Using distance-based linear machine learning model to learn Monte Carlo simulation potentials on small reactive systems

Ziyue Ji[1,*]

[1] School of Chemistry and Chemical Engineering, University of Chinese Academy of Sciences, 19A Yuquan Road, Beijing, 100049 P.R. China

[*]Corresponding author's e-mail: jiziyue15@mails.ucas.ac.cn

Abstract. Distance-based linear machine learning (ML) models have drawn attention due to its simplicity and applications in predicting potential energies of physical and chemical materials and systems. Ammonia borane (AB) is known for its potential applications as a hydrogen-storing medium, and molecular dynamics (MD) and Monte Carlo (MC) simulations are important approaches to study its chemical properties. In this work we applied a distance-based linear ML model to predict the potential energy of an AB system. The dataset for the training of this model is produced by a short-lengthed density functional-molecular dynamics (DFT/MD) simulation. Furthermore, we investigated the soundness of this model by comparing it with a previously published potential function by running MC simulations on the same system with same base conformation. Principal component analysis (PCA) and further density functional calculations are performed on the MC-generated conformations to verify the result. We expect this work cast light on future studies on the design of MC potential functions with further improved machine learning algorithms that applies to a wider variety of different materials and situations.

1. Introduction

Computational simulation of nanoscopic and microscopic physical or chemical systems in various scientific disciplines serves as an important complementary method to experimental methods [1]. One of the most accurate methods among these is the ab initio quantum mechanics calculation based on the density functional theory (DFT). However, DFT calculations are the most time-consuming among all these methods, and usually in small- or medium-scale simulations, this method is combined with molecular dynamics (MD) or Monte Carlo simulation (MC) [2].

There have been many successful designs of MD or MC force fields based on DFT calculation data, and they can be generally classified into partially knowledge-based ones and completely data-driven ones, where the former require at least handcrafted features and the latter do not require any previous physics or chemistry knowledge to design [3]. As one of the recent researches of the data-driven ones, Pihlajamäki et al. [4] applied a simple distance-based linear machine learning (ML) method called extreme minimal learning machine (EMLM) [5] on a multi-layer-protected gold nanoparticle system to predict its potential energy and obtained accurate results. Pihlajamäki et al.'s work depicted the possibility of spending considerably lower computational resources and time to yield rather accurate models. However, multi-layer-protected nanoparticles are systems where weak interaction forces like van der Waals force, hydrogen bond and π-π interactions, play the decisive role in the dynamics of this system [6]. For reactive systems, i. e. systems that are related to chemical reactions, most traditional MD or MC potentials that only consider weak forces no longer apply. There have been a variety of tools and methods to simulate the potential energy of a reactive system, and ReaxFF is one of the most

developed tool that applies mainly on MD/MC simulations of reactive systems [7]. This force field, however, is a knowledge-based model that has many complicated hand-crafted parameters designed by physicists and chemists.

Ammonia borane (AB), and other boron-nitrogen hydrides, have been considered as potentially good hydrogen-storing media for their high hydrogen content and their behavior of decomposing and releasing hydrogen in high temperature [8]. MD/MC simulations of the combustion and other reactions of AB systems, as well as designs of potential functions for these simulations, are thus meaningful. Weismiller et al.'s work [9] is an example of using ReaxFF to design potential functions for AB systems. In this work, we chose AB as an example of small reactive systems, and trained a model with the same ML method in Pihlajamäki et al.'s work on an AB system to study the versatility of the EMLM method in different materials and situations, by comparing this predictor with both the DFT/MD results and Weismiller et al.'s potential predictor on the training set and in further MC simulations.

2. Method

In this work we trained an EMLM model on a Born-Oppenheimer DFT/MD simulation dataset on an AB system. This crystal has the chemical composition of (NH4)2(B12H12) [10], and the original crystallography data of this crystal is downloaded from Crystallography Open Database [11]. The simulation is performed with ASE and GPAW software package. We used Perdew–Burke-Ernzerhof (PBE) functional to calculate its energy and NVT Berendsen velocity scaling to run a 0.1 ps simulation contains of 500 frames of a timestep of 0.2 fs. Many-body-tensor representation (MBTR) based on pairwise distance (k=2) implemented in the software package DScribe is used to map the cartesian coordinates of conformations to a descriptor that is immune to SO(3) transformations, and the parameters {min,max,nx,σ,d,cut-off} of the MBTR transformation are {0, 0.5, 50, 0.01, 0.5, 10−3}. The EMLM model is built with Keras on Tensorflow [12]. We use all data points in this simulation data as the reference points for the EMLM model and trained it on the dataset by 10 000 epochs with different learning rate and batch size.

