Supplementary Information for

Explainable Optimized 3D-MoRSE Descriptors for the Power Conversion Efficiency Prediction of Molecular Passivated Perovskite Solar Cells Through Machine Learning

Xin Ye, Ningyi Cui, Wen Ou, Donghua Liu, Yufan Bao, Bin Ai, Yecheng Zhou**

Guangzhou Key Laboratory of Flexible Electronic Materials and Wearable Devices

School of Materials Science & Engineering, Sun Yat-sen University, Guangzhou 510006, Guangdong, People's Republic of China.

E-mail address: stsab@mail.sysu.edu.cn (B. Ai), zhouych29@mail.sysu.edu.cn (Y. Zhou)

Contents

1. **Fig. S1~S5** The morU descriptors calculated with different scale factors.
2. **Fig. S6** Standard 3D-MoRSE radial basis function $f(s, r_{ij})$ at different scattering parameters.
3. **Fig. S7~S8** The RMSE versus s_L-s heatmap of morE and morIP descriptor sets.
4. **Fig. S9~S10** The bar plot and scatter plot of 100 times fitting with 30% samples as the test set.
5. **Fig. S11** The bar plot and scatter plot of ML result based on morX (0.1,14) with the LOO method.
6. **Fig. S12~S14** The RMSE heatmap to optimized s_L-s with ARD regression and LOO method.
7. **Fig. S15** The SHAP summary plot for ARD models based on morX (0.1,14) and E-C.
8. **Fig. S16** The key descriptor value of morU (0.21, 38).
9. **Fig. S17** Using optimized 3D-MoRSE to predict the adsorption energy and transferred charge of the molecule on a PVK surface.
10. **Fig. S18~S19** The result of morSel-1 and morSel-2
11. **Fig. S20** The PCE prediction with graph neural network model.
12. **Table S1** Molecular descriptor sets and their abbreviations.
13. **Table S2** Prediction R^2, RMSE, and r for all molecular descriptors using ARD regression.
14. **Table S3** The optimized hyperparameters for all ML processes.
15. **Table S4** The result with the LOO method and RF, ARD, ARD, and LASSO algorithms.
16. **Table S5** The result predicted by morIP (0.1,22) and added molecular descriptors.

17. **Table S6** The prediction of values of molecular descriptors by optimized 3D-MoRSE descriptors.

Table S1 Molecular descriptor sets and their abbreviations.

Abbr.	Illustration / weight of atoms (A)	Abbr.	Illustration / weight of atoms (A)
E-C	E-state index and cheminformatics from reference [2]	3DM-U	3D MoRSE, A = 1.0
		3DM-E	3D MoRSE, A=atomic Sanderson electronegativity
morRC	A = atomic covalent radius	3DM-IP	3D MoRSE, A=atomic ionization potential[4]
morU	A = 1.0	morIS	A = atomic intrinsic state[3]
morE	A = atomic Sanderson electronegativity	morIP	A = atomic ionization potential[4]
morC	A = atomic charge	morCa	A = | atomic charge |
morV	A = atomic van der Waals volume	morPI	A = atomic π electrons
morM	A = atomic mass	morZV	A = atomic valence electrons
morP	A = atomic polarizability	morZC	A = (atomic valence electrons) – (atomic charge)

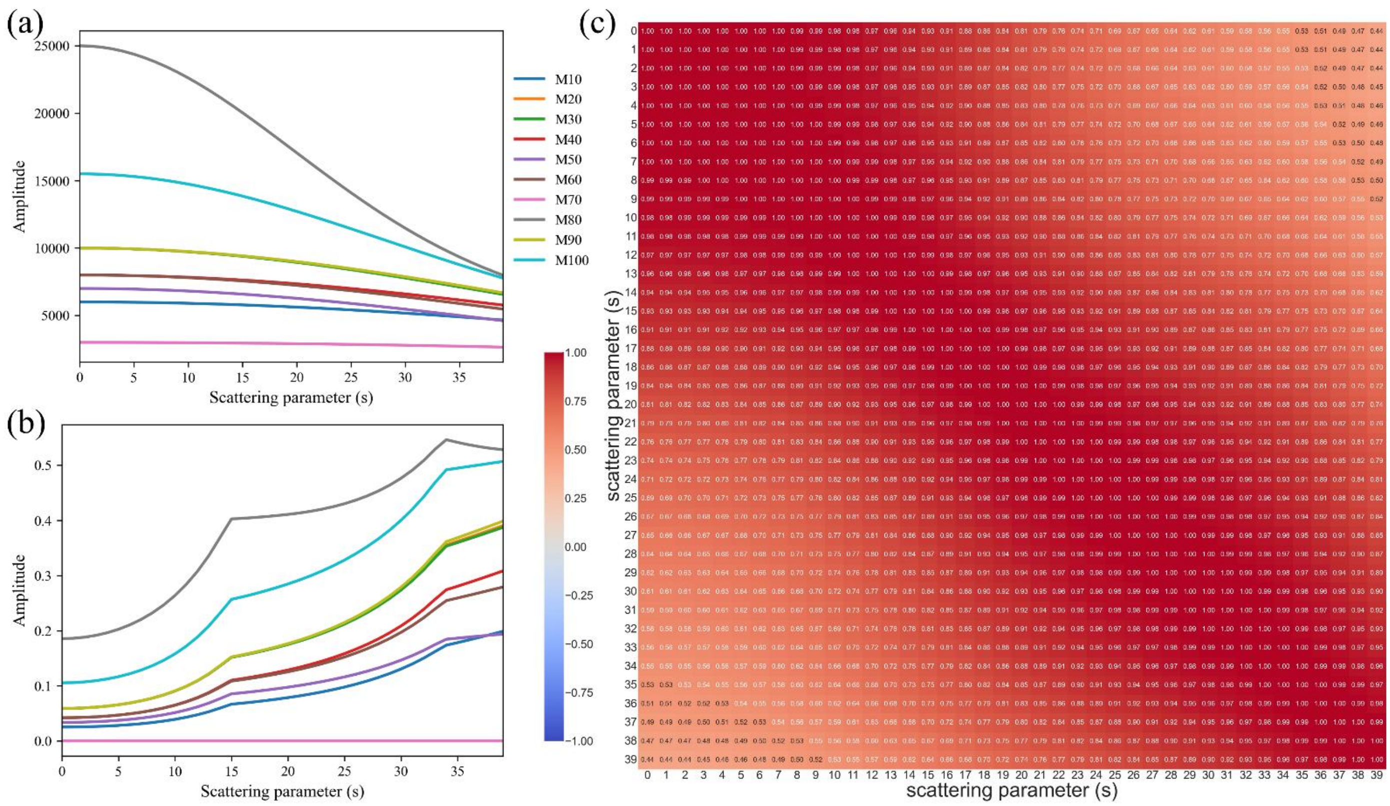

Fig. S1 (a) The morU descriptors calculated with the scale factor of 0.01 on ten example molecules, and (b) scaled with MinMaxScaler. (c) The Pearson correlation coefficient of the whole descriptor set.

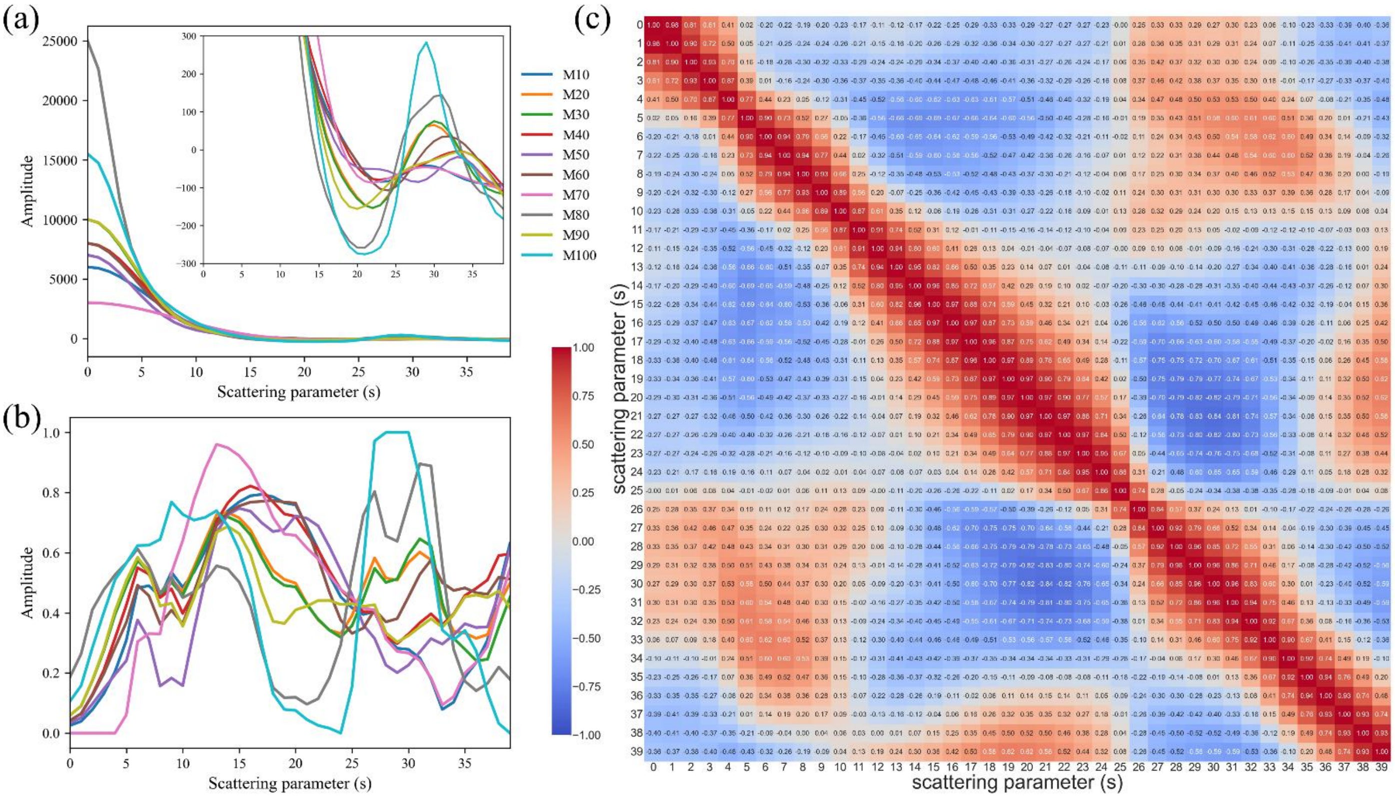

Fig. S2 (a) The morU descriptors calculated with the scale factor of 0.1 on ten example molecules, and (b) scaled with MinMaxScaler. (c) The Pearson correlation coefficient of the whole descriptor set.

