**A deep learning-based energy and force prediction framework for high-throughput quantum chemistry calculations**

Guoxin Wu, Qilei Liu\*, Jian Du, Qingwei Meng, Lei Zhang\*

State Key Laboratory of Fine Chemicals, Frontiers Science Center for Smart Materials Oriented Chemical Engineering, Institute of Chemical Process Systems Engineering,School of Chemical Engineering, Dalian University of Technology, Dalian 116024, China

Qilei Liu (liuqilei@dlut.edu.cn), Lei Zhang (keleiz@dlut.edu.cn)

Abstract

Quantum chemistry (QC) calculations rely on solving the Schrödinger equation and have emerged as a significant computational tool in solvent design and drug discovery. However, the computational cost of conventional QC calculation approaches increases dramatically with the complexity of molecular systems, hindering the large-scale screening of solvents and drugs. In this work, a deep learning-based energy and force prediction (DeepEF) framework is established for high-throughput QC calculations. In this framework, a machine learning potential model is first built by utilizing the atomic self-attention mechanism, and then integrated with a geometry optimization algorithm, GeomeTRIC, for efficient molecular geometry optimizations. Finally, an external test set containing 125 solvents is used to test the prediction accuracy and computational speed of the DeepEF framework.

**Keywords**: quantum chemistry, deep learning, attention mechanism, machine learning potential.

* 1. Introduction

Quantum chemistry (QC) calculations use the principles and methodologies of quantum mechanics to precisely predict electronic structures, energy characteristics, and other attributes of chemical substances, thus yielding physicochemical properties. QC has become an indispensable tool in fields such as solvent design and drug discovery. For example, QC calculations can be used to predict the solubilities and the absorption rates of CO2 in different solvents, which helps to reduce the blindness of experiments and shorten the research and development cycle to some extent. The commonly used QC calculation methods are the density functional theory (DFT) method (Parr, 1980) and the Hartree-Fock method (Matthew & Sutcliffe, 1978). It is often hard to adopt these rigorous methods to perform high-throughput calculations on a large number of molecules. Therefore, the efficient and precise execution of QC calculations has emerged as a worldwide challenge.

