

Supplementary information I

Figures and discussions for explanatory purposes

Transition State Structure Detection with Machine Learning

Yitao Si^{1,2,3,¶}, Yiding Ma^{2,4,¶}, Tao Yu^{2,3}, Yifan Wu⁵, Yingzhe Liu^{2,6}, Weipeng Lai², Zhixiang Zhang², Jinwen Shi¹, Liejin Guo¹, Prezhdo V. Oleg^{5*}, Maochang Liu^{1,7*}

¹*International Research Center for Renewable Energy, State Key Laboratory of Multiphase Flow, Xi'an Jiaotong University, Xi'an 710049, P. R. China*

²*State Key Laboratory of Fluorine & Nitrogen Chemicals, Xi'an Modern Chemistry Research Institute, Xi'an 710065, P. R. China*

³*School of Chemistry and Chemical Engineering, Southeast University, Nanjing 211189, P. R. China*

⁴*Electric Power Research Institute of Guangdong Power Grid Co., Ltd, Guangzhou 510080, P. R. China*

⁵*Departments of Chemistry, and Physics and Astronomy, University of Southern California, Los Angeles, CA 90089, USA*

⁶*Xi'an Key Laboratory of Liquid Crystal and Organic Photovoltaic Materials, Xi'an Modern Chemistry Research Institute, Xi'an 710065, P. R. China*

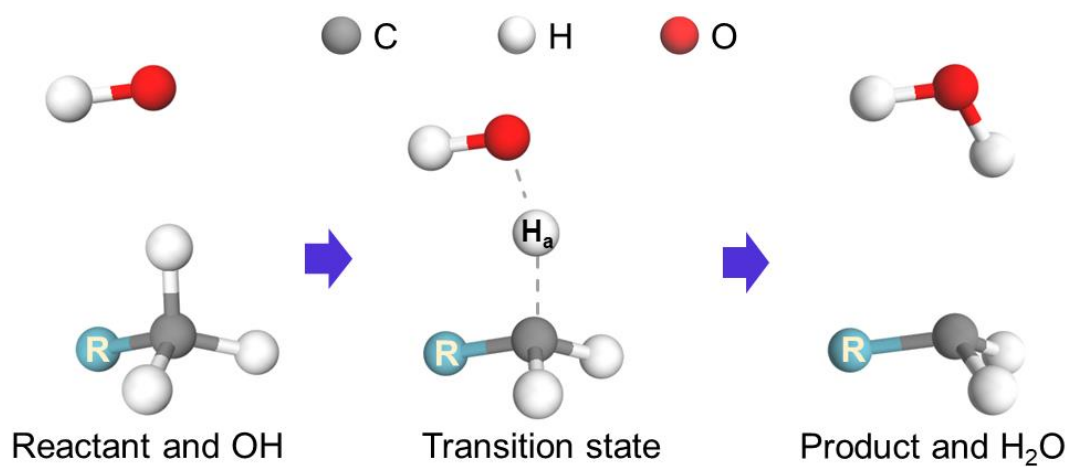
⁷*Suzhou Academy of Xi'an Jiaotong University, Suzhou, Jiangsu 215123, P. R. China*

[¶]The two authors contribute equally.

*Correspondence

E-mail: prezhdo@usc.edu (Oleg V. Prezhdo) and maochangliu@mail.xjtu.edu.cn (Maochang Liu)

Section1: Supplementary Figures



Scheme S1| Transition structures of the hydrogen abstraction reactions discussed in this work.

Spheres in grey, white, red and green represent carbon, hydrogen, oxygen and omitted part of HFCs and HFEs, respectively.

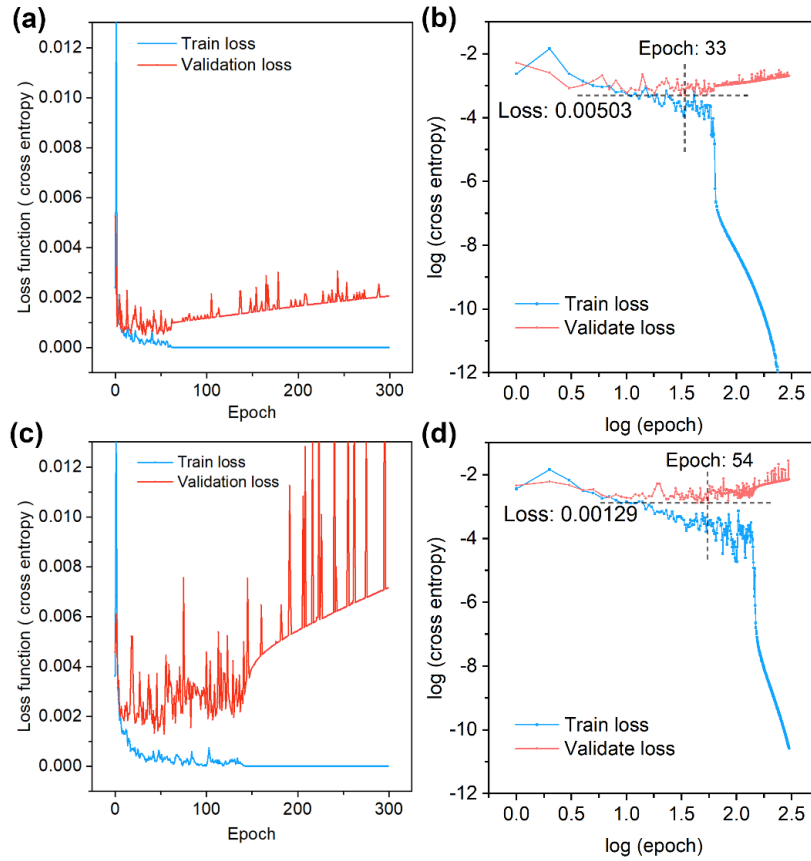


Fig. S1| Loss function evolution trends along with their log-log plot during the training process with varying sample amount. (a) Trends of the HFC model, (b) and log-log plot. (c) Trends of the HFE model, (d) and log-log plot.

Supplementary Notes 1: As illustrated in Fig. S1, the model was trained effectively in the early stages. However, as the number of epochs increased, signs of overfitting occurred. Our further testing results demonstrated that this overfitting is benign.

