

As an alternative to CMD, Manolopoulos and co-workers subsequently introduced ring polymer molecular dynamics (RPMD) as another approximate quantum dynamic approach,<sup>14–16</sup> which shared certain key aspects of the spirit of CMD. In RPMD, the fictitious masses of the ring polymer beads are set to the physical mass of the particle, and each bead is evolved as a dynamical variable with PIMD. One then makes the ad hoc argument that the MD sampling time in PIMD (as measured by integrator timesteps) is related to the actual real time of the quantum dynamics. The RPMD approach thus removes the requirement that any sort of centroid force averaging be carried out as in CMD. It should be noted that a subsequent analysis<sup>17</sup> showed that RPMD has no clear connection to real-time quantum dynamics, but the methodology remains popular among users given its ease of use. In addition, an analysis of an approximate but to date impractical form of quantum dynamics called “Matsubara Dynamics” has suggested that RPMD and CMD can be related to that approximation through further approximations.<sup>18</sup> Also of note is that RPMD is numerically faster than CMD, but the increase in speed is relatively small in comparison to classical MD, which is significantly faster than both.

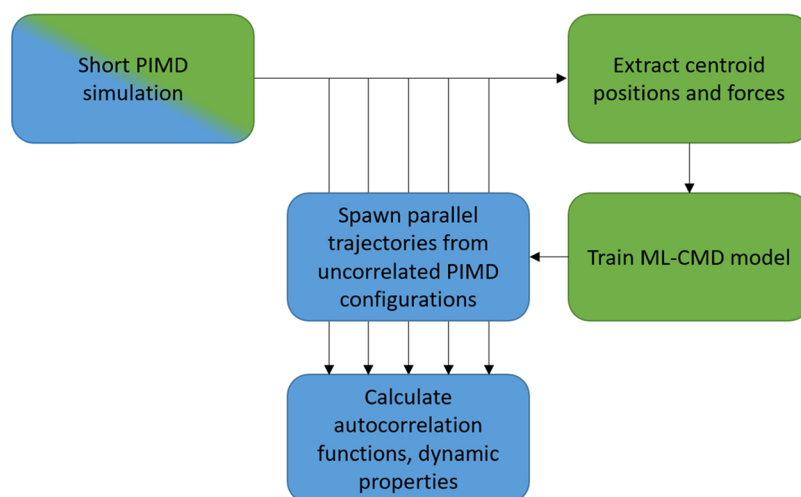
While both RPMD and CMD can capture a range of quantum effects such as incoherent tunneling and zero-point quantization, they are not without their drawbacks. For example, for certain potential energy functions and at low enough temperatures, CMD may exhibit a “curvature problem”, which has a tendency to red shift certain vibrational frequencies for some systems.<sup>19</sup> This behavior can make CMD less suitable for spectrum prediction for certain systems, although this issue generally vanishes at room temperature or higher.<sup>20</sup> On the other hand, RPMD suffers from a spurious resonance problem, in which the centroid dynamics becomes coupled to the harmonic oscillations of the ring polymer, introducing artificial resonances into the spectra.<sup>21</sup> Thermostatted RPMD (TRPMD), in which Langevin thermostats are attached to the ring polymer internal modes, has been introduced as an ad hoc “fix” for this problem.<sup>22</sup>

Beyond these issues, both RPMD and CMD remain relatively computationally expensive due to the need to represent each physical atom in a simulation with dozens or more ring particle replicas (beads). Employing these methods on systems containing many thousands of atoms or more can be infeasible for all but the shortest simulations. For CMD, it has been found to be possible in some cases to generate accurate CMD dynamics with only a partial adiabatic separation of the internal ring polymer modes from the centroid mode.<sup>23</sup> Partially adiabatic CMD (PACMD) simulations can handle larger timesteps as well, allowing them to reach similar levels of efficiency to similar to RPMD. In the case of RPMD, however, a clever ring polymer contraction scheme can be used to increase computational efficiency.<sup>24</sup> In this method, the short-range interactions of the system can be treated using a full ring polymer, while longer-

range interactions can be approximated by interactions calculated across a smaller number of beads (sometimes even just the centroid mode). In order to further bridge the gap between classical MD and PIMD-based methods, approaches have been proposed<sup>25,26</sup> to directly evolve the centroids of quantum particles along a learned centroid force field, but these methods have used pairwise tabulated potentials in their effective quantum force fields which can limit their accuracy in capturing the NQEs.