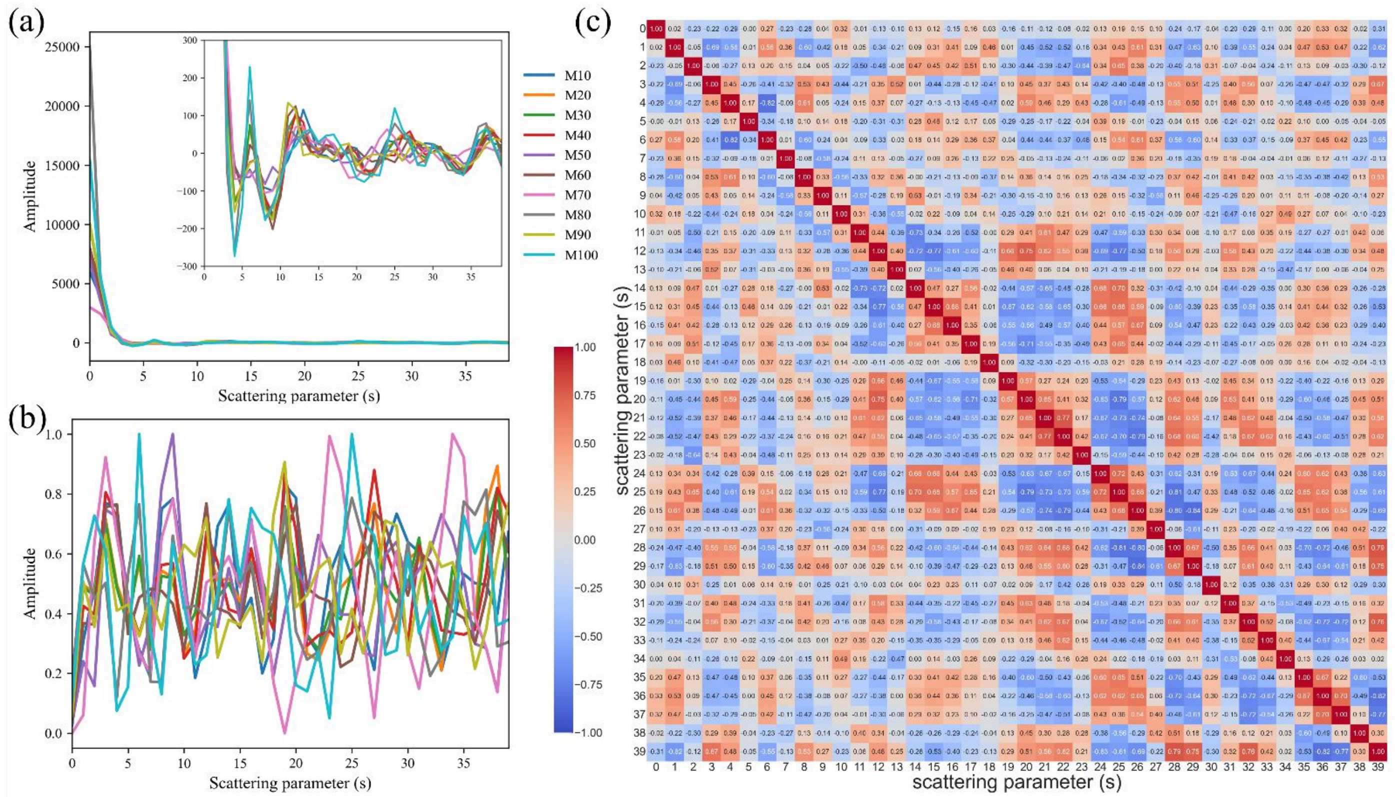

Fig. S3 (a) The morU descriptors calculated with the scale factor of 0.5 on ten example molecules, and (b) scaled with MinMaxScaler. (c) The Pearson correlation coefficient of the whole descriptor set.

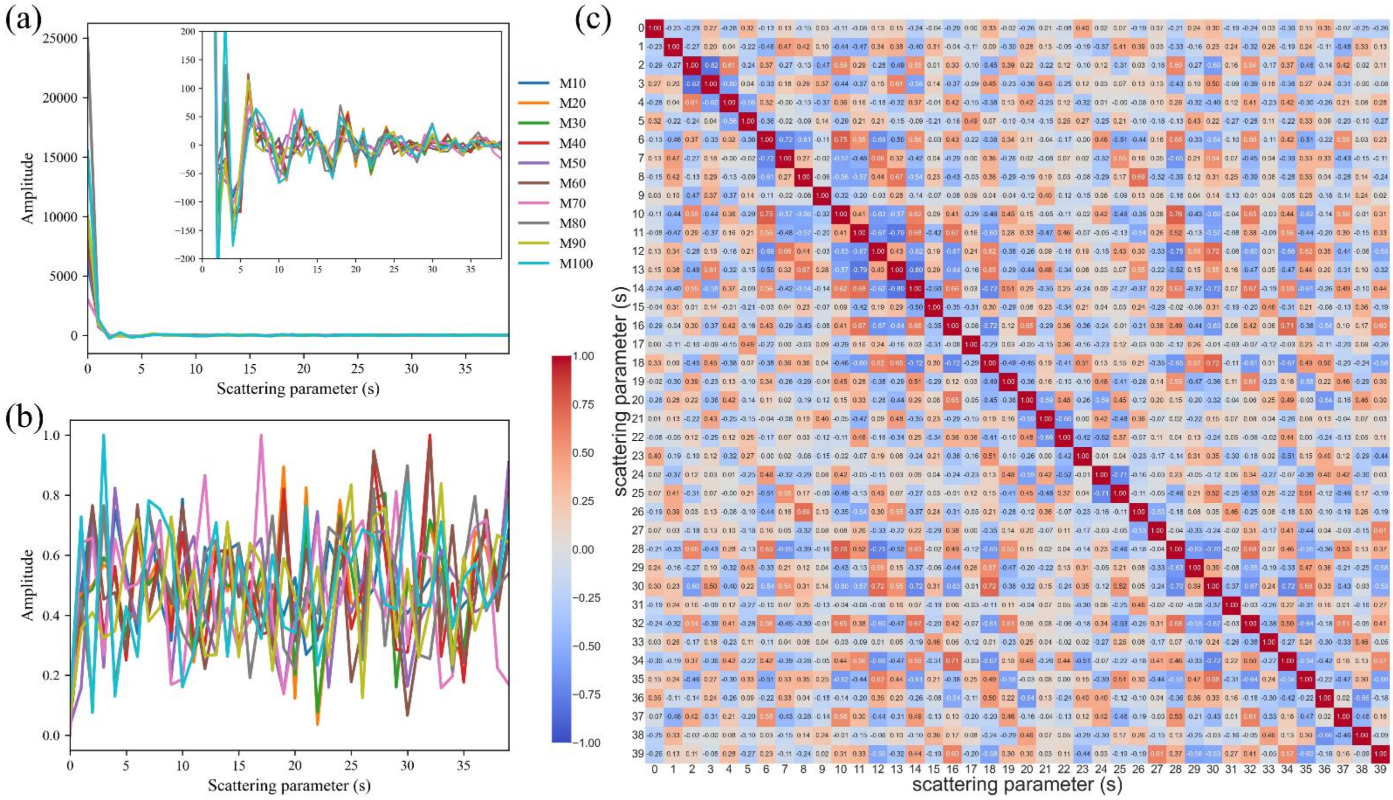

Fig. S4 (a) The morU descriptors calculated with the scale factor of 1 on ten example molecules, and (b) scaled with MinMaxScaler. (c) The Pearson correlation coefficient of the whole descriptor set.