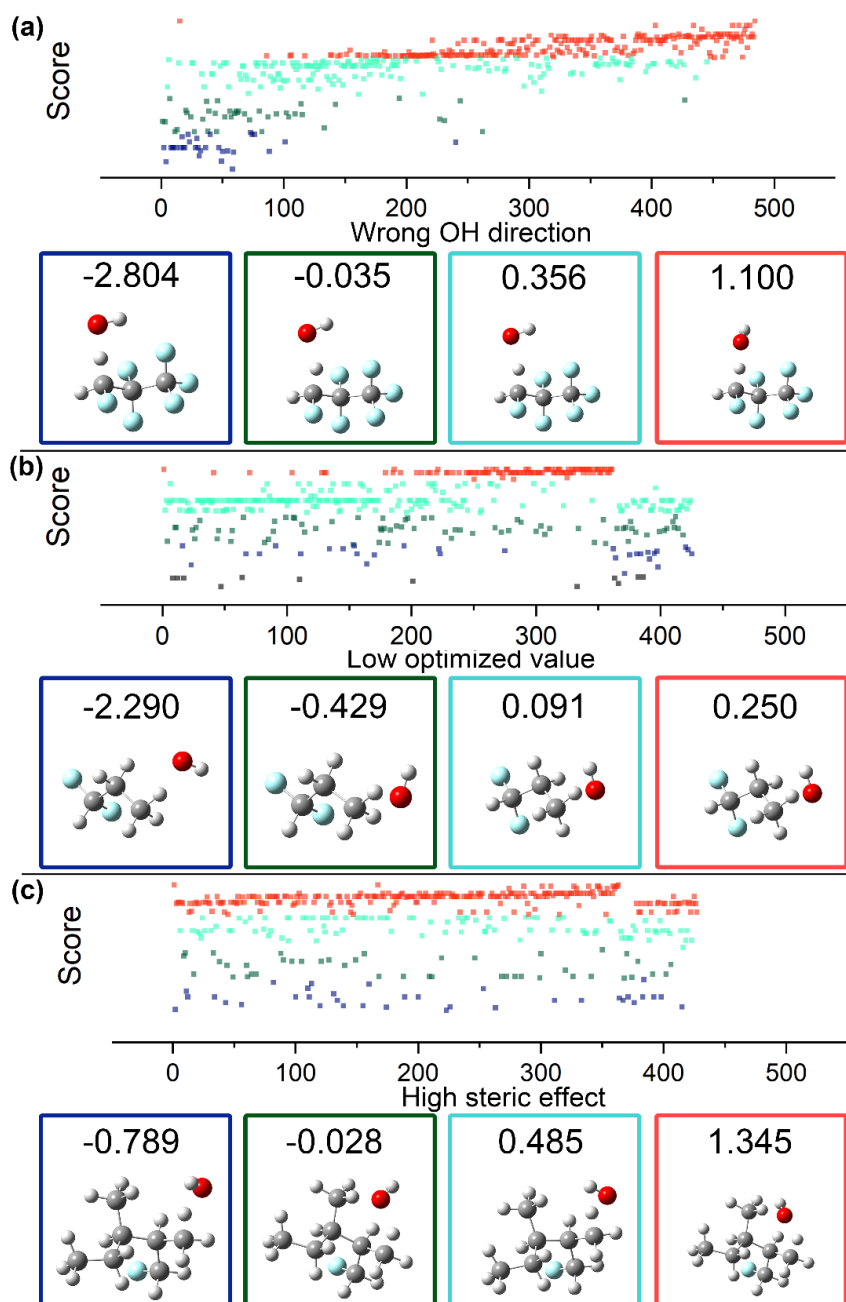


Fig. S2| CNN output logit evolution trends during the GA optimization process. (a) The generation of a transition structure for the $C_2F_5CH_2F$ molecule. (b) The generation of a transition structure for the $CHF_2C_2H_5$ molecule (c) The generation of a transition structure for the $C_2H_5CH(CH_3)CH(CH_2F)CH_3$ molecule. Atoms in red: oxygen; in cyan: fluorine; grey: carbon; white: hydrogen.

Supplementary Notes 2: During the generation of transition structures, we observed that the model's understanding of transition state structures remains somewhat limited. Consequently, in subsequent calculations, we will sample and compute at most five structures with the highest scores.

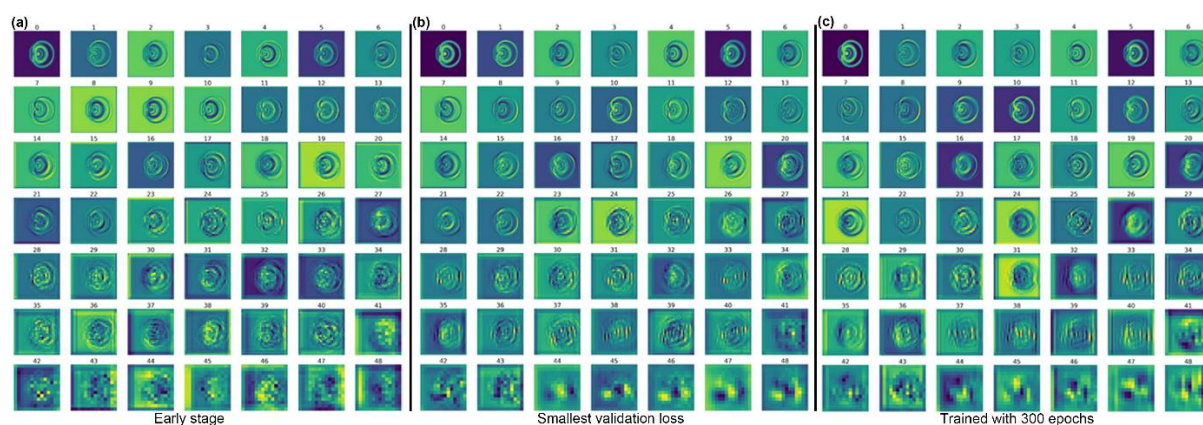


Fig. S3| Feature map for an HFC molecule derived in (a) early training stage, (b) a model with the smallest validation loss, (c) after 300 epochs training.

Supplementary Notes 3: Fig. S3 shows the feature map of the HFC training model from the models derived at different stages. It concluded that the models with more epochs didn't exhibit massive information loss.

