High-Dimensional Neural Network Potentials for Accurate Prediction of Equation of State: A Case Study of Methane

Mostafa Abedi,∗† Jörg Behler,‡¶ and C. Franklin Goldsmith∗†

†School of Engineering, Brown University, Providence, Rhode Island 02906, United States
‡Lehrstuhl für Theoretische Chemie II, Ruhr-Universität Bochum, 44780 Bochum, Germany
¶Research Center Chemical Sciences and Sustainability, Research Alliance Ruhr, 44780 Bochum, Germany

E-mail: mostafa_abedi@brown.edu; franklin_goldsmith@brown.edu
Abstract

Machine learning-based interatomic potentials, such as those provided by neural networks, are increasingly important in molecular dynamics simulations. In the present work, we consider the applicability and robustness of machine learning molecular dynamics to predict the equation of state properties of methane using high-dimensional neural network potentials (HDNNPs). We investigate two different strategies for generating training data: one based upon bulk representations using periodic cells, and another based upon clusters of molecules. We assess the accuracy of the trained potentials by predicting the equilibrium mass density for a wide range of thermodynamic conditions to characterize the liquid phase, supercritical fluid and gas phase, as well as the liquid-vapor coexistence curve. Our results show an excellent agreement with reference phase diagrams, with an average error below $\sim 2\%$ for all studied phases. Moreover, we confirm the applicability of models trained on cluster data set for producing accurate and reliable results.

Keywords

high-dimensional neural network potentials, molecular dynamics, equation of state, methane
Introduction

Molecular dynamics (MD) simulations can provide detailed and comprehensive dynamic information at the atomic level. Such information is helpful not only for interpreting experimental observations, but it also can aid in the design and optimization of novel materials and chemical processes. The quality and predictive power of MD simulations depend on the accuracy of the interatomic potential energy model that is used to compute the forces acting on the nuclei. Numerous potential models have been developed for a broad range of molecular systems at different levels of accuracy and applicability. In particular, machine learning potentials (MLP) are increasingly popular, due to their accuracy and efficiency. Emerging MLP have indeed allowed scientists to push the boundaries of MD simulations to larger systems and longer time scales with \textit{ab initio} accuracy.

In principle, the aim in MLP is to define a mathematical function (or model) that describes the relationships between the atomic structure and the corresponding potential energy of the system. This task begins with the training of the MLP model on a set of reference data through a fitting procedure. The main part of the training process, e.g., in the case of neural networks (NN), is to optimize the weight parameters by minimizing the error between the ML predictions and the reference values. The reference data generally contains atomic configurations and their corresponding potential energies (and interatomic forces) obtained from electronic structure calculations, typically using density functional theory (DFT). High-dimensional neural network potentials (HDNNPs), developed by Behler and Parrinello, are a powerful and promising approach that has been successfully applied for many molecular, condensed, and materials systems. We refer the interested readers to the recent reviews on HDNNPs.

Accurate prediction of equation of state (EOS) is crucial in order to describe the thermodynamic and kinetic properties of molecular systems and chemical reactions. Development of potentials for such purposes has been a challenge, since extensive parameterization is required at a broad range of temperatures and pressures ($T$, $P$). This task can be even more
challenging for systems in which the intermolecular interactions are highly affected by small changes in temperature or pressure.

In this work, we aim to explore the ability of HDNNP to predict the EOS properties accurately. To do so, we choose bulk methane as the test system, and our goal is to compute the mass density with the accuracy of high-level DFT calculations. The methane system is ideal for our purposes, since its phase diagram and EOS data are well known and available. We also compute the radial distribution function (RDF), which is a useful characterization of the local distribution and structure of the molecules in bulk systems, and compare the results to experiments. Finally, we consider the quality of the model trained on a cluster data set as compared to data obtained from bulk calculations. Applying the cluster data sets for MLP development has recently been considered for water systems, liquid electrolytes, hydrocarbons, and metal systems to access high-level electronic structure methods and shown a successful performance for transferring the locally trained ML models to bulk and polymer systems.

