Supporting information (SI)

Predicting the Enthalpy of Formation of Energetic Molecules via Conventional Machine Learning and GNN

Di Zhang[a], Qingzhao Chu[a], Dongping Chen[a,*]

[a] *State Key Laboratory of Explosion Science and Safety Protection, Beijing Institute of Technology, Beijing, 100081, China.*
** Corresponding author: dc516@bit.edu.cn*

In the supplementary material, we present:

Table S1 Prior knowledge-based descriptors in CDS descriptors.

Table S2 E-state fingerprint spectrum.

Table S3 Atom and edge attributes for constructing the molecule graph.

Table S4-S11 The performance of different models with different hyperparameters.

Table S12 Top 10 features and meanings for the CDS-RF model using the SHAP tool.

Table S13 Performance of different work on the QM9 data set.

Figure S1 (a) Training curve, (b) error distribution, and (c) parity plot for CDS-MLP model.

Figure S2 (a) Training curve, (b) error distribution, and (c) parity plot for ECFP-MLP model.

Figure S3 (a) Training curve, (b) error distribution, and (c) parity plot for SOAP-MLP model.

Figure S4 (a) Error distribution, and (b) parity plot for CDS-RF model.

Figure S5 (a) Error distribution, and (b) parity plot for ECFP-RF model.

Figure S6 (a) Error distribution, and (b) parity plot for SOAP-RF model.

Figure S7 EOF correlation on (a) oxygen (nO) and (b) nHbondA count in whole data set.

Figure S8 Top 10 features for the CDS-RF model using the SHAP tool.

Figure S9 Top 10 features on QM9-120k dataset.

Figure S10 (a) Training curve, (b) error distribution, and (c) parity plot for GCN model.

Figure S11 (a) Training curve, (c) error distribution, and (c) parity plot for MPNN model.

Table S1 Prior knowledge-based descriptors in CDS.

Abbreviation	Description	Abbreviation	Description
nHbondA	Number of hydrogen bond acceptor	nH	Number of hydrogen atom
nHbondD	Number of hydrogen bond donor	nC	Number of carbon atom
nAHC	Number of aromatic heterocycle	nN	Number of nitrogen atom
nACC	Number of aromatic carbocycle	nO	Number of oxygen atom
nHC	Number of heterocycle	ob	Oxygen balance
nR	Number of ring	Molecular weight	Molecular weight
nRbond	Number of rotatable bond	MOL volume	Molecular volume
nNO2	Number of nitro group	MinPartialCharge	Minimum value of partial charge
nNNO2	Number of nitramine group	MaxPartialCharge	Maximum value of partial charge
nONO2	Number of nitric ester group	TPSA	Topological polar surface area
nC(NO2)3	Number of nitroform group	PMI1	Principle moments of inertia 1
nC(NO2)2	Number of dinitro group	PMI2	Principle moments of inertia 2
nC(NO2)	Number of single nitro group	PMI3	Principle moments of inertia 3
nCH3	Number of methyl group	NPR1	Normalized principal moments ratios 1
nOCH3	Number of methoxy group	NPR2	Normalized principal moments ratios 2
nNH2	Number of amino group	PBF	Plane of best fit

total energy	Energy calculated by UFF	Eccentricity	Defined by sqrt(PMI3**2 - PMI1**2)/PMI3

Table S2 E-state fingerprint spectrum.