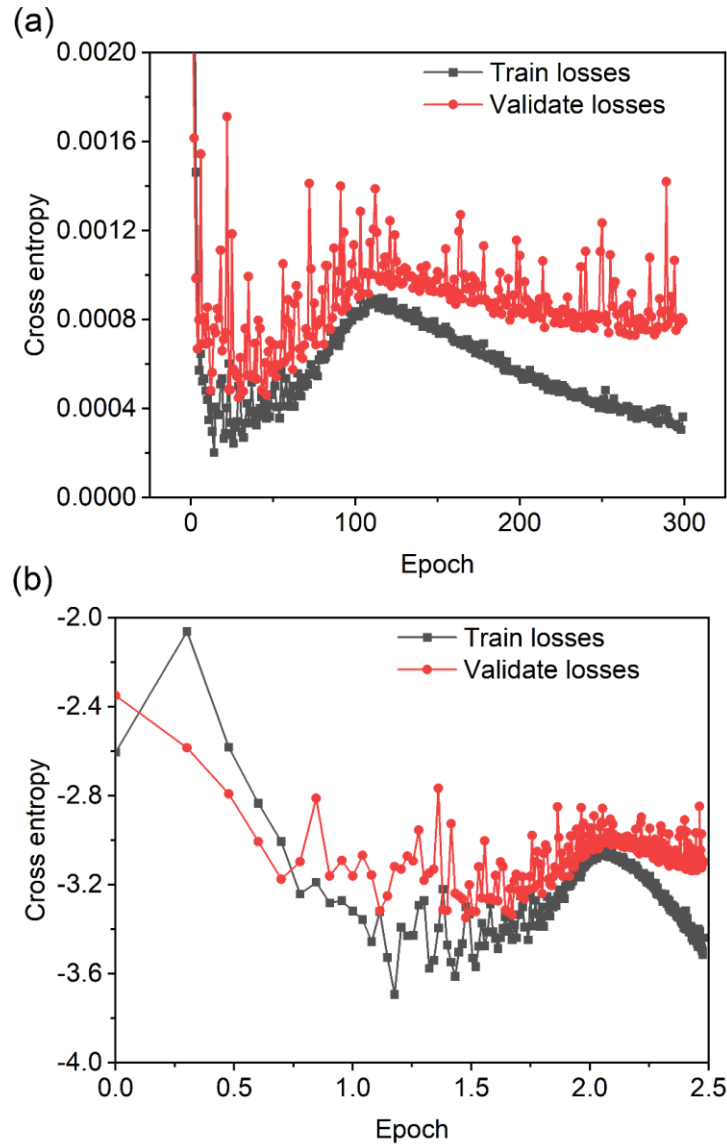


Fig. S4| Loss function evolution trends during the training process with L2 regularizer.

Supplementary Notes S4: Introducing the L2 regularizer effectively diminished the overfitting trend in the tail of training process of the previous models. However, it brought rises of both the training loss function and the validation loss function from the 30th to the 100th epochs. Therefore, given that L2 regularization did not exhibit a net positive benefit, we do not use it as the primary means to deal with overfitting in our study.

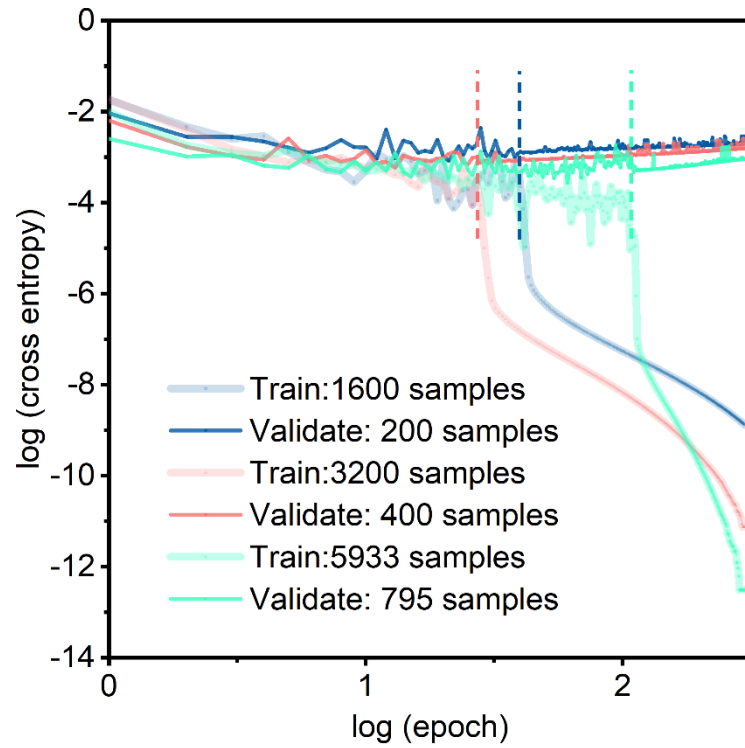


Fig. S5| Log-log plot of loss function evolution during the training process with varying sample size

Supplementary Notes 5: The onset of benign overfitting is not strongly correlated with the number of training samples; however, a larger sample size resulted in a lower validation loss function.

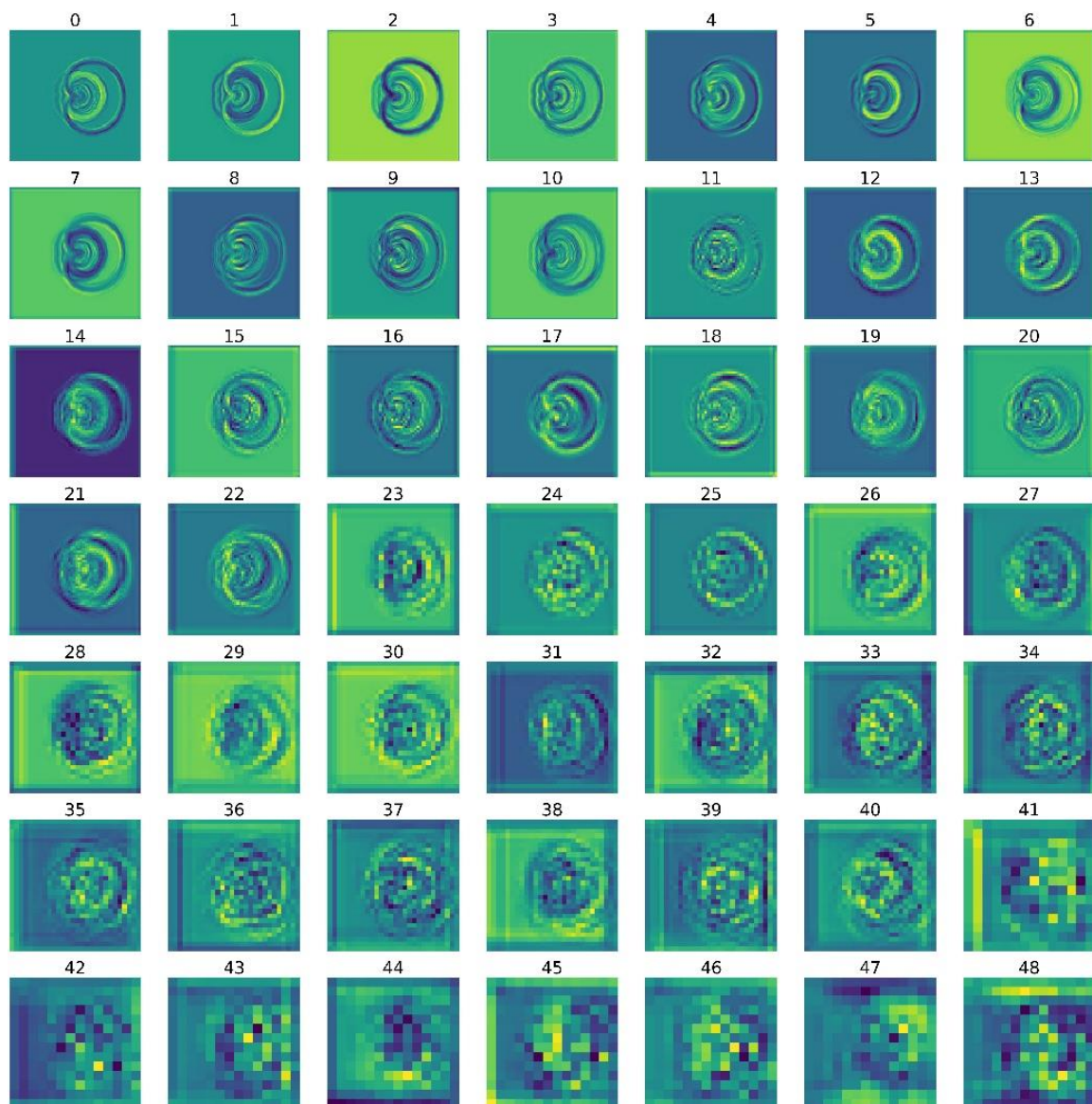


Fig. S6| Feature maps of HFC training models.