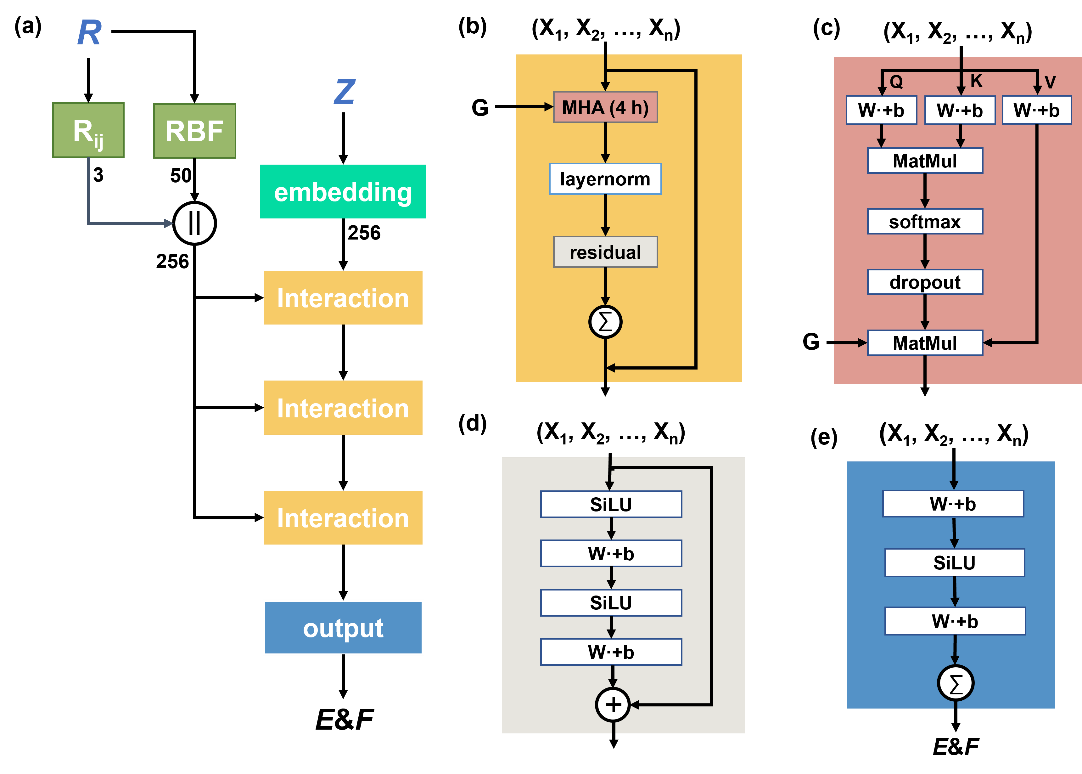
One possible solution to this challenge is developing the machine learning potential (MLP) model, which uses machine learning models to directly represent the functional relationship between atomic coordinates and molecular energies, thus avoiding solving the explicit Schrödinger equation. In this way, the potential energy surface (PES) can be constructed in a fast manner. According to the Born-Oppenheimer approximation (Born & Oppenheimer, 1927), the nucleus mass is much larger than the electron mass. Thus, the molecular energy of the fundamental electronic state can be regarded as a function of nuclear coordinates, which is the theoretical basis of the MLP model. The first MLP for high-dimensional systems is the high-dimensional neural network potential (BPNN) proposed by Behler and Parrinello (2007), which borrows from the atomic contribution method of classical molecular mechanics, that is, the system energy is obtained by adding up the atomic energy output of each atomic network. However, the input of BPNN is an atom-centric symmetric function, where all parameters need to be manually tuned, which is often a painstaking and tedious task. Han et al. (2018) have developed a simple but versatile end-to-end MLP model naming Deep Potential for molecular and atomic systems. The time cost of the MLP model is comparable to that of the empirical force field, but its accuracy is comparable to that of QC calculations. Unke et al. (2021) have added the charge and spin multiplicity to the traditional MLP model, so that the MLP model is able to predict the long-range interaction and the PES of different electronic states. Although MLP models have demonstrated their ability to accurately predict energy and forces for a variety of molecular systems, most of the MLP models are difficult to couple with downstream applications. This limits the users to the secondary development and the extended application of MLP models. The ASE library (Hjorth Larsen et al., 2017) has the capability to use the MLP models for geometry optimizations. However, its current geometry optimization algorithm is characterized by a single convergence condition, requiring that the maximum atomic force value be below a specified threshold. This criterion has the limitation to find the accurate equilibrium geometries. Therefore, it is still a challenge to achieve the seamless integration of the MLP model with external molecular geometry optimization algorithms that have rigorous convergence criteria.

In this work, a deep learning-based energy and force prediction (DeepEF) framework is established for high-throughput QC calculations. In Section 2, a MLP model is developed using an atomic self-attentive neural network for accurate and rapid predictions of molecular energy and forces. Then, in Section 3, the MLP model is coupled with the GeomeTRIC algorithm (Wang & Song, 2016) for molecular geometry optimizations.

* 1. DeepEF framework for QC calculations
     1. Construction of the MLP model

In this work, the QM7-X dataset (Hoja et al., 2021) is used to construct the MLP model. The QM7-X dataset was generated starting from ~7,000 molecular graphs with up to five non-hydrogen atoms (C, N, O, S, Cl) drawn from the GDB13 chemical universe (Blum & Reymond, 2009). After sampling and optimizing the structural and constitutional (stereo) isomers of each graph, 42,000 equilibrium structures are obtained. Each of them was then perturbed in order to obtain 99 additional non-equilibrium structures, leading to a total of about 4.2 million samples. To rigorously evaluate the generalizability of the MLP model, 25 molecular graphs and their conformations are selected from the original 7,000 molecular graphs as the test dataset, and the remaining data are used to train the model in this work.