To quickly summarize, CMD is by now a venerable approach to estimating certain quantum effects in finite temperature systems, for better or for worse. After nearly 30 years, it has largely stood the test of time as a valuable approximation. Yet, the demanding nature of the calculation of the effective centroid force (usually done “on the fly”) has in some ways held back the method from wider use and applicability (e.g., in comparison to RPMD). To shed light on a path to overcome this central challenge—and to capitalize on the rapidly evolving developments in machine learning—in this paper, we introduce machine-learned CMD (ML-CMD) and demonstrate its (arguably remarkable) features in increasing the computational speed and overall efficiency (time to solution) for CMD simulations. The method employs a deep neural network (DNN) trained on PIMD data to act as a force field which calculates the effective centroid forces based on configurations of the path centroids alone. This method retains most of the efficiency of a pairwise force field without its corresponding constraints. ML-CMD employs a DeepMD model,<sup>27</sup> which has been applied to both ab initio data and classical atomistic MD in the past to efficiently predict forces and energies of complicated systems.

DeepMD is a general use DNN method for learning atomic forces and energies based on two main components. The first is a descriptor network which converts the local environment, analogous to a neighbor list in classical MD, of a particle into translationally, rotationally, and permutationally invariant embeddings. This network then passes these embedded features to a second fitting network which considers this environment to predict atomic contributions to energy or force.<sup>28,29</sup> The method can learn a force field for a completely generic representation of a system using mapped forces, including those from PIMD trajectories.<sup>30</sup> In the case of ML-CMD, the training dataset consists of a PIMD trajectory with forces projected to the centroids of the imaginary time paths using the mean squared error in the forces as a loss function. The dataset is then used to learn the centroid forces directly, thus front-loading the work of deriving the effective centroid forces. Naturally, this approach benefits simulations of larger systems the most, but even for simple applications, it can result in significantly faster results. We show later in this paper that ML-CMD can be applied to low-temperature para-hydrogen as well as room-temperature liquid water in order to calculate centroid time correlation functions to a great degree of accuracy, as well as significantly greater efficiency (between one and two orders of magnitude faster) against comparable PI-based methods. The ML-CMD models are also easy to train, require no more additional simulation than CMD, and can be deployed quickly enough to justify their use over CMD and RPMD even for simple systems with no simulation data readily available.



**Figure 1.** Workflow for calculating autocorrelation functions using PI MD methods and for generating ML-CMD models. Blue panels correspond to necessary steps for CMD and RPMD using uncorrelated configurations. Green panels are additional necessary steps for creating an ML-CMD model. The initial PIMD simulation step is shared among both, eliminating the cost of data collection for training ML-CMD force fields when calculating autocorrelation functions.

## 2. METHODS

We trained two ML-CMD models to test the method's ability to capture the static and dynamic properties of systems with significant NQEs. The first is the Silvera–Goldman model<sup>31</sup> of para-hydrogen at 14 K and a density of  $\rho = 0.0235 \text{ \AA}^{-3}$ . A total of 180 particles were simulated, each corresponding to an entire  $\text{H}_2$  molecule. This spherical approximation is justified as hydrogen is in the rotational ground state at the chosen state point. The second system contained 233 water molecules simulated using the qSPC/fw force field at 300 K and atmospheric pressure.<sup>32</sup> The datasets used for DeepMD were generated using i-PI and LAMMPS software packages.<sup>33,34</sup> Each system was simulated using normal mode PIMD with 32 replicas per particle for 250 picoseconds. A total of 8000 frames were taken from the final 200 picoseconds of each simulation, and the coordinates and forces were mapped to the centroid resolution.

The para-hydrogen model was trained using an embedding DNN with three hidden layers containing 10, 20, and 40 neurons, no timestep, and 46 nearest neighbors were considered as the local environment for the descriptor network. The fitting network was composed of three hidden layers of 240 neurons each, with a timestep. The learning rate schedule was exponential, with a starting rate of  $5 \times 10^{-3}$ , ending rate of  $1.76 \times 10^{-7}$ , and 5000 decay steps. The model was trained for 500,000 iterations before validation and testing. The water model was trained using an identical model except for the following differences: first, the number of neighbors considered for the local environment of each atom was increased to 60 hydrogens and 30 oxygens. Second, the model was trained for 100,000 iterations instead of 500,000. Additionally, the water training was batched with a batch size of 10. Both models were trained using DeepMD's `se_e2_a` descriptor which uses pairwise distances and embeds both radial and angular information about the system into the network. In order to obtain the most efficient models possible, the training length and local environment size were systematically reduced until dynamical fidelity was impacted, thus resulting in a model which balances accuracy with speed. Input

files for DeepMD training for both water and para-hydrogen are included in the [Supporting Information](#).

