

DeePKS+ABACUS as a Bridge between Expensive Quantum Mechanical Models and Machine Learning Potentials

Wenfei Li,^{†,⊥} Qi Ou,^{†,⊥} Yixiao Chen,[‡] Yu Cao,[¶] Renxi Liu,[¶] Chunyi Zhang,[§]
Daye Zheng,[†] Chun Cai,^{†,#} Xifan Wu,[§] Han Wang,^{*,||} Mohan Chen,^{*,¶} and
Linfeng Zhang^{*,†,#}

[†]*AI for Science Institute, Beijing 100080, P.R. China*

[‡]*Program in Applied and Computational Mathematics, Princeton University, Princeton,
NJ 08544, USA*

[¶]*HEDPS, CAPT, College of Engineering and School of Physics, Peking University, Beijing
100871, P.R. China*

[§]*Department of Physics, Temple University, Philadelphia, PA 19122, USA*

^{||}*Laboratory of Computational Physics, Institute of Applied Physics and Computational
Mathematics, Huayuan Road 6, Beijing 100088, P.R. China*

[⊥]*Contributed equally to this work*

[#]*DP Technology, Beijing 100080, P.R. China*

E-mail: wang_han@iapcm.ac.cn; mohanchen@pku.edu.cn; linfeng.zhang.zlf@gmail.com

Abstract

Recently, the development of machine learning (ML) potentials has made it possible to perform large-scale and long-time molecular simulations with the accuracy of quantum mechanical (QM) models. However, for high-level QM methods, such as

density functional theory (DFT) at the meta-GGA level and/or with exact exchange, quantum Monte Carlo, etc., generating a sufficient amount of data for training a ML potential has remained computationally challenging due to their high cost. In this work, we demonstrate that this issue can be largely alleviated with Deep Kohn-Sham (DeePKS), a ML-based DFT model. DeePKS employs a computationally efficient neural network-based functional model to construct a correction term added upon a cheap DFT model. Upon training, DeePKS offers closely-matched energies and forces compared with high-level QM method, but the number of training data required is orders of magnitude less than that required for training a reliable ML potential. As such, DeePKS can serve as a bridge between expensive QM models and ML potentials: one can generate a decent amount of high-accuracy QM data to train a DeePKS model, and then use the DeePKS model to label a much larger amount of configurations to train a ML potential. This scheme for periodic systems is implemented in a DFT package ABACUS, which is open-source and ready for use in various applications.

Introduction

Over the past few decades, rapid developments of high speed and massively parallel computing have boosted the exploration of tremendous microscopic phenomena in condensed phases. For such investigations, one of the most widely applied tools is molecular dynamics (MD), which models atomic and molecular systems by numerically solving the Newtonian equations of motion subject to specific boundary conditions.^{1,2} The interatomic energies and forces involved in the Newtonian equations can be either obtained via an empirical force field (EFF),³⁻⁵ or computed from *ab initio* calculations, known as *ab initio* MD (AIMD).⁶⁻¹⁰ Despite the high efficiency of EFF-based MD, its applications are sometimes inhibited due to less satisfying modeled results compared to experiments as well as the presumably questionable transferability when employed to a new system.^{11,12}

AIMD simulation generates trajectories by performing quantum mechanical (QM) calcu-

lations “on-the-fly” as the simulation proceeds. For condensed systems, density functional theory (DFT)¹³ is usually the QM method of choice for AIMD owing to its relatively balanced treatment for the trade-off between efficiency and accuracy. In the framework of DFT, the performance of AIMD simulations rests on the selection of the DFT exchange-correlation (XC) functionals. Liquid water, for example, cannot be quantitatively modeled by AIMD with normal general gradient approximation (GGA) functionals,¹⁴ which lacks proper description of van der Waals (vdW) interactions.^{15,16} Going beyond GGA, meta-GGA or even hybrid meta-GGA functionals that lie on higher rungs of Perdew’s metaphorical Jacob’s ladder offer significantly more accurate predictions, albeit with manifold increased computational cost.^{16,17} Therefore, AIMD simulations with those higher-rung functionals are inevitably limited to fairly small systems with a short simulation time scale (tens of picoseconds), which prevents the quantitative investigation on macroscopic properties of which the converged predictions require much longer simulation time and larger system size. In cases where the nuclear quantum effect is non-negligible, more sophisticated approaches like path-integral MD would be needed for properly describing relevant phenomena, which typically require one or two orders of magnitude more computational resources than classical MD.^{17,18} Moreover, for systems involving strongly-correlated electronic interactions, methods beyond DFT, such as quantum Monte Carlo,¹⁹ will have a more satisfactory accuracy at the price of larger computational cost. See, e.g., Ref. 20, for discussions on the properties of hydrogen and helium under extreme conditions and the influence of different simulation methods.

In the last few years, machine-learning (ML) based potentials have been advanced to circumvent the high computational cost of AIMD without loss of accuracy.^{21–25} One of the representatives is the Deep Potential Molecular Dynamics (DeePMD) scheme,^{25,26} of which the potential energy surface is fitted via a deep neural network to expensive *ab initio* data. Studies have demonstrated that for a wide variety of systems, DeePMD simulations possess accuracy comparable to that of AIMD and efficiency competitive to classical MD simulations. We refer to Ref. 27 for a thorough review of some recent development of DeePMD for

materials science. Notwithstanding the successes achieved via ML-based potentials, obstacles still remain in the scenario that requires comprehensive description offered by high-level QM methods. The training of DeePMD model usually demands thousands of QM-labeled frames, which might become a bottleneck when the QM method of choice is expensive. As shown in Table S1, for example, for a 64-water-molecule system, within the DFT framework, the computational costs using different functionals can differ by nearly three orders of magnitude. Indeed, the bridge that efficiently connects time-consuming QM calculations and ML-based potential energy models is yet to be assembled so as to alleviate or even eliminate such computational bottleneck.

Proposed in 2020, the Deep Kohn-Sham (DeePKS) approach introduces a general framework for generating highly accurate self-consistent energy functionals with remarkably reduced computational cost,^{28,29} which makes it an ideal “bridge” between expensive QM models and DeePMD. While the DeePKS model has been comprehensively tested for isolated molecular systems, we implement it here for periodic systems in an open-source software ABACUS^{30,31} and demonstrate that trained by a considerably small number of data, the DeePKS model reproduces the target energies and forces given by strongly constrained and appropriately normed (SCAN) meta-GGA functional³² for salt water and hybrid SCAN0 functional³³ for pure water at only a few times more expensive computational cost as compared to the Perdew-Burke-Ernzerhof (PBE) GGA functional.¹⁴ The trained DeePKS model is applied in SCF calculations to generate labels for the DeePMD model, with an estimated two orders of magnitude saving in computational time for labeling. The resulting DeePMD model is then employed for MD simulations to compute various structural and dynamical properties of pure and salt water. Excellent agreement is found between the DeePKS-DeePMD predicted results and the previously reported data from SCAN/SCAN0-based DeePMD simulations, which highlights the reliability of the bridging role played by the DeePKS model.

It should be noted that a variety of ML-assisted functionals other than DeePKS have also

been developed, such as NeuralXC³⁴ and OrbNet,³⁵ which share a similar goal as DeePKS, i.e., to lift the accuracy of baseline functionals towards that provided by more accurate methods via a ML-based model while maintaining their efficiency. Other ML-based functionals, including DM21³⁶ and SCAN-L,³⁷ are developed to pave the way toward exact universal functional. Here we apply DeePKS in this work to demonstrate the capability of such ML-based functional in connection with ML-based potentials. From a practical point of view, we see this work as a timely contribution to this rapidly developing field, and we stress that with a series of open-source implementations of the methodology, including DeePKS-kit³⁸ for the training and generation of the DeePKS model, ABACUS for DeePKS-based DFT calculations for periodic systems, as well as DeePMD-kit³⁹ and DP-GEN⁴⁰ for the training and generation of DeePMD models, various applications demanding QM accuracy at a higher level will be made computationally feasible.

Method

DeePKS for isolated systems

Before introducing the DeePKS formalism for periodic systems, we briefly review the case of isolated systems. We consider the many-body Schrödinger equation of N electrons:

$$(\hat{T} + \hat{V}_{ee} + \hat{V}_{\text{ext}})\Psi_0 = E\Psi_0, \quad (1)$$

where \hat{T} , \hat{V}_{ee} , and \hat{V}_{ext} are the operators for kinetic, electron-electron interaction, and external potential, respectively, E is the ground-state energy of the system, and Ψ_0 represents the ground-state N -electron wavefunction.

