# Exploring the chemical space of linear alkane pyrolysis via Deep Potential GENerator

Zeng, Jinzhe; Zhang, Linfeng; Wang, Han; et.al.



Zeng, J., Zhang, L., Wang, H., & Zhu, T. (2020). Exploring the chemical space of linear alkane pyrolysis via Deep Potential GENerator. In Energy & fuels (Vol. 35, Issue 1, pp. 762–769). American Chemical Society. https://doi.org/10.7282/00000187
Document Version: Accepted Manuscript (AM)

Published Version: https://doi.org/10.1021/acs.energyfuels.0c03211

# Explore the Chemical Space of Linear Alkanes Pyrolysis via Deep Potential Generator

Jinzhe Zeng,[†] Linfeng Zhang,[*,‡] Han Wang,[*,¶] and Tong Zhu[*,§,‖]

[†]*Department of Chemistry and Chemical Biology, Rutgers University, Piscataway, NJ 08854, USA*

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

[¶]*Laboratory of Computational Physics, Institute of Applied Physics and Computational Mathematics, Huayuan Road 6, Beijing 100088, People's Republic of China*

[§]*School of Chemistry and Molecular Engineering, East China Normal University, Shanghai 200062, People's Republic of China*

[‖]*NYU-ECNU Center for Computational Chemistry at NYU Shanghai, Shanghai 200062, People's Republic of China*

E-mail: linfengz@princeton.edu; wang_han@iapcm.ac.cn; tzhu@lps.ecnu.edu.cn

**Abstract**

Reactive molecular dynamics (MD) simulation is a powerful tool to study the reaction mechanism of complex chemical systems. Central to the method is the potential energy surface (PES) that can describe the breaking and formation of chemical bonds. The development of PES of both accurate and efficient has attracted significant effort in the past two decades. Recently developed Deep Potential (DP) model has the promise to bring *ab initio* accuracy to large-scale reactive MD simulations. However, for complex chemical reaction processes like pyrolysis, it remains challenging to generate

reliable DP models with an optimal training dataset. In this work, a dataset construction scheme for such a purpose was established. The employment of a concurrent learning algorithm allows us to maximize the exploration of the chemical space while minimize the redundancy of the dataset. This greatly reduces the cost of computational resources required by *ab initio* calculations. Based on this method, we constructed a dataset for the pyrolysis of *n*-dodecane, which contains 35,496 structures. The reactive MD simulation with the DP model trained based on this dataset revealed the pyrolysis mechanism of *n*-dodecane in detail, and the simulation results are in good agreement with the experimental measurements. In addition, this dataset shows excellent transferability to different long-chain alkanes. These results demonstrate the advantages of the proposed method for constructing training datasets for similar systems.

# Introduction

Linear alkanes are major components of many jet fuels. At high temperatures, heavy linear alkanes will decompose into smaller alkanes and then trigger the combustion process.[1–4] Such process is also used to cool down the wall of the scramjet engine. A comprehensive understanding of the chemical kinetics in the pyrolysis process of heavy linear alkanes is crucial to the design of novel engines for the improvement of combustion efficiency. In the past decades, reactive molecular dynamics (MD) simulations based on either empirical potential energy surface (force fields) functions[5] or *ab initio* quantum mechanical (QM) calculations[6–9] have gradually become an indispensable tool for the investigation of the mechanisms of complex chemical reactions such as combustion.[10] Compared with traditional knowledge-based QM calculations such as transition state optimization,[11] reactive MD simulation can provide the dynamical properties of the involved reactions without the need of additional information other than the initial state of the reactants. Moreover, compared with experiments, the reaction conditions and resolution can be controlled in a much easier way by atomistic simulation.

Over the past 20 years, the ReaxFF force field, which was initially designed for combustion simulation has been a great success in the theoretical study of combustion mechanisms.[12–16] However, it has not satisfied all the demands of researchers in the community of combustion study. Due to the employment of empirical energy functions and complex parameterization processes, the accuracy of the ReaxFF is of great concern.[17–19] Reactive MD based on *ab initio* QM calculations (*ab initio* MD simulation, AIMD), despite its much more reliable accuracy, is only suitable for short-time (normally dozens of pico-seconds) simulation of small systems due to its high computational cost. Thus, more and more researchers have devoted themselves to the improvement of existing reactive MD methods and the exploration of new approaches.

Machine learning based tools, especially neural networks (NN), have provided the possibility to develop PES models with the efficiency of the molecular mechanics (MM) method

and the accuracy of the QM method. NN models constitute a very flexible class of mathematical functions, which in principle is able to efficiently approximate a large class of real-valued functions to a satisfactory accuracy. Different types of NN-based PES models have been proposed for materials and bio-molecules. Some examples are found in Refs.[20–41]

In our previous study, an end-to-end NN-based model called Deep Potential - Smooth Edition (DeepPot-SE)[38] has been developed to efficiently represent the PES of organic molecules, metals, semiconductors, and insulators, with an accuracy of *ab initio* QM models. With the assistance of DeepPot-SE, in an earlier study by some of the authors, a Deep Potential (DP) model was built to simulate the combustion process of methane[42] at the MN15/6-31G** level.[43] Benefitted from the efficiency of the DP model, one nano-second reactive MD simulation for a system containing 100 $CH_4$ and 200 $O_2$ molecules was performed and details of the combustion mechanisms were revealed.

The accuracy and transferability of DP models are determined by the quality of its training set, including the QM level at which the training set is labeled and how representative and complete the training snapshots are. As such, one needs to ensure that the training set covers the target chemical space with as small redundancy as possible. Only in this way can we label the training set with high-level QM calculations. Such a task has been very difficult. For example, in our previous study of the combustion process of methane, a training set consisting of half a million structures was constructed to cover all reactants, products, intermediates, and their reactions. To address this issue, we have introduced the Deep Potential GENerator (DP-GEN) scheme,[44] a concurrent-learning algorithm used to generate PES models in a way that minimizes human intervention and reduces the computational cost for data generation and model training. The DP-GEN scheme has demonstrated its success in the modeling of metallic systems,[44,45] chemical reactions at the interface of water and $TiO_2$,[46] transition from molecular to ionic ice at high pressure,[47] etc. Moreover, DP-GEN has been turned into an open-source software platform for the generation of reliable DP models.