Supplementary Notes 6: Fig. S6 shows the feature map of the HFC training model. It can be seen that as the depth of the convolutional layer increases, the extracted feature transformed from details into a whole, and there is no significant information loss in each layer. This is one of the necessary conditions for the effectiveness of the training model.

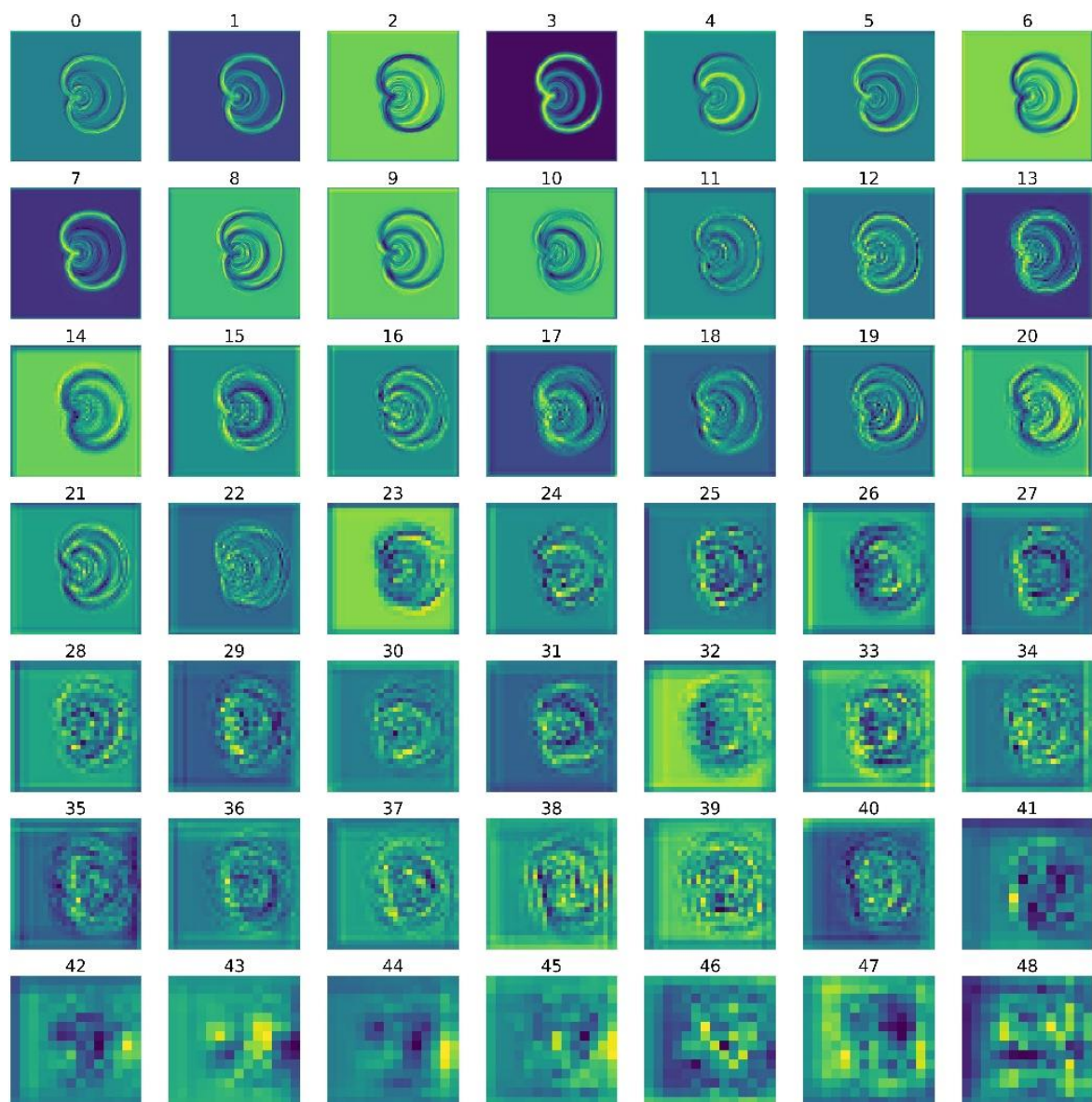


Fig. S7| Feature maps of HFE training models.

Supplementary Notes 7: Fig. S7 shows the feature map of the HFC training model. Similar to Fig. S1, the extracted feature transformed from details into a whole, and there is no significant information loss in each layer.

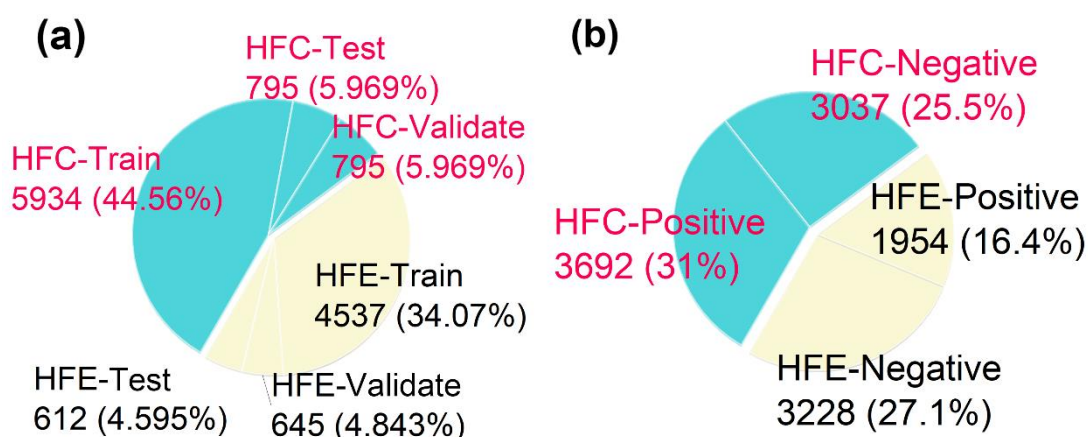


Fig. S8| The composition of the dataset in this work. (A) Classified according to data usage. **(B)** classified according to positive and negative samples (test sets excluded).

Supplementary Notes 8: From Fig. S8, it can be seen that the sample size of HFCs and HFEs is close, and the proportion used for training, validation, and testing is the same. In addition, the number of positive samples used for training HFC transition structure recognition is slightly larger than that of negative samples, while conversely the number of positive samples used for HFE is slightly smaller than that of negative samples. It can be inferred that the transition structure modeling success rate for HFE system is lower than that for HFC system.