**Computational Details**

**Construction of Initial Data Sets**

Two distinct sets of training data were generated. The first set, referred to as the “bulk”, utilized hybrid Gaussian and plane waves (GPW) basis functions; the second set, “cluster”, was performed using Gaussian-type orbital (GTO) basis functions. The bulk reference data set were constructed through several short-run *ab initio* molecular dynamic (AIMD) simulations with constant number of particles, volume, and temperature (*NVT*) under periodic boundary conditions (PBCs) using the Quickstep module of the CP2K program package. All potential energies and interatomic forces were computed by employing the meta-GGA B97M functional and D3(BJ) dispersion correction in combination with the TZVP-MOLOPT-GTH basis set using a plane-wave cutoff energy of 600 Ry.
The cluster data set was obtained using the ORCA program package\textsuperscript{37–39}. The AIMD simulations were performed using the \( \omega \text{B97M-D3(BJ)} \)\textsuperscript{33–35} functional along with the def2-TZVP(-f) basis set\textsuperscript{40,41}. The \( \omega \text{B97M-D3(BJ)} \) is a parametrized and efficient version of the \( \omega \text{B97M-V} \) functional\textsuperscript{42,43}. In order to maintain a constant mass density during the AIMD simulations, the system was confined to a well-defined cubic volume utilizing harmonic repulsive soft walls using the keyword \texttt{Cell} as implemented in the ORCA MD module. By virtue of including these repulsive walls, more sophisticated DFT functionals with GTO basis functions – as would be appropriate for molecular systems – can be used, instead of less accurate methods that are necessary for periodic calculations. We note that both applied DFT functionals have shown excellent performance for describing the non-covalent interactions necessary for fluid simulations\textsuperscript{43}.

The AIMD simulations were performed for multiple uncorrelated configurations constituted by 50 methane molecules for the bulk data set and 25 for the cluster data set, produced using the Packmol package\textsuperscript{44}. The initial configurations were generated at different representative mass densities, which range from a low-density gas to a compressed liquid. A temperature range of 1000, 2000, and 4000 K was used to ensure that the system was not biased towards lower-entropy configurations. This strategy of different densities at elevated temperatures allows us to generate training data that covers a wide range of phase space, both in terms of intramolecular and intermolecular interactions. Each AIMD simulation was run for 2000 (bulk) and 1000 (cluster) steps, with a timestep of 0.5 fs, and the configurations were collected at each timestep; the bulk simulations required more time to converge. We generated 17000 and 18000 initial data points for the bulk and cluster data sets, respectively.

**Development of HDNNPs**

The training and HDNNP weight optimization were performed on both potential energies and interatomic forces using the neural network potential package (N2P2) developed by Singraber and co-workers\textsuperscript{45}. The N2P2 code exploits a multistream Kalman filter\textsuperscript{46,47} optimizer.
that facilitates the training process, which is especially useful when the data set is large and forces are included in the training. The local atomic environments were described within a cutoff value of 12 Bohr (6.35 Å) using a set of optimal many-body atom-centered radial and angular symmetry functions consisting 64/64 and 55/47 (bulk/cluster) functions for H and C atoms, respectively. The chosen cutoff value is large enough to capture all necessary short- and long-range interactions for the bulk methane within the cutoff sphere. The data were split randomly into two sets in which 90% of them were used as a training set for optimizing the HDNNP weight parameters, and the remaining 10% were set aside as an independent test set to inspect the possibility of overfitting and the reliability of the HDNNP models.

**Machine Learning-Based Molecular Dynamics Simulations**

The machine learning molecular dynamics (MLMD) simulations were carried out in a cubic box containing 512 methane molecules under PBCs using the LAMMPS/N2P2 interface. The initial configurations were generated using the `replicate` command of LAMMPS according to the target mass densities. The system was first equilibrated for 1 ns in the isobaric-isothermal NPH (constant number of particles, pressure, and enthalpy) ensemble coupled to a Langevin thermostat to keep the temperature constant. The equilibrated system was then used for a 5 ns production run in the NPT (constant number of particles, pressure, and temperature) ensemble using a Nosé-Hoover (NH) chain thermostat and barostat. The pressure and temperature damping factors were set, respectively, to 1000 and 100 timesteps. The timestep of 1.0 fs was applied to all MLMD simulations, unless specified otherwise, and the data were collected every 20 fs. The sufficiency of the simulation box size was inspected by doubling the box size for several representative simulations and confirming no change in the predicted values. The results confirmed that the selected box size is adequate for obtaining the equilibrated mass density (see Table S1 in the Supporting Information). The reference mass densities were calculated using EOS as implemented
in the CANTERA software,\textsuperscript{53} which is based on the computational equations and methods formulated by Reynolds.\textsuperscript{61}

Results

Evaluation of HDNNPs

Table 1: The RMSE values for energy and forces obtained after 30 epochs of training for the final bulk and cluster data sets.