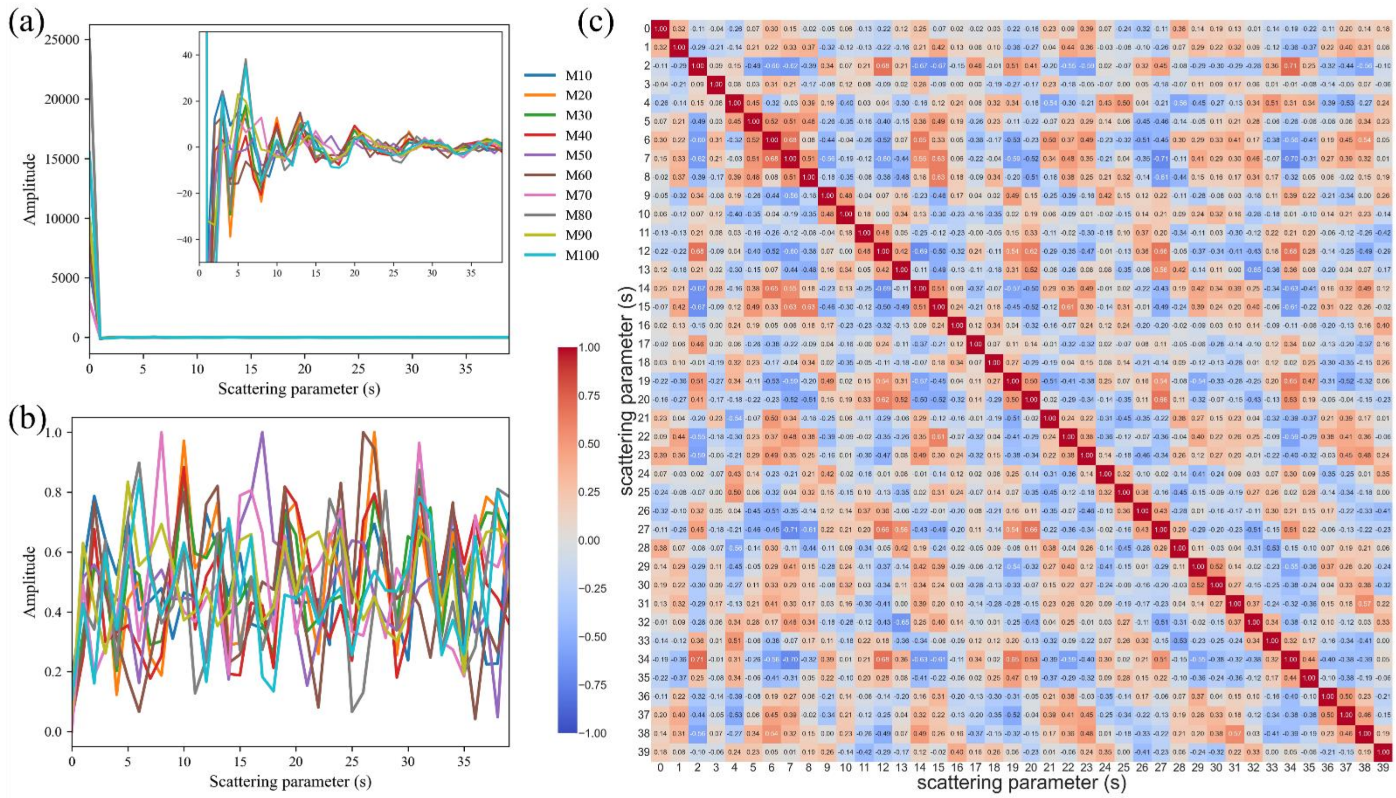

Fig. S5 (a) The morU descriptors calculated with the scale factor of 5 on ten example molecules, and (b) scaled with MinMaxScaler. (c) The Pearson correlation coefficient of the whole descriptor set.

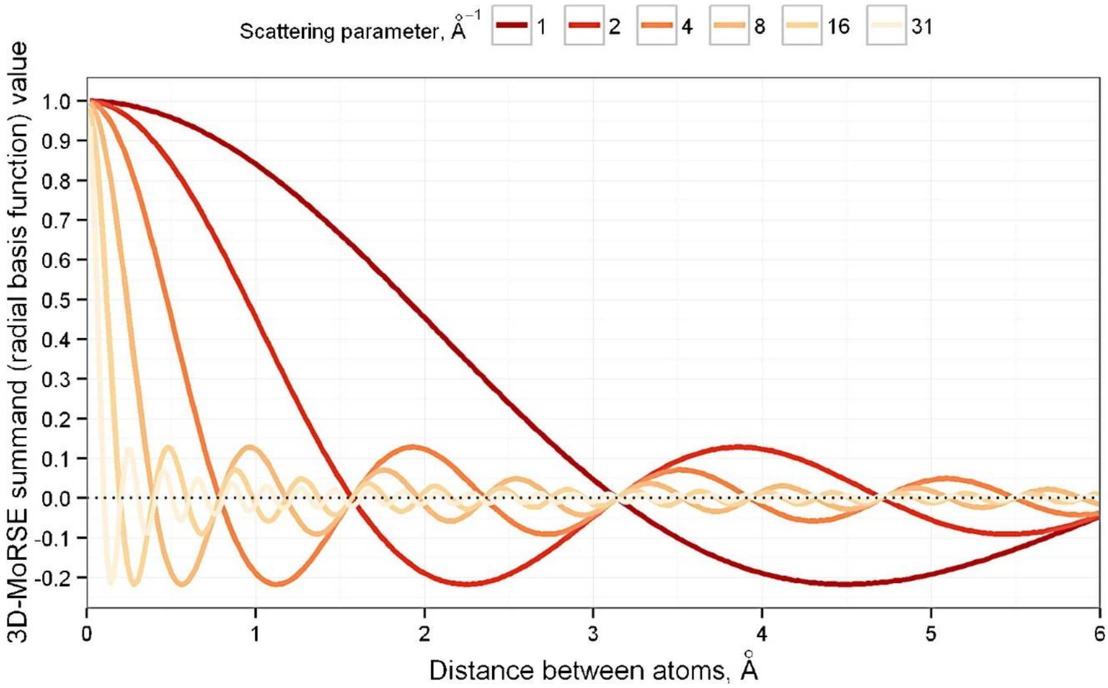

Fig. S6 Standard 3D-MoRSE radial basis function $f(s, r_{ij})$ at different scattering parameters.[1]

Table S2 Prediction R^2, RMSE, and r for all molecular descriptors using ARD regression. For thirteen optimized 3D-MoRSE descriptors, the s_L was consistently set to 0.1 and the s ranged from 0 to 14.

Descriptor	R^2	RMSE (%)	r	Descriptor	R^2	RMSE (%)	r
E-C + PVK	0.72	0.73	0.87	morRC + PVK	0.69	0.77	0.85
morU + PVK	0.74	0.70	0.87	morIS + PVK	0.70	0.76	0.86
morE + PVK	0.76	0.68	0.89	morIP + PVK	0.75	0.69	0.88
morC + PVK	0.69	0.76	0.85	morCa + PVK	0.69	0.76	0.85
morV + PVK	0.69	0.77	0.85	morPI + PVK	0.69	0.77	0.85
morM + PVK	0.66	0.80	0.84	morZV + PVK	0.69	0.76	0.85
morP + PVK	0.70	0.76	0.85	morZC + PVK	0.69	0.76	0.85
PVK only	0.70	0.76	0.86	CPCE only	0.67	0.80	0.84

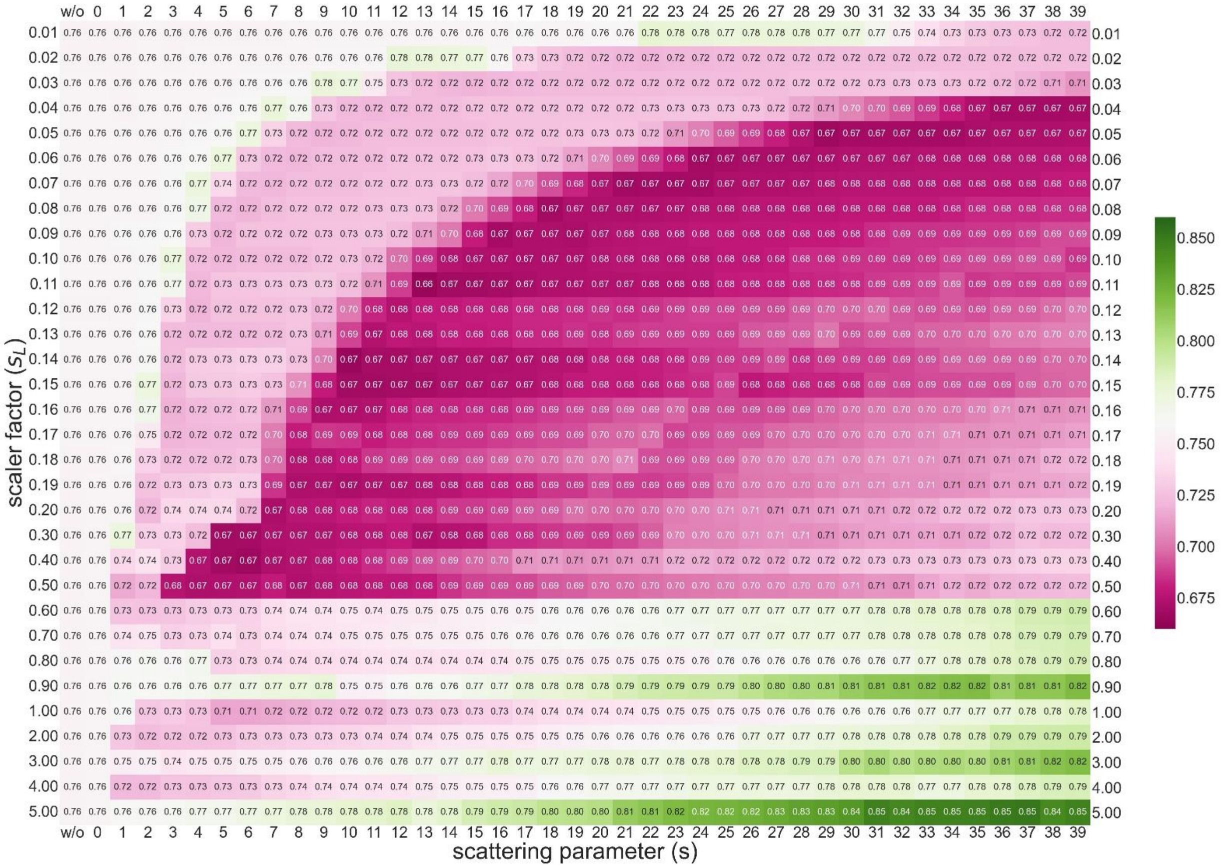

Fig. S7 The RMSE of PCE prediction implemented with ARD regression and morE descriptor set with different dimensions that *s* ranges from 0 to x ($x \leq 39$). The w/o means only PVK descriptors are used.