The resulting force fields were tested against TRPMD and PACMD considering accuracy of static and dynamic properties as well as integration efficiency. All ML-CMD simulations were carried out in LAMMPS with the DeepMD force field add-on using a 0.5 fs timestep. For both RPMD and PACMD, each system was simulated using i-PI and LAMMPS using a 0.25 fs timestep for PACMD and a 0.5 fs timestep for TRPMD and 32 replicas per atom. PACMD frequencies were chosen according to the following equation:

$$\Omega = P^{P(P-1)}/\beta\hbar, \quad (12)$$

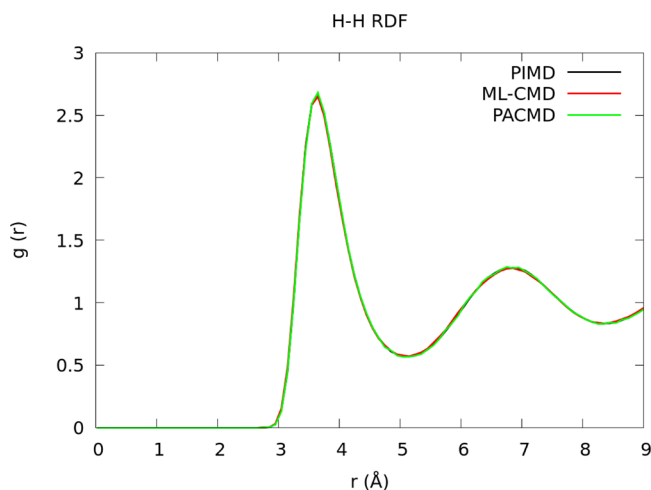
which for the para-hydrogen system is  $349 \text{ cm}^{-1}$  and for the water system is  $7481 \text{ cm}^{-1}$ .<sup>21</sup> CMD calculations for para-hydrogen were performed using a frequency of  $1741 \text{ cm}^{-1}$  and a timestep of 0.05 fs. In order to accurately compute diffusion and time correlation functions for these systems, a parallel scheme developed by Pérez et al. was used.<sup>35</sup> A total of 64 individual frames were selected from an initial PIMD simulation to generate uncorrelated starting points as shown in [Figure 1](#). Each starting point was then used to launch two 6 ps TRPMD, PACMD or ML-CMD trajectories, for a total of 128 simulations for each method. Velocity autocorrelation functions (VACFs) were computed from the final 5 ps of each of the resulting centroid trajectories and averaged to obtain a converged result. Self-diffusion constants were then obtained from the zero-frequency Fourier transform of the Kubo transformed VACF. It is noteworthy that the PIMD trajectory used to start each short simulation can double as a dataset for ML-CMD. In cases where such a scheme is used, one saves even more time as this simulation performs two critical steps instead of just one.

When measuring efficiency, all methods were tested using the same computational environment. Each test simulation was run on 32 cores of a 40 core Cascade-Lake compute node to allow for one core per replica in the full PI simulations. All simulations were carried out in LAMMPS. The total time used to calculate autocorrelation functions was calculated as the sum of all required simulations and training (“time to solution”).

For ML-CMD, this includes the 250 ps PIMD simulation used to generate the dataset which was also used as the starting points for the autocorrelation function calculations, the DeepMD network training time, and the total time to run the 768 total ps of simulation required for the autocorrelation functions. For PACMD and TRPMD, this includes the short PIMD trajectory required to generate starting configurations and the time to simulate the 768 ps of data for the autocorrelation functions.

### 3. RESULTS AND DISCUSSION

Both ML-CMD para-hydrogen and water results show excellent agreement with TRPMD and PACMD. Figures 2