<table>
<thead>
<tr>
<th>NN Architecture</th>
<th>$E^{\text{RMSE}}$ (meV/atom)</th>
<th>$F^{\text{RMSE}}$ (meV/Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bulk</td>
<td>Cluster</td>
</tr>
<tr>
<td></td>
<td>Training</td>
<td>Test</td>
</tr>
<tr>
<td>15-15</td>
<td>0.66</td>
<td>0.64</td>
</tr>
<tr>
<td>15-15-15</td>
<td>0.42</td>
<td>0.43</td>
</tr>
<tr>
<td>20-20</td>
<td>0.45</td>
<td>0.46</td>
</tr>
<tr>
<td>20-20-20</td>
<td>0.47</td>
<td>0.47</td>
</tr>
<tr>
<td>25-25</td>
<td>0.38</td>
<td>0.37</td>
</tr>
<tr>
<td>25-25-25</td>
<td>0.50</td>
<td>0.48</td>
</tr>
<tr>
<td>30-30</td>
<td>0.57</td>
<td>0.59</td>
</tr>
<tr>
<td>30-30-30</td>
<td>0.44</td>
<td>0.45</td>
</tr>
</tbody>
</table>

The stability and accuracy of the trained HDNNP models were evaluated and improved by adding relevant configurations through an active learning procedure, as discussed in Ref. 13. Starting from the initial data sets, we trained eight preliminary models with different atomic NN architectures, including two and three hidden layers with the number of nodes ranging from 15 to 30 (see Table 1). The HDNNP with the lowest energy and forces root mean squared errors (RMSEs) was selected as the reference one, and its performance was assessed by predicting the energy and forces for 1500 new configurations obtained from simulations at different representative low and high ($T$, $P$) conditions. The missing structures were identified and added to the data sets, and the models were retrained. The final HDNNPs were obtained after multiple active learning iterations, at which point the errors were well below the minimum required convergence values of 2 meV/atom (energy) and 200 meV/Å (force). The RMSE results are reported in Table 1. The model with the (30,30,30) NN structure was
selected as the best one for both bulk and cluster data sets. The corresponding training and weight parameters for the selected models can be found in the Supporting Information.

Figure 1: Variation of energy and force errors of the reference HDNNP with respect to other potential models for both bulk, (a) and (b), and cluster, (c) and (d), data sets. The mean ($\mu$) and standard deviation ($\sigma$) values are reported.

Figure 1 represents the variation of energy and force component errors of the final selected potential with respect to other trained models. This procedure resulted in collecting 18255 and 22500 configurations from the final bulk and cluster data sets, respectively.

These results can be further analysed to assess the performance of the two data sets in terms of sensitivity and generalization of the potentials. A comparison of the obtained mean ($\mu$) and standard deviation ($\sigma$) values in Figure 1a and Figure 1c, suggests that the bulk method offers a modest improvement in energy compared to the cluster method, with the cluster method having a slightly longer tail at higher positive errors. However, with regard to prediction of the interatomic force, the difference between the two methods is negligible,
as shown in Figures 1b and Figure 1d. Ultimately, the two methods are quite similar in
their performance, and the overall accuracy of the trained models for both data sets are
comparable, as detailed by the RMSEs in Table 1. Collectively, these results suggest that
the cluster model cannot be immediately dismissed.

Prediction of Mass Density Equation of State

![Graphs showing mass density and pressure relationship](image)

Figure 2: Plots of mass density of methane as a function of pressure at different temperatures
obtained from the NPT MLMD simulations. The dots correspond to the simulation results,
and the solid lines represent the reference EOS values, as obtained from Cantera. Error
bars are the standard deviation from the mean in the related trajectory. The subplots display
the corresponding percent errors with respect to the reference EOS values.