The feature matrix size varies for different molecules due to differences in the number of atoms in each molecule. The self-attention mechanism in deep learning is used to deal with the case of matrices with different sizes (Vaswani et al., 2023). The architecture of the MLP model is shown in **Figure 1**. The input to the model is the atomic cartesian coordinate matrix and the corresponding nuclear charge number vector , where *N* is the number of atoms. The nuclear charge number vector *Z* is mapped through the embedding layer to the elemental information. The structural information *G* can be obtained from the cartesian coordinate matrix *R* via the coordinate vector difference and the radial basis function (Gasteiger et al., 2022) to ensure the molecular rotation and translation invariances. Elemental information *X* and structural information *G* are collected in three interaction layers (**Figure 1(b)**) dominated by a multi-head self-attention network (**Figure 1(c)**) and a residual block (**Figure 1(d)**). Finally, the molecular energy and forces are obtained by integrating the information using the output layer (**Figure 1(e)**). The MLP model developed in this work has achieved a high prediction accuracy in estimating the energies and forces of the molecules in the test set with the determination coefficients of 1.00 and 0.98, as well as the mean absolute errors of 0.048 eV and 0.044 eV/Å, respectively.



**Figure 1**. The machine learning potential model. **a** Atom self-attentional neural network. **b** Interaction layer. **c** Multi-head self-attention network. **d** Residual block. **e** Output layer.

* + 1. Integration of the MLP model and the geometry optimization algorithm

Accurate calculations of QC-based properties are essential for the study and design of new materials, drugs, and so on. Common QC tasks include the optimizations of minimum energy structures and excited states, the calculations of spectral characteristics, thermodynamic properties of reactions, and so on. All of these properties can be inferred from QC calculations of electronic structures by solving the Schrödinger equation. However, as the complexity of the system increases, the computation time can range from hours to days, and the need for efficient computing cannot be satisfied.

The DeepEF framework is able to optimize molecular structures by integrating the developed MLP model and the GeomeTRIC algorithm (Wang & Song, 2016). The convergence criteria of the GeomeTRIC algorithm are:

(1) The root mean square deviation (*RMSD*) from the previous step is less than 1.2 × 10-3 Å.

(2) The maximum atomic displacement from the previous step is less than 1.8×10-3 Å.

(3) The root mean square of the nuclear gradient is less than 3.0 × 10-4 a.u.

(4) The maximum nuclear gradient for any atom is less than 4.5 × 10-4 a.u.

(5) The energy change from the previous step is less than 1.0 × 10-6 a.u.

Compared with the Gaussian software, the convergence criteria of GeomeTRIC has a large probability to find the molecular equilibrium geometry. The equilibrium geometry optimized by DeepEF and the equilibrium geometry optimized by DFT (B3LYP/6-31G\*) are compared by *RMSD* (as shown in **Eq. (1)**), which represents the root mean square of the coordinate differences between two molecular structures.

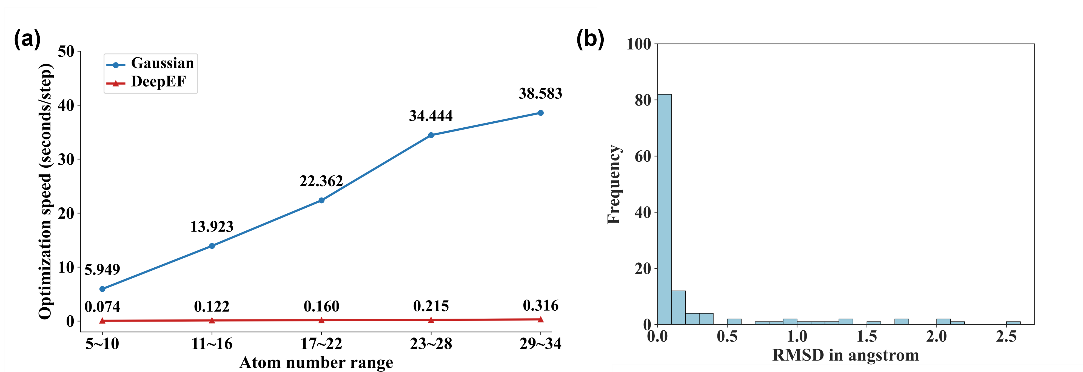
|  |  |
| --- | --- |
|  | (1) |

The optimized structures can be then sent to the Gaussian software (Frisch et al., 2016) to make calculations of single point energy, dipole, HUMO/LUMO energy, chemical shift, and other QC-based properties.