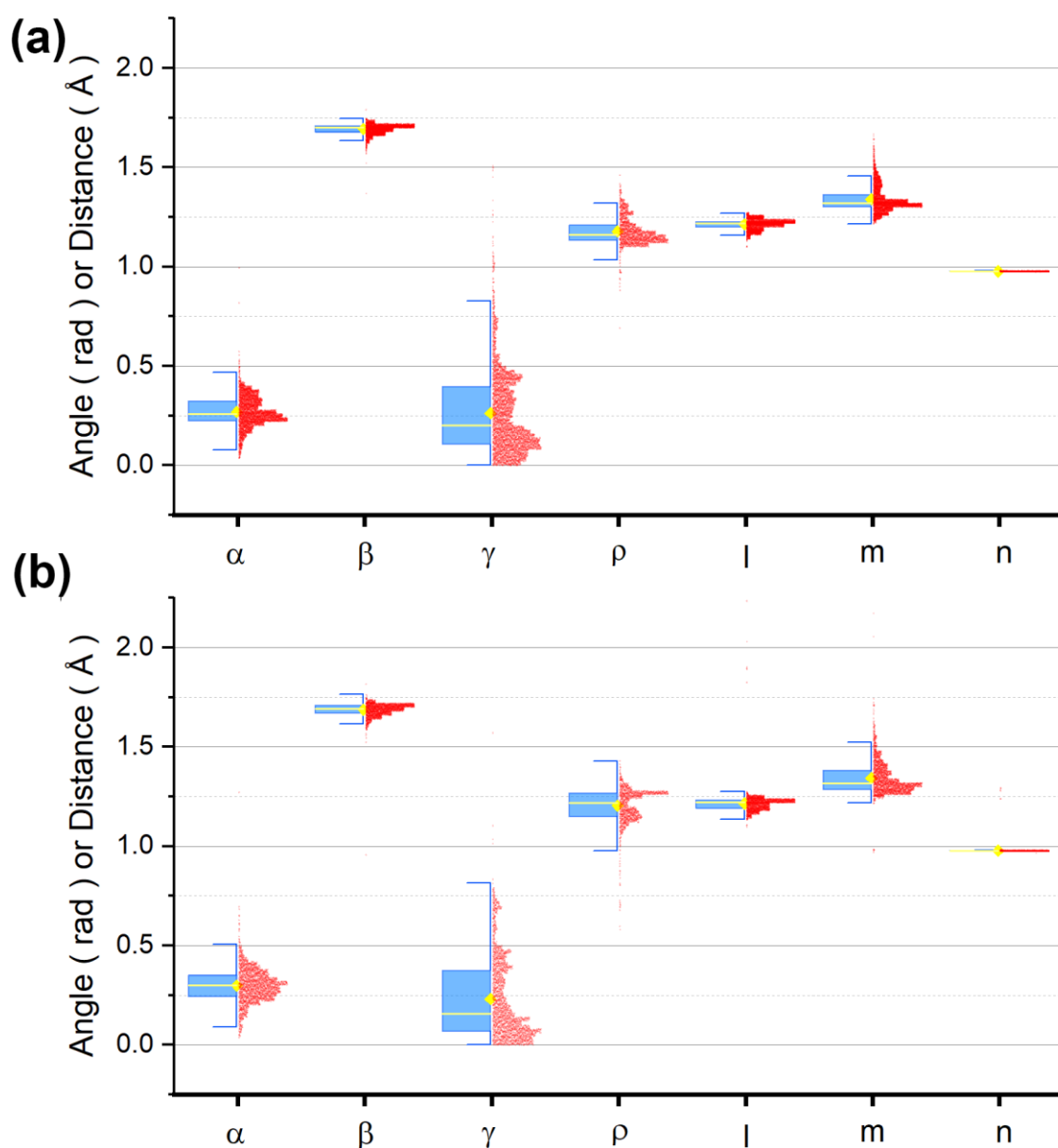


Fig. S9| Statistical distribution of geometric parameters. (A) for HFC system. (B) for HFE system.

Supplementary Notes 9: From Fig. S9, it can be observed that among the angle geometric parameters, the most localized parameter is β , corresponding to the bond angle of water molecules. The most dispersed parameter is γ . In the distance geometric parameters, the value of n is almost a constant, while the distribution of l is more concentrated than that of m . This phenomenon is one of the manifestations where the transition structure is closer to the reactant.

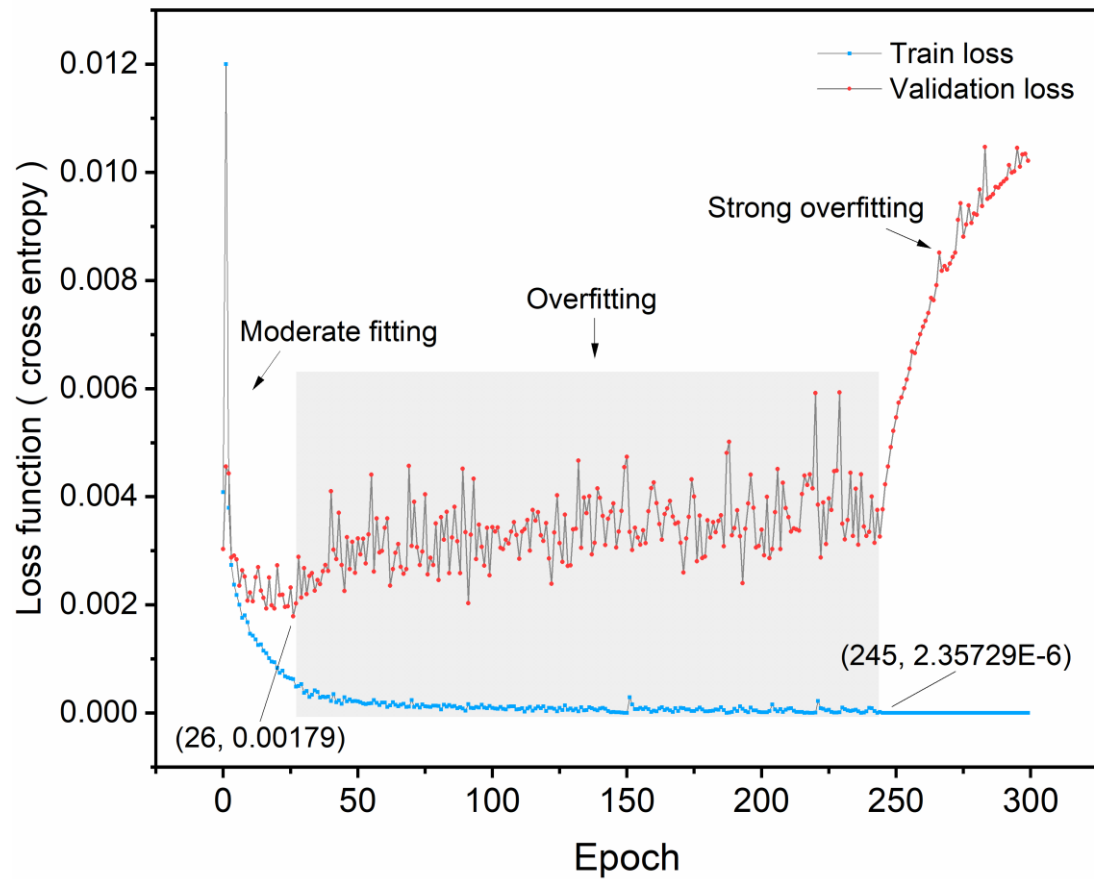


Fig. S10| Training process when HFCs and HFEs were trained together.