Index	Type	Index	Type	Index	Type
1	-Li	28	=N-[a]	55	-GeH$_3$
2	-Be-	29	aNa[a]	56	-GeH$_2$-
3	>Be<[-2]	30	>N-[a]	57	>GeH-
4	-BH-	31	-N<<[a]	58	>Ge<
5	>B-	32	aaNs[a]	59	-AsH$_2$
6	>B<[-1]	33	>N<[+1][a]	60	-ASH-
7	-CH$_3$[a]	34	-OH[a]	61	>AS-
8	=CH$_2$[a]	35	=O[a]	62	->As=
9	-CH$_2$-[a]	36	-O-[a]	63	->As<
10	≡CH[a]	37	aOa[a]	64	-she
11	=CH-[a]	38	-F-[a]	65	=Se
12	aCHa[a]	39	-SiH$_3$	66	-Se-
13	>CH-[a]	40	-SiH$_2$-	67	aSea
14	=C=[a]	41	>SiH-	68	>Se=
15	≡C-[a]	42	>Si<	69	≥Se=
16	=C<[a]	43	-PH$_2$	70[a]	-Br
17	aCa-[a]	44	-PH-	71	-SnH$_3$
18	aaCa[a]	45	>P-	72	-SnH$_2$-
19	>C<[a]	46	->P=	73	>SnH-
20	-NH3[+1][a]	47	->P<	74	>Sn<
21	=N-[a]	48	-SH	75[a]	-I
22	-NH$_2$-[+1][a]	49	=S	76	-PbH$_3$
23	=NH-[a]	50	-S	77	-PbH$_2$-
24	-NH-[a]	51	aSa	78	>PbH-
25	aNHa[a]	52	>S=	79	>Pb<
26	≡N[a]	53	≥S≤		
27	>NH-[+1][a]	54	-Cl[a]		

[a]These fingerprints are selected in this study.

Table S3 Atom and edge attributes for constructing the molecule graph.

Graph-level	Feature	Description	Size
Atom	atom type	Type of atom(ex.C,N,O)	9
	degree	Number of neighbors (ex.0,1,2,3,4)	9
	formal charge	Integer electronic charge assigned	8

		to atom (ex.-3, -2, -1, 0, 1, 2, 3)	
	hybridization type	*s, sp, sp^2, sp^3*	7
	is_in_a_ring	Whether the atom is in a ring	1
	aromaticity	Whether the atom is part of an aromatic system	1
	atomic mass	Mass of atom, scaled	1
	vdw_radius	van der Waals Radius, scaled	1
	covalent_radius	covalent radius, scaled	1
	chirality_type	Chirality of atom (ex.CHI_UNSPECIFIED, CHI_TETRAHEDRAL_CW, CHI_TETRAHEDRAL_CCW)	4
	n_hydrogens	Number of bonded hydrogens 0, 1, 2, 3, 4	6
Bond	bond type	Single, double, triple, aromatic	4
	conjugated	Whether the bond is conjugated	1
	in ring	Whether the bond is part of ring	1
	stereo type	None,any,Z/E	4

Table S4 The performance of CDS-RF model with different hyperparameters.

	Parameters	**Performance**		
	n estimators	R^2	MAE	RMSE
1	200	0.970	11.605	16.231
2	150	0.970	11.606	16.229
3	100	0.970	11.637	16.295

Table S5 The performance of ECFP-RF model with different hyperparameters.

	Parameters	**Performance**		
	n estimators	R^2	MAE	RMSE
1	200	0.899	19.532	30.562
2	150	0.898	19.569	30.628
3	100	0.897	19.671	30.838

Table S6 The performance of SOAP-RF model with different hyperparameters.

Parameters	**Performance**

	n estimators	R^2	MAE	RMSE
1	200	0.968	12.268	17.222
2	150	0.968	12.294	17.270
3	100	0.968	12.319	17.280

Table S7 The performance of CDS-MLP model with different hyperparameters.

	Parameters		**Performance**		
	hidden layer sizes	learning rate	R^2	MAE	RMSE
1	256,128	0.001	0.986	7.750	11.073
2	256,256	0.001	0.986	8.034	11.379
3	128,128	0.001	0.986	8.105	11.495
4	256,128	0.01	0.984	8.721	12.153
5	256,256	0.01	0.983	8.825	12.508
6	128,128	0.01	0.983	8.603	12.384

Table S8 The performance of ECFP-MLP model with different hyperparameters.