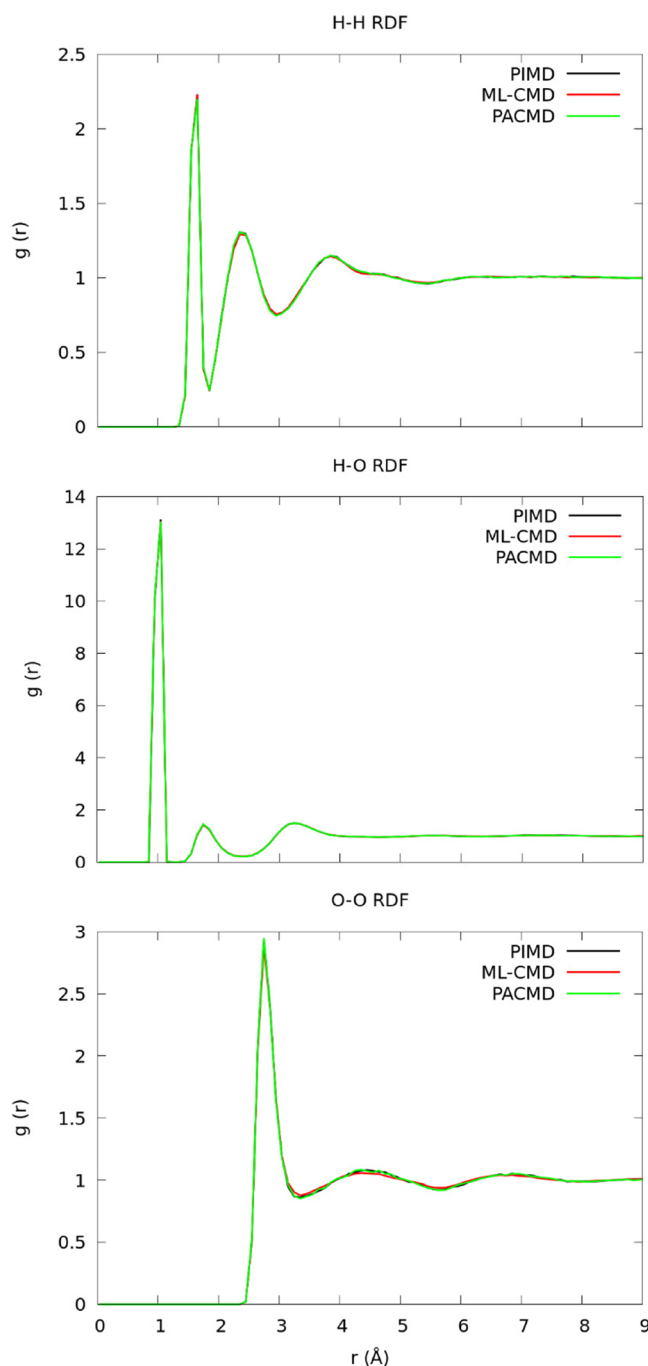


**Figure 2.** Plot of the centroid radial distribution function ( $g(r)$ ) for 14 K para-hydrogen. PIMD, ML-CMD, and PACMD are compared. As the Silvera–Goldman potential represents one hydrogen molecule as a single particle, no peak corresponding to the H–H bond is present.

and 3 show centroid radial distribution functions for all three models as well as PIMD. Of note are the water peaks corresponding to the O–H bond and H–O–H angles, which ML-CMD captures very accurately despite there being no explicit bonded or angular interactions in the force field. It is, however, not surprising that the static properties of ML-CMD align with full PI methods as PIMD is well known to be ideal for sampling the equilibrium positions of quantum particles. In addition, a previous attempt at calculating effective centroid forces from PIMD trajectories for CMD has shown similar levels of accuracy for para-hydrogen.<sup>25</sup>

The approximate quantum dynamics of the ML-CMD models are also well in line with the PACMD and TRPMD results. Figure 4 shows the VACF for all three types of simulation for para-hydrogen as well as CMD simulations with greater adiabatic separation. ML-CMD generally agrees better with PACMD than TRPMD, which is expected as the latter method is not designed to approximate centroid dynamics. CMD simulations in which the adiabatic separation is not partial show even better agreement with ML-CMD. This suggests that for systems with significant NQEs, ML-CMD is not only faster than PA-CMD but more accurate as well.

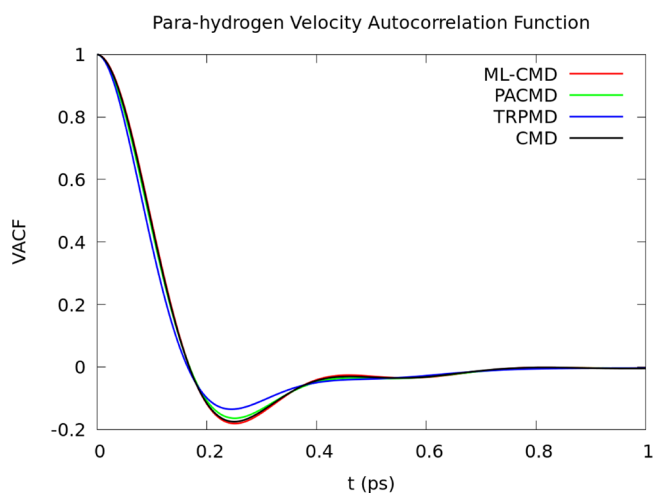
Figure 5 shows the VACFs of water. In this case, ML-CMD matches PACMD and TRPMD nearly perfectly. The fact that RPMD and PACMD converge better for room-temperature water than for 14 K para-hydrogen is not surprising. RPMD



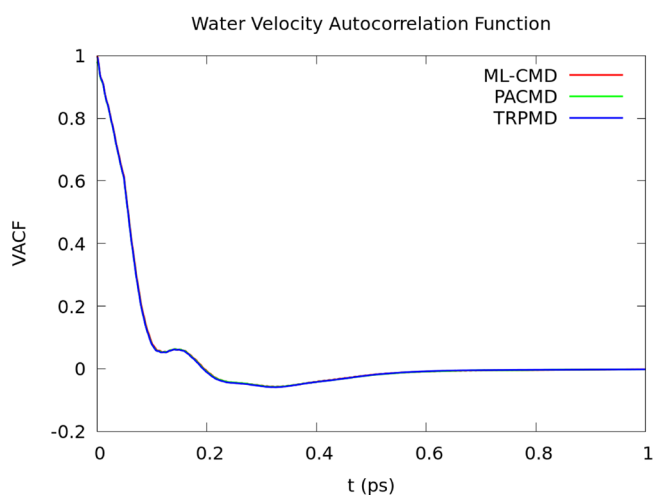
**Figure 3.** Plot of the centroid radial distribution functions ( $g(r)$ ) for hydrogen–hydrogen (top), hydrogen–oxygen (middle), and oxygen–oxygen (bottom) of 300 K water. PIMD, MLCMD, PACMD, and TRPMD are compared.