In the standard Kohn-Sham scheme,^{13,41} one employs an auxiliary non-interacting system under an effective external potential \hat{V}_{KS} , which yields the same ground-state electron density as the original interacting system. The auxiliary system can thus be represented by a single

Slater determinant of a set of one-particle eigenstates $\{\phi_i\}$, obtained by self-consistently solving the single particle Hamiltonian:

$$\hat{h}_i \phi_i = \epsilon_i \phi_i, \quad (2)$$

where $\hat{h}_i = \hat{T} + \hat{V}_{\text{KS}}$.

Conventionally, the effective potential \hat{V}_{KS} is partitioned into three components:

$$\hat{V}_{\text{KS}} = \hat{V}_{\text{ext}} + \hat{V}_{\text{H}} + \hat{V}_{\text{XC}}, \quad (3)$$

namely, the external potential of the original interacting systems \hat{V}_{ext} , the Hartree potential, namely the static Coulomb potential produced by the electron density of the system \hat{V}_{H} , and the exchange-correlation potential \hat{V}_{XC} , which captures the remaining electron-electron interactions.

The exact form of the exchange-correlation potential still remains elusive. The major task in Kohn-Sham DFT is thus to devise better approximations of the exchange-correlation functional. In the traditional Kohn-Sham scheme,^{13,41} \hat{V}_{XC} is spatially local, while in the generalized Kohn-Sham scheme,⁴² \hat{V}_{XC} includes non-local contributions.

The DeePKS scheme seeks a Hamiltonian in the generalized Kohn-Sham framework by connecting a baseline method and a reference method through a neural network model. Typically, baseline methods are chosen to be lower level methods that are computationally efficient but lack the desired level of accuracy for the problem under consideration; while the reference methods are high level methods that are accurate but computationally expensive.

The basic idea is to partition the Hamiltonian into two parts:

$$\hat{h}_{\text{DeePKS}} = \hat{h}_{\text{baseline}} + \hat{V}^\delta. \quad (4)$$

The first part is the Hamiltonian of the baseline method, while the second part is the

correction potential provided by the neural network model. As a result, the total energy is also partitioned into two parts:

$$E_{\text{DeePKS}} = E_{\text{baseline}} + E_{\delta}. \quad (5)$$

In this work, we solve the Hamiltonian in the basis of numerical atomic orbitals $\{\chi_{\mu}\}$,^{31,43–46} and the neural network contribution term V_{δ} is constructed based on projected density matrices:

$$D_{nlmm'}^I = \sum_{\mu\nu} \rho_{\mu\nu} \langle \chi_{\mu} | \alpha_{nlm}^I \rangle \langle \alpha_{nlm'}^I | \chi_{\nu} \rangle, \quad (6)$$

where $\rho_{\mu\nu}$ is the density matrix of the system, and $\{|\alpha\rangle\}$ is a set of localized orbitals centered on atoms, labeled by atomic index I , and quantum numbers nlm .

To preserve the rotational invariance, we further take the eigenvalues of blocks of projected density matrices with the same indices I , n and l to obtain a series of descriptors:

$$\mathbf{d}_{nlm}^I = \text{Eig}(D_{nlmm'}^I). \quad (7)$$

In some cases, this eigenvalue decomposition step introduces discontinuities due to the sorting of eigenvalues, and may cause convergence problems when applying the model in SCF calculations. To circumvent this issue, there is an option to further symmetrize the descriptors:

$$\mathbf{d}_{nlm}^{I,\text{symm}} = g_{\text{symm}}(\mathbf{d}_{nlm}^I). \quad (8)$$

where g_{symm} is a symmetrization function which is invariant under the permutation of its arguments. In our current implementation, g_{symm} is chosen to be thermal averaging, and details of the symmetrization step as well as the neural network structure can be found in the Appendix of Ref. 38.

The descriptors are then grouped into vectors according to the atomic index I , and the

correction energy term becomes a summation of atomic contributions:

$$E_\delta = \sum_I F_{\text{NN}}(\mathbf{d}^I|\omega), \quad (9)$$

where ω is the vector of parameters for the deep neural network F_{NN} . By calculating a set of reference systems, we have the target energies E_{target} , the baseline energies E_{baseline} , as well as the descriptors generated by the baseline method \mathbf{d}^I . The training of F_{NN} is then carried out by using the energy difference $E_{\text{target}} - E_{\text{baseline}}$ as the label.

With the expression for E_δ , the corresponding matrix elements of the correction potential are given by:

$$\begin{aligned} \hat{V}_{\mu\nu}^\delta &= \frac{\partial E_\delta}{\partial \rho_{\mu\nu}} \\ &= \sum_{Inlmm'} \frac{\partial E_\delta}{\partial D_{nlmm'}^I} \frac{\partial D_{nlmm'}^I}{\partial \rho_{\mu\nu}} \\ &= \sum_{Inlmm'} \frac{\partial E_\delta}{\partial D_{nlmm'}^I} \langle \chi_\mu | \alpha_{nlm}^I \rangle \langle \alpha_{nlm'}^I | \chi_\nu \rangle. \end{aligned} \quad (10)$$

Solving the Hamiltonian $\hat{h}_{\text{DeePKS}} = \hat{h}_{\text{baseline}} + \hat{V}^\delta$ gives a set of ground state wavefunctions $\{\phi_i|\omega\}$ and ground state energy $E_{\text{DeePKS}} = E_{\text{baseline}}[\{\phi_i|\omega\}] + E^\delta[\{\phi_i|\omega\}, \omega]$. However, in general there is a discrepancy between the E_{DeePKS} here and the target energy E_{target} . The origin of this discrepancy comes from the fact that the ground state of \hat{h}_{DeePKS} is different from that of the initial baseline method $\hat{h}_{\text{baseline}}$.

As a result, the training of DeePKS adopts an iterative strategy, where the vector of model parameters ω is updated through training, followed by solving the new \hat{h}_{DeePKS} to get a new set of descriptors and labels. The process is repeated until convergence is achieved.

For later iterations, we can also calculate the total force under \hat{h}_{DeePKS} , given by:

$$\begin{aligned}\mathbf{F}_{\text{DeePKS}}[\{\phi_i|\omega\}] &= \mathbf{F}_{\text{baseline}}[\{\phi_i|\omega\}] - \frac{\partial E^\delta[\{\phi_i|\omega\}]}{\partial \mathbf{X}} \\ &= \mathbf{F}_{\text{baseline}}[\{\phi_i|\omega\}] - \sum_{Inlmm'} \frac{\partial E_\delta}{\partial D_{nlmm'}^I} \sum_i \frac{d}{d\mathbf{X}} [f_i \langle \phi_i | \alpha_{nlm}^I \rangle \langle \alpha_{nlm'}^I | \phi_i \rangle].\end{aligned}\quad (11)$$

where f_i is the occupation number of orbital ϕ_i . As our goal is to reproduce the total energies and forces of the target method, we also include force term in the loss function $L(\omega)$, and the optimization problem now becomes:

$$\min_{\omega} L(\omega), \quad L(\omega) = |E_{\text{target}} - E_{\text{DeePKS}}[\{\phi_i|\omega\}]|^2 + \lambda |\mathbf{F}_{\text{target}} - \mathbf{F}_{\text{DeePKS}}[\{\phi_i|\omega\}]|^2, \quad (12)$$

where the weighting factor λ is adjusted in the iterative training process to balance the error in energy and force. Its value typically falls in the range of 1 to 50. Here the notation $\{\phi_i|\omega\}$ is used to emphasize that the eigenstates ϕ_i of the DeePKS Hamiltonian $\hat{h}_{\text{DeePKS}} = \hat{h}_{\text{baseline}} + \hat{V}^\delta$ depend on the expression of \hat{V}^δ , hence on the neural network parameters ω .

We refer to Ref. 29 and Ref. 38 for more details of the DeePKS formalism and the training strategy, including the treatment of force labels, as well as the construction of the neural network.

DeePKS for Periodic Systems

For periodic systems, the external potential possesses the translational symmetry:

$$\hat{V}_{\text{ext}}(\mathbf{r} - \mathbf{R}) = \hat{V}_{\text{ext}}(\mathbf{r}), \quad (13)$$

where \mathbf{R} is the lattice vector used to label the unit cells in the periodic lattice. According to the Bloch theorem, the Kohn Sham eigenstates of the system are expressed as:

$$\phi_{i\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\mathbf{r}} u_{i\mathbf{k}}(\mathbf{r}), \quad (14)$$

where \mathbf{k} is the reciprocal space lattice vector and the quantum number i labels the band index. The Bloch wavefunction $u_{i\mathbf{k}}(\mathbf{r})$ has the same periodicity of the external potential \hat{V}_{ext} and can be solved by diagonalization of the Hamiltonian $H(\mathbf{k})$.

