nature computational science

Article

https://doi.org/10.1038/s43588-023-00563-7

Accurate transition state generation with an object-aware equivariant elementary reaction diffusion model

In the format provided by the authors and unedited

Table of Contents

List of Tables

List of Figures

Abbreviation

The following is the list of abbreviation utilized in the main paper.

1. OA-ReactDiff: Object-aware SE(3) GNN for generating sets of 3D molecules in elementary reactions under the diffusion model
2. RMSD: Root mean square deviation.
3. SE(3): Special Euclidean group in 3D space.
4. TS: Transition state.
5. MAE: Mean absolute error.

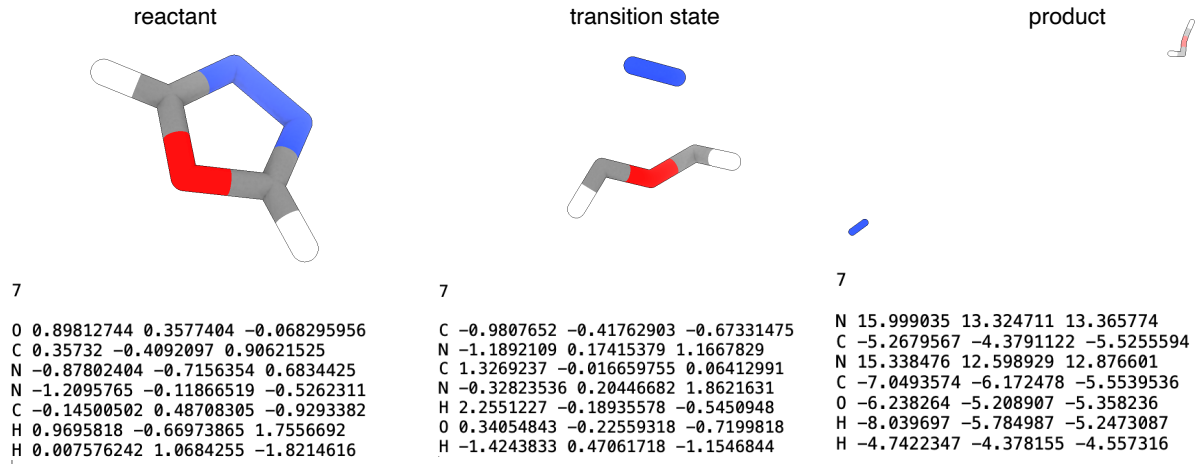
1 Physical symmetries and constraints in an elementary reaction.

An elementary reaction that consists of n fragments as reactant and m fragments as product can be described as $\{\mathrm{R}^{(1)}, ..., \mathrm{R}^{(n)}, \mathrm{TS}, \mathrm{P}^{(1)}, ..., \mathrm{P}^{(\mathrm{m})}\}$. This reaction requires the following symmetries:

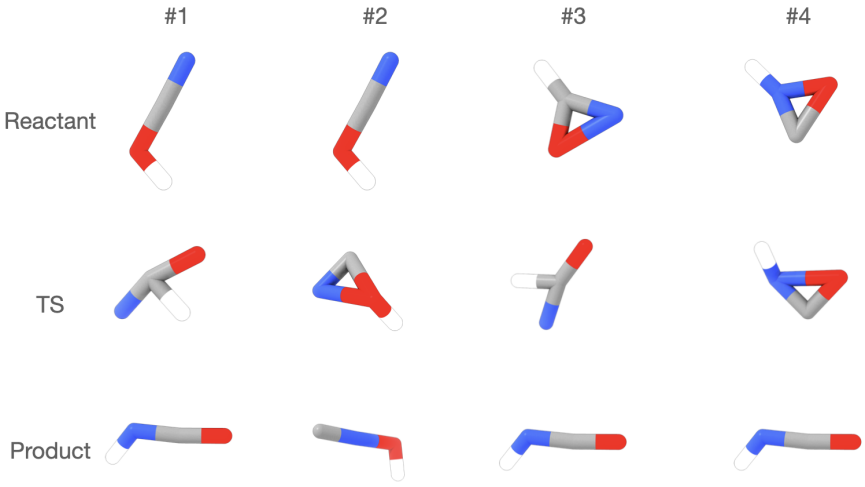
1. *Permutation symmetry among atoms in a fragment.* For any fragment in $\mathrm{R}^{(i)}, \mathrm{TS}, \mathrm{P}^{(\mathrm{j})}$, change of atom ordering preserves the reaction.
2. *Permutation symmetry among fragments in reactant and product.* The change of ordering in $\{\mathrm{R}^{(1)}, ..., \mathrm{R}^{(n)}\}$ and $\{\mathrm{P}^{(1)}, ..., \mathrm{P}^{(m)}\}$ preserve the reaction.
3. *Rotation and translation symmetry for each fragment.* Rotation and translation operations on any fragment (i.e., $\mathrm{R}^{(i)}, \mathrm{TS}, \mathrm{P}^{(\mathrm{j})}$) preserve the reaction.