In the present work, we extended the capability of DP-GEN to complex gas-phase reactive systems. Taking the pyrolysis of *n*-dodecane as an example, the exploration and labeling algorithms were developed and their performance was systemically benchmarked. We also evaluated the transferability of the DP model by applying the PES trained on n-dodecane to study the pyrolysis of *n*-decane, *n*-tetradecane, and *n*-icosane. The details of the methodology are introduced in Section 2. The results and discussions are presented in Section 3. Finally, the conclusion is given in Section 4.

# Methodology

## Construction of DP

As shown in Figure 1, the whole workflow for the construction of DP model is divided into 3 modules: Initialization, Concurrent Learning, and Finalization. An initial dataset is obtained in the *Initialization* module to kick-off the *Concurrent learning* section. In the *Concurrent learning* section, the chemical space of the reaction is gradually explored as the reaction proceeds. When the chemical space is sufficiently explored, a final dataset will be obtained and the final DP can be trained based on it. Taking the *n*-dodecane pyrolysis as an example, details of these modules are introduced below.

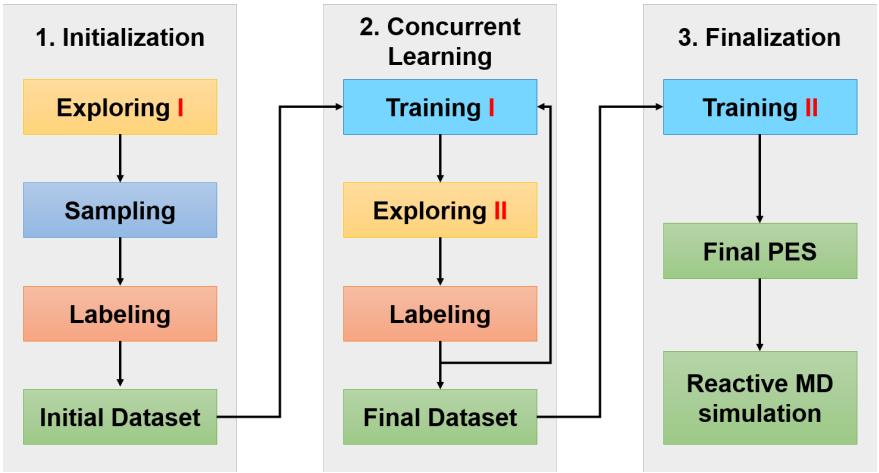


Figure 1: The DP-GEN workflow for the construction of reliable DP models.

***Initialization***. The main purpose of the initialization module is to create an initial dataset for the concurrent learning module. It roughly contains 3 steps:

*Exploring I.* As the starting configuration, a box containing 40 $n$-dodecane molecules with the density of 0.25g/cm$^3$ was constructed by using the Amorphous Cell module in the Materials Studio 2017 program suite.[48] Then, a 1 ps NVT MD simulation was performed by using the LAMMPS software[49] and the ReaxFF PES model developed by Chenoweth et al. (CHO-2008).[50] The NVT ensemble was employed and the temperature was maintained at 3000K with the Berendsen thermostat. The time step was set to 0.1 fs and the atomic coordinates were recorded in every 10 fs. We also built boxes containing $n$-decane, $n$-tetradecane, and $n$-icosane, respectively, with the same method and density. Then, a molecular cluster was created for each atom (defined as the center atom) in the trajectory, which contains the center atom and any molecular species within a 3.5 Å distance from it. Since the pyrolysis reaction is dominated by bond breaking and short-range collision between different spices, 3.5 Å is a reasonable cutoff threshold that balances the accuracy and computational cost.

*Sampling.* From the previous step, we might obtain thousands of molecular clusters, which exhibit heavy redundancy. In this step, these clusters were firstly classified according to the bond-type of the center atom and then the $k$-means clustering algorithm was used to remove the redundancy. Details of relevant methods can be found in our previous study.[42] Finally, an initial dataset containing 590 molecular clusters was obtained.

*Labeling.* The potential energy and atomic forces of each molecular cluster in the dataset were calculated with Gaussian 16[51] at the MN15/6-31G** level.[43] The MN15 functional was chosen because it is specifically designed to have broad accuracy for multi-reference and single-reference systems. When compared with 82 other density functionals, MN15 gives the second smallest mean unsigned error for 54 inherently multiconfigurational systems. The spin multiplicity was set to 1 or 2 according to the number of electrons in the cluster. The stability of DFT wavefunction was checked during self-consistent field calculation, and if instability was found, further optimization was performed until a stable solution was achieved.

It is worth mentioning that the size and the content of the initial dataset was not very critical. It was only used as a start point of the concurrent learning section. However, the employment of initial data set might effectively reduce the number of iterations of concurrent learning.

***Concurrent Learning***. To explore the chemical space more efficiently, the concurrent learning algorithm was used, which contains a number of iterations. Each iteration is composed of three steps:

*Training I.* In this step, based on the dataset from the previous step (or the initial dataset at the beginning), 4 DeepPot-SE models were trained at the same time with the same network architectures but different random seeds for their parameters. In the DeepPot-SE model, the total energy of the system is expressed as a sum of atomic energies. And the atomic energy of a given atom depends on its local chemical environment. A local environment matrix and an embedding network were used to generate the molecular descriptors which can preserve the translational, rotational and permutational symmetry. Details of this method can be found in Ref 38. The DeePMD-kit package[52] was used for this step. The batch number, which denotes the number of training steps, was set to a relatively medium value (400,000) to improve the efficiency without too much loss of the accuracy. Other technical details were consistent with Ref 42.

*Exploring II.* In this step, four reactive MD simulations of *n*-dodecane at different temperatures (1500, 2000, 2500, and 3000K, respectively) were performed starting from the initial configuration of the system, driven by the DP model from the previous iteration. The Nose-Hoover thermostat[53] was used to sample the NVT ensemble. The simulation time gradually increases from 0.1 ps to 1 ns as the number of iterations increases. To enhance the transability of the model, we added an extra iteration to perform a 1 ns MD simulations of *n*-decane under 2000 K. For each atom in the system, its atomic force was also evaluated by the other three DP models. And the relative force model deviation between these four

models was calculated by

$$E_{f_i} = \frac{|D_{f_i}|}{|f_i| + l}, \tag{1}$$

where $f_i$ denotes the force on atom $i$, $D_{f_i}$ denotes the corresponding model deviation, and $l$ is a constant (1 eV/Å in this study) used to make sure that an atom having a small absolute model deviation will also have a small relative model deviation. The employment of relative model deviation instead of the absolute one is crucial for studying combustion reactions, since in such a situation the atomic forces have a very large range of values.