To obtain the matrix elements of $H(\mathbf{k})$ in the atomic basis $\{\chi_\mu\}$, we first calculate:

$$H_{\mu\nu}(\mathbf{R}) = \langle \chi_{\mu\mathbf{R}} | \hat{h} | \chi_{\nu\mathbf{0}} \rangle, \quad (15)$$

where $\chi_{\mu\mathbf{R}}$ is the periodic image of atomic basis χ_μ in the unit cell \mathbf{R} , namely $\chi_{\mu\mathbf{R}}(\mathbf{r}) = \chi_\mu(\mathbf{r} - \mathbf{R})$.

The Hamiltonian for single \mathbf{k} -point $H(\mathbf{k})$ is then given by:

$$H(\mathbf{k}) = \sum_{\mathbf{R}} e^{-i\mathbf{k}\mathbf{R}} H(\mathbf{R}). \quad (16)$$

All physical quantities are obtained as an average over single k points. For example, the electron density is given by:

$$\begin{aligned} \rho(\mathbf{r}) &= \frac{1}{N_{\mathbf{k}}} \sum_{i\mathbf{k}} f_{i\mathbf{k}} \phi_{i\mathbf{k}}^*(\mathbf{r}) \phi_{i\mathbf{k}}(\mathbf{r}) \\ &= \frac{1}{N_{\mathbf{k}}} \sum_{\mu\nu} \sum_{\mathbf{R}} \sum_{\mathbf{k}} \rho_{\mu\nu}(\mathbf{k}) \chi_{\mu\mathbf{R}}^*(\mathbf{r}) \chi_{\nu\mathbf{0}}(\mathbf{r}) e^{-i\mathbf{k}\mathbf{R}} \\ &= \sum_{\mu\nu} \sum_{\mathbf{R}} \rho_{\mu\nu}(\mathbf{R}) \chi_{\mu\mathbf{R}}^*(\mathbf{r}) \chi_{\nu\mathbf{0}}(\mathbf{r}), \end{aligned} \quad (17)$$

where we define the real-space density matrix as:

$$\rho_{\mu\nu}(\mathbf{R}) = \frac{1}{N_{\mathbf{k}}} \sum_{\mathbf{k}} \rho_{\mu\nu}(\mathbf{k}) e^{-i\mathbf{k}\mathbf{R}}. \quad (18)$$

Similarly, the projected density matrix used to construct descriptors is calculated as:

$$\begin{aligned}
D_{nlmm'}^I &= \frac{1}{N_{\mathbf{k}}} \sum_{\mu\nu} \sum_{\mathbf{k}} \sum_{\mathbf{R}'} \rho_{\mu\nu}(\mathbf{k}) \langle \chi_{\mu\mathbf{R}} | \alpha_{nlm\mathbf{R}'}^I \rangle \langle \alpha_{nlm'\mathbf{R}'}^I | \chi_{\nu\mathbf{0}} \rangle e^{-i\mathbf{k}\mathbf{R}} \\
&= \sum_{\mathbf{R}\mathbf{R}'} \sum_{\mu\nu} \rho_{\mu\nu}(\mathbf{R}) \langle \chi_{\mu\mathbf{R}} | \alpha_{nlm\mathbf{R}'}^I \rangle \langle \alpha_{nlm'\mathbf{R}'}^I | \chi_{\nu\mathbf{0}} \rangle.
\end{aligned} \tag{19}$$

Then, the contribution of \hat{V}_δ to the real-space Hamiltonian is derived as:

$$\hat{V}_{\mu\nu}^\delta(\mathbf{R}) = \sum_{Inlmm'} \frac{\partial E_\delta}{\partial D_{nlmm'}^I} \frac{\partial D_{nlmm'}^I}{\partial \rho_{\mu\nu}(\mathbf{R})} \tag{20}$$

$$= \sum_{Inlmm'} \sum_{\mathbf{R}'} \frac{\partial E_\delta}{\partial D_{nlmm'}^I} \langle \chi_{\mu\mathbf{R}} | \alpha_{nlm\mathbf{R}'}^I \rangle \langle \alpha_{nlm'\mathbf{R}'}^I | \chi_{\nu\mathbf{0}} \rangle. \tag{21}$$

As noted in the previous section, we also include force label in the training process, with target function given in Eq. 12. The total force in the case of multiple k points is given by:

$$\begin{aligned}
\mathbf{F}_{\text{DeePKS}}[\{\phi_i|\omega\}] &= \mathbf{F}_{\text{baseline}}[\{\phi_i|\omega\}] - \frac{\partial E^\delta[\{\phi_i|\omega\}]}{\partial \mathbf{X}} \\
&= \mathbf{F}_{\text{baseline}}[\{\phi_i|\omega\}] - \sum_{Inlmm'} \frac{\partial E_\delta}{\partial D_{nlmm'}^I} \sum_{i\mathbf{k}} \frac{d}{d\mathbf{X}} [f_{i\mathbf{k}} \langle \phi_{i\mathbf{k}} | \alpha_{nlm}^I \rangle \langle \alpha_{nlm'}^I | \phi_{i\mathbf{k}} \rangle] \\
&= \mathbf{F}_{\text{baseline}}[\{\phi_i|\omega\}] - \sum_{Inlmm'} \frac{\partial E_\delta}{\partial D_{nlmm'}^I} \sum_{\mathbf{R}\mathbf{R}'} \rho_{\mu\nu}(\mathbf{R}) \frac{d}{d\mathbf{X}} [\langle \chi_{\mu\mathbf{R}} | \alpha_{nlm\mathbf{R}'}^I \rangle \langle \alpha_{nlm'\mathbf{R}'}^I | \chi_{\nu\mathbf{0}} \rangle].
\end{aligned} \tag{22}$$

Computational Details

As mentioned in previous sections, the training of DeePKS adopts an iterative strategy, which alternates between training the neural network $F_{\text{NN}}(\mathbf{d}^I|\omega)$ and solving SCF with $\hat{H}_{\text{DeePKS}} = \hat{H}_{\text{baseline}} + \hat{V}^\delta$. While the training step is performed within DeePKS-kit,³⁸ the package *per se* does not contain the functionality of solving SCF. Instead, an existing SCF software is invoked for such purpose.

We have implemented the DeePKS method in the ABACUS^{30,31} package, which supports both numerical atomic orbitals and plane-wave basis with the periodic boundary conditions.

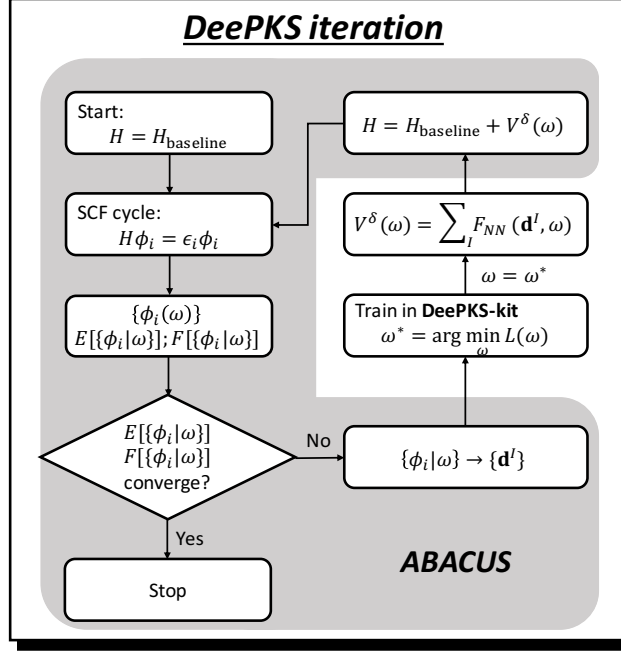


Figure 1: Flowchart of the DeePKS iterative training implemented within the ABACUS density functional theory package.

The ABACUS package can be freely downloaded online.⁴⁷ For the numerical atomic orbitals that form the projectors, the radial parts of $\{|\alpha\rangle\}$ are chosen to be the spherical bessel functions, namely:

$$\alpha_{nlm}(\mathbf{r}) = f_{nl}(r)Y_{lm}(\theta, \phi), \quad (23)$$

where

$$f_{nl}(r) = \begin{cases} j_{nl}(q_n r) & (r \leq r_c) \\ 0 & (otherwise). \end{cases} \quad (24)$$

Here r_c is the radius cutoff, and q_n is chosen to ensure $j_{nl}(q_n r_c) = 0$. A kinetic energy cutoff is imposed to determine the upper bound for the value of q_n , hence the number of spherical bessel functions. Typically, the kinetic energy cutoff is set to be the same as the underlying SCF calculations.