Supplementary Table 1. Ablation studies comparing OA-ReactDiff performance on RMSD evaluation with different models. Vanilla SE(3) LEFTNet[2] is shown to demonstrate the importance of preserve object-wise symmetry in elementary reaction. EGNN[1] is shown to reflect the importance of vanilla SE(3) model.

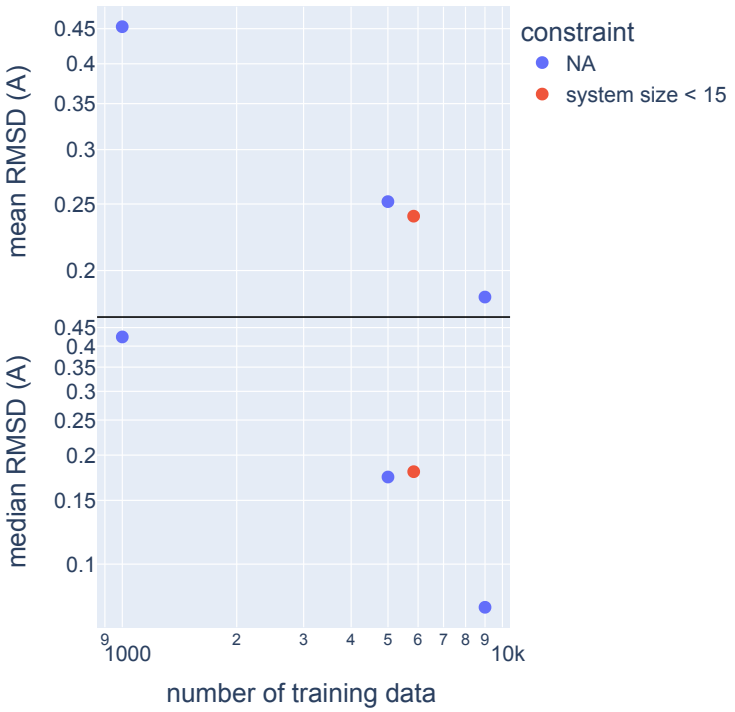
Approach	RMSD (Å)	
	mean	median
Object-aware SE(3) LEFTNet	0.183	0.076
Vanilla SE(3) LEFTNet	0.638	0.620
Object-aware SE(3) EGNN	0.372	0.360



Supplementary Figure 1. Mutil-molecular elementary reactions sampled from OA-ReactDiff by specifying reactant and product. Here, the atom mapping and fragment alignment are randomized to demonstrate the capability of OA-ReactDiff not relying on these factors. The generated TS structure only has a RMSD of 0.03 Å compared to the DFT (ωB97x/6-31G(d)) optimized true TS. Atoms are colored as follows: gray for C, blue for N, red for O, and white for H.



Supplementary Figure 2. Elementary reactions sampled from OA-ReactDiff by only specifying the chemical composition of interest. Here, we consider a system that contains one C, H, N, and O is chosen. This chemical composition is absent in the Transition1x dataset, and thus is completely new to the trained OA-ReactDiff model. Atoms are colored as follows: gray for C, blue for N, red for O, and white for H.