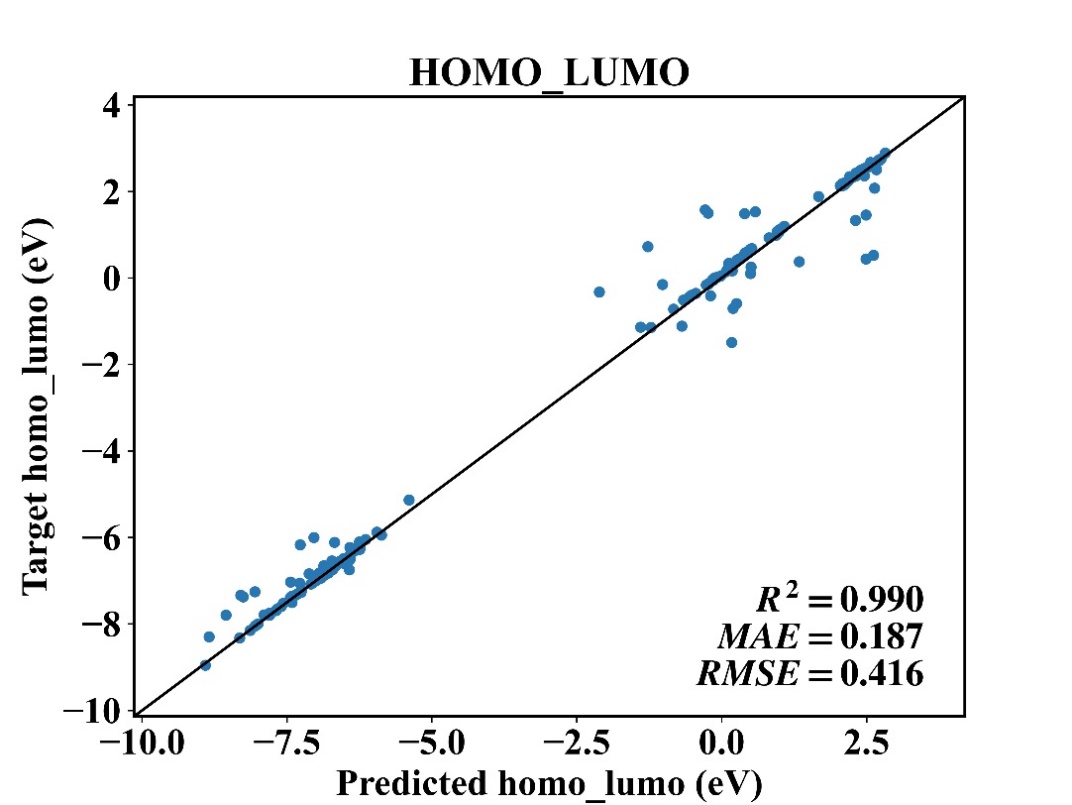
* 1. Results and discussions

Geometry optimization is a core task in computational chemistry, which plays a crucial role in understanding and predicting molecular properties. The geometry of a molecule not only affects its energy and stability, but also directly affects some important properties such as reactivity, spectral property, dipole, and so on. In many applications, from new drug development to material design, accurate molecular structure is one of the keys to success. By utilizing DeepEF for geometry optimizations, the need for costly DFT methods is circumvented.

Here, 125 solvent molecules are screened from the solvent database (Krasnov et al., 2022) outside the training dataset. The geometries of 125 solvent molecules are optimized by DeepEF and compared using the Gaussian software (Frisch et al., 2016). The optimization results are shown in **Figure 2**. **Figure 2(a)** shows that the calculation speed for each geometry optimization step using our MLP model is about two orders of magnitude faster than that using the DFT method in the Gaussian software. As the number of atoms increases, the optimization time of our MLP model hardly changes at each step, exhibiting a linear relationship between the optimization time and the number of atoms, while the optimization time of the DFT method increases exponentially with the increase of the atom number. **Figure 2(b)** shows that the *RMSD*s of 70% molecules are less than 0.1 Å, indicating that our MLP model has a high prediction accuracy in predicting molecular energy and forces, and our DeepEF framework enables the geometry optimizations in a rapid and accurate manner. After optimizing the molecular structures, some QC tasks are performed by the Gaussian software to obtain the molecular QC-based properties. A script has been provided by DeepEF to convert molecular “.xyz” files into the Gaussian input files. The highest occupied molecular orbital and lowest unoccupied molecular orbital (HOMO/LUMO) of 125 solvent molecules are computed. The calculation results are shown in **Figure 3**, and the HOMO/LUMO properties calculated from the DeepEF optimized structures are closely consistent with those calculated from the DFT optimized structure.