In this work, we used a radius cutoff of 5 Bohr, and the kinetic energy cutoff is set to be 100 Ry, with $l = 0, 1, 2$. This resulted in a total number of 15 spherical bessel functions per l channel, giving an overall of 135 descriptors per atom. More details on the spherical bessel

functions and the evaluation of orbital overlaps $\langle \alpha_{nlm} | \chi_{\mu} \rangle$ can be found in Ref. 31 and Ref. 30.

The implemented DeePKS iterative training process is summarized in Fig. 1. The steps in the grey region are those carried out by ABACUS. In each iteration, ABACUS reads the neural network model file provided by DeePKS-kit and calculates the desired matrix elements $\hat{V}_{\mu\nu}^{\delta}$, then solves the DeePKS Hamiltonian \hat{H}_{DeePKS} and outputs the descriptors and labels in the format that is readable by DeePKS-kit.

We notice that a practical challenge for testing the performance of the DeePKS+ABACUS scheme is that we need extensively generated high-level electronic structure data for benchmark purposes. As such, we chose two representative datasets that have already been well benchmarked in recent works and used to train DeePMD potential models for important applications. The first dataset¹⁷ contains water snapshots from both classical and Feynman path-integral molecular dynamics calculations with energy and force labels at the SCAN0 level. A Deep Potential model was generated from this dataset and used to calculate several properties of water, and later used to investigate the many-body effects in the X-ray absorption spectra of liquid water.⁴⁸ The second dataset⁴⁹ was generated via a concurrent learning approach⁵⁰ and used to train Deep Potentials for modeling the structural properties of sodium chloride solutions at different concentrations at the level of the SCAN functional. Additional computational details for these two datasets can be found in the Supporting Information.

We used the SCAN0 AIMD trajectories of 64 water molecules to compare the sample efficiency of the DeePKS model and the DeePMD model. Next, for both datasets, we chose the PBE functional as the baseline model and used only a small group of samples to obtain reliable DeePKS models at production level. We tested the validity of the resultant DeePKS models by relabeling a much larger group of samples from the same datasets with DeePKS and carried out DeePMD training. The DeePMD models were in turn applied to run MD simulations using LAMMPS.⁵¹ Structural and thermodynamic properties, including the

radial distribution function (RDF), bulk density, and others, were calculated and compared with existing results.

Result and Discussion

DeePKS learning curves with respect to training samples

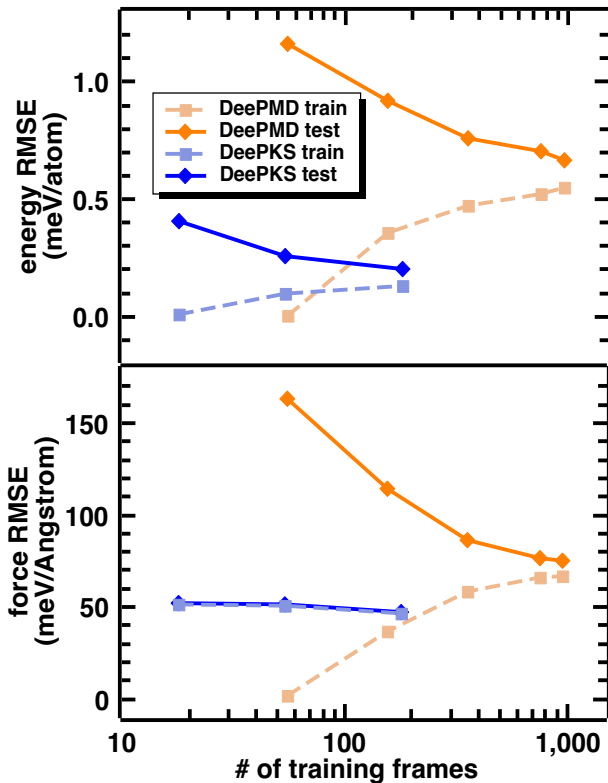


Figure 2: Learning curves for energy (upper panel) and force (lower panel) given by DeePMD (orange) and DeePKS (blue) with respect to the number of training frames. Dashed line with squares indicates train set error; solid line with diamonds indicates test set error.

To explore the capability and generalizability of the DeePKS model, we construct three training sets by randomly (subject to uniform distribution) picking 18, 54, and 180 frames from the previously reported SCAN0 AIMD trajectories of 64 water molecules,¹⁷ and train the PBE-based DeePKS model with SCAN0 energy and force labels. Similar uniform sampling from these SCAN0 trajectories, with larger sampling sizes (55, 155, 355, 755, and 955

frames), is applied to the training of the DeePMD model so as to make a comparison between these two models. The learning curves of DeePMD and DeePKS with respect to the size of the training set are given in Fig. 2. It can be seen that with significantly fewer frames, the DeePKS model provides more accurate predictions as compared to DeePMD model. The generalization gap of DeePKS model is also notably smaller than that of DeePMD model. It is shown in Table S1 that the SCAN0 SCF result for 64 water molecules, which takes more than a day to be obtained, can be accurately reproduced within a quarter of an hour by applying the DeePKS model, which corresponds to more than two orders of magnitude savings in time. The take-home message conveyed by Fig. 2 and Table S1 is that the training process of the DeePMD potential, which originally demands more than a thousand expensive SCAN0 jobs, can be effectuated with around one hundred SCAN0 jobs plus a thousand significantly faster DeePKS jobs. In other words, the DeePKS model can serve as a bridge that connects the expensive *ab initio* calculations such as SCAN0 DFT and the machine learning potentials, and remarkably reduces the effort required in the MD simulations at higher rung of the Jacob’s ladder.

Modelling liquid water

For systems consisting of 64 water molecules, we perform SCF calculations on 1022 unique structures randomly (subject to uniform distribution) picked from previous DeePMD (with SCAN0 label) and SCAN0 AIMD modelling results (from Ref. 17) with the DeePKS model trained via 180 training samples. 1000 out of 1022 SCF calculations reach the convergence threshold, and these 1000 converged energies and forces are applied as labels for DeePMD training. The resulting Deep Potential model is then employed in LAMMPS for molecular dynamics simulation of 512 water molecules. Various structural properties and the diffusion coefficient are explored and compared with previously reported results. All structural properties are obtained via 30 ps NpT ensemble simulations at 1 bar and 330 K with a time step of 0.5 fs with the first 10 ps discarded for equilibrium, while the diffusion coefficient is

obtained via 300 ps *NVE* ensemble simulations with the cell size fixed at the value obtained from the *NpT* simulation.

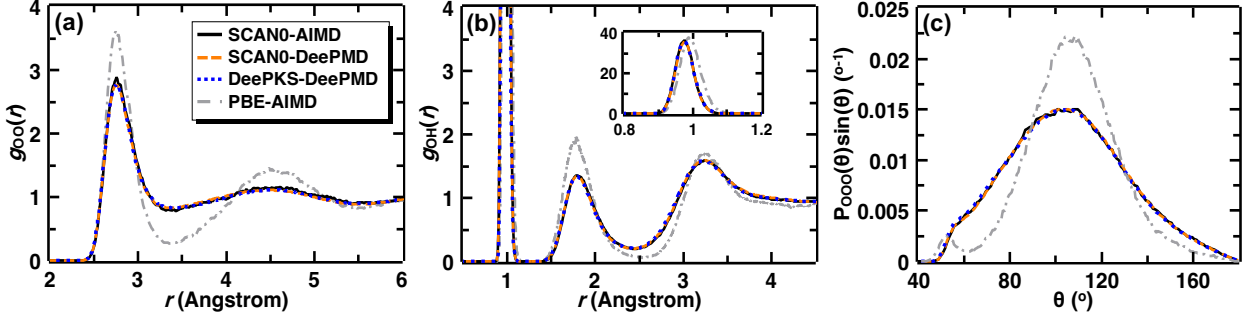


Figure 3: RDFs (a) $g_{OO}(r)$, (b) $g_{OH}(r)$, and (c) bond angle distribution $P_{OOO}(\theta)$ given by DeePKS-DeePMD (blue dotted line), SCAN0-AIMD (black solid line) from Ref. 17, SCAN0-DeePMD (orange dashed line) from Ref. 17, and PBE-AIMD (gray dotted-dashed line) from Ref. 16.