In each iteration, a large number of molecular clusters were extracted from the trajectory. Based on the relative model deviation of the atomic force of the central atom, these clusters were classified into three different groups named "Accurate" (the relative force deviation is less than 20%), "Candidate" (the relative force deviation is between 20% and 45%), and "Failed" (the relative force deviation is greater than 45%). To minimize data redundancy, in each iteration, at most 1000 molecular clusters were randomly selected from the "Candidate" group and added to the training dataset.

In the *Labeling* step, the PES and atomic forces of the candidate clusters were calculated using the same method detailed in the previous section. It has to be emphasized that in each iteration, the MD simulations should start from the initial configurations. When the iteration procedure stops depends on the specific problem. In this work, we stopped the iteration when MD reached 1 ns, with all the clusters extracted from the trajectory at this point correctly predicted by the DP model (with an accuracy of 99.5% and a failure rate of 0.0).

In other words, the dataset at this point should cover the chemical space we need from 1-ns simulations, during which the reactants should have been consumed and then equilibrium should have been reached. In the end of the *Concurrent learning* module, we got the final training dataset.

***Finalization***. In this section, a DP model was trained (in the *Training II* step) based on the final dataset. To guarantee the accuracy of the model, the batch number of the

training process was set to 4,000,000 while other parameters were as same as those in the *Training I* step.

All three sections were integrated into the DP-GEN software[45] and are fully automatic and user-friendly.

### The Production MD Simulation

With the final DP model, a 1-ns reactive MD simulation at 2000 K was re-performed to study the pyrolysis process of *n*-dodecane. The NVT ensemble was sampled by the Nose-Hoover thermostat[53] with a time step of 1 fs. To evaluate the transferability of the DP model, 1-ns reactive MD simulations were also performed for the pyrolysis of *n*-decane, *n*-tetradecane, and *n*-icosane, based on the same final DP model.

### Analysis

The analysis of MD trajectories, which contain thousands of species and reaction pathways, has become a major obstacle to the application of reactive MD simulation in large-scale systems. In the current study, the ReacNetGenerator method[54] developed in our previous work was used. ReacNetGenerator can automatically detect reaction events from the trajectory and construct the reaction network. Details of this method can be found in Ref 54.

## Results

### The concurrent learning process

Details of the concurrent learning process for pyrolysis of *n*-dodecane are listed in Table 1. To obtain a minimal dataset based on which the DP model can accurately simulate the pyrolysis of n-dodecane, we pre-set an concurrent learning process containing 37 iterations. In each iteration,four MD simulations of same length but different temperatures (1500, 2000,

Table 1: Details of the concurrent learning process for the pyrolysis of $n$-dodecane. The length of trajectories, the number of frames, and the percentages of accurate, candidate, and failed data points for each iteration are given in the table. In order to maximize the exploration of chemical space, four MD simulations were performed at 1500, 2000, 2500, and 3000 K, respectively, in each iteration of $n$-dodecane. The NVT ensemble was used.

| Iteration | Length (ps) | # of frames | Accurate(%) | Candidate(%) | Failed(%) |
|---|---|---|---|---|---|
| 0 | 0.2 | 84 | 95.32 | 4.25 | 0.44 |
| 1 | 0.2 | 84 | 99.66 | 0.33 | 0.00 |
| 2 | 0.2 | 84 | 99.88 | 0.12 | 0.00 |
| 3 | 0.4 | 164 | 99.70 | 0.30 | 0.00 |
| 4 | 0.8 | 324 | 99.76 | 0.24 | 0.00 |
| 5 | 1.6 | 644 | 99.39 | 0.59 | 0.03 |
| 6 | 3.2 | 1284 | 97.06 | 2.81 | 0.13 |
| 7 | 3.2 | 1284 | 99.77 | 0.23 | 0.00 |
| 8 | 3.2 | 1284 | 99.94 | 0.06 | 0.00 |
| 9 | 3.2 | 1284 | 99.85 | 0.15 | 0.00 |
| 10 | 3.2 | 1284 | 99.97 | 0.03 | 0.00 |
| 11 | 6.4 | 2564 | 99.96 | 0.04 | 0.00 |
| 12 | 12.8 | 5124 | 98.77 | 0.42 | 0.81 |
| 13 | 12.8 | 5124 | 90.54 | 8.29 | 1.17 |
| 14 | 12.8 | 5124 | 46.13 | 35.10 | 18.77 |
| 15 | 12.8 | 5124 | 81.12 | 18.00 | 0.88 |
| 16 | 25.6 | 10244 | 86.70 | 9.51 | 3.79 |
| 17 | 25.6 | 10244 | 67.45 | 27.77 | 4.77 |
| 18 | 25.6 | 10244 | 80.25 | 19.12 | 0.62 |
| 19 | 25.6 | 10244 | 72.65 | 25.45 | 1.89 |
| 20 | 25.6 | 10244 | 96.40 | 3.55 | 0.05 |
| 21 | 25.6 | 10244 | 95.88 | 4.08 | 0.04 |
| 22 | 25.6 | 10244 | 49.04 | 17,71 | 33.25 |
| 23 | 25.6 | 10244 | 95.08 | 4.81 | 0.11 |
| 24 | 51.2 | 20484 | 79.83 | 18.90 | 1.27 |
| 25 | 51.2 | 20484 | 82.97 | 10.75 | 6.27 |
| 26 | 51.2 | 20484 | 46.58 | 39.30 | 14.12 |
| 27 | 51.2 | 20484 | 91.19 | 8.52 | 0.30 |
| 28 | 102.4 | 40964 | 45.18 | 24.13 | 30.69 |
| 29 | 102.4 | 40964 | 93.95 | 5.76 | 0.29 |
| 30 | 102.4 | 40964 | 91.97 | 7.58 | 0.45 |
| 31 | 102.4 | 40964 | 96.00 | 3.93 | 0.07 |
| 32 | 204.8 | 40964 | 60.01 | 21.48 | 18.51 |
| 33 | 204.8 | 40964 | 89.61 | 10.07 | 0.32 |
| 34 | 500 | 100004 | 64.96 | 22.61 | 12.43 |
| 35 | 1000 | 200004 | 95.94 | 3.94 | 0.12 |
| 36 | 1000 ($n$-decane) | 50001 | 98.79 | 1.12 | 0.00 |

2500, and 3000K, respectively) were performed to maximize the exploration of the chemical space. It should be noted that since pyrolysis is a highly non-equilibrium process, we always start from the initial state for MD simulations during DP-GEN. Usually, when we extend the length of the simulations, the accuracy of the DP model will firstly decrease, but then increase at later iterations. For example, when we extend the trajectory from 51.2 ps to 102.4 ps, the accuracy ratio of the DP model dropped from 91.19% to 45.18%, but after another three iterations, the accuracy ratio increased back to 96.00%. Sudden accuracy decreases can be observed occasionally. For example, in the $3^{rd}$ round of the 51.2 ps, the accuracy of the model suddenly decreased to 46.59%. But this is exactly what we want, because it means that we have explored new chemical spaces and new training data is required.

