TSeC: an efficient transition state search tool driven by machine learning potential

Kun Tang,a Lei Zhang,a Qingwei Meng,a Jian Du,a Qilei Liua,\*

aState 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)

Abstract

The density functional theory (DFT) has achieved success in the study of reactive systems by combining diverse transition state search algorithms in various electronic structure calculation packages. However, due to the complexity and high cost of the DFT-based computation of the complex reactive systems, there has been continuous development in machine learning potential (MLP) in recent years. It has been demonstrated that MLP is feasible to approximate the potential derived from electronic structure methods on the basis of certain data, and it has a computational cost comparable to that of semi-empirical methods. In this work, an efficient transition state search (TSeC) tool driven by MLP is developed for reactive system research. Multiple transition state search algorithms are combined with the MLP model to achieve efficient evaluations of the reaction properties of organic reaction systems within the scope of traditional DFT research.

**Keywords**: machine learning potential, transition state, reaction barrier, deep learning.

* 1. Introduction

With the continuous progress of machine learning modelling techniques, the machine learning potential (MLP) is now recognized for its capability to approximate the computational accuracy of density functional theory (DFT), a widely employed method in molecular simulations. Additionally, MLP exhibits a computational speed several orders of magnitude faster than DFT (Schreiner et al., 2022a). This provides a cheap and accessible way to study chemical reactions, as the search process for chemical reaction paths and transition states require iterative calculations of molecular energies and atomic forces. But the prerequisite for doing so is that the MLP needs to well describe the potential energy surface (PES) of the transition state regions, depending on whether the database used for training the model is sampled from these regions. Recently, Schreiner et al. (2022b) have constructed such a database (Transition 1x) and developed the corresponding MLP model (a model trained on Transition 1x database), making it feasible that the MLP can be applied to study the properties of reactive systems (Schreiner et al., 2022a). However, the known transition state search algorithms have more or less individual shortcomings and more effective transition state search algorithms need to be combined with MLP.

Searching for transition states in chemical reactions is an inherently challenging task, primarily due to two key factors. First, the molecule system involves multiple atoms and numerous degrees of freedom, rendering the PES a high-dimensional surface. Navigating this space to pinpoint transition state positions is exceedingly difficult due to the vast array of potential configurations that must be considered. Second, in the realm of large molecule systems or intricate reactions, the computational cost associated with determining the energy and structure of transition states can be prohibitively high. This necessitates the use of advanced and computationally cheap computational methods to complete this task more quickly and accurately.

Researchers have proposed numerous transition state search algorithms. For example, the Berny algorithm used in Gaussian software (Frisch et al., 2016), which is based on the initial guess structure. It needs to calculate the Hessian matrix and check the eigenvalue at each step. In addition, there is an efficient Dimer algorithm (Henkelman and Jónsson, 1999) that does not require time-consuming calculation of the Hessian matrix at each step, but only calculates the first derivative of energy (i.e., force) to determine the direction of the searching steps. Recently, Liu et al. (2023) has proposed a transition state search algorithm, GENiniTS-RS, based on the reactive sites, which significantly accelerates the identification of transition states for specific reaction types and demonstrates a high success rate. Besides, the nudged elastic band (NEB) algorithm (Henkelman and Jónsson, 2000) is also an efficient and widely used dual ended transition state search method. It constructs an energy elastic band on the reaction coordinates by interpolating between the reactants and products structures. However, the original NEB algorithm could not provide accurate transition state structures. To this end, an improved version of the climbing image-nudged elastic band (CI-NEB) algorithm (Henkelman et al., 2000) has been proposed. The above methods do not use Hessian matrices to determine the direction of the reaction coordinates, so it is necessary to further identify and refine the transition state structures rigorously. But completing this task requires time-consuming calculations of the complete Hessian and iterative diagonalization for molecular systems. Therefore, Hermes et al. (2022) have proposed the Sella algorithm, a novel efficient algorithm for the optimization of molecular structures to saddle points using a redundant internal coordinate system. This algorithm converges to saddle point structures in fewer effective steps than that of popular quantum chemistry codes.

In this work, an efficient MLP-based transition state search (TSeC) framework is developed to predict reaction barriers of organic reactions in an automatic, fast manner. In Section 2, the TSeC framework that combines the MLP model and advanced transition state search algorithms is discussed in detail. In Section 3, a benchmark test comprising 55 initial guesses of transition states generated by the GENiniTS-RS algorithm is introduced to highlighting the applications of TSeC tool in reactive system research.