We first analyze the RDFs, which correspond to the probability of finding a given pair of atoms as a function of distance in real space, with DeePKS-DeePMD simulations. The resulting oxygen-oxygen and oxygen-hydrogen RDFs, $g_{OO}(r)$ and $g_{OH}(r)$ are shown in Fig. 3(a) and (b), respectively. The previously computed results via SCAN0-AIMD, SCAN0-DeePMD, and PBE-AIMD are also shown for comparison. It can be seen from Figs. 3(a) and (b) that the RDFs given by DeePKS-DeePMD simulations are in good agreement with both the SCAN0-AIMD and the SCAN0-DeePMD results, including the significantly less overstructured peaks of $g_{OO}(r)$ and the slightly shortened O-H covalent bond length (indicated by the first peak of $g_{OH}(r)$) as compared to the PBE result. Similar observations are found in the bond angle distribution ($P_{OOO}(\theta)$) analysis as shown in Fig. 3(c), which quantifies the three-body correlations in water. While PBE predicts a narrower bond angle distribution, DeePKS is able to quantitatively reproduce the distribution predicted by SCAN0. Overall, for liquid water, the overstructuring issue in PBE functional is remarkably alleviated with the trained DeePKS model, which provides almost identical structural properties as compared to SCAN0 results.

Next, we explore the bulk density and H-bonds of liquid water with DeePKS-DeePMD

simulations. Note that the diffusion coefficient is computed for both water and deuterated water so as to provide a more comprehensive comparison with the SCAN0 results from Ref. 17. By including the description of nondirectional van der Waals (vdW) interaction on intermediate length-scales, SCAN0 predicts a more disordered and compact water structure, leading to a higher bulk density and weakened H-bond strength. As shown in Table 1, the bulk density predicted via DeePKS-DeePMD (1.024 g/cm^3) is consistent with that predicted by SCAN0-DeePMD (1.030 g/cm^3), and is notably larger than that predicted via PBE-AIMD (0.850 g/cm^3). The average number of H-bonds per water molecule computed with DeePKS-DeePMD is 3.58, which is identical to that given by SCAN-DeePMD and notably smaller than the PBE-AIMD result (3.77 according to Ref. 16). The weakened H-bond strength is also evidenced by the more dominant region between the first peak and the second peak of $g_{\text{OO}}(r)$ predicted by SCAN0 (as shown in Fig. 3(a)), which mainly comprises non-H-bonded molecules that occupy interstitial space between H-bonded ones.

The dynamic property of liquid water we examine in this work with DeePKS-DeePMD is the diffusion coefficient, for both normal and deuterated water. The diffusion of liquid water depends on the formation and breakage of H-bonds through thermal fluctuations. Weakened H-bond strength predicted by SCAN0 and DeePKS (as illustrated above) escalates the tendency of H-bond-breaking and consequently increases the diffusion coefficient. It can be seen in Table 1 that D predicted by DeePKS-DeePMD is in excellent agreement with the one predicted by SCAN0-DeePMD, which is one order of magnitude larger than the PBE-AIMD result. The same consistency is also observed for the case of deuterated water. The good agreement on these structural and dynamical properties highlight the fact that the intermediate-ranged vdW interactions in liquid water, which are intrinsically missed in PBE functional, are successfully captured via the trained DeePKS model using the PBE functional as its baseline, and properties of liquid water with expensive hybrid meta-GGA (SCAN0) quality can now be much more efficiently predicted within the time comparable to a few PBE jobs.

Table 1: Bulk density (ρ), average number of H-bonds per water molecule (N_{HB}), and diffusion coefficients (D) predicted by DeePKS-DeePMD, SCAN0-DeePMD, and PBE-AIMD simulations at 330K with 512 water molecules.^a

Method	$\rho/\text{g}\cdot\text{cm}^{-3}$	N_{HB}	$D/\text{\AA}\cdot\text{ps}^{-1}$	deuterated $D/\text{\AA}\cdot\text{ps}^{-1}$
DeePKS-DeePMD	1.024 ± 0.010	3.58	0.254 ± 0.024	0.234 ± 0.019
SCAN0-DeePMD ¹⁷	1.030	3.58	0.251	0.223
PBE-AIMD ¹⁶	0.850 ± 0.016	3.77	0.018 ± 0.002	NA
exp ^b	0.997^{52}	3.58^{53}	0.24^{54}	0.20^{54}

^aAll error bars correspond to one standard deviation.

^bExperimental values are measured at T=300K.

Modelling salt water and high-pressure water

While the investigation of liquid water with DeePKS DeePMD simulations demonstrates accurate computational results compared to SCAN0 counterparts, modelling the electrolyte structure is rather cumbersome due to the sparsity of ions that requires significantly longer simulation time for statistical convergence. Here, we train a PBE-based DeePKS model for salt water with previously conducted SCAN SCF calculations on various concentrations of NaCl solution as labels. The composition of the DeePKS training set can be found in Table S2. SCF calculations with such DeePKS model are then carried out on 5676 frames (varying in concentration as shown in Tabel S2), of which 5406 converged results (including energy and force) are utilized as labels for DeePMD training. With the trained potential, we investigate structural properties and densities for salt water with different concentrations as well as pure water under different pressure via DeePKS-DeePMD in LAMMPS.

Comparison with SCAN-AIMD simulations

We first compare the RDFs of 1:62 NaCl solution with previously reported SCAN-AIMD and SCAN-DeePMD results. A 2-ns DeePKS-DeePMD simulation is carried out with the *NVT* ensemble using one cubic cell, which consists of one NaCl ion pair and 62 water molecules at 300 K; the setup is consistent with the SCAN-DeePMD simulation conditions in Ref. 49. Fig. 4 exhibits four RDFs $g_{\text{OO}}(r)$, $g_{\text{OH}}(r)$, $g_{\text{ONa}}(r)$, and $g_{\text{OCl}}(r)$, predicted by DeePKS-DeePMD,

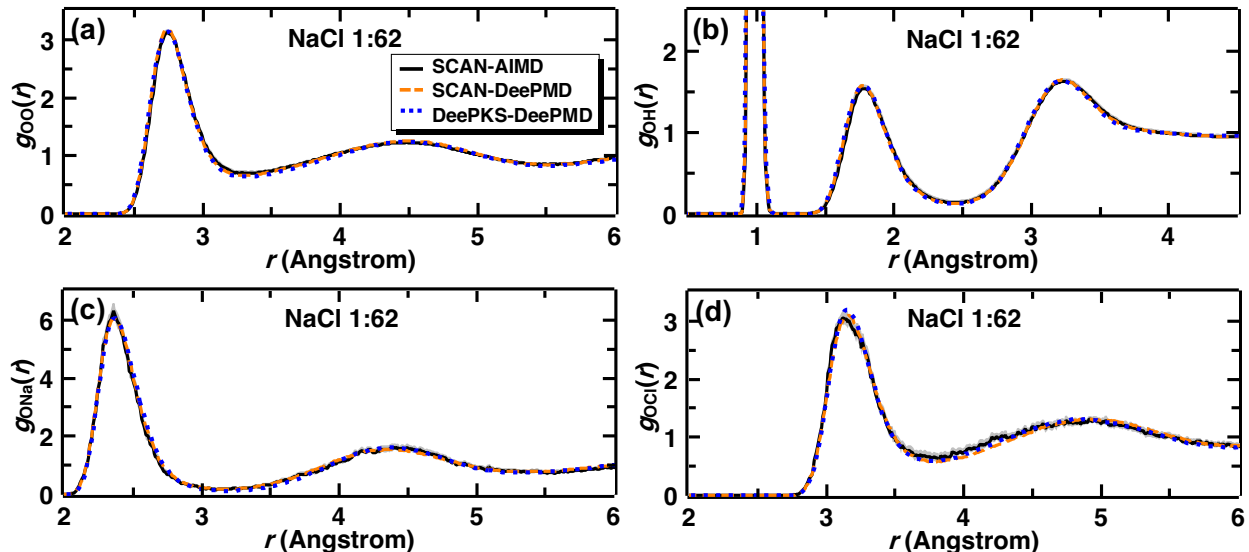


Figure 4: RDFs (a) $g_{OO}(r)$, (b) $g_{OH}(r)$, (c) $g_{ONa}(r)$, and (d) $g_{OCl}(r)$ given by DeePKS-DeePMD (blue dotted line), and SCAN-AIMD (black solid line) as well as SCAN-DeePMD (orange dashed line) from Ref. 49. Gray shaded area corresponds to one standard deviation from DeePKS-DeePMD statistics with 100 ps time interval.