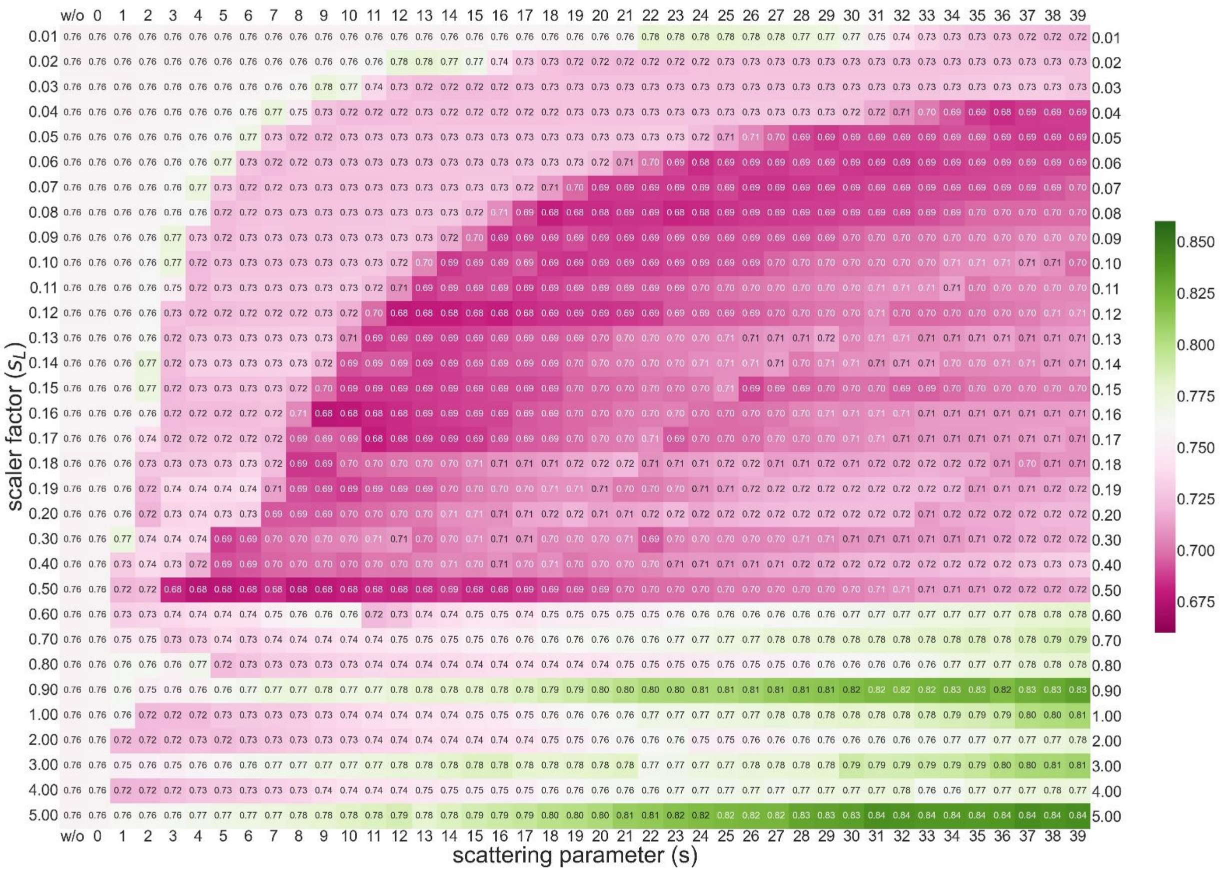

Fig. S8 The RMSE of PCE prediction implemented with ARD regression and morIP descriptor set with different dimensions that *s* ranges from 0 to x ($x \leq 39$). The w/o means only PVK descriptors are used.

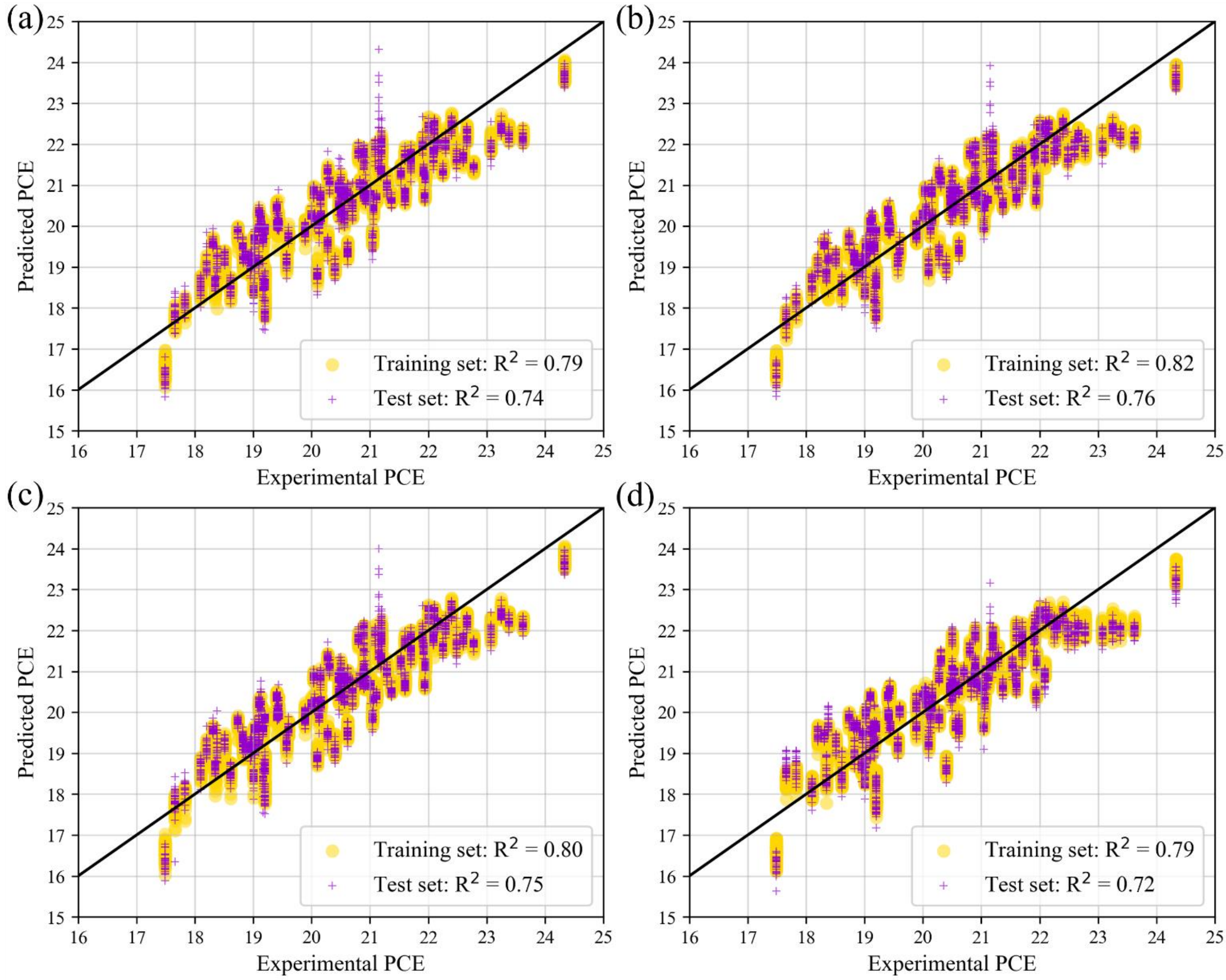

Fig. S9 ML-predicted PCE versus experimental PCE with ARD regression based on (a) morU (0.1,14), (b) morE (0.1,14), (c) morIP (0.1,14), and (d) E-C.

Table S3 The optimized hyperparameters for all ML processes.