* 1. TSeC framework

The execution of transition state search algorithms of the TSeC framework is shown in **Figure 1**. Here, a hybrid transition state search strategy is developed by combining CI-NEB (Henkelman et al., 2000), GENiniTS-RS (Liu et al., 2023), and Sella (Hermes et al., 2022) algorithms and integrated with the Transition 1x MLP model (Schreiner et al., 2022a; Schreiner et al., 2022b), which is used as a calculator for energies and forces in the atomic simulation environment (ASE) package (Larsen et al., 2017). If the user only inputs the SMILES of products or reactants, it will be transferred to GENiniTS-RS to obtain the transition state initial guess of this reaction as long as the reaction exists in the template library. If the SMILES is not in the reaction template library, it is necessary to prepare the structures of reactants and products for the CI-NEB algorithm. Then the Sella algorithm is used for further refinement of the saddle point obtained from the two technical routes of CI-NEB and GENiniTS-RS. Finally, the Sella algorithm is used to perform the intrinsic reaction coordinate (IRC) calculation of the saddle point obtained by Sella refinement, and determine the final IRC path and corresponding reaction barrier. This also confirms the correctness of the searched reaction transition states.



**Figure 1**. The overview of the TSeC framework. **a** There are two input methods for searching reaction transition states: CI-NEB requires manual preparation of appropriate three-dimensional structures of reactants and products, and GENiniTS-RS only requires the inputs of the SMILES representations of reactants or products to automate the search. **b** Transition state refinement and intrinsic reaction coordinate calculation. **c** Research scenarios of reactive systems applied by the TSeC framework.

* + 1. Transition 1x MLP model

Transition 1x is a dataset providing a collection of molecular configurations on and along reaction paths for approximately 10k reactions. Schreiner et al. (2022b) published the dataset and trained the Transition 1x MLP model, which is built upon the PaiNN (Schütt et al., 2021) architecture that employs equivariant message passing graph neural networks. The reactions in the database involve four types of elements (C, H, O, N), and the molecular structures are sampled from reactive and high energy regions (non-equilibrium molecular structure) by the NEB algorithm to ensure that the MLP trained in this dataset can accurately understand the reaction process.

* + 1. Reaction path and transition state search algorithms

As shown in **Figure 1**, for reactions within the reaction templates of the GENiniTS-RS algorithm, it can be started from the SMILES representations of products or reactants. For other types of reactions, the CI-NEB algorithm started from the three-dimensional structures of reactants and products has been integrated into the TSeC framework. In addition, the Sella algorithm is also integrated for further refinements of transition state and acquisitions of IRC paths.

* + - 1. GENiniTS-RS algorithm

GENiniTS-RS has a reactive site coordinate database containing a few common organic chemical reactions, which are sufficient for ordinary organic retrosynthesis research. Searching for reaction transition states through GENiniTS-RS mainly involves the following three steps.

1. The roughly initial three-dimensional geometries of reactive complex are automatically and fast generated from the SMILES strings of reactants or products through the molecular mechanics method.
2. Adjust the coordinates of the reaction complex through a series of algorithms to generate a transition state initial guess structure. First, the reaction sites are matched in point-to-point coordinates, followed by the translation and rotation of the molecular branches, and the distance geometry algorithm is used to avoid atomic collisions to ensure the rationality of the molecular structure. Finally, applying molecular mechanics and semi-empirical DFT calculations to obtain the initial guess transition state.
3. Perform a conformational isomer search on the obtained initial guess transition state structure to find the stable conformation with the lowest energy.
   * + 1. CI-NEB algorithm

Considering the limitations of the reaction templates of GENiniTS-RS, there is a problem of absence of initial structural reactive site coordinates, it remains necessity to provide the option to manually search for transition states from the three-dimensional structure of reactants and products. Thus, a dual-ended algorithm, CI-NEB, is employed in TSeC. The CI-NEB algorithm comes from the AutoNEB module, which is provided in the ASE package (Larsen et al., 2017) employing the basic NEB method and a CI-NEB method. First of all, the user needs to supply at minimum two end-points (the initial and final states). Subsequently, additional images will be interpolated between these two end-points dynamically, and in each step, a small force is applied through “nudged” atom pushing, adjusting the band structure along the energy gradient direction, and then minimizing the energy to ensure that the lowest energy point on the path is found. After this process is completed, the climbing image is performed to locate the saddle point on the minimal energy path (MEP). In this work, the spring constant between images on the path is set to 0.1 eV·Å-2, the maximum force along the NEB path is set to 0.025 eV·Å-1, and the max images along the NEB path when done is set to 20.