SCAN-AIMD, and SCAN-DeePMD. Note that the SCAN-AIMD simulation only runs for 100 ps due to the highly time-consuming SCF calculations with the SCAN functional. We therefore compute the statistical deviation with 100 ps time interval based on DeePKS-DeePMD trajectories (indicated by the gray shaded area in Fig. 4). All RDFs predicted by DeePKS-DeePMD simulations are in accordance with the SCAN-DeePMD results and statistically matches with the SCAN-AIMD results, which conceptually proves the reliability of our trained DeePKS model.

Comparison with SCAN-DeePMD simulations for high-pressure water

Structural differences between the high-pressure water and salt water have been comprehensively illustrated in Ref. 49. Here, with the aforementioned trained Deep Potential based on DeePKS energies and forces, we first investigate the structural properties of water under high pressure via DeePMD simulations using the NpT ensemble with 512 water molecules for 2 ns at 333 K and four different pressures (1 bar, 1 kbar, 2 kbar, and 3 kbar). The integration time step is 0.5 fs, with the first 100 ps of trajectory discarded for equilibrium. As shown in

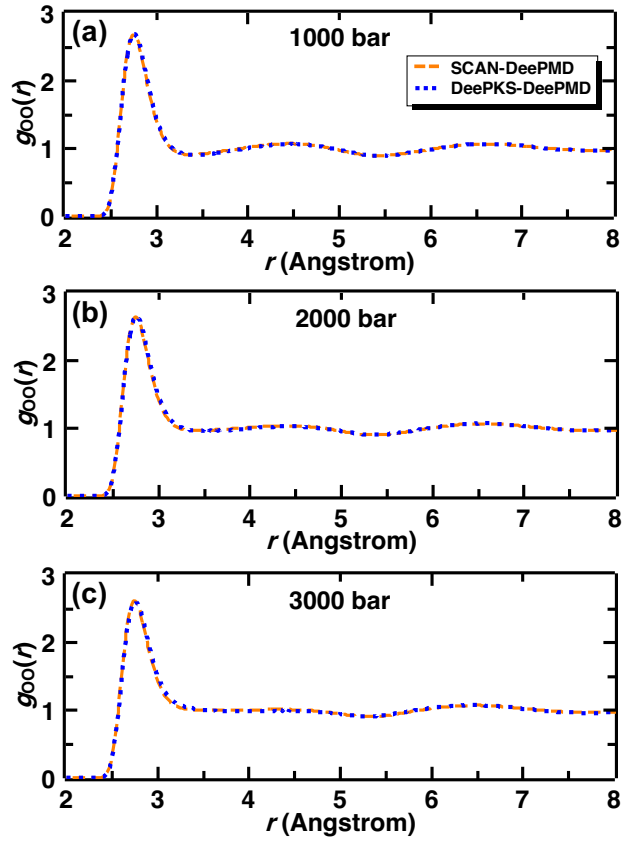


Figure 5: Oxygen-oxygen RDF [$g_{OO}(r)$] for pure water at pressure equals (a) 1000 bar, (b) 2000 bar, and (c) 3000 bar given by DeePKS-DeePMD (blue dotted line) from this work and SCAN-DeePMD (orange dashed line) from Ref. 49.

Fig. 5, the oxygen-oxygen RDFs predicted SCAN-DeePMD under all three high pressures are accurately reproduced by DeePKS-DeePMD simulations. (The comparison at 1 bar will be shown in the following part.) It can be seen in Figure S1(a) that as pressure increases, the second and the third coordination shells move inwards, leading to a more compact structure. The diminishing feature of the second shell at 3000 bar is consistent with the fact that the tetrahedral network inside liquid water is significantly distorted under high pressure. Bulk densities of liquid water have also been calculated as shown in Fig. 6. Excellent agreement is again observed between SCAN-DeePMD and DeePKS-DeePMD simulations.

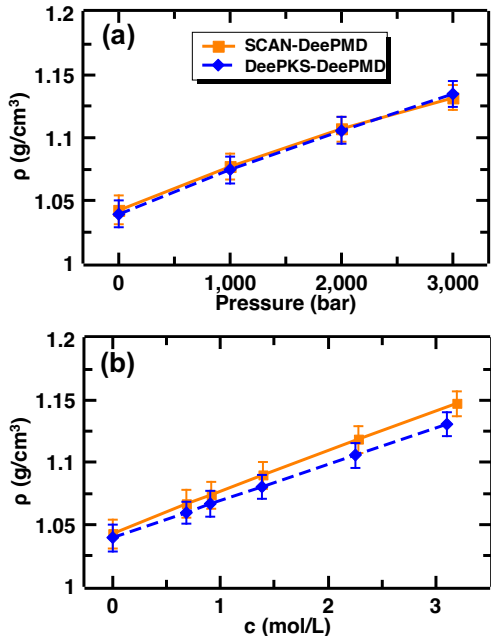


Figure 6: Bulk densities for (a) pure water at different pressures and (b) pure water and NaCl solutions with various concentrations at 1 bar predicted by DeePKS-DeePMD (blue diamond) from this work and SCAN-DeePMD (orange square) from Ref. 49. The simulation temperature is 330K. Error bars correspond to one standard deviation.

Comparison with SCAN-DeePMD simulations for salt water with various concentrations

In this part, we examine our DeePKS model by comparing the structural properties given by DeePKS-DeePMD with those predicted by SCAN-DeePMD for pure and salt water with

various concentrations. Simulation conditions for DeePKS-DeePMD are kept the same as last section and the pressure is fixed at 1 bar. The numbers of NaCl ion pairs and water molecules contained in the periodic cubic cell for DeePMD simulations for each investigated concentration are listed in Table S3.

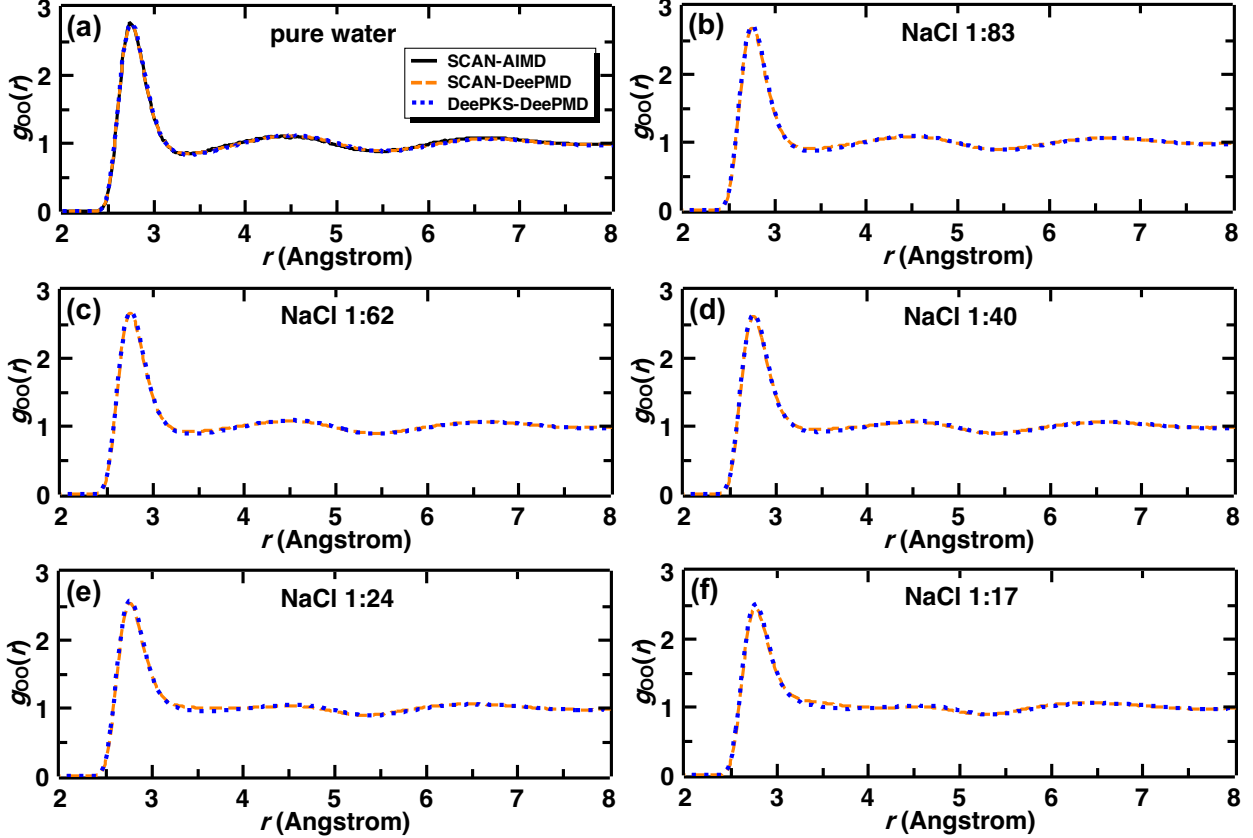


Figure 7: Oxygen-oxygen RDF [$g_{OO}(r)$] for (a) pure water and NaCl solutions with concentration (b) 1:83, (c) 1:62, (d) 1:40, (e) 1:24, and (f) 1:17 given by DeePKS-DeePMD (blue dotted line) from this work and SCAN-DeePMD (orange dashed line) from Ref. 49. SCAN-AIMD result for pure water reported in Ref. 49 is also displayed for comparison (black solid line in (a)).