M-Des	Algorithm	Hyperparameters
E-C	RF	min_samples_leaf=1, min_samples_split=6, max_depth=7, n_estimators=200
	SVR	C=18, gamma=0.21, epsilon=0.2
	ARD	default
	LASSO	alpha=0.012
morX (0.1,14) (X=U, E, IP)	RF	min_samples_leaf=1, min_samples_split=2, max_depth=5, n_estimators=200
	SVR	C=28, gamma=0.12, epsilon=0.74
	ARD	default
	LASSO	alpha=0.004
morX (0.5,14)	ARD	default
morU (0.21,38)	RF	min_samples_leaf=1, min_samples_split=2, max_depth=10, n_estimators=200
	SVR	C=18, gamma=0.15, epsilon=0.15
	ARD	default
	LASSO	alpha=0.003
morE (0.38,16)	RF	min_samples_leaf=1, min_samples_split=2, max_depth=10, n_estimators=200
	SVR	C=27, gamma=0.30, epsilon=0.22
	ARD	default
	LASSO	alpha=0.002
morIP (0.40,22)	RF	min_samples_leaf=1, min_samples_split=2, max_depth=10, n_estimators=200
	SVR	C=37, gamma=0.03, epsilon=0.13
	ARD	default
	LASSO	alpha=0.003

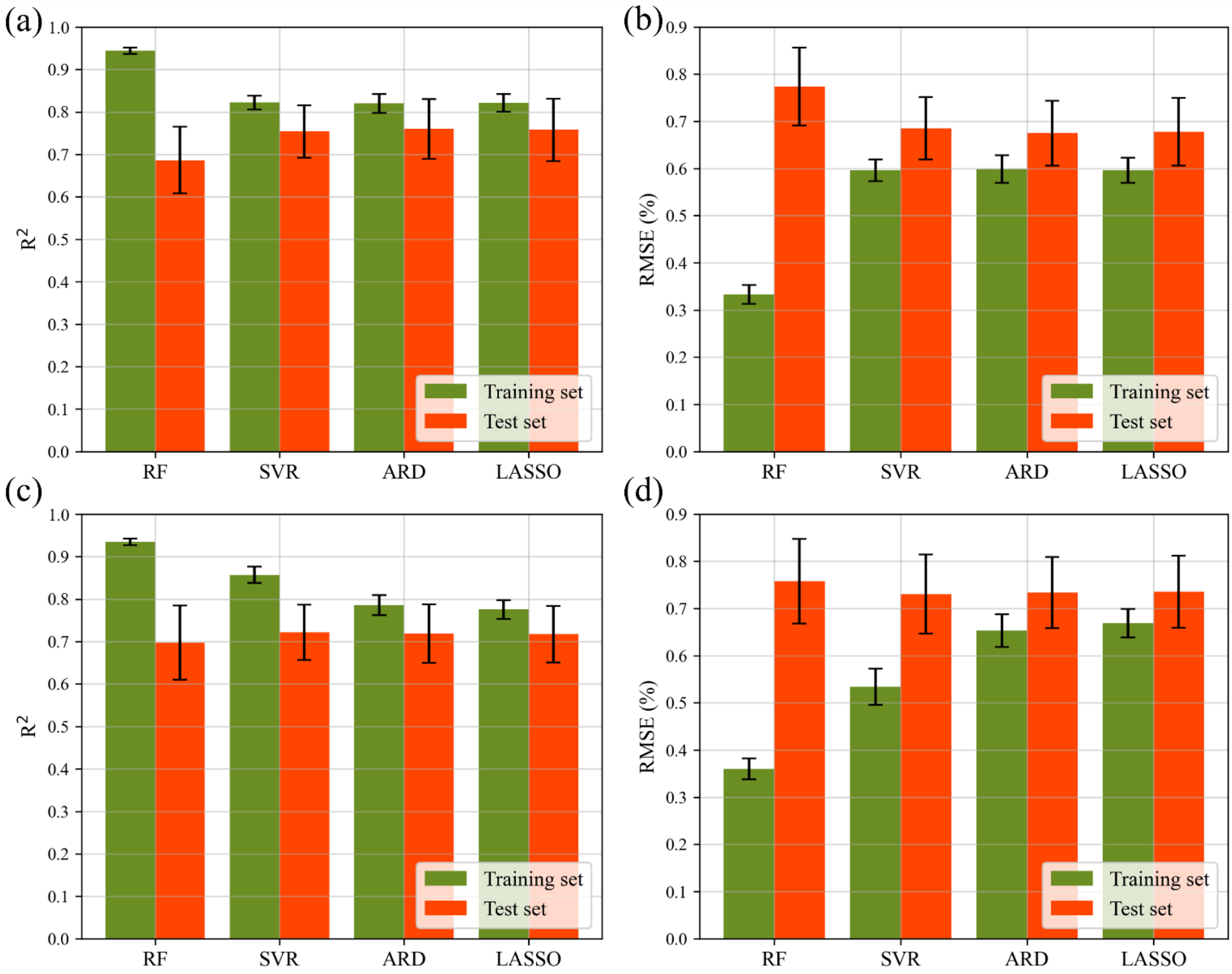

Fig. S10 The training and test RMSE and R^2 based on (a-b) morE (0.1,14) and (c-d) E-C. The test set sampled by 30% from the entire database is mutually exclusive with the training set.

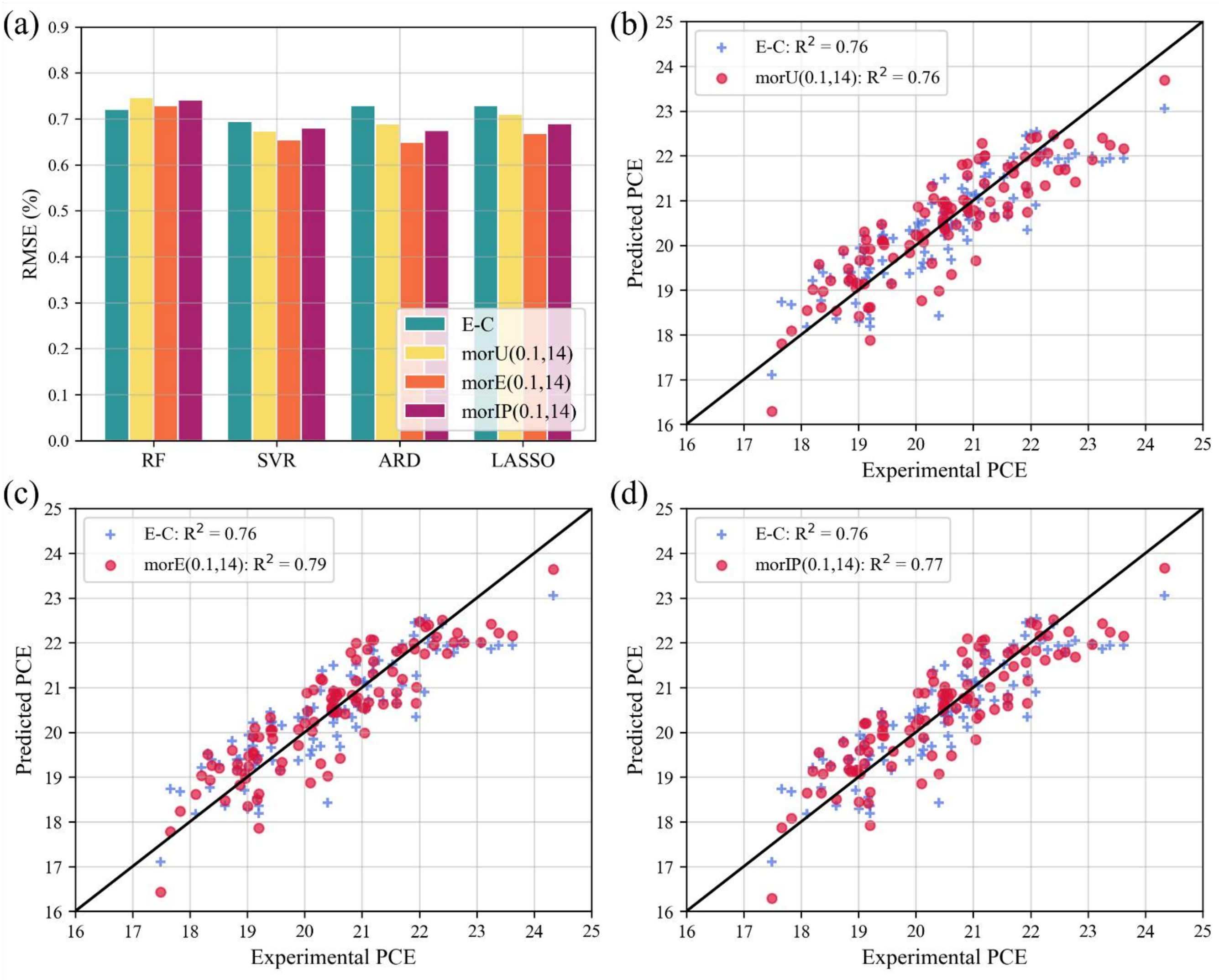

Fig. S11 (a) The RMSE of ML based on different algorithms including RF, SVR, ARD, and LASSO with the LOO method. (b-d) ML-predicted PCE versus experimental PCE based on the LOO method, where ARD regression for optimized 3D-MoRSE descriptors and SVR for E-C.