	Parameters		**Performance**		
	hidden layer sizes	learning rate	R^2	MAE	RMSE
1	256,128	0.001	0.933	17.946	24.991
2	256,256	0.001	0.933	18.052	25.038
3	128,128	0.001	0.926	19.111	26.247
4	256,128	0.01	0.931	18.195	25.429
5	256,256	0.01	0.927	18.835	26.057
6	128,128	0.01	0.924	19.240	26.678

Table S9 The performance of SOAP-MLP model with different hyperparameters.

	Parameters		**Performance**		
	hidden layer sizes	learning rate	R^2	MAE	RMSE
1	256,128	0.001	0.984	8.124	12.061
2	256,256	0.001	0.984	8.270	12.075
3	128,128	0.001	0.984	8.552	12.137
4	256,128	0.01	0.986	8.000	11.632
5	256,256	0.01	0.986	7.997	11.456
6	128,128	0.01	0.986	7.951	11.323

Table S10 The performance of GCN model with different hyperparameters.

	Parameters		**Performance**		
	hidden layer sizes	learning rate	R^2	MAE	RMSE
1	256,256	0.001	0.983	8.912	12.759
2	512,512	0.001	0.987	7.597	11.051
3	128,128	0.001	0.980	9.794	13.707
4	256,256	0.01	0.986	8.088	11.372
5	512,512	0.01	0.986	8.067	11.611
6	128,128	0.01	0.986	7.933	11.352
7	256,256,256	0.001	0.989	7.014	10.333
8	512,512,512	0.001	0.990	6.537	9.6493

Table S11 The performance of MPNN model with different hyperparameters.

	Parameters		**Performance**		
	hidden layer sizes	learning rate	R^2	MAE	RMSE
1	256,256	0.001	0.989	6.773	9.911
2	512,512	0.001	0.990	6.504	9.595
3	128,128	0.001	0.989	6.874	10.079
4	256,256	0.01	0.990	6.420	9.377
5	512,512	0.01	0.990	6.426	9.474
6	128,128	0.01	0.989	6.725	10.000
7	256,256,256	0.001	0.992	5.379	8.450
8	512,512,512	0.001	0.992	5.243	8.419

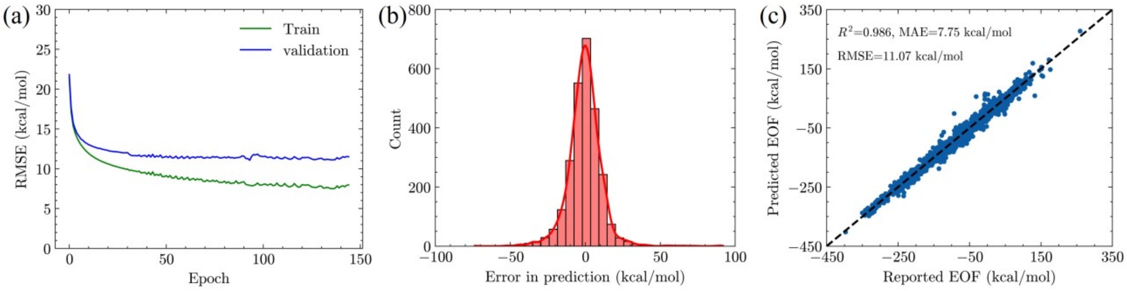

Figure S1 (a) Training curve, (b) error distribution, and (c) parity plot for CDS-MLP model.

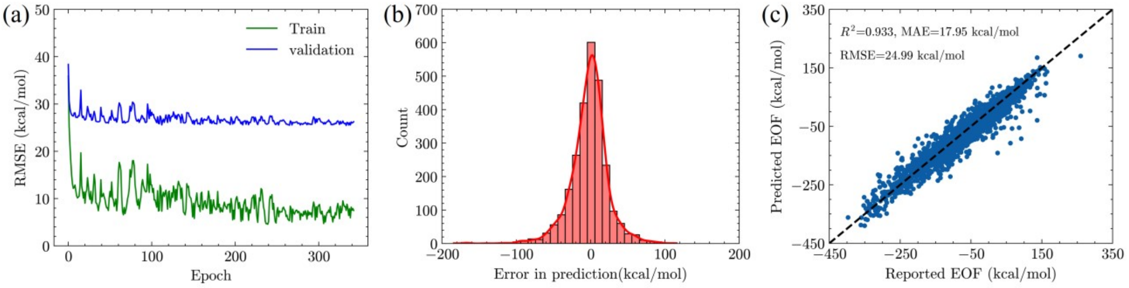

Figure S2 (a) Training curve, (b) error distribution, and (c) parity plot for ECFP-MLP model.