To check the performance of the concurrent learning procedure, we did not iterate the trajectory of length 6.4 ps (iteration 11) and 500 ps (iteration 34). As can be seen from Table 1, the accuracy of the DP model in iteration 11 is as high as 99.96%, thus it is safe to extend the simulation length in iteration 12. However, the accuracy of the DP model in iteration 34 is only 64.96%. Fortunately, after adding 1000 data to the training set and retraining, the model performed well in iteration 35 (with accuracy 95.95%), and its accuracy reached 99.59% in iteration 36, and the failure rate was 0.00%. This may be because the simulation almost converged at 500ps.

To be more rigorous, one can keep iterating a specific length of MD simulation until the accuracy of the DP model satisfies a user-defined criterion (such as 95%) and then extends the simulation time. The whole concurrent learning procedure contains 37 iterations including 59,200,000 training batches, 114,040,000 MD simulation timesteps. A total of 37,000 structures were sent to the Labeling process. However, the QM calculation of very few structures cannot reach convergance. Thus after the last iteration, the final dataset was obtained which contains 35,496 structures.

Table 2 shows the computing cost in the concurrent learning section. The *Training* and *Exploring* steps were performed with 4 NVIDIA V100 GPUs. On a single V100 GPU card,

the training process consumes 0.03257s per batch and the MD simulation consumes 0.03056s per timestep.[55,56] The *Labeling* steps were even more efficient as DFT calculation for small molecular cluster is very fast.

Table 2: Computing cost of the *n*-dodecane workflow.

| Task | Hardware | Cost |
|---|---|---|
| Training | NVIDIA Tesla V100-PCIE-32GB | 535.60 card·hours |
| Exploring | NVIDIA Tesla V100-PCIE-32GB | 968.07 card·hours |
| Labeling | Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz | 5174.79 core·hours |

Compare with our previous study,[42] the explosion of the concurrent learning procedure in this work can greatly reduce the redundancy of the data set, saving the computational cost of Labeling and Training. The whole concurrent learning procedure in this work can be finished within 2 weeks with a couple of CPU servers and 4 NVIDIA V100 GPU cards.

## The predictive power of the DP model

To check the accuracy of the final DP model, the predicted atomic forces (based on the final training set) were compared with that calculated by DFT. As shown in Fig. 2, the overall correlation between DFT and DP results is quiet well. The mean absolute error (MAE) is 0.42 eV/Å and the root mean squared error is 0.75 eV/Å. Compared with the range of the force, the MAE and RMSE values are very small.

## Pyrolysis mechanisms of *n*-dodecane

Fig. 3 shows the evolution of the number of molecular species that contain different numbers of carbon atoms during the simulation. After the simulation started, the number of *n*-dodecanes dropped sharply, and almost all of them were pyrolyzed at about 30ps. Meanwhile, the number of small species such as C1, C2, C3 increased rapidly, with the largest number of C3. As intermediates, the number of species such as C5 to C11 was maintained within a dozen during the simulation.

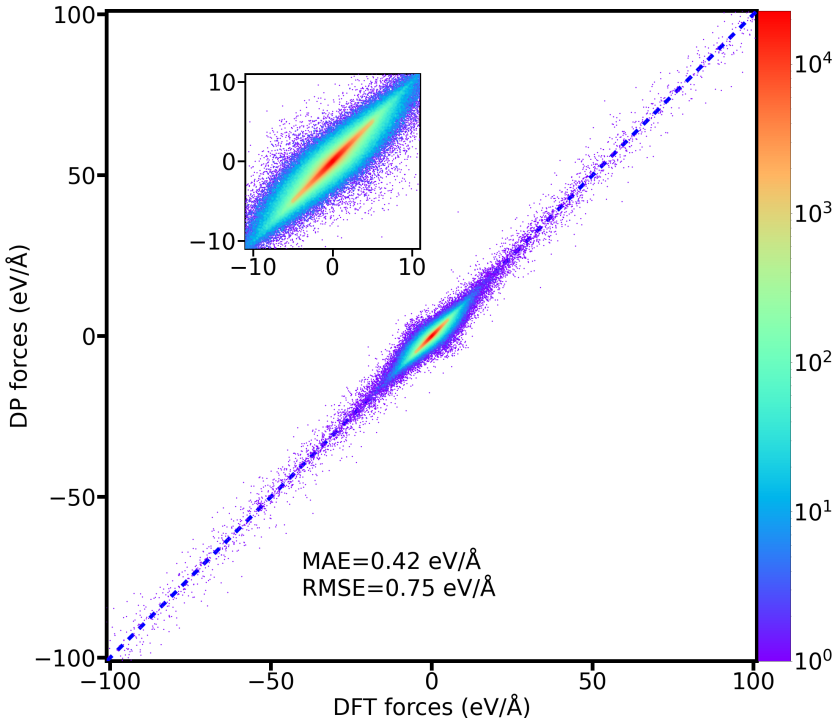


Figure 2: Correlation between atomic forces predicted by the DP model and that calculated by DFT in the final training set. The color bar indicates the density of data. Forces greater than 100 eV/Å are not shown in the figure.