Supplementary Notes 10: Fig. S10 shows the training process when HFCs and HFEs were trained together.

The final performance of this model is inferior to that of separately trained modeling, so it has not been applied.

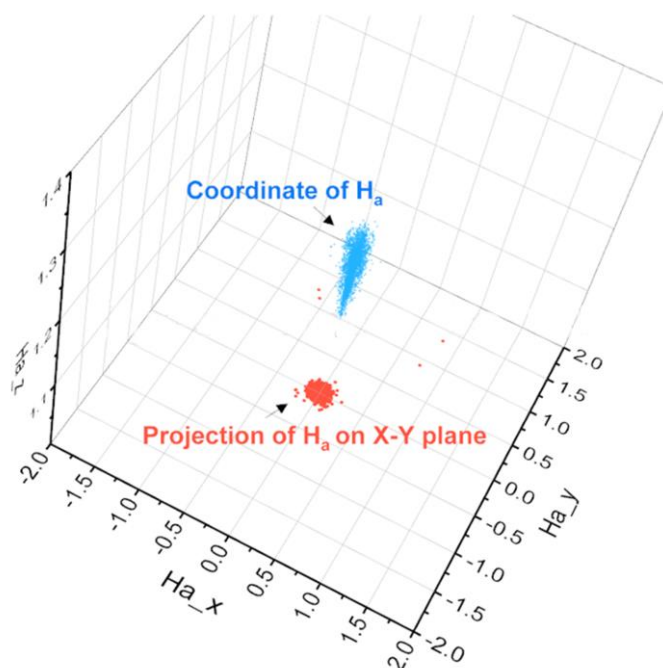


Fig. S11| H_a atom position distribution test results.

Supplementary Notes 11: Fig. S11 reflects the tested coordinate distribution of active hydrogen atoms in the transition structures before the formal proceeding of this work. In the figure, the x and y coordinates of H_a atoms are mainly distributed between positive and negative 0.25 Å. Therefore, in our transition structure modeling program, the x and y coordinates of the H_a atom are perturbed at (0, 0) by the hydroxyl orientation during the initial guess generation.

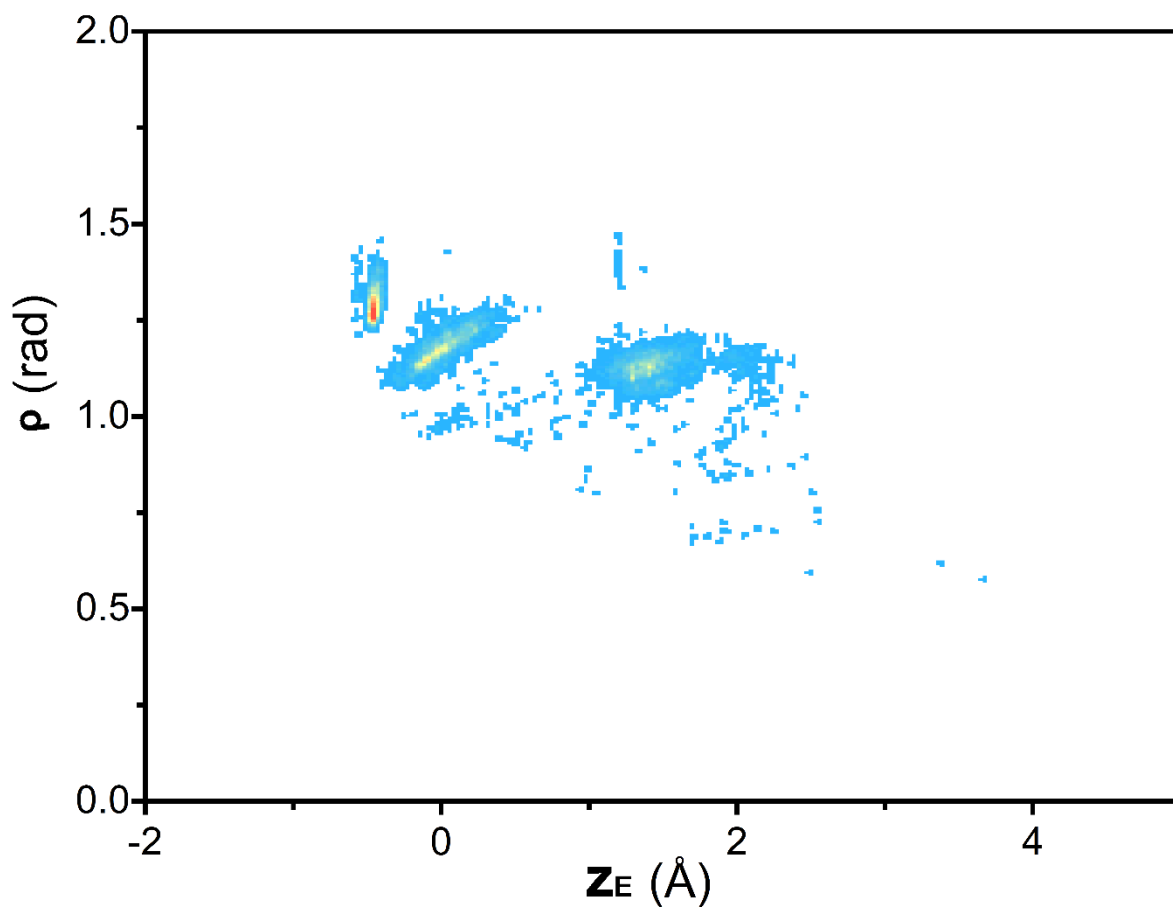


Fig. S12| The trend of ρ variation with the height of E atoms.

Supplementary Notes 12: Denoted in Fig. S12 is the trend of ρ variation with the height of reference E atoms. It can be observed that as the height of the reference E atoms increases, the value of ρ presents a downward trend. When the height of the reference E atom is less than 0 Å, ρ is larger than 1, and when the reference atom height is less than 0 Å, ρ is always smaller than 1.25.



Fig. S13| Curve direction changes with different $a_{1,j}$ parameter.

Supplementary Notes 13: The parameter r_j corresponds to the line width of curves. The parameter y_j determines the difference between the biggest ρ ($\rho_{\max}$) and the smallest ρ ($\rho_{\min}$). Considering that y_j can be either positive or negative, it is essential to interpret y_j with r_j . Combining the line width and the value of $\rho_{\max}$ minus $\rho_{\min}$, different y_j contributes to different curves. How different $a_{1,j}$ values correspond to different curve directions is illustrated in Fig. S13. Taken together, the unique mapping from molecular coordinates to bitmap curves is established.