and CMD are known to converge as the system approaches the classical limit, e.g., heavier nuclei or higher temperatures.<sup>23</sup>

Table 1 next shows the calculated self-diffusion constants for both water and para-hydrogen for all three types of simulation. ML-CMD shows excellent agreement in both cases; it appears to be slightly more diffusive although this is well within the margin of error. The neural network architecture of ML-CMD's force field seems to directly contribute to the dynamical accuracy of the method. A previous attempt for accelerating CMD using force matching of the centroid forces



**Figure 4.** Plot of the normalized centroid VACFs for 14 K para-hydrogen using ML-CMD, PACMD, CMD, and TRPMD simulations. Each VACF was averaged over 128 individual trajectories.



**Figure 5.** Plot of the normalized centroid VACFs for 300 K using ML-CMD, PACMD, and TRPMD simulations. Each VACF was averaged over 128 individual trajectories.

**Table 1. Self-Diffusion Constants for Para-Hydrogen and Liquid Water<sup>a</sup>**

System	Para-hydrogen 14 K ( $\text{\AA}^2/\text{ps}$ )	Water 300 K ( $\text{\AA}^2/\text{ps}$ )
ML-CMD	$0.30 \pm 0.03$	$0.32 \pm 0.03$
PACMD	$0.29 \pm 0.03$	$0.31 \pm 0.03$
TRPMD	$0.28 \pm 0.03$	$0.31 \pm 0.04$

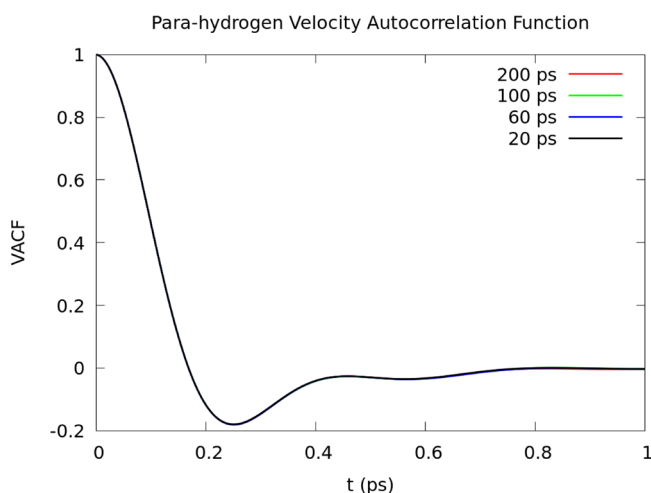
<sup>a</sup>All values were calculated using the zero frequency Fourier transform of the Kubo transformed VACF.

resulted in less-accurate values for self-diffusion at lower temperatures.<sup>25</sup>

An important factor in the training and validation of DNNs is the amount of data required to achieve a converged model. The ratio of training data to extrapolation informs not only the feasibility of the method but also its quality. We trained three additional para-hydrogen models using 20, 60, and 100 ps of PIMD reference data to test this. Table 2 shows diffusion coefficients for each of these models. Figure 6 shows VACFs comparing the models to the model trained on the full 200 ps reference dataset.

**Table 2. Diffusion Coefficients for ML-CMD Para-Hydrogen Trained on Datasets of Varying Lengths Compared to the Base Model, Which Was Trained on 200 ps of PIMD Simulation**

Reference simulation length (ps)	Diffusion coefficient ( $\text{\AA}^2/\text{ps}$ )
20	$0.31 \pm 0.03$
60	$0.30 \pm 0.03$
100	$0.31 \pm 0.03$
200	$0.30 \pm 0.03$



**Figure 6.** Plot of the normalized centroid VACFs for 14 K para-hydrogen using several trained ML-CMD models. Each model was trained with a different amount of reference PIMD data, and VACFs were calculated in the same manner as the preceding figures.

These results show that the method works with much less data than the full 200 ps trajectory. The diffusion constant values are well within a standard error of each other, and the autocorrelation functions match perfectly, even with 10 percent of the total training data used. Future applications of ML-CMD are thus likely to be deployed even more quickly and efficiently than the ones presented here.