For the principal component analysis (PCA) we used singular value decomposition (SVD) method provided by Tensorflow software package.

3. Results and Discussion

The MD simulation potential trajectory is shown in Fig. 1. We can see from this trajectory that there is severe energy fluctuation during the simulation, and this might because of inappropriate choose of the NVT Berendsen scaling method. The training loss of the dataset regarding epochs is shown in Fig. 2, and the average loss after the training process is less than 4 eV.

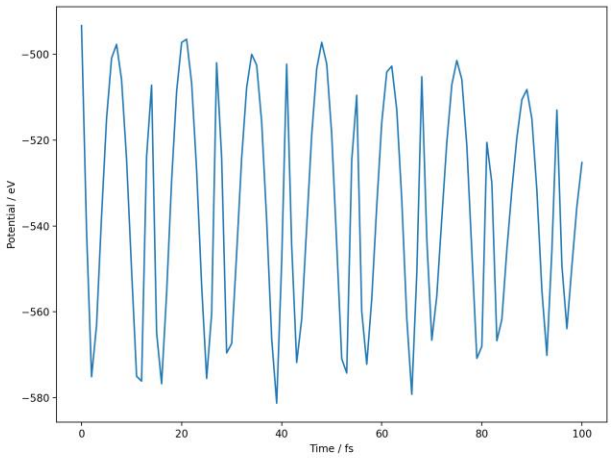


Fig. 1 The MD simulation potential trajectory.

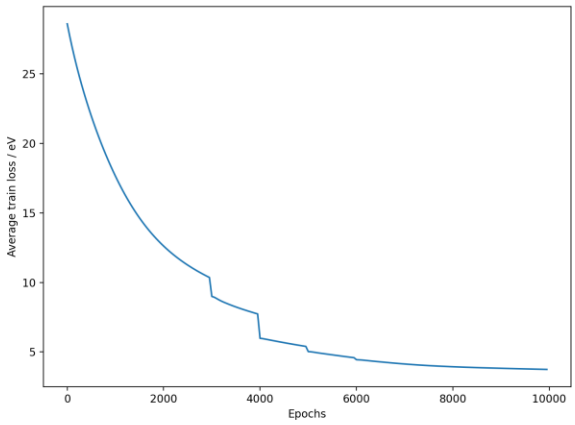


Fig. 2 The average train loss regarding training epochs. The pre-training history is not shown here. Training rates on different epochs: 5e-7 with batch size 5 on 0 to 3000 epochs, 3e-7 with batch size 2 on 3000 to 4000 epochs, 2e-7 with batch size 1 on 4000 to 5000 epochs, 3e-7 with batch size 1 on 5000 to 6000 epochs, and 4e-7 with batch size 1 on 6000 to 10000 epochs.

On the train set, we compared the results of Weismiller et al.'s model and our model, with the PBE calculation result as the baseline, in Fig.3. This figure reveals that although Weismiller et al.'s model yielded a result that is mostly linear to the PBE one, some data points seem to distribute on different lines. This is probably due to the intrinsic of this hybrid model that combines the advantages of different DFT functionals. On the other hand, our model's training result is almost linear with the train dataset PBE energy, but for conformations with higher energy the train loss seems to be larger.

We performed two MC simulations beginning from the last frame of the MD one as the base conformation, one with Weismiller et al.'s model and the other used our model, each by 50 000 frames. We set the MC step and energy barrier to make the acceptance rate of each MC step falls in 40% to 60%. After the MC simulation completes, we used PCA analysis to show the MBTR range of each conformation in those MD/MC simulations in Fig. 4. In this figure we can see that Weismiller et al.'s model and our model go thought different MC trajectories, and their trajectories have intersections, which implies that these two energy model have different local energy minimums and different saddle points. Also, our model leads the MC conformation to wander around a specific local minimum, while Weismiller et al.'s model have not arrived at such a local minimum. This result shows that the two model have different predictions and behaviors in applications.