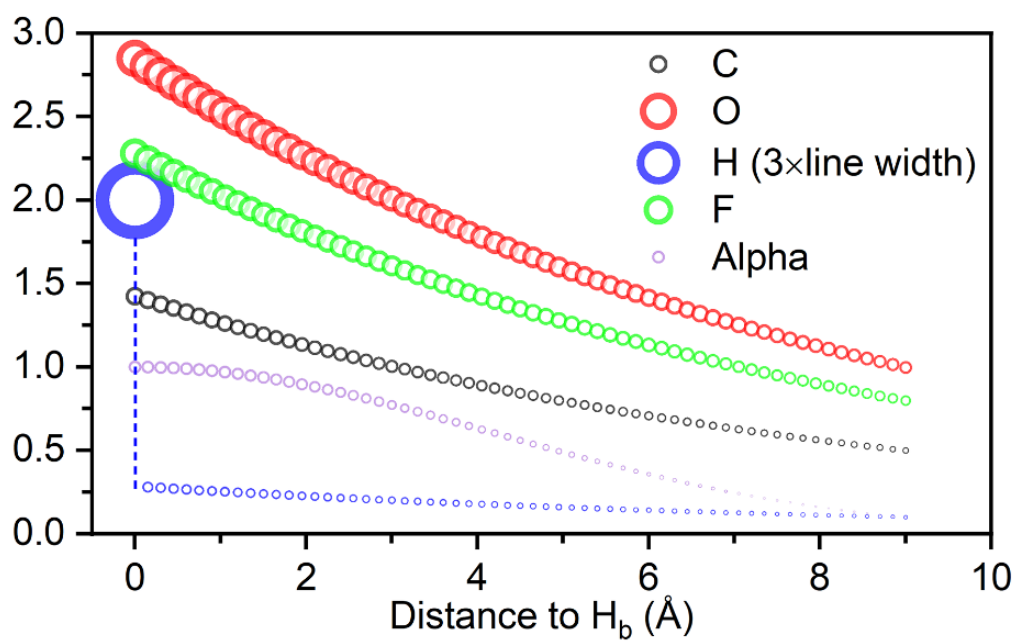


Fig. S14| Line width and transparency changes with the atom's distance to the Hb atom.

Supplementary Notes 14: Specially, the line width and transparency value of Hb were fixed to 2.0 and 1.0.

Section II: Supplementary Tables

Table S1| An HFC model trained after 300 epochs.

Train	Validate	Test
Using cuda:0 device. 64 64 256 128 512 256 1024 512 Testing started Test Error: Accuracy: 100.0000%, Avg loss: 0.000000 Testing Finished	Using cuda:0 device. 64 64 256 128 512 256 1024 512 Testing started Test Error: Accuracy: 99.3703%, Avg loss: 0.001413 Testing Finished	Using cuda:0 device. 64 64 256 128 512 256 1024 512 Testing started Test Error: Accuracy: 99.3703%, Avg loss: 0.005569 Testing Finished

Supplementary Notes 15: during the training of this model, a random seed was identified.

Table S2| An HFC model trained for the smallest validation loss function.

Train	Validate	Test
Using cuda:0 device.	Using cuda:0 device.	Using cuda:0 device.
64 64	64 64	64 64
256 128	256 128	256 128
512 256	512 256	512 256
1024 512	1024 512	1024 512
Testing started	Testing started	Testing started
Test Error:	Test Error:	Test Error:
Accuracy: 99.7135%, Avg loss: 0.000209	Accuracy: 99.4962%, Avg loss: 0.000401	Accuracy: 99.2443%, Avg loss: 0.001164
Testing Finished	Testing Finished	Testing Finished

Supplementary Notes 16: during the training of this model, a random seed consistent with that in Table S1 was identified.

Table S3| Records during the training process

HFCs			
Epoch	Train loss	Validation loss	Validation
0	0.002395	0.005249	0.8728
1	0.014426	0.002571	0.9496
2	0.002389	0.000841	0.9861
3	0.001403	0.001056	0.9736
4	0.001004	0.001424	0.9710
5	0.000911	0.002105	0.9559
6	0.000958	0.000782	0.9786
7	0.000623	0.001462	0.9824
8	0.000668	0.000823	0.9849
9	0.000584	0.000603	0.9887
10	0.000471	0.000825	0.9824
11	0.000566	0.000734	0.9836
12	0.000391	0.000748	0.9874
13	0.000648	0.002278	0.9458
14	0.000838	0.000976	0.9824
15	0.000458	0.000772	0.9836
16	0.000494	0.000724	0.9861
17	0.000327	0.000571	0.9887
18	0.000535	0.000901	0.9836
19	0.000266	0.001037	0.9874
20	0.000232	0.001298	0.9798
21	0.000278	0.001132	0.9824
22	0.000686	0.00167	0.9786
23	0.000455	0.001204	0.9786
24	0.00041	0.000824	0.9849
25	0.000363	0.000881	0.9849
26	0.000294	0.000698	0.9849
27	0.000166	0.000843	0.9899
28	0.00029	0.001601	0.9761
29	0.000317	0.000721	0.9849
30	0.000232	0.000535	0.9887
31	0.000158	0.001029	0.9861
32	0.000115	0.000794	0.9874
33	0.000173	0.000503	0.9937
34	0.000262	0.000815	0.9861
35	0.000314	0.001127	0.9849
36	0.000227	0.000845	0.9861
37	0.000243	0.000959	0.9899
38	0.000302	0.001093	0.9836
39	0.000181	0.000782	0.9924
40	0.000659	0.000876	0.9861
41	0.00033	0.000634	0.9861
42	0.000166	0.001472	0.9773
43	0.000225	0.001044	0.9849

44	0.000136	0.001108	0.9887
45	0.000114	0.000839	0.9861
46	0.000194	0.00054	0.9912
47	0.000316	0.000583	0.9887
48	0.000219	0.00079	0.9899
49	0.000133	0.000684	0.9912
50	9.13E-05	0.001459	0.9887
51	0.000241	0.000663	0.9924
52	0.000207	0.000681	0.9887
53	0.000245	0.000539	0.9924
54	0.000155	0.000777	0.9912
55	0.000102	0.001139	0.9861
56	0.000244	0.000873	0.9937
57	0.000221	0.000661	0.9899
58	2.91E-05	0.000758	0.9924
59	2.58E-05	0.000807	0.9937
60	9.58E-05	0.000686	0.9912
61	3.00E-05	0.000743	0.9924
62	1.48E-05	0.001139	0.9924
63	5.86E-07	0.000984	0.9924
64	2.28E-07	0.000991	0.9924
65	1.62E-07	0.000999	0.9924
66	1.28E-07	0.001003	0.9924
67	1.05E-07	0.001008	0.9924
68	8.81E-08	0.001015	0.9924
69	7.59E-08	0.00102	0.9924
70	6.65E-08	0.001025	0.9924
71	5.84E-08	0.00103	0.9924
72	5.19E-08	0.001037	0.9924
73	4.66E-08	0.00104	0.9924
74	4.19E-08	0.001181	0.9924
75	3.80E-08	0.001049	0.9924
76	3.45E-08	0.001055	0.9924
77	3.13E-08	0.001191	0.9924
78	2.86E-08	0.001198	0.9924
79	2.63E-08	0.001069	0.9924
80	2.42E-08	0.001253	0.9924
81	2.24E-08	0.001346	0.9924
82	2.04E-08	0.001085	0.9924
83	1.89E-08	0.001093	0.9924
84	1.76E-08	0.001094	0.9924
85	1.64E-08	0.001098	0.9924
86	1.51E-08	0.00124	0.9924
87	1.40E-08	0.001108	0.9924
88	1.30E-08	0.001114	0.9924
89	1.21E-08	0.001118	0.9924
90	1.13E-08	0.00126	0.9924
91	1.07E-08	0.001128	0.9924
92	9.86E-09	0.001132	0.9924