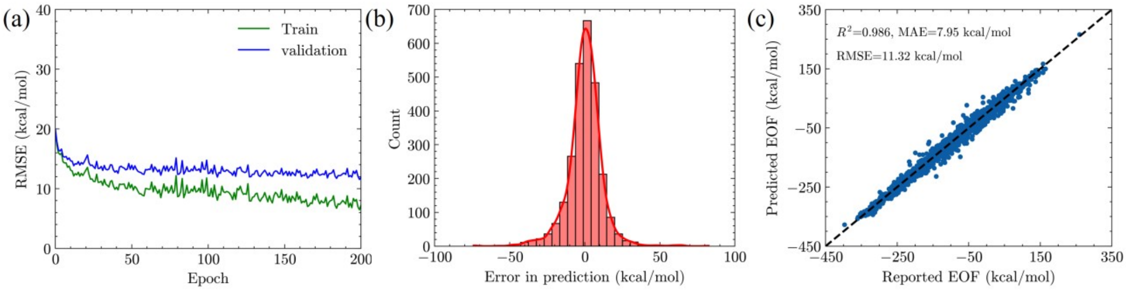

Figure S3 (a) Training curve, (b) error distribution, and (c) parity plot for SOAP-MLP model.

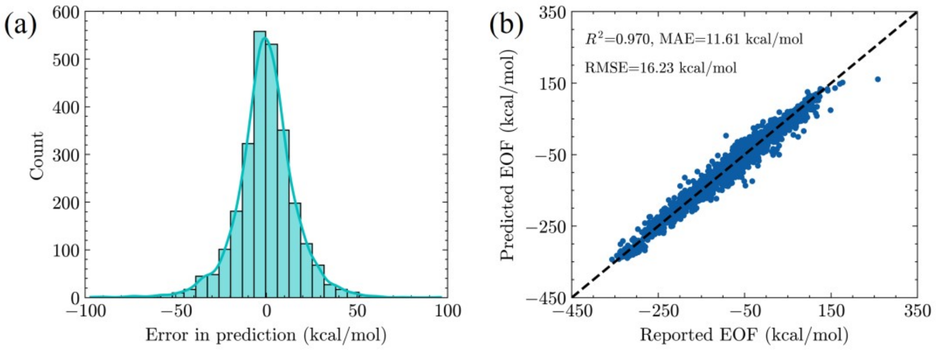

Figure S4 (a) Error distribution, and (b) parity plot for CDS-RF model.

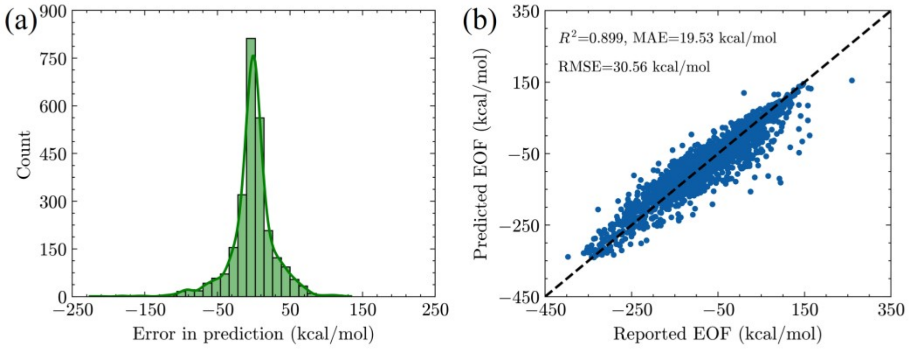

Figure S5 (a) Error distribution, and (b) parity plot for ECFP-RF model.