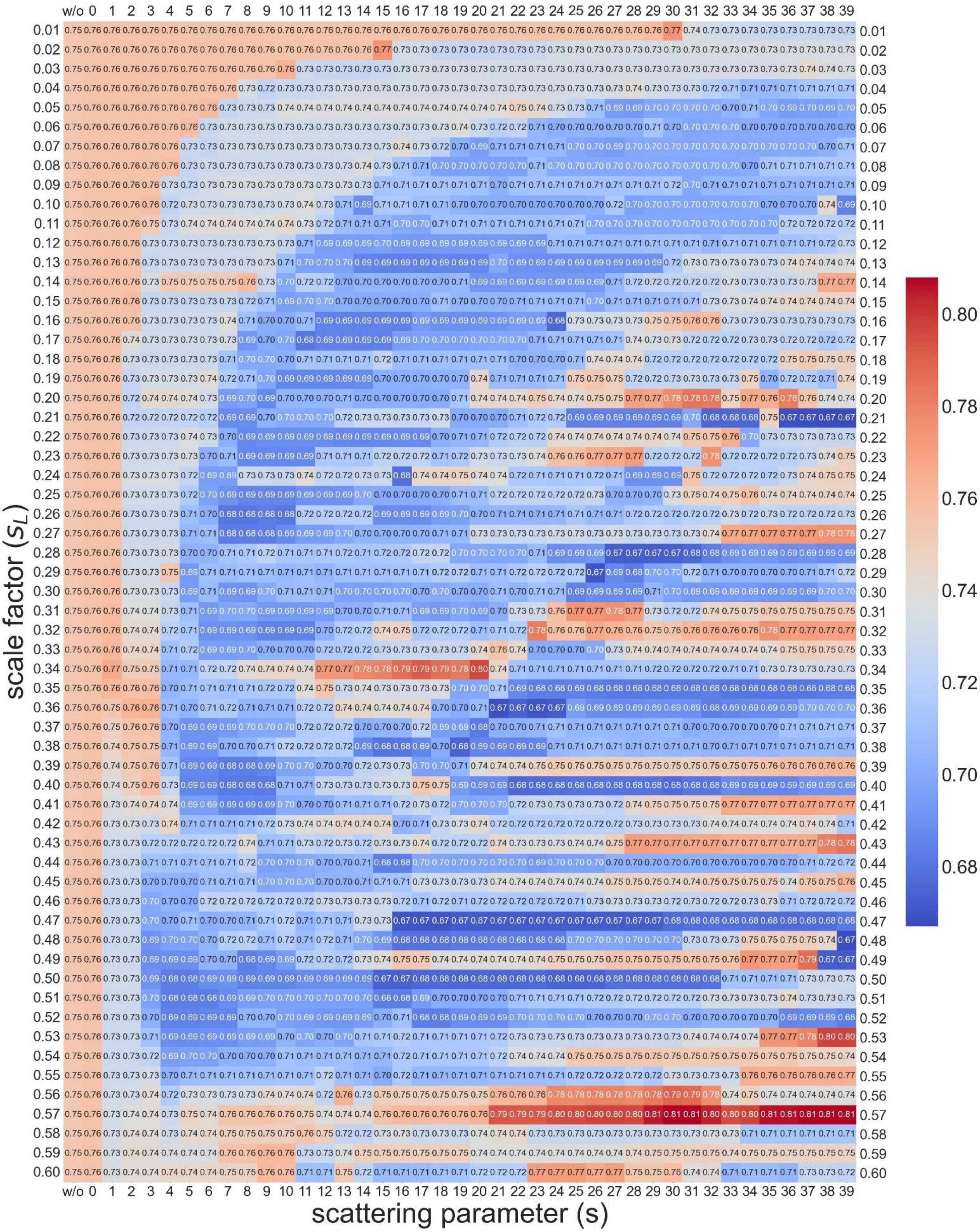

Fig. S12 The RMSE of PCE prediction implemented with ARD regression and morU descriptor set with s_L from 0.01 to 0.60 and dimensions from 0 (w/o) to 40 (s = 39).

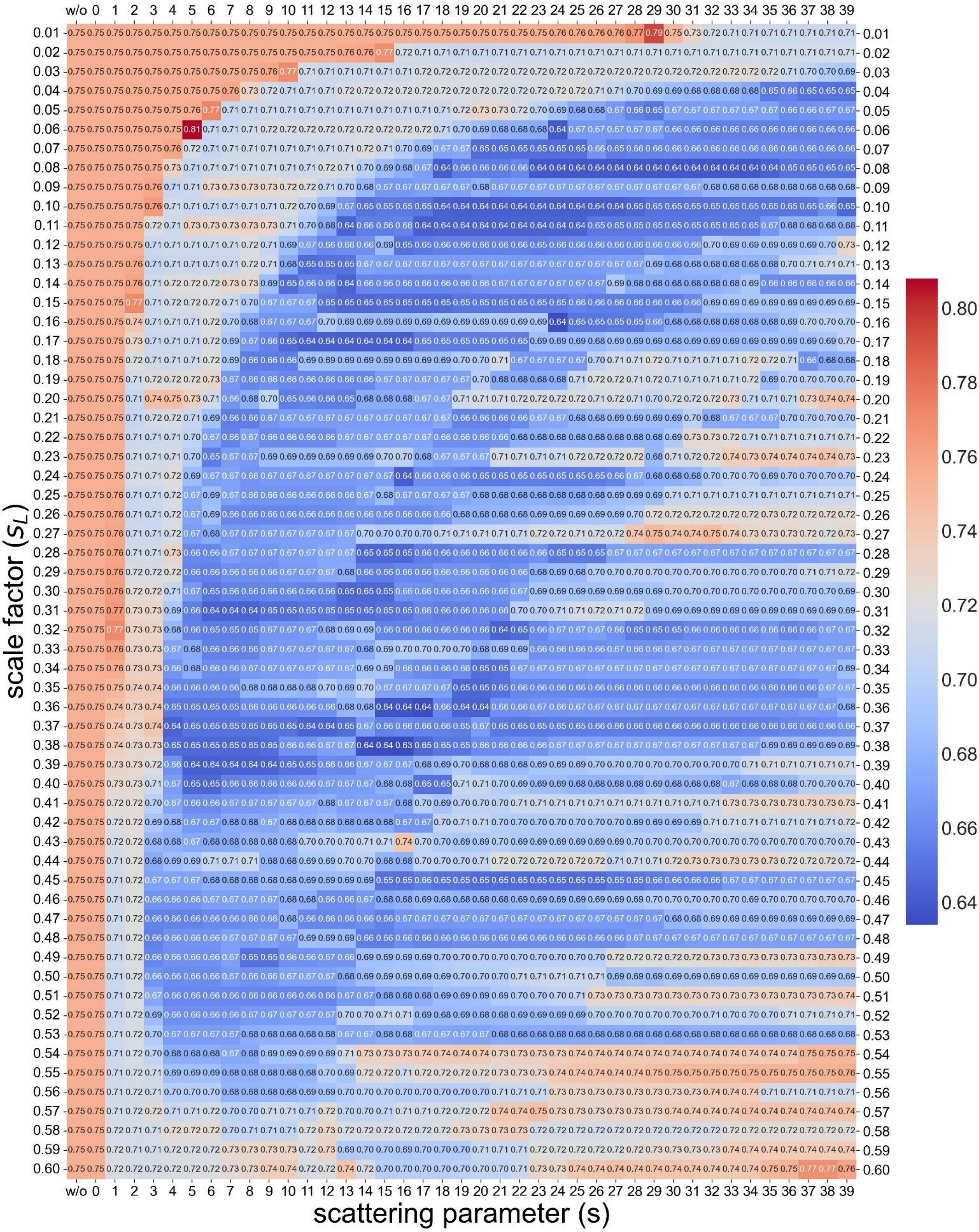

Fig. S13 The RMSE of PCE prediction implemented with ARD regression and morE descriptor set with s_L from 0.01 to 0.60 and dimensions from 0 (w/o) to 40 (s = 39).