To characterize the liquid phase, supercritical fluid, and gas phase of bulk methane, we
performed extensive MLMD simulations for a broad range of \((T, P)\) conditions, \(100 \text{ K} \leq T \leq
800 \text{ K}\) and \(1 \text{ bar} \leq P \leq 1000 \text{ bar}\), using our selected HDNNP models. One of the challenging
issues with simulations at low-density conditions is to obtain the converged mass density.
This problem can usually be overcome by running the simulations for sufficiently long time,
which is computationally expensive. An alternative solution is to minimize the volume
(and subsequently the mass density) fluctuations by freezing the intramolecular vibrational degrees of freedom. Accordingly, we performed rigid-body dynamics simulations for several selected conditions in the gas phase. We believe that this strategy is reasonable, since our system is non-reactive, i.e. bonds are neither broken nor formed, and also at such low-density states the intermolecular interactions are very week. Figure 2 summarizes the calculated mass density as a function of pressure obtained from the $NPT$ simulations. The solid lines represent the reference data, as obtained from CANtera. We obtained excellent average percent errors of 1.5\% and 2.2\% for all considered thermodynamic conditions using the bulk- and cluster-based models, respectively, with the maximum error of $\sim 7\%$, as shown in the subplots, which report the absolute percent error values. The errors from statistical uncertainties are not considered. The results clearly show the potential of HDNNPs to predict the phase behaviour and bulk properties of molecular systems; additionally, these results also demonstrate the capability potentials trained on cluster data.
Figure 3: Plots of compressibility factor of the methane gas phase and supercritical fluid as a function of pressure at different temperatures (400 K, 600 K, and 800 K) obtained from the NPT MLMD simulations. The dots correspond to the simulation results, and the solid lines represent the reference EOS values. Error bars are the standard deviation from the mean in the related trajectory. The subplots display the corresponding percent errors with respect to the reference EOS values.

An alternate way to present the gas- and supercritical-phase data is through the compressibility factor, \( Z = PM/\rho RT \), where \( P \) is pressure, \( M \) is the molar mass, and \( R \) is the gas constant. An ideal gas has a value of \( Z = 1 \); compressibility factors less than unity indicate a reduced molar volume due to the dominance of attractive forces, and a compressibility factor greater than unity indicate strong repulsive effects. The results for the compressibility factor are presented in Figure 3. The average errors of 1.5% (bulk) and 2.3% (cluster) are obtained, similar to ones reported for the mass density. Such low deviations of \( Z \) from the reference data can be interpreted as an indication for sufficient capturing of both the repulsive and attractive parts of phase space in our data sets.
Figure 4: Plot of methane liquid-vapor coexistence (LVC) curve obtained from the \textit{NPT} MLMD simulations. Error bars are the standard deviation from the mean in the related trajectory. The solid red lines show the reference EOS values, as obtained from \textsc{Cantera}. The red star represents the EOS critical point.

To complete our simulations for the phase behavior of methane, we used the HDNNP models to reproduce the liquid-vapor coexistence (LVC) curve. The simulations were carried out at 7 different temperatures, ranging from 120 K to 180 K in increments of 10 K. For each temperature, the corresponding pressures in gas and liquid regions were specified in order to describe the LVC line as precise as possible. We performed the MLMD simulations for 14 selected \((T, P)\) conditions (see Table S5). To maintain the stability of simulations, we applied a small timestep of 0.25 fs and the \textit{NPT} simulations were run for 2.0 ns. In Figure 4, we report the mass density values and demonstrate the LVC curve modeled. The predicted coexistence curve is in good agreement with the reference EOS values from \textsc{Cantera}, with modest errors at the highest temperatures. These deviations can be attributed to the fact that characterizing such conditions near the critical point is challenging from both potential model and simulation protocol perspectives. The LVC of methane has been previously calculated
by several studies using \textit{ab initio} and force field Monte Carlo simulations.\textsuperscript{55-57} Wohlfahrt et al.\textsuperscript{58} reported the most accurate LVC of water employing a HDNNP at the DFT level. We provide the raw data including predicted mass densities and corresponding EOS values for all studied thermodynamic conditions in Tables S2-4 of the Supporting Information.

![Graph](image)

Figure 5: The RDFs, $g(r)$, of bulk methane for the C–C pair obtained from the \textit{NPT} MLMD simulations at 91.6 K and 0.133 bar (100 mmHg). The experimental data (black dots) are taken from the X-ray diffraction study of Habenschuss et. al.\textsuperscript{59}

Finally, we apply our HDNNP models to compute the radial distribution functions using the visual molecular dynamics (VMD) software\textsuperscript{60} with a bin size of 0.1 Å for the radial sampling. Habenschuss et al.\textsuperscript{59} reported experimental carbon-carbon RDFs for liquid methane using x-ray diffraction measurements at $T = 92 \pm 1$ K and $P = \sim 0.133$ bar (near the triple point). In order to reproduce the experimental data, we conducted the \textit{NPT} MLMD simulations at the same ($T$, $P$) condition using our trained models, and we then calculated the corresponding RDF. Figure 5 shows the RDFs for the C-C pair obtained from the simulations using the bulk- and cluster-based models. As it can be seen, our results agree well with the experimental data (black dots). The first, second, and third peaks (maxima) for the simulated RDFs are located at 4, 7.7, and 11 Å, which are close to the experimentally obtained values of 4, 7.5, and 10.7 Å\textsuperscript{59}. The results were further analyzed in terms of the cu-
accumulative number. The simulations predict that 12/12 and 54.3/55.5 (bulk/cluster) methane molecules reside in the first and second coordination shells, respectively, which are in excellent agreement with the reported experimental values of $\simeq 12$ and $\simeq 55.5$. The RDF results indeed indicate the accuracy and reliability of our trained HDNNPs for prediction of local bulk density.