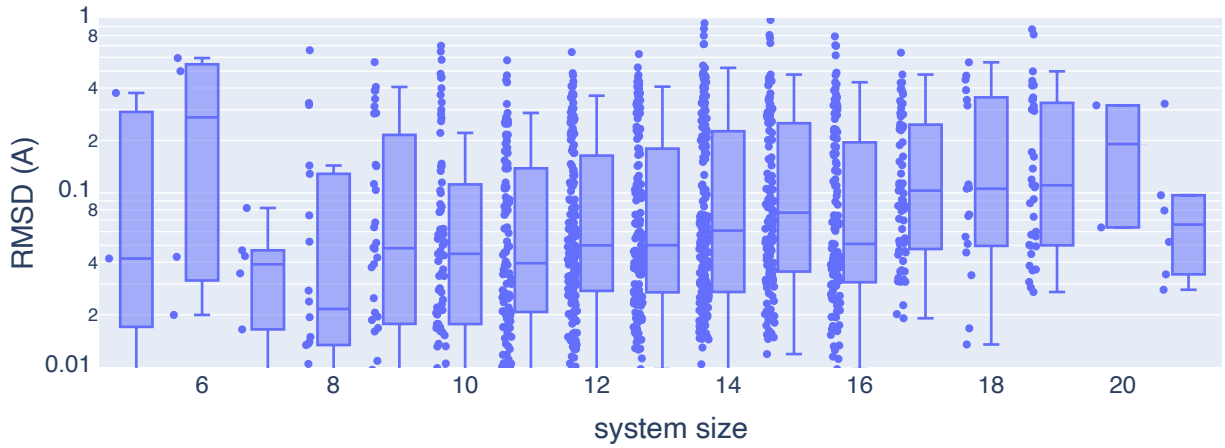
Supplementary Figure 3. RMSD vs. the number of training data. Mean (top) and median (bottom) RMSD for OA-ReactDiff models trained on different number of training data that are either randomly sampled (blue) or sampled under the constraint that only systems within 15 atoms in size (red) from the original 9000 training reactions.

Supplementary Table 2. Resource utilization of OA-ReactDiff inference. Runtime and GPU memory consumption with different batch sizes.

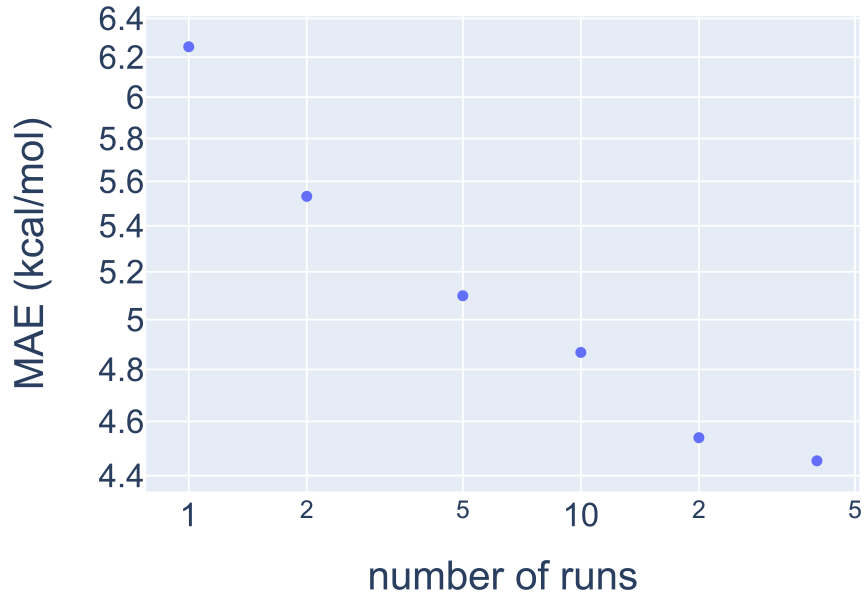
batch size	1	2	4	8	16	32	64	128	256
walltime (sec)	17.1	18.9	28.0	44.7	87.1	171.8	322.9	582.1	1106.1
runtime per sample (sec)	17.1	9.5	7.0	5.6	5.4	5.4	5.1	4.6	4.3
GPU memory (GB)	1.1	1.3	1.4	1.6	2.2	3.6	5.7	11.2	21.8

Supplementary Table 3. Performance of OA-ReactDiff. RMSD at different number of training data and constraints for OA-ReactDiff.

number of sample	1000	5000	5734	9000
mean RMSD (Å)	0.453	0.252	0.240	0.183
median RMSD (Å)	0.424	0.174	0.180	0.076
constraints in training data	–	–	system size $<$ 15 atoms	–