Fig. 4 shows the main reaction paths of $n$-dodecane pyrolysis detected from the trajectory. The $\beta$-C-C scission of $C_{12}H_{26}$ was observed most frequently which can form alkenes and alkyls. Species contains 2 to 10 carbons were all observed with C3 species the most abundant one. The H abstract reaction was also observed which produce the $C_{12}H_{25}$ radical. These alkenes and alkyls further underwent a new round of $\beta$ scission to eventually generate butyl, propyl and ethyl radicals. These radicals are the dominant sources of ethylene and methane, which are primary pyrolysis products. The reaction network shown in Fig. 4 agrees well with the experiments[1,3] except that ethane is also a main product in our simulation. This may be because the density in our simulation is higher than that of experiments, thus the $C_2H_5$ radical can capture H radicals more easily. From the trajectory, we found that the

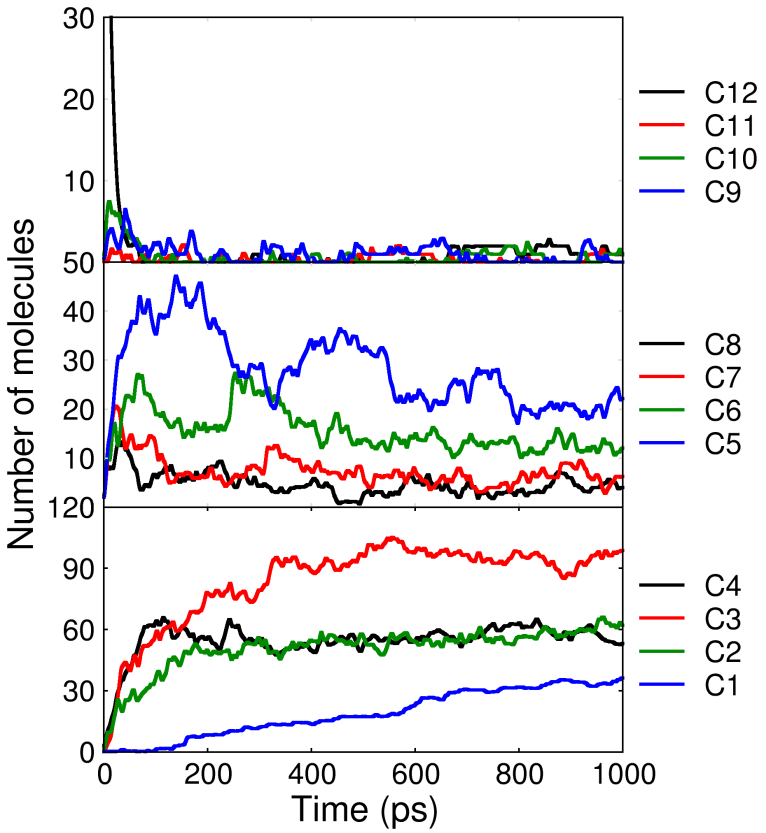


Figure 3: Time evolution of species that contains different number of carbon atoms during DPMD simulation of pyrolysis of $n$-dodecane at 2000K. There are 40 $n$-dodecane molecules in the beginning of the simulation. These curves are smoothed to make them look better and clearer.

carbon-carbon breaking reaction occured more easily than H-abstraction reactions, which is also consistent with experiments.[1,3]

## Transferability of the DP model/dataset

Transferability is one of the key requirements for the DP model, and it is determined by the coverage of the dataset on the target chemical space. In the previous sections, we have greatly reduced the redundancy of the $n$-dodecane dataset through the concurrent learning algorithm. To evaluate the transferability of the DP model (and dataset), we simulated the pyrolysis of $n$-decane, $n$-tetradecane, and $n$-icosane under 2000 K based on the four DP models obtained in the last *Training I* step of $n$-dodecane. The length of all simulations was

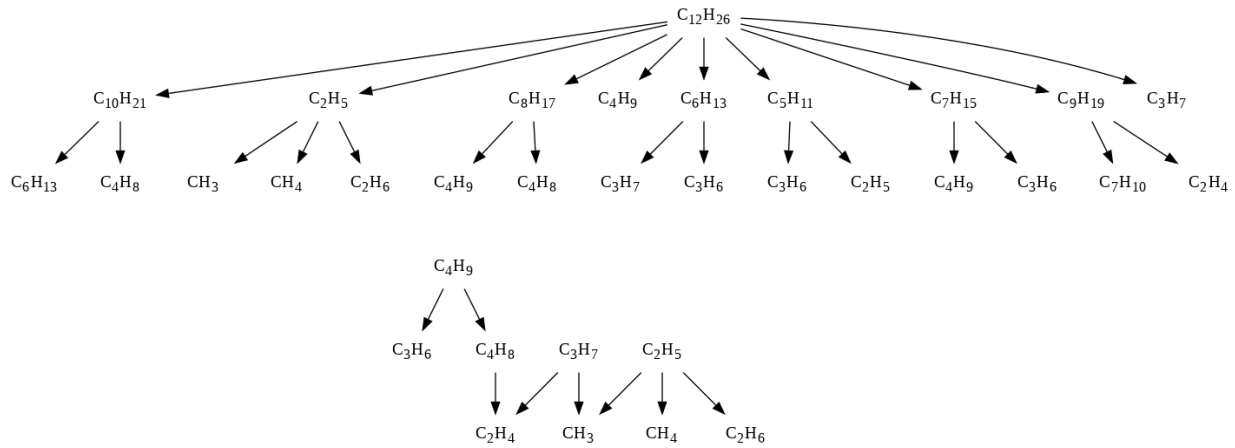


Figure 4: Main paths extracted from the MD trajectory *n*-dodecane pyrolysis.

1ns and the deviations of these four DP models were calculated.

Table 3: Concurrent learning and model deviations of DP models in the pyrolysis simulations of different systems.

| System | Accurate (%) | Candidates (%) | Failed (%) |
|---|---|---|---|
| *n*-decane | 99.74 | 0.26 | 0.00 |
| *n*-dodecane | 99.69 | 0.31 | 0.00 |
| *n*-tetradecane | 99.72 | 0.28 | 0.00 |
| *n*-isocane | 99.60 | 0.40 | 0.00 |

It can be seen from Table 3 that the DP model trained on the *n*-dodecane data set can be perfectly used for the simulation of pyrolysis of other long-chain alkanes.

Detail mechanisms of the pyrolysis of *n*-decane, *n*-tetradecane, and *n*-icosane can be found in the supplement material. These results are in agreement with experiments,[57–60] which indicate that the DP model and the dataset has strong transferability in long-chain alkanes. It is worth mentioning that long-chain alkanes are the main components of many fuels. Based on the dataset of this work, and further expanding it with DP-GEN, we can study the pyrolysis and even combustion of other complex fuels.