* + - 1. Sella algorithm

Sella is a novel algorithm for the automated optimization of molecules to saddle points in a basis of redundant internal coordinates. Iterative diagonalization of the Hessian is necessary for the refinement of the rough transition state structures found by the above algorithms. Sella provides a method to incorporate the information obtained from iterative differentiation algorithm into the construction of approximate Hessian matrices, which accelerates the refinement task. In this study, the preliminary transition state structure, obtained through the above two algorithms, is served as the initial input for the Sella algorithm calculation. Following refinement, the first-order saddle point structure undergoes additional IRC calculations to delineate the reaction pathway. And ultimately save the reaction kinetics information and optimization trajectory file of the reaction.

* 1. Results and discussions

To showcase the efficiency of the transition state search algorithm and the Transition 1x MLP utilized in TSeC, 55 transition state initial guess of the reactions derived from the GENiniTS-RS reaction templates are employed for benchmark test. **Table 1** compares the success rate, average time consumption, and average number of steps calculated using the DFT method (the calculation level is b3lyp/6-31g(d,p) em=gd3bj) in the Gaussian software (Frisch et al., 2016) combined with its transition state search algorithm, the MLP model combined with Sella algorithm, and the DFT method (b3lyp/6-31g(d,p) em=gd3bj) combined with Sella algorithm for these 55 reactions.

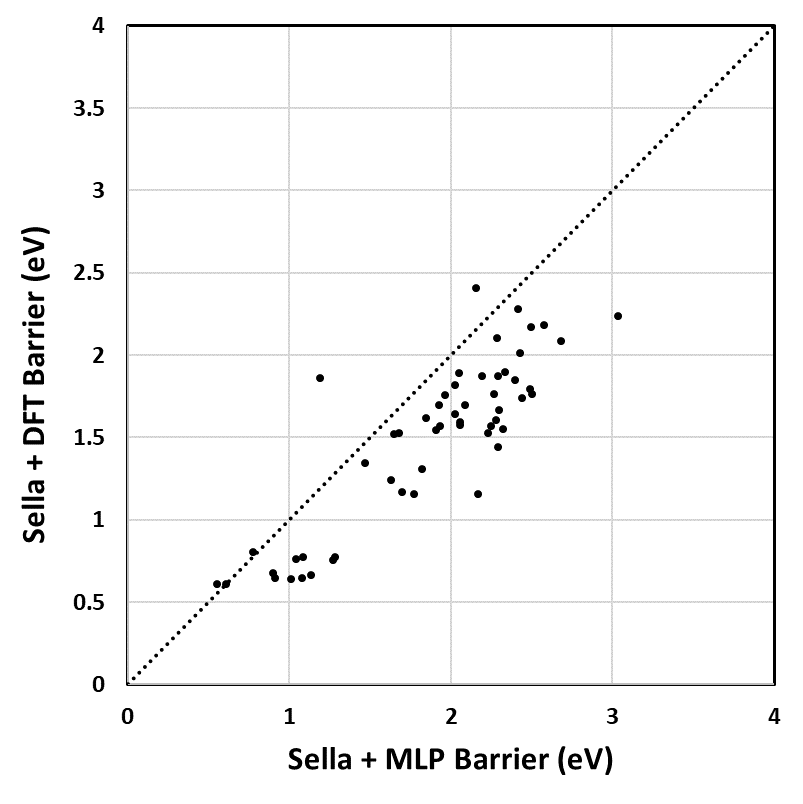
From **Table 1**, it is evident that, under the same transition state search algorithm, MLP significantly expedites this task and exhibits a high success rate. This further substantiates the efficacy of the Transition 1x MLP. In the case of using DFT calculation, Sella algorithm is also more efficient than the search algorithm in Gaussian (Higher success rate and lower average single step computation time), and its Python implementation also facilitates the combination with MLP.