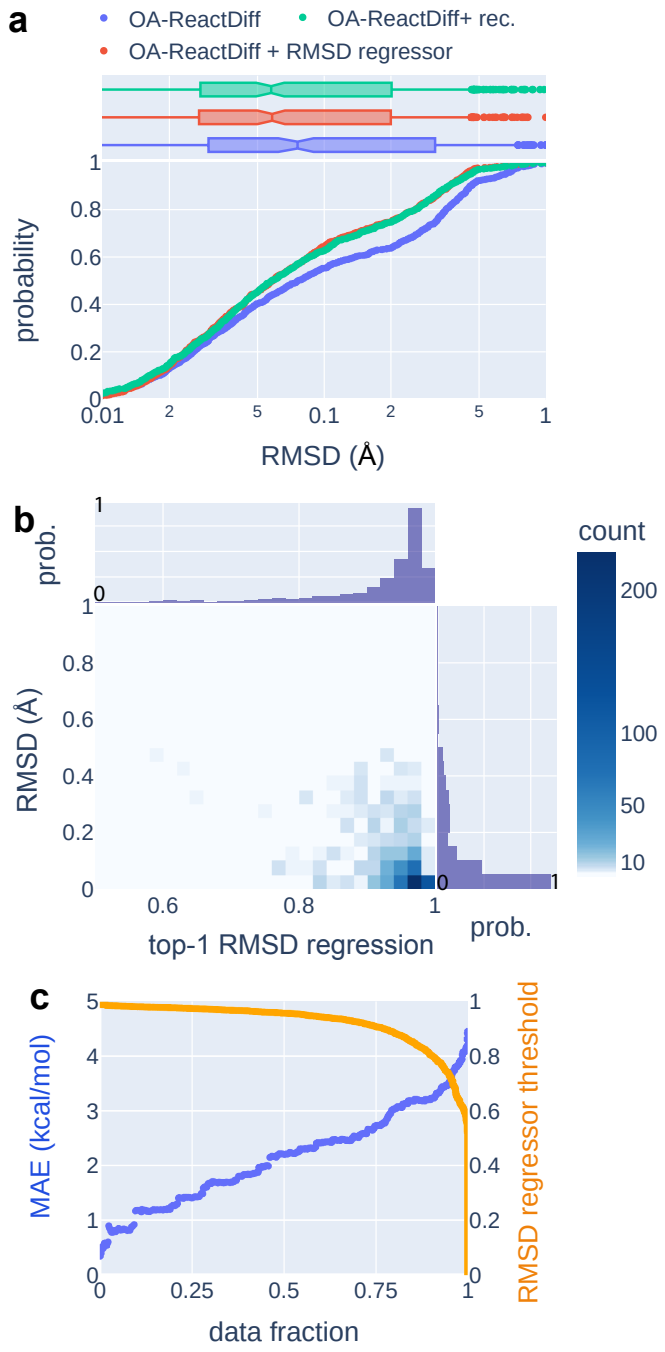
Supplementary Figure 4. Distributions of RMSD binned by system size. The RMSD is computed between top-1 confidence TS structure generated by OA-ReactDiff and the true TS structure for the 1073 test reactions. A standard box plot with median as the horizontal bar, quartile 1 (Q1) to quartile 3 (Q3) as box edge, and the whiskers correspond to the edges +/- 1.5 times the interquartile range (or, IQR= Q3-Q1) is shown with all RMSDs at that system size displayed on its left hand side.



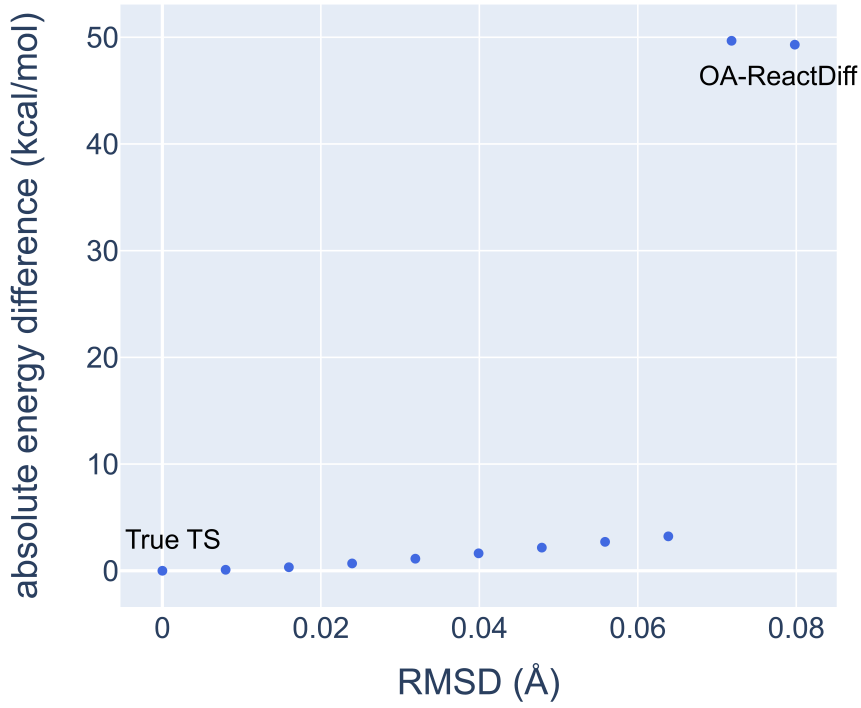
Supplementary Figure 5. Mean absolute energy difference vs. number of runs for OA-ReactDiff sampling.. A log-log axis is used to shown the near power law dependence. The results are shown on 1073 test elementary reactions.