# Conclusion

In this work, a dataset generation approach for the pyrolysis simulation of long-chain alkanes was established. The employment of the concurrent learning algorithm allows us to maximize the exploration of the chemical space while minimize the redundancy of the dataset. This greatly reduces the cost of computational resources of the *Labeling* and *Training* process of the DP model.

Based on this method, we constructed a dataset for the pyrolysis of *n*-dodecane, which contains only 35,496 structures. (The dataset, DP model and input files of DP-GEN and DeePMD-kit can be downloaded at https://github.com/tongzhugroup/NNREAX). The reactive MD simulation with the DP model trained based on this dataset revealed the pyrolysis mechanism of *n*-dodecane in detail, and the simulation results are in good agreement with the experimental measurements. In addition, this dataset has excellent transferability for different long-chain alkanes, and can be reliably used for the pyrolysis simulation of *n*-decane, *n*-tetradecane, and *n*-icosane.

It should be noted that in exploring steps, both the density of the system and the temperature in MD simulation were increased to improve the sampling efficiency. Although these are widely used strategies in reactive MD simulations, this can make it difficult to cover the chemical space of reactions at low temperatures and pressures. To solve this issue, in our future work we will try to employ the molecular generative model and/or enhanced sampling MD algorithms in the exploring step. One might concern the accuracy of the QM level used in the labeling steps. The MN15 functional was chosen because it was specifically designed to have broad accuracy for multi-reference and single-reference systems.[43] In fact, the main advantage of the current method is to ensure that the reference dataset covers the target chemical space while it has as small redundancy as possible. The smaller dataset makes it possible to perform calculations with a higher level of QM method. However, ultra-high computational demands make it still difficult to apply *ab initio* methods with ideal accuracies, such as MRCI, to label the reference dataset of such complex reaction systems. To achieve

this, further method development is still needed. In addition to pyrolysis, this method can also be readily used in the reactive MD simulation of combustion. In addition, it is worth to point that while pyrolysis and combustion are usually thought to be dominated by free radical reactions, under certain conditions, one may also want to investigate the influence of the excited electronic state of the species on the reaction. Although MD simulations involving excited states are highly non-trivial, some recent work has made promising progress in relatively small systems.[61–67] Through further development, these methods are expected to be used in more complex pyrolysis or combustion systems.

The algorithms developed in this work has been integrated into the DP-GEN (https://github.com/deepmodeling/dpgen) software platform, which is user-friendly and efficient. This method can be used not only for the simulation of reaction mechanisms of fuel pyrolysis or combustion, but also for constructing general datasets for similar target systems. Related research is currently being carried out in our laboratory.

In Eq. (12) averaging over networks is missing, i.e. $\langle\cdot\rangle_{\mathrm{net}_i}$. How is $\delta_{\mathrm{Force}}$ defined?

## Acknowledgement

The authors would like to thank Mr. Jin Xiao and Miss Liqun Cao for their help in building the initial structures of the simulation system. This work was supported by the National Natural Science Foundation of China (Grants No. 91641116). J.Z. was partially supported by the National Innovation and Entrepreneurship Training Program for Undergraduate, China (201910269080), and the Excellence Fellowship for doctoral study provided by Rutgers University. The work of L.Z. was supported in part by the Center Chemistry in Solution and at Interfaces (CSI) at Princeton University funded by the DOE Award DE-SC0019394. The work of H.W. is supported by the National Science Foundation of China under Grant No. 11871110, the National Key Research and Development Program of China under Grants No. 2016YFB0201200 and No. 2016YFB0201203, and Beijing Academy of Ar-

tificial Intelligence (BAAI). We also acknowledge the support of computational resources from the ECNU Multifunctional Platform for Innovation (No. 001).

## Supporting Information Available

Supporting Information Available: Time evolution of species that contains different number of carbon atoms during DPMD simulation of pyrolysis of n-decane at 2000K (Figure S1). Time evolution of species that contains different number of carbon atoms during DPMD simulation of pyrolysis of n-tetradecane at 2000K (Figure S2). Time evolution of species that contains different number of carbon atoms during DPMD simulation of pyrolysis of n-isosane at 2000K (Figure S3). Main reaction paths extracted from the MD trajectory of the n-decane pyrolysis (Figure S4). Main reaction paths extracted from the MD trajectory of the n-tetradecane pyrolysis (Figure S5). Main reaction paths extracted from the MD trajectory of the n-isosane pyrolysis (Figure S6).

## References

(1) Banerjee, S.; Tangko, R.; Sheen, D. A.; Wang, H.; Bowman, C. T. An experimental and kinetic modeling study of n-dodecane pyrolysis and oxidation. Combustion and Flame **2016**, 163, 12–30.

(2) Malewicki, T.; Brezinsky, K. Experimental and modeling study on the pyrolysis and oxidation of n-decane and n-dodecane. Proceedings of the Combustion Institute **2013**, 34, 361–368.

(3) Zeng, M.; Yuan, W.; Li, W.; Zhang, Y.; Wang, Y. Investigation of n-dodecane pyrolysis at various pressures and the development of a comprehensive combustion model. Energy **2018**, 155, 152–161.

(4) Liu, G.; Han, Y.; Wang, L.; Zhang, X.; Mi, Z. Supercritical thermal cracking of n-dodecane in presence of several initiative additives: Products distribution and kinetics. Energy & Fuels **2008**, 22, 3960–3969.

(5) Meuwly, M. Reactive molecular dynamics: From small molecules to proteins. Wiley Interdisciplinary Reviews: Computational Molecular Science **2019**, 9, e1386.

(6) Andreoni, W.; Curioni, A. New advances in chemistry and materials science with CPMD and parallel computing. Parallel Computing **2000**, 26, 819–842.

(7) Hutter, J.; Iannuzzi, M.; Schiffmann, F.; VandeVondele, J. cp2k: atomistic simulations of condensed matter systems. Wiley Interdisciplinary Reviews: Computational Molecular Science **2014**, 4, 15–25.

(8) Kresse, G.; Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. Computational materials science **1996**, 6, 15–50.

(9) Kresse, G.; Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. Physical review B **1996**, 54, 11169.

(10) Wang, Q.-D.; Wang, J.-B.; Li, J.-Q.; Tan, N.-X.; Li, X.-Y. Reactive molecular dynamics simulation and chemical kinetic modeling of pyrolysis and combustion of n-dodecane. Combustion and Flame **2011**, 158, 217–226.