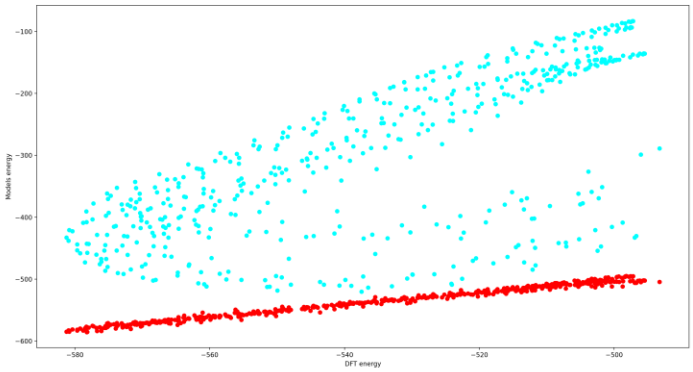
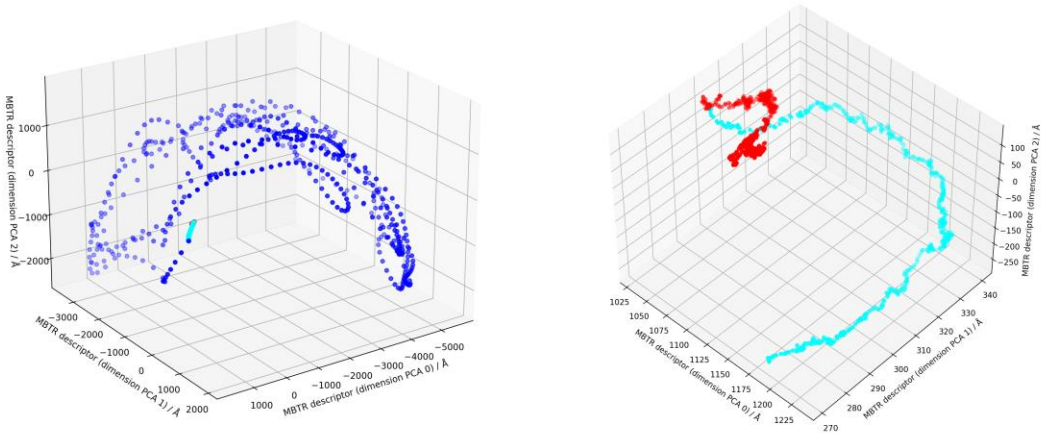


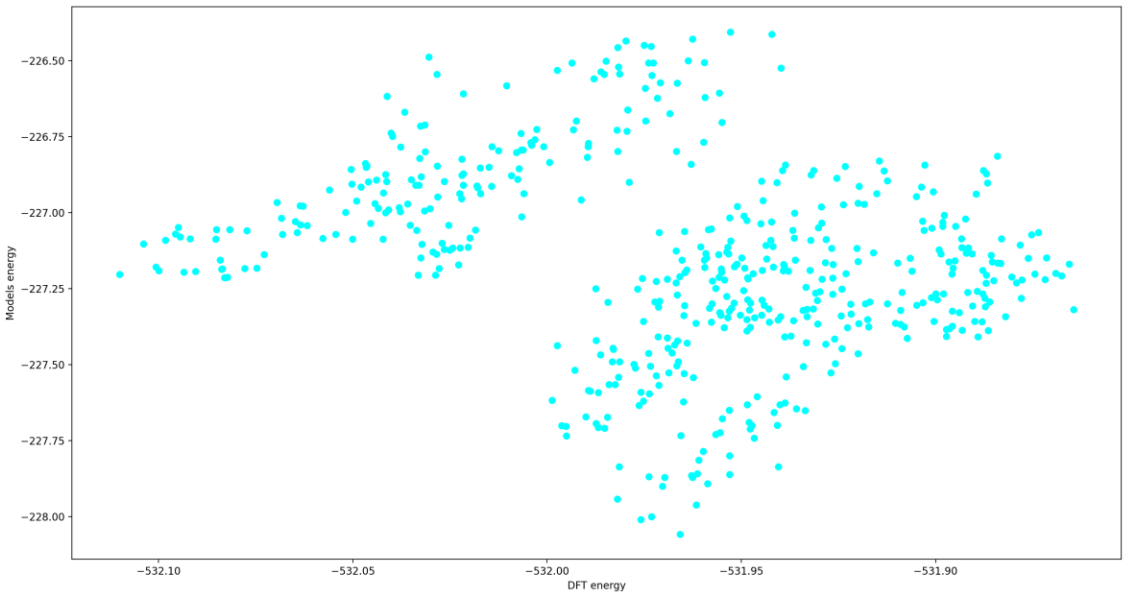
Fig. 3 Weismiller et al.'s model's prediction compared to PBE energy (cyan points) and the training result of this work (red points). X axis: PBE energy on a specific conformation. Y axis: The corresponding model's predicted energy on the same conformation.