Supplementary Table 4. Barrier height error of OA-ReactDiff evaluated by different methods. All energies are reported in the unit of kcal/mol.

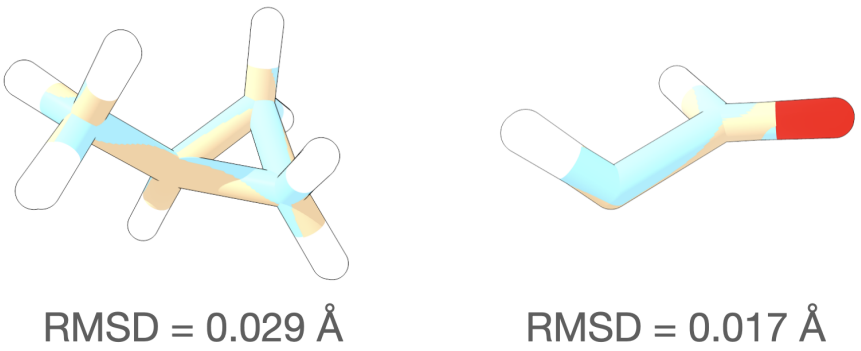
Approach	ωB97x		B3LYP		PBE		data fraction
	mean	median	mean	median	mean	median	
OA-ReactDiff + rec.	4.4	1.6	4.3	1.6	4.9	1.7	1.0
OA-ReactDiff + rec. (p>0.5)	3.1	1.4	3.0	1.4	3.7	1.6	0.86



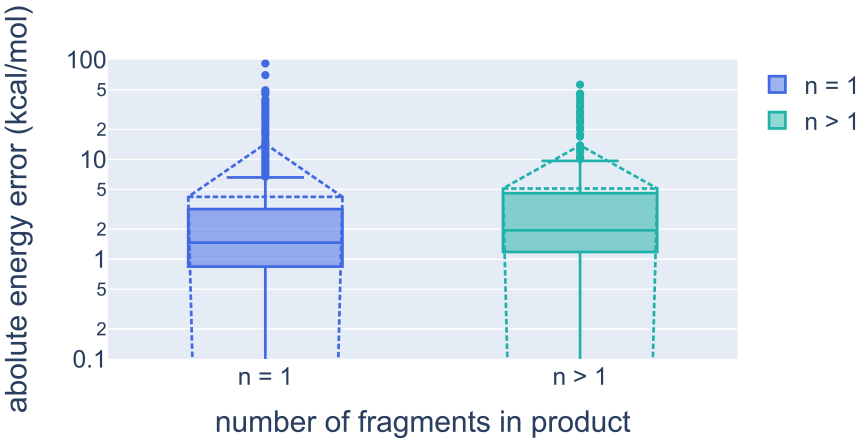
Supplementary Figure 6. Performance of using a RMSD regressor as the confidence model. a. Cumulative probability for RMSD between the true TS structures and OA-ReactDiff samples on 1073 set-aside test reactions. The OA-ReactDiff samples are evaluated under one-shot generation (blue), the top-1 confidence sample via classifier-based recommender (green), and the top-1 confidence sample via RMSD regressor out of 40 generated samples for each reaction (red). A log scale of the RMSD is presented for better visibility of the low-RMSD region. b. 2D density map for the RMSD vs. top-1 RMSD regressor confidence for OA-ReactDiff generated samples. A log-scale color gradient is applied to the color bar to reveal low-density areas, which would otherwise be difficult to distinguish. A histogram of relative probability (prob.) is shown for each axis. c. MAE of $|\Delta E_{\mathrm{TS}}|$ (blue, left y-axis) and the corresponding confidence threshold (orange, right y-axis) as a function of the fraction of data considered in the 1073 TS structures selected via the RMSD regressor.