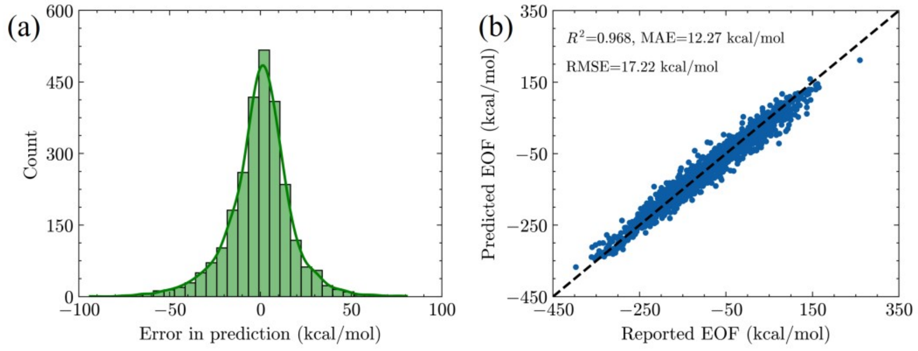

Figure S6 (a) Error distribution, and (b) parity plot for SOAP-RF model.

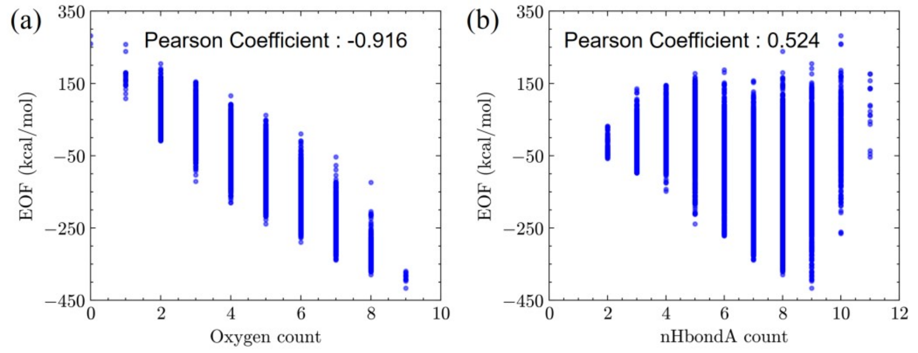

Figure S7 EOF correlation on (a) oxygen (nO) and (b) nHbondA count in whole data set.

To further understand the significant impact of nO on the EOF, we verify it from the following two aspects.

Firstly, to eliminate the influence of the ranking tool on the results, we use the SHAP (SHapley Additive exPlanations) tool to sort the importance of each feature in the CDS-RF model. The impact of each feature on the EOF is shown in Figures S8. The meaning of the first 10 features is shown in Table S12. Although the principles of feature sorting in SHAP and this study are different, nO still has the greatest impact on the model, which supports the reliability of the results from another perspective.

Secondly, it is generally believed that nN and nC should have a greater impact. However, nN and nC have a smaller influence in our results. We speculate that this is due to the uniqueness of the current data set. To support our hypothesis, we conduct an auxiliary verification using a publicly available QM9 dataset[1], consisting of 134 k organic small molecules containing CHONF, encompassing geometric, energetic, electronic, and thermodynamic properties. After excluding molecules that cannot be processed and species containing the element F, we train our RF model using the same CDS descriptors on the remaining 120,056 molecules. We then perform feature importance ranking and obtain the results shown in Figure S9. The most influential factors are molecule weight, oxygen balance, nO, nC, nH, and nN, where the top two factors are closely related to the quantities of C, O, and N. Additionally, nC and nN exhibit high ranking. The results from the QM9 dataset demonstrate substantial differences in feature importance ranking obtained from different datasets. Therefore, we believe that the significant impact of nO is attributed to the uniqueness of the data set in our study.