**Discussion**

The present results demonstrate the performance of HDNNPs for capturing interatomic interactions at low and high temperature and pressure domains of methane. The models were trained on bulk and cluster data sets derived from DFT energy and atomic forces calculations. Using the bulk-based model, we achieved mean absolute errors (MAEs) of 0.26 meV/atom and 30.50 meV/Å for the potential energies and atomic forces, respectively, compared to the reference DFT. Our MLMD simulations revealed an average error of 1.5% for this model when compared to EOS for the condensed and gas phases. The cluster-based model, with MAEs of 0.20 meV/atom (energies) and 31.50 meV/Å (forces), has produced promising and accurate results (within 2% error) and proven to be a successful and significant approach for developing ML potentials, thanks to the local, atom-centered descriptors. The strategy of training potential models on local cluster data is potentially beneficial for molecular systems and chemical reactions where high levels of accuracy are necessary. We note that in this study, we observed no noticeable difference in accuracy between the two applied meta-GGA and range-separated hybrid meta-GGA DFT functionals, and they both describe the system well.

One of the main bottlenecks in MLP development is the computational cost for generating data set through *ab initio* calculations. The number of structures in each configuration plays a major role in this regard, and the optimal number of molecules can be identified. In this work, to construct the cluster data set, we used 25 methane molecules and ran 18
trajectories at different selected temperatures and mass densities for 1000 steps. For the
bulk calculations, in principle, the system must be large enough to avoid any finite size
effects of the simulation box, and more simulation time is required to equilibrate the system.
We use 50 methane molecules and ran 17 trajectories for 2000 steps. The first 1000 steps
were discarded as the pre-equilibration portion. In terms of calculation time, our AIMD
simulations required \( \sim 700,000 \) CPU-hours for the bulk data and \( \sim 500,000 \) CPU-hours for
the cluster data. In other words, generating the cluster data was nearly 30% more efficient,
despite being a higher-level of theory. Furthermore, it must be noted that applying hybrid
meta-GGA functionals for the bulk calculations would be prohibitively expensive from the
perspective of computational resources.

The two methods were also evaluated regarding computational cost for the HDNNP train-
ing. The size of cluster data set is 1.5X smaller than the bulk one, and in the absence of PBC,
it is expected that the training becomes faster. Our analysis revealed overall time of 15,000
and 10,000 CPU hours to train all eight models (Table 1) on the bulk and cluster data sets,
respectively. These results show that using cluster data sets for HDNNP developments can
increase the computational efficiency appreciably, in addition to a significant improvement
of the level of theory.

The presented results have indicated that HDNNPs can be readily applied to describe
the phase behaviour and thermophysical properties of homogeneous systems at high levels
of accuracy with an acceptable range of error.

**Conclusion**

In this work, we have explored and evaluated the validity of HDNNP for describing the phase
behavior of molecular systems. We performed extensive MLMD simulations to characterize
the liquid phase, supercritical fluid, gas phase, and LVC curve of bulk methane, in terms of
mass density, using our MLP models trained on DFT data. We also assessed the possibility
and validation of applying the potential model trained on local cluster data set compared to the conventional bulk one. Our results show strong agreement between the predicted and the reference EOS data with the average errors of $\sim 2\%$. The LVC interface was reproduced with reasonable precision. Furthermore, we confirmed the capability of the cluster-based model for providing accurate and reliable results. We further evaluated our potentials in terms of predicting the local density and describing molecular distribution by computing the RDFs and a perfect agreement was achieved with the experimental data.

Acknowledgement

This work was supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences and Biosciences, Gas-Phase Chemical Physics Program under Grant No. DE-SC0019489, with Dr. Wade Sisk as the program manager, and by the Office of Naval Research under Grant No. N00014-21-1-2291, with Dr. Chad Stoltz as the program manager. M.A. is thankful to Dr. Bjarne Kreitz for fruitful and helpful discussions.

References


(39) Neese, F. Software update: The ORCA program system-Version 5.0. WIREs Computational Molecular Science 2022, 12, e1606.