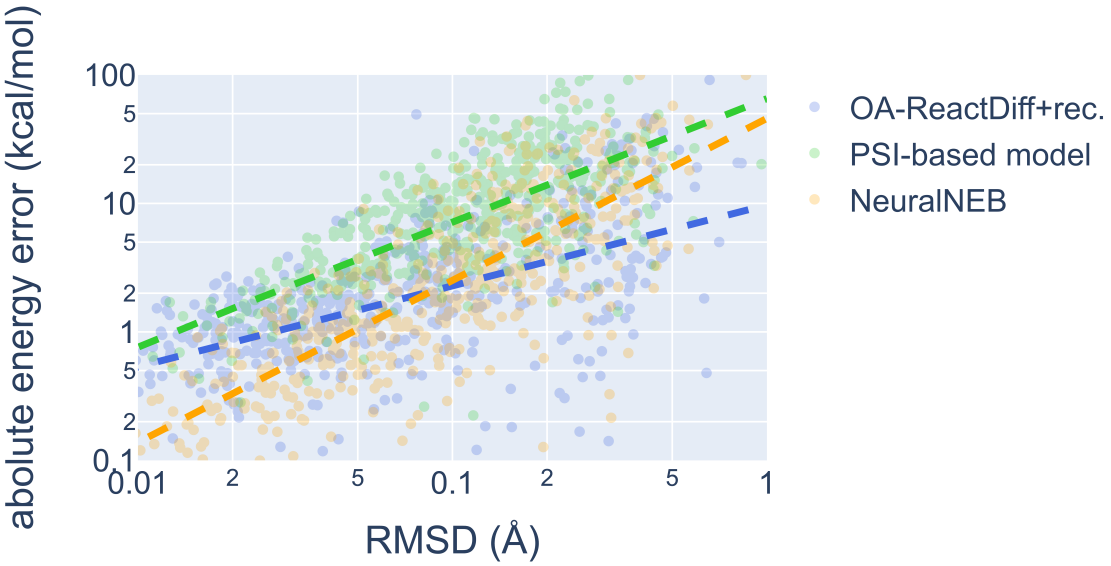
Supplementary Figure 7. Absolute energy difference vs. RMSD for the ten interpolated structure between true (left) and OA-ReactDiff TS (right) for $C_4H_6O_2$. The abrupt change in energy difference indicates a change in converged electronic state for self-consistent field calculation.



Supplementary Figure 8. Overlapping OA-ReactDiff and true TS structures of $C_6H_{10}O$ separated as two fragments and their corresponding RMSD. Atoms are colored as follows: C in the true TS structure are in tan and those in the OA-ReactDiff sample are in skyblue; O for red, and H for white.



Supplementary Figure 9. **Box plot for absolute energy difference of OA-ReactDiff + rec. TS structures grouped by single (i.e., n=1) and multi (i.e., n > 1) product cases.**. The median is shown as the horizontal bar, quartile 1 (Q1) to quartile 3 (Q3) as box edge, and the whiskers corresponding to the edges +/- 1.5 times the interquartile range (or, IQR= Q3-Q1). The dashed lines are shown for the mean and standard deviation. The results are shown on 1073 test elementary reactions.



Supplementary Figure 10. **Absolute energy difference vs. RMSD.** The corresponding linear fit in a log-log plot for OA-ReactDiff + rec. (blue), PSI-based model[3] (green), and NeuralNEB[4] (orange)
.

References

[1] Satorras, V. G., Hoogeboom, E. & Welling, M. E(n) equivariant graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, 9323–9332 (2021).

[2] Du, W. *et al.* A new perspective on building efficient and expressive 3D equivariant graph neural networks. *arXiv:2304.04757* (2023).

[3] Choi, S. Prediction of transition state structures of gas-phase chemical reactions via machine learning. *Nature Communications* **14**, 1168, DOI: 10.1038/s41467-023-36823-3 (2023).

[4] Schreiner, M., Bhowmik, A., Vegge, T., Jørgensen, P. B. & Winther, O. Neuralneb—neural networks can find reaction paths fast. *Machine Learning: Science and Technology* **3**, 045022, DOI: 10.1088/2632-2153/aca23e (2022).