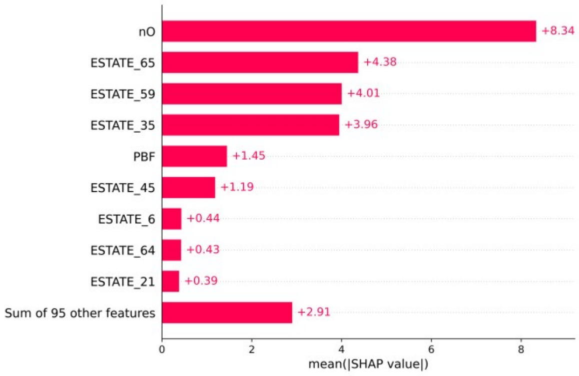

Figure S8 Top 10 features for the CDS-RF model using the SHAP tool.

Table S12 Top 10 features and meanings for the CDS-RF model using the SHAP tool.

Features	**Meaning**
nO	The number of O
ESTATE_65	The sum of the electrical topological state index of aO
ESTATE_59	The sum of the electrical topological state index of -N<<
ESTATE_35	The sum of the electrical topological state index of -CH3
PBF	Plane of best fit
ESTATE_45	The sum of the electrical topological state index of aCa
ESTATE_6	The number of -CH-
ESTATE_64	The sum of the electrical topological state index of -O-
ESTATE_21	The number of =N-

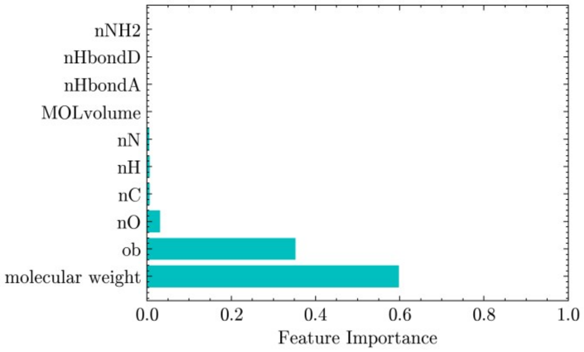

Figure S9 Top 10 features on QM9-120k dataset.

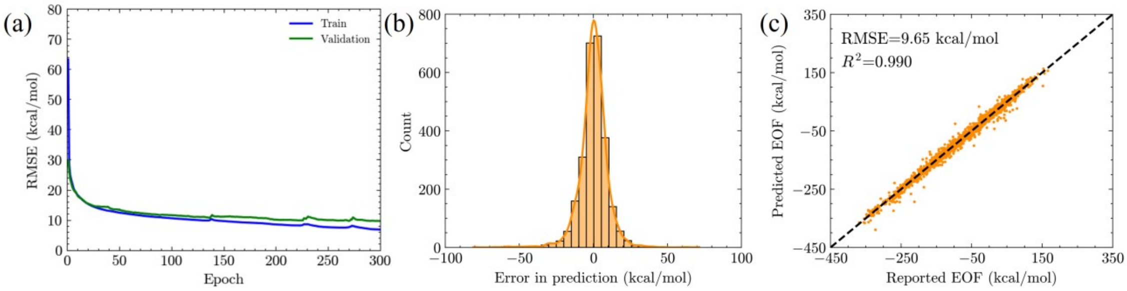

Figure S10 (a) Training curve, (b) error distribution, and (c) parity plot for GCN model.