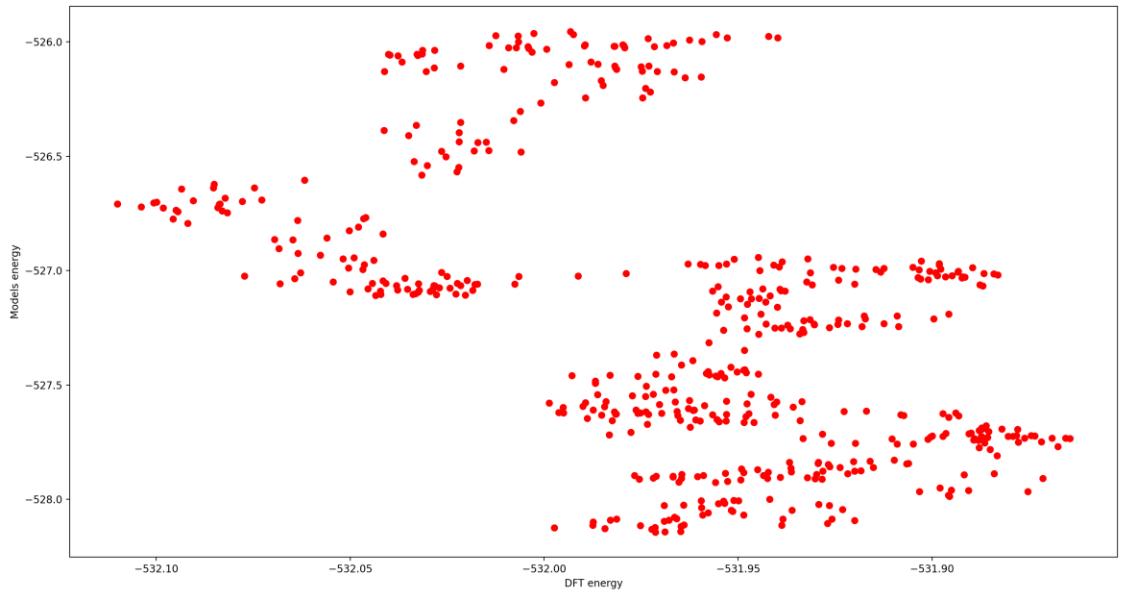
a b

Fig. 4 System conformation in the MD simulation (blue points), MC simulation with Weismiller et al.'s model (cyan points), MC simulation with the model of this work (red points). 4a: MD compared to two MC simulations. 4b: Comparison between two MC simulations.

Finally, to examine whether our model behaves correctly on the conformations generated by MC, we performed PBE calculations on the first 15000 frames of the second MC simulation (the one guided by our model). The result is shown in Fig. 5. Weismiller et al.'s model yield data that are distributed into two clusters and shows some extent of linear distribution in both clusters of data points. Meanwhile, the data from our model have a similar behavior of clustering but one of the clusters does not show apparent linear distribution. Further training of the model, and further experiments are needed to explain this behavior of the ML model.



a



b

Fig. 5 Predicted energy on MC-generated conformations, compared with PBE ones. X axis: PBE energy on a specific conformation. Y axis: The corresponding model's predicted energy on the same conformation. 5a: Result of Weismiller et al.'s model. 5b: Result of our model.