Beyond accurate dynamics, the most important aspect of ML-CMD is its speed over methods such as full CMD, PACMD, RPMD, and TRPMD. In order to be useful as a replacement for these methods, it must be faster than them even when considering the time to collect the PIMD data and train the DNN force field. Table 3 shows integration speeds for the ML-CMD, PACMD, and TRPMD as well as the total time required to collect converged VACFs. ML-CMD performs over

**Table 3. Integration Speed and Velocity Autocorrelation Function (VACF) Calculation Times for ML-CMD, PACMD, and TRPMD<sup>a</sup>**

System	Para-hydrogen speed (ps/h)	Para-hydrogen VACF calculation time (h)	Water speed (ps/h)	Water VACF calculation time (h)
ML-CMD	734.4	21	144.5	49
PACMD	9.8	90	6.1	137
TRPMD	17.9	55	10.7	84

<sup>a</sup>All measurements performed using 32 cascade-lake cores. VACF calculation time for PACMD and ML-CMD includes an initial PIMD simulation followed by 768 ps of integration. For ML-CMD, this also includes DeepMD neural network training time.

70 times faster than PACMD for para-hydrogen and over 20 times faster for water. Additionally, the time to data collection is shorter than both PI-based methods by a considerable amount, even when considering the amount of excess PIMD data used to train the models in the present case. These results are encouraging, as the correlation function calculations require only 768 total picoseconds of integration, less than four times the total amount of simulation time used to train the model in the first place. In cases where more or larger simulations are required such as calculating long time correlation functions for complex free energy surfaces, ML-CMD becomes even more efficient, as the initial simulations and model training only need to be performed once. Furthermore, each ML-CMD model can be shared for use in multiple simulations either individually or through a publicly available repository of models which can be contributed by anyone using the method.

It is also worth discussing the flexibility of the machine learning algorithm chosen for ML-CMD. DeepMD is simple to deploy and yields excellent performance, but it is one of many similar neural network-based methods which can predict forces and energies in molecular systems. One of the most important aspects of any of these methods is how they encode translational, rotational, and permutational invariances into the input features of the model. This allows for the network to work in scalar space with the same symmetries present in the real chemical environment. Recently, several models have been developed which replace these invariant networks with equivariant ones which can directly encode vectored information such as forces. These models have been shown to require far less data than symmetry invariant models and can produce better results with up to one thousandth the number of training examples.<sup>36,37</sup> Applying such a method would naturally speed up the training of ML-CMD models and should be considered for future study.

#### 4. CONCLUSIONS

In this work, we have presented ML-CMD, a machine learning approach for the calculation of the effective centroid potential in CMD simulations. Over the past 30 years, regular CMD has been limited by the need to repeatedly calculate the centroid effective force “on-the-fly” through adiabatic separation of the centroid and noncentroid imaginary time PI modes. Such calculations represent a significant computational overhead for larger systems. Instead, in the present work, we have shown that by training a neural network to first learn the centroid effective potential from a PIMD simulation, one can greatly increase the efficiency and time to solution of a CMD simulation without sacrificing accuracy.

We have demonstrated that for both room-temperature water and 14 K para-hydrogen, ML-CMD provides highly accurate results which closely match both the static and dynamic properties given by full PACMD and TRPMD simulations. The ML-CMD simulations are also many times faster than either of the latter approaches, thereby extending the range of systems for which CMD will be applicable. While the initial PIMD simulations and DNN training steps somewhat constrain the overall speed at which the ML-CMD models can be deployed, the PIMD simulations are also necessary to provide the initial conditions for trajectories in both PACMD and TRPMD.

This paper provides yet another example where machine learning promises to transform the field of molecular

simulation, in this case by making CMD simulations feasible and accurate for a wider range of systems. Likewise, one can expect that future advances in machine learning, e.g., to better treat heterogenous systems, rare events, etc., will also provide clear benefits to the ML-CMD approach developed in this work.

#### ■ ASSOCIATED CONTENT

##### Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jctc.2c00706>.

The files used to train the para-hydrogen and water DeepMD models for use in replicating the data from this manuscript, as well as for training additional ML-CMD models (PDF)

#### ■ AUTHOR INFORMATION

##### Corresponding Author

Gregory A. Voth – Department of Chemistry, Chicago Center for Theoretical Chemistry, James Franck Institute, and Institute for Biophysical Dynamics, The University of Chicago, Chicago, Illinois 60637, United States; [orcid.org/0000-0002-3267-6748](https://orcid.org/0000-0002-3267-6748); Email: [gavoth@uchicago.edu](mailto:gavoth@uchicago.edu)

##### Authors

Timothy D. Loose – Department of Chemistry, Chicago Center for Theoretical Chemistry, James Franck Institute, and Institute for Biophysical Dynamics, The University of Chicago, Chicago, Illinois 60637, United States  
Patrick G. Sahrman – Department of Chemistry, Chicago Center for Theoretical Chemistry, James Franck Institute, and Institute for Biophysical Dynamics, The University of Chicago, Chicago, Illinois 60637, United States; [orcid.org/0000-0002-4781-9561](https://orcid.org/0000-0002-4781-9561)

Complete contact information is available at: <https://pubs.acs.org/doi/10.1021/acs.jctc.2c00706>

##### Notes

The authors declare no competing financial interest.