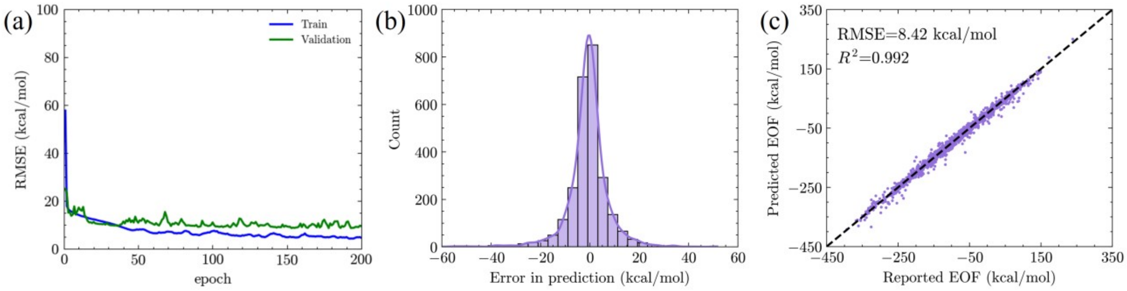

Figure S11 (a) Training curve, (c) error distribution, and (c) parity plot for MPNN model.

In order to validate the reliability of the MPNN model, we conduct experiments on the publicly available QM9 dataset and compare the performance of our MPNN model with other models on the same QM9 dataset. The QM9 dataset comprises 130,829 molecules containing CHONF with no more than 9 heavy atoms. This dataset exists in two versions, as provided by Faber et al. [2] and Ramakrishnan et al. [1] We conduct tests on the U0 (internal energy at 0 K) from both versions of the QM9 dataset and compare the results with widely used models such as Chemprop[3, 4] and Megnet[5]. The results are shown in Table S13.

For the U0 from Faber et al. [2], our MPNN model yield an MAE of 0.0494 eV, lower than

Gabriel et al.'s[6] results but slightly higher than Megnet[5] and SchNet[7, 8]. The MAE in Faber's[2] work ranges from 0.0421 to 1.08 eV, and our error falls within the same range. Overall, our findings are within a reasonable range.

Regarding the U0 from Ramakrishnan et al. [1], after processing, we obtain a total of 120,056 data points. The RMSE of this dataset is 2.47 Hartree, which is close to the reported RMSE of Chemprop[3, 4] as 2.44 Hartree. In summary, the prediction performance obtained from various works on the QM9 dataset are generally consistent with the performance of our MPNN model.

Table S13 Performance of different work on the QM9 data set.

Data set	Model	R^2	MAE	RMSE	Number
Faber et al.[2]	Faber [2]	-	0.0421~1.08 eV	-	118k
	Megnet-simple [5]	-	0.012 eV	-	118k
	SchNet[7, 8]	-	0.014 eV	-	100k
	Gabriel [6]	-	0.0573~0.084 eV	-	101k
	This work	0.989	0.0494 eV	0.1034 eV	130k
Ramakrishnan et al.[1]	Chemprop [3, 4]	-	1.08 Hartree	2.44 Hartree	130k
	This work	0.996	0.91 Hartree	2.47 Hartree	120k

Reference

1. R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. von Lilienfeld, *Sci. Data*, 2014, **1**, 140022.
2. F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley and O. A. von Lilienfeld, *J. Chem. Theory Comput.*, 2017, **13**, 5255-5264.
3. E. Heid, K. P. Greenman, Y. Chung, S. C. Li, D. E. Graff, F. H. Vermeire, H. Wu, W. H. Green and C. J. McGill, *J. Chem. Inf. Model*, 2023, **64**, 14229-14242.
4. K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen and R. Barzilay, *J. Chem. Inf. Model*, 2019, **59**, 3370-3388.
5. C. Chen, W. Ye, Y. Zuo, C. Zheng and S. P. Ong, *Chem. Mater.*, 2019, **31**, 3564-3572.
6. G. A. Pinheiro, J. Mucelini, M. D. Soares, R. C. Prati, J. L. F. Da Silva and M. G. Quiles, *J. Phys. Chem. A*, 2020, **124**, 9854-9866.
7. K. T. Schütt, M. Gastegger, A. Tkatchenko, K. R. Müller and R. J. Maurer, *Nat. Commun.*, 2019, **10**, 5024.
8. K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko and K.-R. Müller, *J. Chem. Phys.*, 2018, **148**, 241722.