It can be seen in Fig. 7, Fig S2, and Fig S3 that for pure water and each investigated concentration, the oxygen-oxygen, oxygen-sodium, and oxygen-chloride RDFs predicted by DeePKS-DeePMD are nearly coincident with the SCAN-DeePMD results. As clearly shown in Fig S1, as the concentration increases, the population of interstitial water between the first and second peaks increases and the third coordination shell moves inwards with a diminishing feature of the second coordination shell, which is evinced by the pressure-like effect of salt

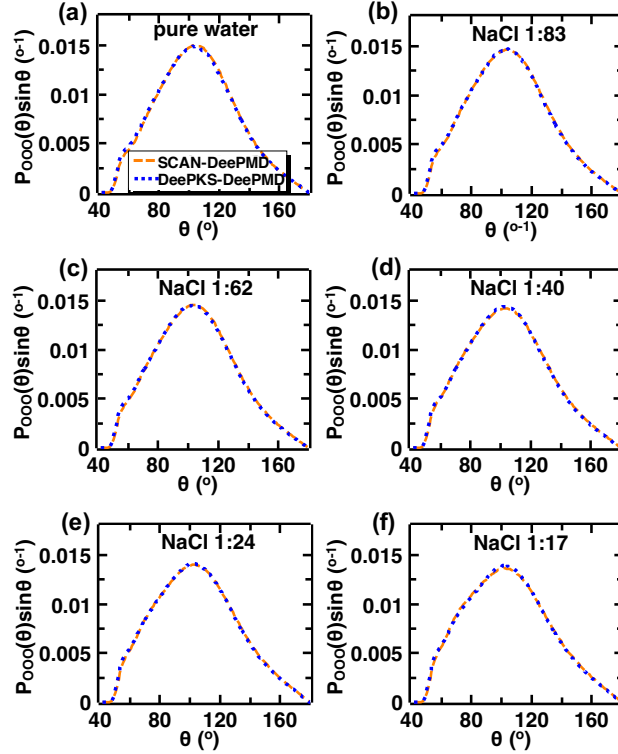


Figure 8: Bond angle distribution $[P_{\text{OOO}}(\theta)]$ for (a) pure water and NaCl solutions with concentration (b) 1:83, (c) 1:62, (d) 1:40, (e) 1:24, and (f) 1:17 given by DeePKS-DeePMD (blue dotted line) from this work and SCAN-DeePMD (orange dashed line) from Ref. 49

water in the reciprocal space. The outward moving trend of the second shell with the increase of the concentration, which is highlighted in Ref. 49 as the major difference between the structures of salt and high-pressure water, is also successfully captured via DeePKS-DeePMD simulation as shown in Fig. S1. The ion-ion RDFs, which require considerably long simulation time to converge due to rather few ion pairs, are also calculated with the DeePKS-DeePMD trajectories. According to Fig. S4, $g_{\text{NaNa}}(r)$ and $g_{\text{ClCl}}(r)$ predicted via DeeKS-DeePMD are qualitatively in line with the SCAN-DeePMD results, with slight deviations that are presumably caused by the insufficiency of the simulation time. The bond angle distributions for pure and salt water with various concentrations are also explored (Fig. 8) and delicate consistency is observed between DeePKS-DeePMD and SCAN-DeePMD results. With the increase of NaCl concentration, $P_{\text{OOO}}(\theta)$ shifts towards smaller angles from a tetrahedral distribution, which is mainly induced by the distribution of the first solvation shells of Na^+ as elucidated in Ref. 49. The calculated bulk densities via DeePKS-DeePMD for NaCl solutions also closely match with those predicted via SCAN-DeePMD as shown in Fig. 6(b). The slight discrepancies at high concentration are presumably due to the fact that the SCAN stress is not included as a label during the DeePKS training process. We shall leave this to our future investigations.

Conclusion and Outlook

In this work, we have bridged expensive high-level *ab initio* calculations and deep neural network potentials with the DeePKS model implemented in the open-source DFT software ABACUS. With less than two hundred frames of the training set labeled by hybrid meta-GGA or meta-GGA functionals, we have shown that the GGA-based DeePKS model is able to quantitatively reproduce the target energies and forces for pure and salt water systems with orders of magnitude savings in time (depending on the choice of the target method). The trained DeePKS model has then been applied in the DeePMD simulations for pure and

salt water, i.e., two prototypical systems that are known to be poorly described via GGA functionals. The resulting structural and dynamical properties are in excellent agreement with the previously reported data obtained via hybrid meta-GGA or meta-GGA AIMD and DeePMD methods, which underlines the reliability of the DeePKS model in connection with the DeePMD simulation.

With the fully open-source implementation of the DeePKS+ABACUS methodology, we are expecting the spring up of extensive applications that require both high accuracy and computational efficiency. It is worth mentioning that even though the DeePKS model is trained on one or two specific periodic systems in this work, its transferability and generalizability should not be disregarded. Looking forward, it would be intriguing to develop the DeePKS model that is applicable to a class of systems such as electrolytes and inorganic semiconductors, enabling a generally accurate description which is hitherto challenging on such systems due to limited computational resources.

Acknowledgments

The work of M.C. was supported by the National Science Foundation of China under Grant No. 12122401, 12074007, and 12135002. The work of H.W. was supported by the National Science Foundation of China under Grant No.11871110 and 12122103. The work of C.Z. and X.W. were supported by National Science Foundation through Award No. DMR-2053195. The work of Y.C. was supported by the Computational Chemical Sciences Center: Chemistry in Solution and at Interfaces (CSI) funded by DOE Award DE-SC0019394 and a gift from iFlytek to Princeton University.

References

- (1) Alder, B. J.; Wainwright, T. E. Studies in Molecular Dynamics. I. General Method. *J. Chem. Phys.* **1959**, *31*, 459–466.

- (2) Tuckerman, M. E.; Berne, B. J.; Martyna, G. J. Molecular Dynamics Algorithm for Multiple Time Scales: Systems with Long Range Forces. *J. Chem. Phys.* **1991**, *94*, 6811–6815.
- (3) van Duin, A. C. T.; Dasgupta, S.; Lorant, F.; Goddard, W. A. ReaxFF: A Reactive Force Field for Hydrocarbons. *J. Phys. Chem. A* **2001**, *105*, 9396–9409.
- (4) MacKerell Jr., A. D.; Banavali, N.; Foloppe, N. Development and Current Status of the CHARMM Force Field for Nucleic Acids. *Biopolymers* **2000**, *56*, 257–265.
- (5) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General AMBER Force Field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.
- (6) Car, R.; Parrinello, M. Unified Approach for Molecular Dynamics and Density-Functional Theory. *Phys. Rev. Lett.* **1985**, *55*, 2471.
- (7) Marx, D.; Hutter, J. *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods*; Cambridge University Press, 2009.
- (8) 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, PMID: 12069631.
- (9) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
- (10) Del Ben, M.; Schönherr, M.; Hutter, J.; VandeVondele, J. Bulk Liquid Water at Ambient Temperature and Pressure from MP2 Theory. *J. Phys. Chem. Lett.* **2013**, *4*, 3753–3759.