#### ■ ACKNOWLEDGMENTS

This material is based on work supported by the National Science Foundation (NSF Grant CHE-2102677). Simulations were performed using computing resources provided by the University of Chicago Research Computing Center (RCC). T.D.L. was also supported by the National Science Foundation Graduate Research Fellowship (DGE-1746045). We thank Dr. Chenghan Li and Dr. Won Hee Ryu for helpful conversations on path integral simulations.

#### ■ REFERENCES

- (1) Warshel, A.; Kato, M.; Pisiakov, A. V. Polarizable Force Fields: History, Test Cases, and Prospects. *J. Chem. Theory Comput.* **2007**, *3*, 2034–2045.
- (2) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D., Jr. CHARMM General Force Field: A Force Field for Drug-like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields. *J. Comput. Chem.* **2009**, *31*, 671–690.

- (3) Tuckerman, M. E. Ab Initio Molecular Dynamics: Basic Concepts, Current Trends and Novel Applications. *J. Phys. Condens. Matter* **2002**, *14*, R1297–R1355.
- (4) Carloni, P.; Rothlisberger, U.; Parrinello, M. The Role and Perspective of Ab Initio Molecular Dynamics in the Study of Biological Systems. *Acc. Chem. Res.* **2002**, *35*, 455–464.
- (5) Warshel, A.; Levitt, M. Theoretical Studies of Enzymic Reactions: Dielectric, Electrostatic and Steric Stabilization of the Carbonium Ion in the Reaction of Lysozyme. *J. Mol. Biol.* **1976**, *103*, 227–249.
- (6) Feynman, R. P.; Hibbs, A. R. *Quantum Mechanics and Path Integrals*; Courier Corporation, 1965.
- (7) Berne, B. J.; Thirumalai, D. On the Simulation of Quantum Systems: Path Integral Methods. *Annu. Rev. Phys. Chem.* **1986**, *37*, 401–424.
- (8) Markland, T. E.; Ceriotti, M. Nuclear Quantum Effects Enter the Mainstream. *Nat. Rev. Chem.* **2018**, *2*, No. 0109.
- (9) Cao, J.; Voth, G. A. A New Perspective on Quantum Time Correlation Functions. *J. Chem. Phys.* **1993**, *99*, 10070–10073.
- (10) Cao, J.; Voth, G. A. The Formulation of Quantum Statistical Mechanics Based on the Feynman Path Centroid Density. II. Dynamical Properties. *J. Chem. Phys.* **1994**, *100*, 5106–5117.
- (11) Cao, J.; Voth, G. A. The Formulation of Quantum Statistical Mechanics Based on the Feynman Path Centroid Density. I. Equilibrium Properties. *J. Chem. Phys.* **1994**, *100*, 5093–5105.
- (12) Jang, S.; Voth, G. A. Path Integral Centroid Variables and the Formulation of Their Exact Real Time Dynamics. *J. Chem. Phys.* **1999**, *111*, 2357–2370.
- (13) Jang, S.; Voth, G. A. A Derivation of Centroid Molecular Dynamics and Other Approximate Time Evolution Methods for Path Integral Centroid Variables. *J. Chem. Phys.* **1999**, *111*, 2371–2384.
- (14) Craig, I. R.; Manolopoulos, D. E. Chemical Reaction Rates from Ring Polymer Molecular Dynamics. *J. Chem. Phys.* **2005**, *122*, No. 084106.
- (15) Miller, T. F.; Manolopoulos, D. E. Quantum Diffusion in Liquid Para -Hydrogen from Ring-Polymer Molecular Dynamics. *J. Chem. Phys.* **2005**, *122*, 184503.
- (16) Miller, T. F.; Manolopoulos, D. E. Quantum Diffusion in Liquid Water from Ring Polymer Molecular Dynamics. *J. Chem. Phys.* **2005**, *123*, 154504.
- (17) Jang, S.; Sinitskiy, A. V.; Voth, G. A. Can the Ring Polymer Molecular Dynamics Method Be Interpreted as Real Time Quantum Dynamics? *J. Chem. Phys.* **2014**, *140*, 154103.
- (18) Hele, T. J. H.; Willatt, M. J.; Muolo, A.; Althorpe, S. C. Communication: Relation of Centroid Molecular Dynamics and Ring-Polymer Molecular Dynamics to Exact Quantum Dynamics. *J. Chem. Phys.* **2015**, *142*, 191101.
- (19) Witt, A.; Ivanov, S. D.; Shiga, M.; Forbert, H.; Marx, D. On the Applicability of Centroid and Ring Polymer Path Integral Molecular Dynamics for Vibrational Spectroscopy. *J. Chem. Phys.* **2009**, *130*, 194510.