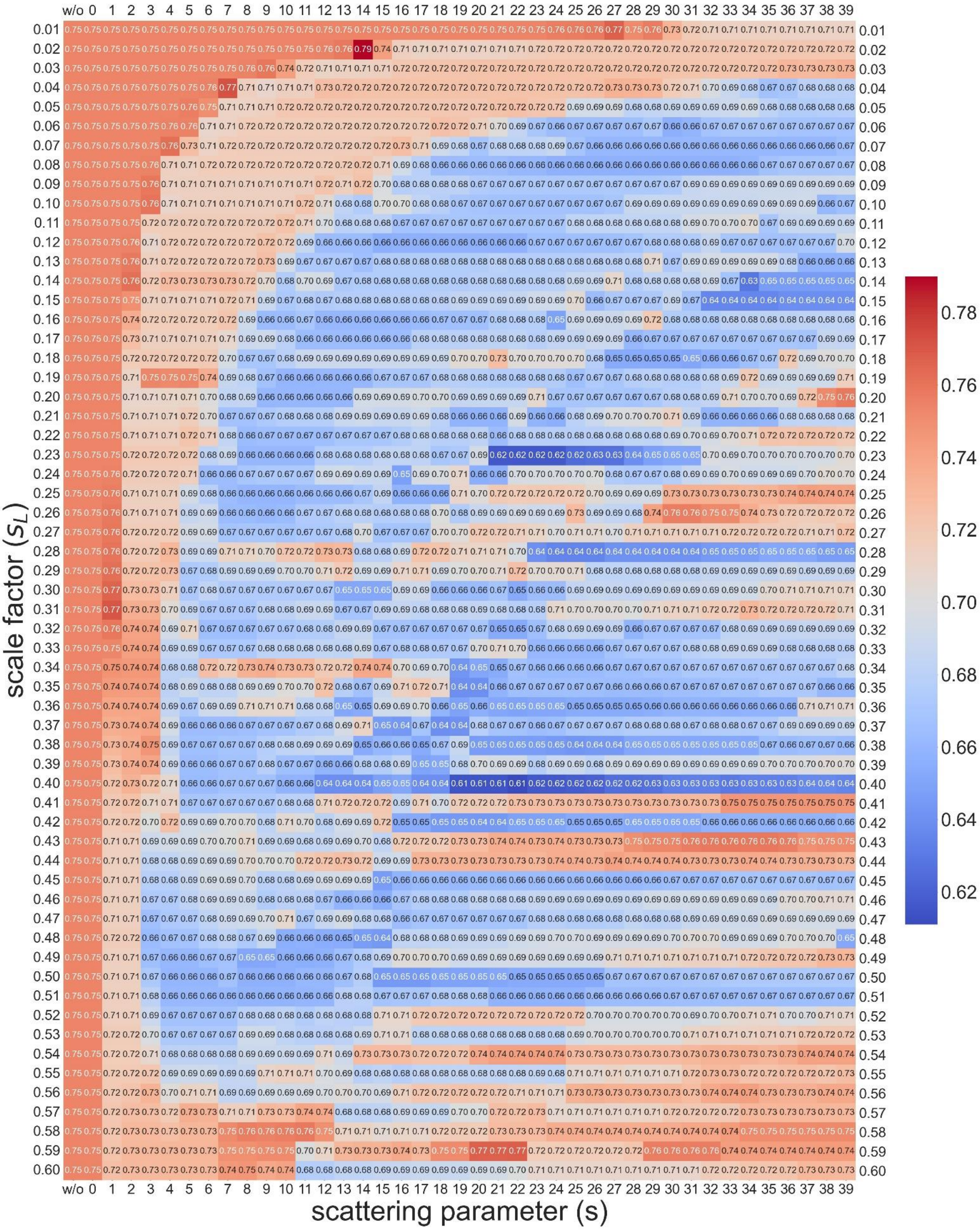

Fig. S14 The RMSE of PCE prediction implemented with ARD regression and morIP descriptor set with s_L from 0.01 to 0.60 and dimensions from 0 (w/o) to 40 (s = 39).

Table S4 The result with RF, ARD, ARD, and LASSO algorithms for morX with the best-optimized s and s_L, for morX(0.1,14), and for E-C. The LOO method and data split method (Test set: 30%) were utilized, respectively.

Algorithm	M-Des	LOO			Test set: 30%		
		R^2	RMSE (%)	r	R^2	RMSE (%)	r
RF	morU (0.1,14)	0.72	0.75	0.85	0.68	0.79	0.84
	morE (0.1,14)	0.74	0.73	0.86	0.68	0.77	0.85
	morIP (0.1,14)	0.73	0.74	0.85	0.68	0.78	0.85
	morU (0.21,38)	0.70	0.78	0.84	0.65	0.82	0.83
	morE (0.38,16)	0.71	0.77	0.84	0.66	0.81	0.83
	morIP (0.40,22)	0.70	0.78	0.84	0.66	0.80	0.83
	E-C	0.74	0.72	0.86	0.70	0.76	0.85
SVR	morU (0.1,14)	0.77	0.67	0.88	0.73	0.71	0.87
	morE (0.1,14)	0.79	0.66	0.89	0.75	0.69	0.88
	morIP (0.1,14)	0.77	0.68	0.88	0.74	0.70	0.87
	morU (0.21,38)	0.77	0.68	0.88	0.69	0.77	0.85
	morE (0.38,16)	0.79	0.65	0.89	0.72	0.73	0.86
	morIP (0.40,22)	0.78	0.66	0.89	0.73	0.72	0.87
	E-C	0.76	0.69	0.87	0.72	0.73	0.87
ARD	3DM-U	0.46	1.04	0.74			
	3DM-E	0.24	1.24	0.69			
	3DM-IP	0.18	1.28	0.68			
	morU (0.1,14)	0.76	0.69	0.87	0.74	0.70	0.88
	morE (0.1,14)	0.79	0.65	0.89	0.76	0.67	0.89
	morIP (0.1,14)	0.77	0.68	0.88	0.76	0.68	0.88
	morU (0.21,38)	0.78	0.67	0.88	0.72	0.73	0.86
	morE (0.38,16)	0.80	0.63	0.90	0.76	0.67	0.89
	morIP (0.40,22)	0.82	0.61	0.90	0.74	0.70	0.87
	E-C	0.74	0.73	0.86	0.72	0.73	0.86
LASSO	morU (0.1,14)	0.75	0.71	0.87	0.73	0.72	0.87
	morE (0.1,14)	0.78	0.67	0.88	0.76	0.68	0.88
	morIP (0.1,14)	0.76	0.69	0.87	0.74	0.70	0.88
	morU (0.21,38)	0.77	0.69	0.88	0.72	0.72	0.87
	morE (0.38,16)	0.79	0.66	0.89	0.74	0.70	0.88
	morIP (0.40,22)	0.80	0.64	0.89	0.75	0.68	0.88
	E-C	0.74	0.73	0.86	0.72	0.74	0.86

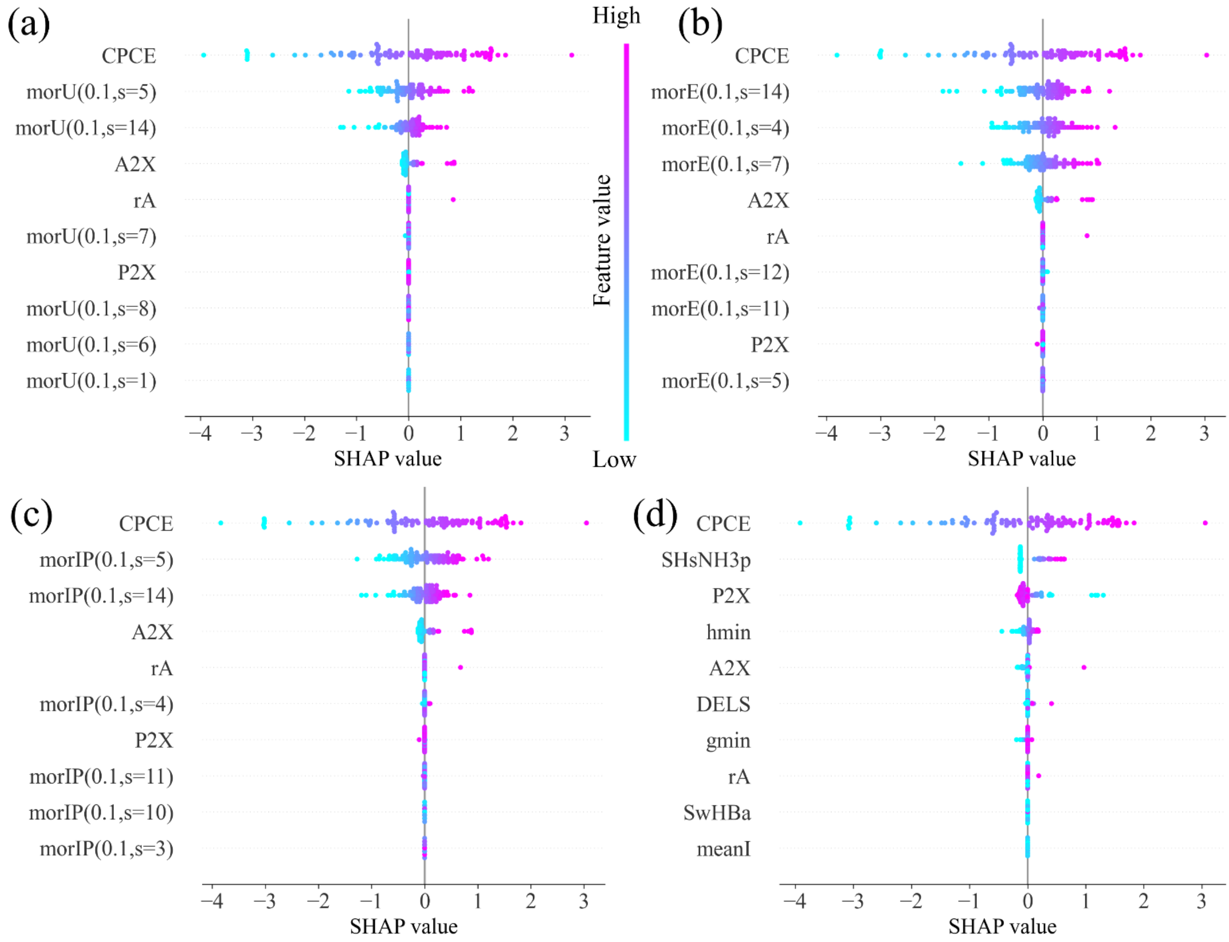

Fig. S15 The SHAP summary plot for (a) morU (0.1,14), (b) morE (0.1,14), (c) morIP (0.1,14), and E-C descriptor sets implemented in ML with ARD regression. The SHAP values for each sample are obtained by one fitting in the LOO process.