- (11) MacKerell Jr, A. D. Empirical Force Fields for Biological Macromolecules: Overview and Issues. *J. Comput. Chem.* **2004**, *25*, 1584–1604.
- (12) Vellore, N. A.; Yancey, J. A.; Collier, G.; Latour, R. A.; Stuart, S. J. Assessment of the Transferability of a Protein Force Field for the Simulation of Peptide-Surface Interactions. *Langmuir* **2010**, *26*, 7396–7404.
- (13) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, A1133.
- (14) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865.
- (15) DiStasio Jr, R. A.; Santra, B.; Li, Z.; Wu, X.; Car, R. The Individual and Collective Effects of Exact Exchange and Dispersion Interactions on the *Ab Initio* Structure of Liquid Water. *J. Chem. Phys.* **2014**, *141*, 084502.
- (16) Chen, M.; Ko, H.-Y.; Remsing, R. C.; Andrade, M. F. C.; Santra, B.; Sun, Z.; Selloni, A.; Car, R.; Klein, M. L.; Perdew, J. P.; Wu, X. *Ab Initio* Theory and Modeling of Water. *Proc. Natl. Acad. Sci.* **2017**, *114*, 10846–10851.
- (17) Zhang, C.; Tang, F.; Chen, M.; Xu, J.; Zhang, L.; Qiu, D. Y.; Perdew, J. P.; Klein, M. L.; Wu, X. Modeling Liquid Water by Climbing up Jacob’s Ladder in Density Functional Theory Facilitated by Using Deep Neural Network Potentials. *J. Phys. Chem. B* **2021**, *125*, 11444–11456, PMID: 34533960.
- (18) Ko, H.-Y.; Zhang, L.; Santra, B.; Wang, H.; E, W.; DiStasio Jr, R. A.; Car, R. Isotope Effects in Liquid Water via Deep Potential Molecular Dynamics. *Mol. Phys.* **2019**, *117*, 3269–3281.
- (19) Ceperley, D.; Alder, B. Quantum monte carlo. *Science* **1986**, *231*, 555–560.

- (20) McMahon, J. M.; Morales, M. A.; Pierleoni, C.; Ceperley, D. M. The Properties of Hydrogen and Helium under Extreme Conditions. *Reviews of modern physics* **2012**, *84*, 1607.
- (21) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.
- (22) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.
- (23) Schütt, K.; Kindermans, P.-J.; Felix, H. E. S.; Chmiela, S.; Tkatchenko, A.; Müller, K.-R. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. *Adv. Neural Inf. Process. Syst.* **2017**, 992–1002.
- (24) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chem. Sci.* **2017**, *8*, 3192–3203.
- (25) 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*, 143001.
- (26) Zhang, L.; Han, J.; Wang, H.; Saidi, W.; Car, R.; E, W. End-to-End Symmetry Preserving Inter-Atomic Potential Energy Model for Finite and Extended Systems. *Adv. Neural Inf. Process. Syst.* **2018**, 4436–4446.
- (27) Wen, T.; Zhang, L.; Wang, H.; Weinan, E.; Srolovitz, D. J. Deep Potentials for Materials Science. *Mater. Futures* **2022**,
- (28) Chen, Y.; Zhang, L.; Wang, H.; E, W. Ground State Energy Functional with Hartree–Fock Efficiency and Chemical Accuracy. *J. Phys. Chem. A* **2020**, *124*, 7155–7165.

- (29) Chen, Y.; Zhang, L.; Wang, H.; E, W. DeePKS: a Comprehensive Data-Driven Approach towards Chemically Accurate Density Functional Theory. *J. Chem. Theory Comput.* **2020**, *17*, 170–181.
- (30) Li, P.; Liu, X.; Chen, M.; Lin, P.; Ren, X.; Lin, L.; Yang, C.; He, L. Large-Scale *Ab Initio* Simulations Based on Systematically Improvable Atomic Basis. *Comput. Mat. Sci.* **2016**, *112*, 503–517.
- (31) Chen, M.; Guo, G.-C.; He, L. Systematically Improvable Optimized Atomic Basis Sets for *Ab Initio* Calculations. *J. Phys.: Condens. Matt.* **2010**, *22*, 445501.
- (32) Sun, J.; Ruzsinszky, A.; Perdew, J. P. Strongly Constrained and Appropriately Normed Semilocal Density Functional. *Phys. Rev. Lett.* **2015**, *115*, 036402.
- (33) Hui, K.; Chai, J.-D. SCAN-Based Hybrid and Double-Hybrid Density Functionals from Models without Fitted Parameters. *J. Chem. Phys.* **2016**, *144*, 044114.
- (34) Dick, S.; Fernandez-Serra, M. The Properties of Hydrogen and Helium under Extreme Conditions. *Nat. Commun.* **2020**, *11*, 3509.
- (35) Qiao, Z.; Welborn, M.; Anandkumar, A.; Manby, F. R.; Miller III, T. F. OrbNet: Deep Learning for Quantum Chemistry Using Symmetry-Adapted Atomic-Orbital Features. *J. Chem. Phys.* **2020**, *153*, 124111.
- (36) Kirkpatrick, J. et al. Pushing the Frontiers of Density Functionals by Solving the Fractional Electron Problem. *Science* **2021**, *374*, 1385–1389.
- (37) Pokharel, C.; Furness, J. W.; Yao, Y.; Blum, V.; Irons, T. J. P.; Teale, A. M.; Sun, J. Exact Constraints and Appropriate Norms in Machine Learned Exchange-Correlation Functionals. *arXiv preprint arXiv:2205.14241* **2022**,
- (38) Chen, Y.; Zhang, L.; Wang, H.; E, W. DeePKS-Kit: a Package for Developing Machine

- Learning-Based Chemically Accurate Energy and Density Functional Models. *arXiv preprint arXiv:2012.14615* **2020**,
- (39) Wang, H.; Zhang, L.; Han, J.; E, W. DeePMD-kit: A Deep Learning Package for Many-Body Potential Energy Representation and Molecular Dynamics. *Comput. Phys. Commun.* **2018**, *228*, 178 – 184.
 - (40) Zhang, Y.; Wang, H.; Chen, W.; Zeng, J.; Zhang, L.; Wang, H.; Weinan, E. DP-GEN: A Concurrent Learning Platform for the Generation of Reliable Deep Learning Based Potential Energy Models. *Comput. Phys. Commun.* **2020**, 107206.
 - (41) Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136*, B864.
 - (42) Becke, A. D. A New Mixing of Hartree–Fock and Local Density-Functional Theories. *J. Chem. Phys.* **1993**, *98*, 1372–1377.
 - (43) Chen, M.; Guo, G.-C.; He, L. Electronic Structure Interpolation via Atomic Orbitals. *J. Phys.: Condens. Matt.* **2011**, *23*, 325501.
 - (44) Blum, V.; Gehrke, R.; Hanke, F.; Havu, P.; Havu, V.; Ren, X.; Reuter, K.; Scheffler, M. Ab Initio Molecular Simulations with Numeric Atom-Centered Orbitals. *Comput. Phys. Commun.* **2009**, *180*, 2175–2196.
 - (45) Soler, J. M.; Artacho, E.; Gale, J. D.; García, A.; Junquera, J.; Ordejón, P.; Sánchez-Portal, D. The SIESTA Method for *Ab Initio* Order-*N* Materials Simulation. *J. Phys.: Condens. Matt.* **2002**, *14*, 2745–2779.
 - (46) Ozaki, T. Variationally Optimized Atomic Orbitals for Large-Scale Electronic Structures. *Phys. Rev. B* **2003**, *67*, 155108.
 - (47) <https://github.com/deepmodeling/abacus-develop>.

- (48) Tang, F.; Li, Z.; Zhang, C.; Louie, S. G.; Car, R.; Qiu, D. Y.; Wu, X. Many-Body Effects in the X-Ray Absorption Spectra of Liquid Water. *Proc. Natl. Acad. Sci.* **2022**, *119*, e2201258119.
- (49) Zhang, C.; Yue, S.; Panagiotopoulos, A. Z.; Klein, M. L.; Wu, X. Dissolving Salt is not Equivalent to Applying a Pressure on Water. *Nat. Commun.* **2022**, *13*, 822.
- (50) Zhang, L.; Lin, D.-Y.; Wang, H.; Car, R.; E, W. Active Learning of Uniformly Accurate Interatomic Potentials for Materials Simulation. *Phys. Rev. Mater.* **2019**, *3*, 023804.
- (51) Plimpton, S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *J. Comput. Phys.* **1995**, *117*, 1–19.
- (52) Linstrom, P. J.; Mallard, W. G. The NIST Chemistry WebBook: A Chemical Data Resource on the Internet. *J. Chem. Eng. Data* **2001**, *46*, 1059–1063.
- (53) Soper, A. K.; Bruni, F.; Ricci, M. A. Site–Site Pair Correlation Functions of Water from 25 to 400°C: Revised Analysis of New and Old Diffraction Data. *J. Chem. Phys.* **1997**, *106*, 247–254.
- (54) Mills, R. Self-Diffusion in Normal and Heavy Water in the Range 1–45°. *J. Phys. Chem.* **1973**, *77*, 685–688.