93	9.21E-09	0.00114	0.9924
94	8.60E-09	0.00128	0.9924
95	8.05E-09	0.001147	0.9924
96	7.55E-09	0.00115	0.9924
97	7.09E-09	0.001158	0.9924
98	6.63E-09	0.001302	0.9924
99	6.22E-09	0.001166	0.9924
100	5.84E-09	0.001171	0.9924

HFES			
Epoch	Train loss	Validation loss	Validation
0	0.003615	0.004563	0.8944
1	0.014492	0.0061	0.8711
2	0.006811	0.004733	0.8773
3	0.003145	0.003321	0.9519
4	0.002678	0.003454	0.9239
5	0.001835	0.002268	0.9503
6	0.002037	0.001989	0.9596
7	0.001595	0.002394	0.9503
8	0.001302	0.002044	0.9627
9	0.001421	0.001903	0.9565
10	0.001282	0.001822	0.9596
11	0.001292	0.002422	0.9519
12	0.001468	0.002205	0.9503
13	0.001401	0.002547	0.9534
14	0.000823	0.00183	0.9705
15	0.00111	0.002117	0.9581
16	0.000988	0.002096	0.9550
17	0.000606	0.003448	0.9565
18	0.000699	0.005209	0.9565
19	0.000663	0.005213	0.9534
20	0.000707	0.003183	0.9332
21	0.000602	0.002216	0.9627
22	0.000752	0.001735	0.9705
23	0.000418	0.001736	0.9689
24	0.000493	0.002099	0.9736
25	0.000684	0.002114	0.9596
26	0.000407	0.001967	0.9736
27	0.000333	0.003758	0.9534
28	0.000488	0.002459	0.9627
29	0.000503	0.002365	0.9612
30	0.000488	0.001669	0.9705
31	0.000426	0.002576	0.9658
32	0.000378	0.001685	0.9736
33	0.000274	0.001893	0.9736
34	0.000177	0.002239	0.9783
35	0.000272	0.001889	0.9689
36	0.000234	0.002963	0.9581
37	0.000384	0.003647	0.9581

38	0.000256	0.002061	0.9674
39	0.000355	0.003064	0.9643
40	0.000272	0.001905	0.9720
41	0.000385	0.001712	0.9674
42	0.000584	0.00222	0.9658
43	0.000347	0.001328	0.9752
44	0.000206	0.001983	0.9689
45	0.000189	0.001867	0.9752
46	0.000247	0.003726	0.9658
47	0.000404	0.001412	0.9705
48	0.000632	0.001703	0.9705
49	0.000359	0.001581	0.9767
50	0.000288	0.00194	0.9581
51	0.000454	0.002438	0.9643
52	0.000165	0.002407	0.9674
53	0.000258	0.001806	0.9705
54	0.000346	0.001293	0.9783
55	0.000281	0.002823	0.9627
56	0.000348	0.004444	0.9627
57	0.000379	0.002882	0.9643
58	0.000252	0.00256	0.9658
59	0.000247	0.004062	0.9503
60	0.000299	0.003246	0.9581
61	0.000178	0.003357	0.9658
62	0.000143	0.003116	0.9627
63	0.000151	0.001896	0.9736
64	8.70E-05	0.002667	0.9720
65	0.000149	0.002814	0.9612
66	0.000327	0.001736	0.9736
67	0.000426	0.002825	0.9643
68	0.000448	0.003411	0.9550
69	0.000391	0.002996	0.9643
70	0.000383	0.002599	0.9643
71	0.000255	0.003986	0.9658
72	0.000278	0.002129	0.9705
73	8.98E-05	0.002418	0.9736
74	3.39E-05	0.00211	0.9736
75	0.000134	0.007558	0.9534
76	0.000197	0.003417	0.9674
77	0.000186	0.002589	0.9643
78	0.000128	0.002473	0.9736
79	0.000305	0.003237	0.9689
80	0.000225	0.003538	0.9596
81	6.54E-05	0.002692	0.9689
82	9.87E-05	0.003385	0.9705
83	0.000246	0.001549	0.9674
84	0.000574	0.003213	0.9596
85	0.000303	0.002712	0.9643
86	0.00032	0.003338	0.9674

87	0.000174	0.003271	0.9643
88	0.000141	0.002948	0.9627
89	9.00E-05	0.002834	0.9674
90	5.49E-05	0.003174	0.9658
91	9.16E-05	0.002979	0.9658
92	3.29E-05	0.002885	0.9658
93	5.41E-05	0.002819	0.9627
94	7.75E-05	0.002279	0.9720
95	7.83E-05	0.003028	0.9643
96	2.50E-05	0.002813	0.9689
97	2.13E-05	0.003154	0.9689
98	1.92E-05	0.002927	0.9658
99	5.32E-05	0.002328	0.9674
100	1.85E-05	0.004576	0.9689

Supplementary Notes 17: Given our implementation of an early stopping strategy, the significance of subsequent data is limited. To simplify the presentation, we have provided only the training results for the first 100 epochs.

Section III: Supplementary Discussion

Supplementary Notes 18: Having determined the atmospheric lifetime and radiative forcing efficiency of a compound, the GWP can be calculated as follows:

$$\text{GWP}_x = \frac{44000RF\tau(1 - e^{-\frac{x}{\tau}})}{m\text{Ref}_{RFx}}$$

where, GWP_x is the Global Warming Potential referenced to x years /pW m⁻² yr kg⁻¹; RF is the Radiative Forcing /W m⁻² ppb⁻¹; τ is the Atmospheric Lifetime /yr; m is the Molecular molar mass /g mol⁻¹; and Ref_{RFx} is the Reference term, with the values of 0.1858, 0.66919, and 2.31845 for $x = 20, 100$, and 500, respectively.

If high precision is not required, a simple Transition State Theory (TST) can be used to predict the reaction rate. For a more accurate prediction, a variety of variational TST methods can be used, and knowledge of the transition state structure is crucial in this process.