(11) Truhlar, D. G.; Garrett, B. C.; Klippenstein, S. J. Current Status of Transition-State Theory. The Journal of Physical Chemistry **1996**, 100, 12771–12800.

(12) Van Duin, A. C.; Dasgupta, S.; Lorant, F.; Goddard, W. A. ReaxFF: a reactive force field for hydrocarbons. The Journal of Physical Chemistry A **2001**, 105, 9396–9409.

(13) Senftle, T. P.; Hong, S.; Islam, M. M.; Kylasa, S. B.; Zheng, Y.; Shin, Y. K.; Junkermeier, C.; Engel-Herbert, R.; Janik, M. J.; Aktulga, H. M., et al. The ReaxFF reac-

tive force-field: development, applications and future directions. npj Computational Materials **2016**, 2, 1–14.

(14) van Duin, A. C.; Zeiri, Y.; Dubnikova, F.; Kosloff, R.; Goddard, W. A. Atomistic-scale simulations of the initial chemical events in the thermal initiation of triacetonetriperoxide. Journal of the American Chemical Society **2005**, 127, 11053–11062.

(15) Strachan, A.; Kober, E. M.; van Duin, A. C.; Oxgaard, J.; Goddard III, W. A. Thermal decomposition of RDX from reactive molecular dynamics. The Journal of chemical physics **2005**, 122, 054502.

(16) Strachan, A.; van Duin, A. C.; Chakraborty, D.; Dasgupta, S.; Goddard III, W. A. Shock waves in high-energy materials: the initial chemical events in nitramine RDX. Physical Review Letters **2003**, 91, 098301.

(17) Wang, E.; Ding, J.; Qu, Z.; Han, K. Development of a Reactive Force Field for Hydrocarbons and Application to Iso-octane Thermal Decomposition. Energy & fuels **2018**, 32, 901–907.

(18) Ashraf, C.; van Duin, A. C. Extension of the ReaxFF combustion force field toward syngas combustion and initial oxidation kinetics. The Journal of Physical Chemistry A **2017**, 121, 1051–1068.

(19) Bertels, L. W.; Newcomb, L. B.; Alaghemandi, M.; Green, J. R.; Head-Gordon, M. Benchmarking the Performance of the ReaxFF Reactive Force Field on Hydrogen Combustion Systems. The Journal of Physical Chemistry A **2020**, 124, 5631–5645.

(20) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. Physical review letters **2007**, 98, 146401.

(21) Chen, J.; Xu, X.; Xu, X.; Zhang, D. H. A global potential energy surface for the H2+

OH↔ H2O+ H reaction using neural networks. The Journal of Chemical Physics **2013**, 138, 154301.

(22) Jiang, B.; Guo, H. Permutation invariant polynomial neural network approach to fitting potential energy surfaces. The Journal of chemical physics **2013**, 139, 054112.

(23) Shao, K.; Chen, J.; Zhao, Z.; Zhang, D. H. Communication: Fitting potential energy surfaces with fundamental invariant neural network. 2016.

(24) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. Chemical science **2017**, 8, 3192–3203.

(25) Unke, O. T.; Meuwly, M. A reactive, scalable, and transferable model for molecular energies from a neural network approach based on local information. The Journal of chemical physics **2018**, 148, 241708.

(26) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet–A deep learning architecture for molecules and materials. The Journal of Chemical Physics **2018**, 148, 241722.

(27) Unke, O. T.; Meuwly, M. PhysNet: a neural network for predicting energies, forces, dipole moments, and partial charges. Journal of chemical theory and computation **2019**, 15, 3678–3693.

(28) Takamoto, S.; Izumi, S.; Li, J. TeaNet: universal neural network interatomic potential inspired by iterative electronic relaxations. arXiv preprint arXiv:1912.01398 **2019**,

(29) Xie, T.; Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Physical review letters **2018**, 120, 145301.

(30) Zhang, Y.; Hu, C.; Jiang, B. Embedded Atom Neural Network Potentials: Efficient and Accurate Machine Learning with a Physically Inspired Representation. The Journal of Physical Chemistry Letters **2019**, 10, 4962–4967.

(31) Bartók, A. P.; Csányi, G. Gaussian approximation potentials: A brief tutorial introduction. International Journal of Quantum Chemistry **2015**, 115, 1051–1057.

(32) Shao, Y.; Hellström, M.; Mitev, P. D.; Knijff, L.; Zhang, C. PiNN: A Python Library for Building Atomic Neural Networks of Molecules and Materials. Journal of Chemical Information and Modeling **2020**,

(33) Khorshidi, A.; Peterson, A. A. Amp: A modular approach to machine learning in atomistic simulations. Computer Physics Communications **2016**, 207, 310–324.

(34) Hong, Y.; Yin, Z.; Guan, Y.; Zhang, Z.; Fu, B.; Zhang, D. Exclusive Neural Network Representation of the Quasi-Diabatic Hamiltonians Including Conical Intersections. The Journal of Physical Chemistry Letters **2020**,

(35) Zhang, Y.; Hu, C.; Jiang, B. Bridging the Efficiency Gap Between Machine Learned Potentials with ab initio Accuracy and Classical Force Fields. arXiv preprint arXiv:2006.16482 **2020**,

(36) Zhang, Y.; Hu, C.; Jiang, B. Embedded atom neural network potentials: Efficient and accurate machine learning with a physically inspired representation. The Journal of Physical Chemistry Letters **2019**, 10, 4962–4967.

(37) Zhang, L.; Han, J.; Wang, H.; Car, R.; E, W. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. Physical review letters **2018**, 120, 143001.

(38) Zhang, L.; Han, J.; Wang, H.; Saidi, W.; Car, R.; E, W. End-to-end symmetry pre-

serving inter-atomic potential energy model for finite and extended systems. Advances in Neural Information Processing Systems. 2018; pp 4436–4446.

(39) Kang, P.-L.; Shang, C.; Liu, Z.-P. Glucose to 5-hydroxymethylfurfural: Origin of site-selectivity resolved by machine learning based reaction sampling. Journal of the American Chemical Society **2019**, 141, 20525–20536.

(40) Ma, S.; Shang, C.; Liu, Z.-P. Heterogeneous catalysis from structure to activity via SSW-NN method. The Journal of Chemical Physics **2019**, 151, 050901.