**Table 1**. The success rate, average time consumption, and average number of steps calculated using the DFT method in the Gaussian software, the MLP model combined with Sella algorithm, and the DFT method combined with Sella algorithm.

|  |  |  |  |
| --- | --- | --- | --- |
|  | DFT in Gaussian | Sella+MLP | Sella+DFT |
| Success rate | 87% | 98% | 100% |
| Average time consuming (s) | 1699 | 216.4 | 4,861.6 |
| Average steps | 29.83 | 154.2 | 146.53 |

To demonstrate the effectiveness of the Sella algorithm in refining the initial guess structure of the transition states, 54 reactions that successfully searched for transition states are selected. Using the DFT calculation results as the benchmark. The root mean squared error (RMSE) of the transition state structures generated by MLP model combined with CI-NEB is 0.887 Å, and the RMSE of the transition state structure refined by the MLP model combined with GENiniTS-RS + Sella algorithm is 0.879 Å. This indicates that Sella has improved the geometric structure accuracy of the transition state to a certain extent.

The Sella algorithm is employed to conduct IRC searches using MLP and DFT as potential, respectively. The differences in the reaction barriers derived from these two potentials are then compared in **Figure 2**. Based on the results obtained using the DFT potential, the mean average error (MAE) of the potential barriers calculated by Sella + MLP on this dataset is 0.423 eV. This indicates that there is still some room for improvement in Transition 1x MLP on the dataset derived from the template reactions of GENiniTS-RS. For reactions containing a large number of heavy atoms, new MLPs need to be trained.



**Figure 2**. Comparison of reaction barriers found with Sella using MLP as potential on the x-axis vs. DFT on the y-axis.

* 1. Conclusions

In this paper, a computational tool (TSeC) is developed for transition state and MEP searches. It combines the MLP and advanced transition state search algorithms. On the test dataset, the combination of MLP + Sella algorithm in TSeC achieves a 98% success rate in transition state searches and MAE of 0.423 eV on reaction barrier. Additionally, TSeC integrates the CI-NEB algorithm, which can switch between algorithm modes based on various scenario requirements, such as the absence of reaction templates, providing broad applicability. TSeC can also be employed for the swift generation of extensive reaction kinetics databases, providing data for better performing MLPs. For instance, the Transition 1x database exclusively encompasses heavy atoms C, N, O, rendering it incapable of predicting reactions involving halogens in GENiniTS-RS. Consequently, the need arises to train dedicated MLPs on new datasets in the future.

Finally, although there are various search algorithms for reaction transition states, it is still difficult to achieve chemical reactions *in silico* without relying on human intervention. GENiniTS-RS currently relies on templates. In the future, it is imperative to develop new generative artificial intelligence models to enable the generation of transition state initial guesses without the need for prior experience.

References

Q. Liu, K. Tang, L. Zhang, J. Du, and Q. Meng, 2023. Computer‐assisted synthetic planning considering reaction kinetics based on transition state automated generation method. AIChE Journal, 69(7), e18092.

M. Schreiner, A. Bhowmik, T. Vegge, P. B. Jørgensen, and O. Winther, 2022a. NeuralNEB—neural networks can find reaction paths fast. Machine Learning: Science and Technology, 3(4), 045022.

M. Schreiner, A. Bhowmik, T. Vegge, J. Busk, and O. Winther, 2022b. Transition1x - a dataset for building generalizable reactive machine learning potentials. Scientific Data, 9(1), 779.

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

K. T. Schütt, O. T. Unke, and M. Gastegger, 2021. Equivariant message passing for the prediction of tensorial properties and molecular spectra. Paper presented at the International Conference on Machine Learning.

G. Henkelman, B. P. Uberuaga, and H. Jónsson, 2000. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. The Journal of Chemical Physics, 113(22), 9901-9904.

G. Henkelman, and H. Jónsson, 2000. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. The Journal of Chemical Physics, 113(22), 9978-9985.

G. Henkelman, and H. Jónsson, 1999. A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives. The Journal of Chemical Physics, 111(15), 7010-7022.

E. D. Hermes, K. Sargsyan, H. N. Najm, and J. Zádor, 2022. Sella, an Open-Source Automation-Friendly Molecular Saddle Point Optimizer. Journal of Chemical Theory and Computation, 18(11), 6974-6988.

A. H. Larsen, J. J. 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.