**Figure 2**. The geometry optimization results of 125 molecules using the DeepEF framework and the gaussian software. **a** The calculation speed for each geometry optimization step using the MLP model in DeepEF and the DFT (B3LYP/6-31G\*) method in Gaussian. **b** The *RMSD* between the DeepEF optimized molecular geometries and the Gaussian optimized molecular geometries.



**Figure 3**. The HOMO/LUMO calculation results of 125 molecules using the MLP model in DeepEF and the DFT (B3LYP/6-31G\*) method in Gaussian.

* 1. Conclusions

In this work, a DeepEF framework is proposed for efficient QC calculations. With the developed MLP model, the DeepEF framework has a powerful ability to fast optimize the structures of various small organic molecules at the cost of minor prediction error. The computational speed of our MLP model is two orders of magnitude faster than that of the traditional DFT method (B3LYP/6-31G\*), and the *RMSD*s of 70% optimized molecular geometries are less than 0.1 Å. In the future work, our DeepEF framework will be integrated with the conductor-like screening model-based computer-aided molecular design technique to achieve the high-throughput design of promising solvents and drug candidates.

Acknowledgement

The authors are grateful for financial supports of National Natural Science Foundation of China [22208042], “the Fundamental Research Funds for the Central Universities” [DUT22YG218], and China Postdoctoral Science Foundation [2022M710578].

References

R. G. Parr, 1980. Density functional theory of atoms and molecules, Dordrecht.

O. T. Unke, S. Chmiela, M. Gastegger, K. T. Schütt, H. E. Sauceda, and K.-R. Müller, 2021. SpookyNet: Learning force fields with electronic degrees of freedom and nonlocal effects. Nature Communications, 12(1), 7273.

M. J. Frisch, G. W. Trucks, H. B. Schlegel, et al., 2016. Gaussian 16 Rev. C.01. Wallingford, CT.

M. Born, and R. Oppenheimer, 1927. Zur Quantentheorie der Molekeln. 389(20), 457-484.

L.-P. Wang, and C. Song, 2016. Geometry optimization made simple with translation and rotation coordinates. The Journal of Chemical Physics, 144, 214108.

L. Krasnov, S. Mikhaylov, M. V. Fedorov, and S. Sosnin, 2022. BigSolDB: Solubility dataset of compounds in organic solvents and water in a wide range of temperatures. Zenodo.

L. C. Blum, and J.-L. Reymond, 2009. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. Journal of the American Chemical Society, 131(25), 8732-8733.

J. Hoja, L. Medrano Sandonas, B. G. Ernst, A. Vazquez-Mayagoitia, R. A. DiStasio Jr, and A. Tkatchenko, 2021. QM7-X, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules. Scientific Data, 8(1), 1-11.

J. Han, L. Zhang, R. Car, and E. Weinan, 2018. Deep Potential: A general representation of a many-body potential energy surface. Communications in Computational Physics, 23(3), 629-639.

J. Gasteiger, J. Groß, and S. Günnemann, 2020. Directional message passing for molecular graphs. arXiv preprint arXiv:2003.03123.

J. Behler, and M. Parrinello, 2007. Generalized neural-network representation of high-dimensional potential-energy surfaces. Physical Review Letters, 98(14), 146401.

J. A. D. Matthew, and B. T. Sutcliffe, 1978. The Hartree–Fock method for atoms: A numerical approach. Physics Bulletin, 29(4), 178.

A. Vaswani, N. Shazeer, N. Parmar, et al., 2017. Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008.

A. Hjorth Larsen, J. Jørgen Mortensen, J. Blomqvist, et al., 2017. The atomic simulation environment—a Python library for working with atoms. Journal of Physics: Condensed Matter, 29(27), 273002.