(41) Ma, S.; Huang, S.-D.; Liu, Z.-P. Dynamic coordination of cations and catalytic selectivity on zinc–chromium oxide alloys during syngas conversion. Nature Catalysis **2019**, 2, 671–677.

(42) Zeng, J.; Cao, L.; Xu, M.; Zhu, T.; Zhang, J. Z. Complex reaction processes in combustion unraveled by neural network-based molecular dynamics simulation. Nature Communications **2020**, 11, 1–9.

(43) Haoyu, S. Y.; He, X.; Li, S. L.; Truhlar, D. G. MN15: A Kohn–Sham global-hybrid exchange–correlation density functional with broad accuracy for multi-reference and single-reference systems and noncovalent interactions. Chemical science **2016**, 7, 5032–5051.

(44) Zhang, L.; Lin, D.-Y.; Wang, H.; Car, R.; E, W. Active learning of uniformly accurate interatomic potentials for materials simulation. Physical Review Materials **2019**, 3, 023804.

(45) Zhang, Y.; Wang, H.; Chen, W.; Zeng, J.; Zhang, L.; Wang, H.; E, W. DP-GEN: A concurrent learning platform for the generation of reliable deep learning based potential energy models. Computer Physics Communications **2020**, 253, 107206.

(46) Calegari Andrade, M. F.; Ko, H.-Y.; Zhang, L.; Car, R.; Selloni, A. Free energy of proton transfer at the water–TiO2 interface from ab initio deep potential molecular dynamics. Chem. Sci. **2020**, 11, 2335–2341.

(47) Zhang, L.; Chen, M.; Wu, X.; Wang, H.; E, W.; Car, R. Deep neural network for the dielectric response of insulators. Phys. Rev. B **2020**, 102, 041121.

(48) BIOVIA, D. S. BIOVIA Material Studio 2017. 2017; San Diego: Dassault Systèmes.

(49) Aktulga, H. M.; Fogarty, J. C.; Pandit, S. A.; Grama, A. Y. Parallel reactive molecular dynamics: Numerical methods and algorithmic techniques. Parallel Computing **2012**, 38, 245–259.

(50) Chenoweth, K.; Van Duin, A. C.; Goddard, W. A. ReaxFF reactive force field for molecular dynamics simulations of hydrocarbon oxidation. The Journal of Physical Chemistry A **2008**, 112, 1040–1053.

(51) Frisch, M. J. et al. Gaussian˜16 Revision A.03. 2016; Gaussian Inc. Wallingford CT.

(52) Wang, H.; Zhang, L.; Han, J.; E, W. DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. Computer Physics Communications **2018**, 228, 178–184.

(53) Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. Physical review A **1985**, 31, 1695.

(54) Zeng, J.; Cao, L.; Chin, C.-H.; Ren, H.; Zhang, J. Z. H.; Zhu, T. ReacNetGenerator: an automatic reaction network generator for reactive molecular dynamics simulations. Phys. Chem. Chem. Phys. **2020**, 22, 683–691.

(55) Lu, D.; Wang, H.; Chen, M.; Lin, L.; Car, R.; E, W.; Jia, W.; Zhang, L. 86 PFLOPS Deep Potential Molecular Dynamics simulation of 100 million atoms with ab initio accuracy. Computer Physics Communications **2021**, 259, 107624.

(56) Jia, W.; Wang, H.; Chen, M.; Lu, D.; Lin, L.; Car, R.; E, W.; Zhang, L. Pushing the Limit of Molecular Dynamics with Ab Initio Accuracy to 100 Million Atoms with Machine Learning. Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. 2020.

(57) Billaud, F.; Freund, E. N-Decane pyrolysis at high-temperature in a flow reactor. Industrial & engineering chemistry fundamentals **1986**, 25, 433–443.

(58) Jia, Z.; Wang, Z.; Cheng, Z.; Zhou, W. Experimental and modeling study on pyrolysis of n-decane initiated by nitromethane. Combustion and Flame **2016**, 165, 246–258.

(59) Jia, Z.; Huang, H.; Zhou, W.; Qi, F.; Zeng, M. Experimental and modeling investigation of n-decane pyrolysis at supercritical pressures. Energy & fuels **2014**, 28, 6019–6028.

(60) Song, C.; Lai, W. C.; Schobert, H. H. Condensed-phase pyrolysis of n-tetradecane at elevated pressures for long duration. Product distribution and reaction mechanisms. Industrial & engineering chemistry research **1994**, 33, 534–547.

(61) Chen, W.-K.; Liu, X.-Y.; Fang, W.-H.; Dral, P. O.; Cui, G. Deep learning for nonadiabatic excited-state dynamics. The journal of physical chemistry letters **2018**, 9, 6702–6708.

(62) Hu, D.; Xie, Y.; Li, X.; Li, L.; Lan, Z. Inclusion of machine learning kernel ridge regression potential energy surfaces in on-the-fly nonadiabatic molecular dynamics simulation. The journal of physical chemistry letters **2018**, 9, 2725–2732.

(63) Westermayr, J.; Gastegger, M.; Menger, M. F.; Mai, S.; González, L.; Marquetand, P. Machine learning enables long time scale molecular photodynamics simulations. Chemical science **2019**, 10, 8100–8107.

(64) Westermayr, J.; Faber, F. A.; Christensen, A. S.; von Lilienfeld, O. A.; Marquetand, P. Neural networks and kernel ridge regression for excited states dynamics of CH2NH:

From single-state to multi-state representations and multi-property machine learning models. Machine Learning: Science and Technology **2020**, 1, 025009.

(65) Koner, D.; Bemish, R. J.; Meuwly, M. The C (3P)+ NO (X2Π)→ O (3P)+ CN (X2Σ+), N (2D)/N (4S)+ CO (X1Σ+) reaction: Rates, branching ratios, and final states from 15 K to 20 000 K. The Journal of Chemical Physics **2018**, 149, 094305.

(66) Koner, D.; Unke, O. T.; Boe, K.; Bemish, R. J.; Meuwly, M. Exhaustive state-to-state cross sections for reactive molecular collisions from importance sampling simulation and a neural network representation. The Journal of chemical physics **2019**, 150, 211101.

(67) Pezzella, M.; Koner, D.; Meuwly, M. Formation and Stabilization of Ground and Excited-State Singlet O2 upon Recombination of 3P Oxygen on Amorphous Solid Water. The Journal of Physical Chemistry Letters **2020**, 11, 2171–2176.

# Graphical TOC Entry