- (20) Paesani, F.; Voth, G. A. A Quantitative Assessment of the Accuracy of Centroid Molecular Dynamics for the Calculation of the Infrared Spectrum of Liquid Water. *J. Chem. Phys.* **2010**, *132*, No. 014105.
- (21) Habershon, S.; Fanourgakis, G. S.; Manolopoulos, D. E. Comparison of Path Integral Molecular Dynamics Methods for the Infrared Absorption Spectrum of Liquid Water. *J. Chem. Phys.* **2008**, *129*, No. 074501.
- (22) Rossi, M.; Ceriotti, M.; Manolopoulos, D. E. How to Remove the Spurious Resonances from Ring Polymer Molecular Dynamics. *J. Chem. Phys.* **2014**, *140*, 234116.
- (23) Hone, T. D.; Rosky, P. J.; Voth, G. A. A Comparative Study of Imaginary Time Path Integral Based Methods for Quantum Dynamics. *J. Chem. Phys.* **2006**, *124*, 154103.
- (24) Markland, T. E.; Manolopoulos, D. E. An Efficient Ring Polymer Contraction Scheme for Imaginary Time Path Integral Simulations. *J. Chem. Phys.* **2008**, *129*, No. 024105.
- (25) Hone, T. D.; Izvekov, S.; Voth, G. A. Fast Centroid Molecular Dynamics: A Force-Matching Approach for the Predetermination of the Effective Centroid Forces. *J. Chem. Phys.* **2005**, *122*, No. 054105.
- (26) Fletcher, T.; Zhu, A.; Lawrence, J. E.; Manolopoulos, D. E. Fast Quasi-Centroid Molecular Dynamics. *J. Chem. Phys.* **2021**, *155*, 231101.
- (27) Wang, H.; Zhang, L.; Han, J.; Weinan, E. DeePMD-Kit: A Deep Learning Package for Many-Body Potential Energy Representation and Molecular Dynamics. *Comput. Phys. Commun.* **2018**, *228*, 178–184.
- (28) Zhang, L.; Han, J.; Wang, H.; Car, R.; E, W. Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. *Phys. Rev. Lett.* **2018**, *120*, No. 143001.
- (29) Zhang, L.; Han, J.; Wang, H.; Saidi, W. A.; Car, R.; E, W. End-to-End Symmetry Preserving Inter-Atomic Potential Energy Model for Finite and Extended Systems. In *Advances in Neural Information Processing Systems*; Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R., Ed.; Curran Associates, Inc., 2018.
- (30) Li, C.; Voth, G. A. Using Machine Learning to Greatly Accelerate Path Integral Ab Initio Molecular Dynamics. *J. Chem. Theory Comput.* **2022**, *18*, 599–604.
- (31) Silvera, I. F.; Goldman, V. V. The Isotropic Intermolecular Potential for H<sub>2</sub> and D<sub>2</sub> in the Solid and Gas Phases. *J. Chem. Phys.* **1978**, *69*, 4209–4213.
- (32) Paesani, F.; Zhang, W.; Case, D. A.; Cheatham, T. E.; Voth, G. A. An Accurate and Simple Quantum Model for Liquid Water. *J. Chem. Phys.* **2006**, *125*, 184507.
- (33) Kapil, V.; Rossi, M.; Marsalek, O.; Petraglia, R.; Litman, Y.; Spura, T.; Cheng, B.; Cuzzocrea, A.; Meißner, R. H.; Wilkins, D. M.; et al. I-PI 2.0: A Universal Force Engine for Advanced Molecular Simulations. *Comput. Phys. Commun.* **2019**, *236*, 214–223.
- (34) Thompson, A. P.; Aktulga, H. M.; Berger, R.; Bolintineanu, D. S.; Brown, W. M.; Crozier, P. S.; 't Veld, P. J.; Kohlmeyer, A.; Moore, S. G.; Nguyen, T. D.; et al. LAMMPS - a Flexible Simulation Tool for Particle-Based Materials Modeling at the Atomic, Meso, and Continuum Scales. *Comput. Phys. Commun.* **2022**, *271*, No. 108171.
- (35) Pérez, A.; Tuckerman, M. E.; Müser, M. H. A Comparative Study of the Centroid and Ring-Polymer Molecular Dynamics Methods for Approximating Quantum Time Correlation Functions from Path Integrals. *J. Chem. Phys.* **2009**, *130*, 184105.
- (36) Haghghatari, M.; Li, J.; Guan, X.; Zhang, O.; Das, A.; Stein, C. J.; Heidar-Zadeh, F.; Liu, M.; Head-Gordon, M.; Bertels, L.; et al. NewtonNet: A Newtonian Message Passing Network for Deep Learning of Interatomic Potentials and Forces. *Digit. Discov.* **2022**, *1*, 333.
- (37) Batzner, S.; Musaelian, A.; Sun, L.; Geiger, M.; Mailoa, J. P.; Kornbluth, M.; Molinari, N.; Smidt, T. E.; Kozinsky, B. E(3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials. *Nat. Commun.* **2022**, *13*, 2453.