There are still many details in this work that is left with much room for improvement, since this work is limited in terms of both time and computational resources. Larger training set, and smoothier MD trajectories, might be necessary for training a more accurate potential model. The MC step size should be re-modified to be larger in order to make the conformations escape from local minimums to obtain a more intelligible comparison between the MD and MC trajectories. For systems involves chemical reactions, using TDDFT instead of DFT/MD seems to be a more accurate calculation method, although

requires much more computational resources. Also, investigation of EMLM models trained with B3LYP or other fine DFT functionals instead of PBE is needed for a clearer evaluation of the EMLM method. Bond angles, and dihedral bond angle, are important in estimating energy of chemical system so the results of k=3 and k=4 for MBTR should also be verified. Last but not least, designing a new ML algorithm based on EMLM is also a good direction for future researches.

4. Conclusions

In this work we performed a short DFT/MD simulation on a small ammonia borane system and successfully trained a potential predictor for MC simulation with the EMLM machine learning model. We investigated its performance by comparing it to a previously published ReaxFF model on the training set, and by performing MC simulations with both this model and the ReaxFF model from the last frame of the MD trajectory. Further verification are completed with PBE calculations on the newly-generated structures in the MC simulation. From the data and graphs we can draw an initial conclusion that using EMLM models to learn the potentials of reactive systems is at least feasible, if not favorable. However due to restricted access to computational tools and the time schedule of this work, the data are not clearly elucidated and this topic is left to be further investigated. We hope that future researches make this topic clear and accordingly new potential functions or new ML methods be developed for these materials and situations.

References

[1] Schleder, G. R., Padilha, A. C. M., Acosta, C. M., Costa, M. & Fazzio, A. (2019). From DFT to machine learning: recent approaches to materials science–a review. Journal of Physics: Materials, 2(3), 032001.

[2] Elliott, J. A. (2011). Novel approaches to multiscale modelling in materials science. International Materials Reviews, 56(4), 207-225.

[3] Kang, P.L., Shang, C., & Liu, Z.P. (2020). Large-Scale Atomic Simulation via Machine Learning Potentials Constructed by Global Potential Energy Surface Exploration. Accounts of Chemical Research, 53(10), 2119-2129.

[4] Pihlajamäki, A., Hämäläinen, J., Linja, J., Nieminen, P., Malola, S., Kärkkäinen, T., & Häkkinen, H. (2020). Monte Carlo Simulations of Au38(SCH3)24 Nanocluster Using Distance-Based Machine Learning Methods. The Journal of Physical Chemistry A, 124(23), 4827-4836.

[5] Kärkkäinen, T. (2019). Extreme minimal learning machine: Ridge regression with distancebased basis. Neurocomputing, 342, 33–48.

[6] Tsukuda, T. & Häkkinen, H. (2015). Protected metal clusters: from fundamentals to applications; Elsevier: Amsterdam, Netherlands.

[7] van Duin, A. C. T., Dasgupta, S., Lorant, F. & Goddard, W. A. (2001). ReaxFF: A reactive force field for hydrocarbons, Journal of Physical Chemistry A 105, 9396-9409.

[8] Stephens, F., Pons, V., & Tom Baker, R. (2007). Ammonia–borane: the hydrogen source par excellence?. Dalton Trans., 2613-2626.

[9] Weismiller, M., Duin, A., Lee, J., & Yetter, R. (2010). ReaxFF Reactive Force Field Development and Applications for Molecular Dynamics Simulations of Ammonia Borane Dehydrogenation and Combustion. The Journal of Physical Chemistry A, 114(17), 5485-5492.

[10] Tiritiris, I. & Schleid, T. (2003), Die Dodekahydro-closo-Dodekaborate M2[B12H12] der schweren Alkalimetalle (M+ = K+, Rb+, NH4+, Cs+) und ihre formalen Iodid-Addukte M3I[B12H12] (≡ MI • M2[B12H12]). Z. anorg. allg. Chem., 629: 1390-1402.

[11] antanas@kurmis (2016). Information card for entry 1533618, https://www.crystallography.net/cod/1533618.html

[12] Abadi, M. et al. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. CoRR, abs/1603.04467.