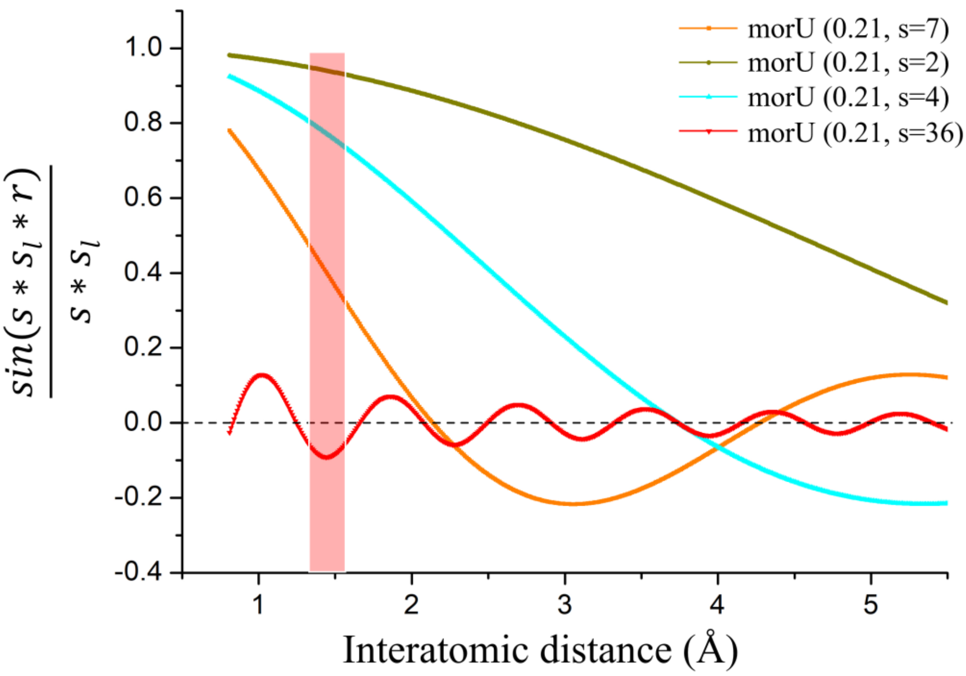

Fig. S16 The descriptor value of morU (0.21, s=7), morU (0.21, s=2), morU (0.21, s=4) and morU (0.21, s=36). As indicated by SHAP analysis, higher morU (0.21, s=7), morU (0.21, s=2), morU (0.21, s=4) and lower morU (0.21, s=36) should realize high PCE, the first region is shaded red.

Table S5 The result was predicted by adding molecular descriptors to morIP (0.40,22) and with the LOO method and ARD algorithm.

Molecular descriptors	R^2	RMSE	r
SHBd, SwHBa, SHsNH3p, SHCsats +morIP(0.40,22)	0.78	0.66	0.89
meanI, gmax, gmin, hmax, hmin+morIP(0.40,22)	0.79	0.65	0.89
DELS, fragC+morIP(0.40,22)	0.81	0.62	0.90
TPSA, AvgIpc, MolLogP+morIP(0.40,22)	0.81	0.62	0.90
BCUT2D_MRHI, BCUT2D_CHGLO, Kappa3+morIP(0.40,22)	0.77	0.67	0.88
DM, RE+morIP(0.40,22)	0.81	0.61	0.90
All in E-C+morIP(0.40,22)	0.75	0.71	0.87
All in RDKit+morIP(0.40,22)	0.79	0.65	0.89

Table S6 The molecular descriptors associated with passivation and the best prediction of their values by optimized 3D-MoRSE descriptors, when the scale factor was set from 0.1 to 0.5. ARD algorithm and LOO method were used. The DM and RE were calculated at the level b3lyp/Def2TZVP by Gaussian16[5].

Molecular descriptors	Illustration	3D-MoRSE	R^2	RMSE	r
SHBd	The sums of E-states of H-bond donors	morE (0.5,38)	0.60	8.14	0.78
SwHBa	The sums of E-states of weak H-bond acceptors	morZV (0.5,34)	0.75	8.15	0.87
SHsNH3p	The sums of E-states of H-bonds from NH_3^+	morIP (0.3,32)	0.62	0.13	0.70
SHCsats	The sums of E-states of H bonded to saturated C	morRC (0.4,29)	0.87	1.13	0.93
meanI	Mean intrinsic state values I	morIS (0.1,31)	0.95	0. 19	0.97
gmax	Maximum E-state	morZC (0.1,30)	0.47	2.90	0.69
gmin	Minimum E-state	morP (0.3,33)	0.66	1.17	0.81
hmax	Maximum H E-state	morP (0.3,31)	0.80	0.17	0.89
hmin	Minimum H E-state	morP (0.4,29)	0.87	0.10	0.93
DELS	The sum of all atom intrinsic state differences	morZV(0.3,22)	0.94	6. 47	0.97
fragC	The complexity of the material	morZC (0.3,38)	0.91	70.45	0.96
TPSA	Topological polar surface area	morZC (0.4,28)	0.37	21.40	0.65
AvgIpc	Average information content of the characteristic polynomial coefficients of the adjacency matrix	morZV (0.5,36)	0.67	0.09	0.82
MolLogP	Molecular lipid-water partition coefficient	morP (0.1,35)	0.99	4.77	1.00
BCUT2D_MRHI	Crippen MR eigenvalue high	morM (0.3,30)	0.65	0.90	0.83
BCUT2D_CHGLO	Gasteiger charge low	morU (0.3,39)	0.91	0.08	0.95
Kappa3	Analysis of the third-order neighborhoods of atoms in the molecular graph	morIP (0.3,18)	0.67	906.12	0.82
DM	Dipole moment	morE (0.3,36)	0.20	6.18	0.46
RE	Reorganization energy	morV (0.5,39)	0.69	0.56	0.83

Adsorption Energy and Transferred Charge Prediction

Adsorption energy and work function have a certain impact on the passivation effect[6, 7]. Among passivation molecules in the dataset, 86 molecules are small enough to put on the PVK surface model described in our previous work, and their adsorption energies (E_{ads}) and transferred charge (N_{tran}) have been calculated by the previous method[7]. Intermolecular 3D-MoRSE has shown good performance in the prediction of electronic couplings[8]. Therefore, intermolecular optimized 3D-MoRSE may have a good performance on the prediction of E_{ads} and N_{tran}. The expression of intermolecular optimized 3D-MoRSE is $\sum_{i=1}^{N_i}\sum_{j=1}^{N_j} A_i A_j f(s, r_{ij})$, where $f(s, r_{ij}) = \sin(s * s_L * r_{ij})/(s * s_L * r_{ij})$. N_i represents the number of atoms in molecules, N_j represents the number of atoms on the PVK surface. As for E_{ads}, based on morRC (33,37), R^2 of 0.87 and RMSE of 0.21 have been obtained; Based on morM (13,27), R^2 of 0.86 and RMSE of 0.21 have been obtained. Therefore, a covalent radius has a direct impact on E_{ads}. As for N_{tran}, based on morM (31,11), R^2 of 0.87 and RMSE of 0.04 have been obtained. The morM descriptor performs well on both of them, which may be due to morM gives Pb atom greater weight. The interaction of Pb with passivation molecules is the most significant[7].

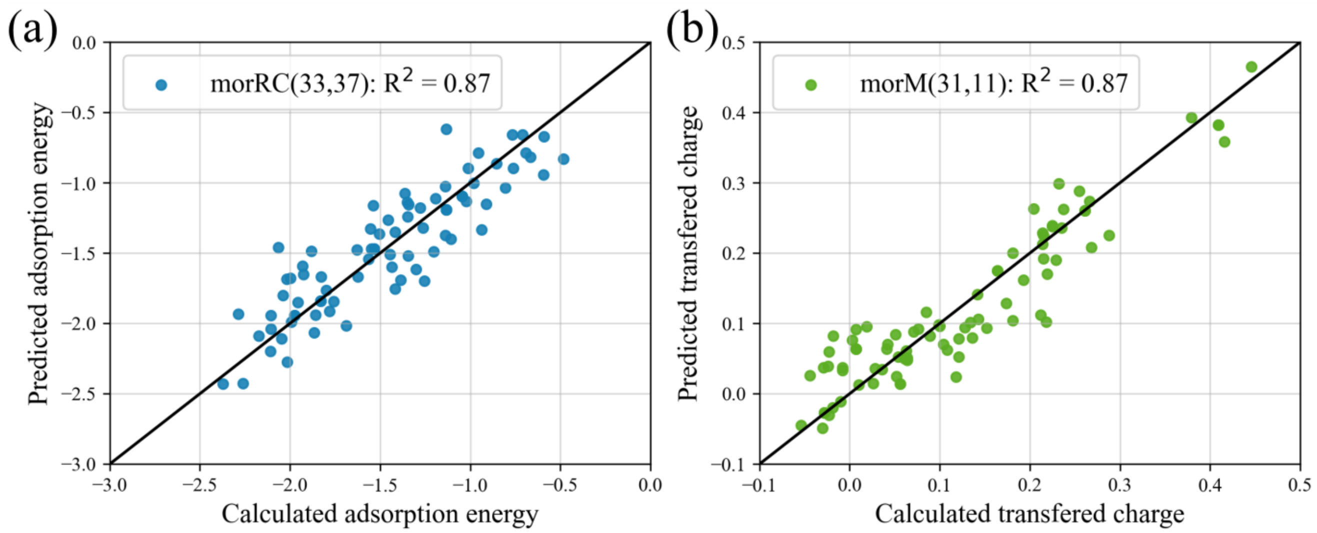

Fig. S17 ARD model predicted value versus the calculated value of (a) E_{ads} and (b) N_{tran} based on the LOO method.

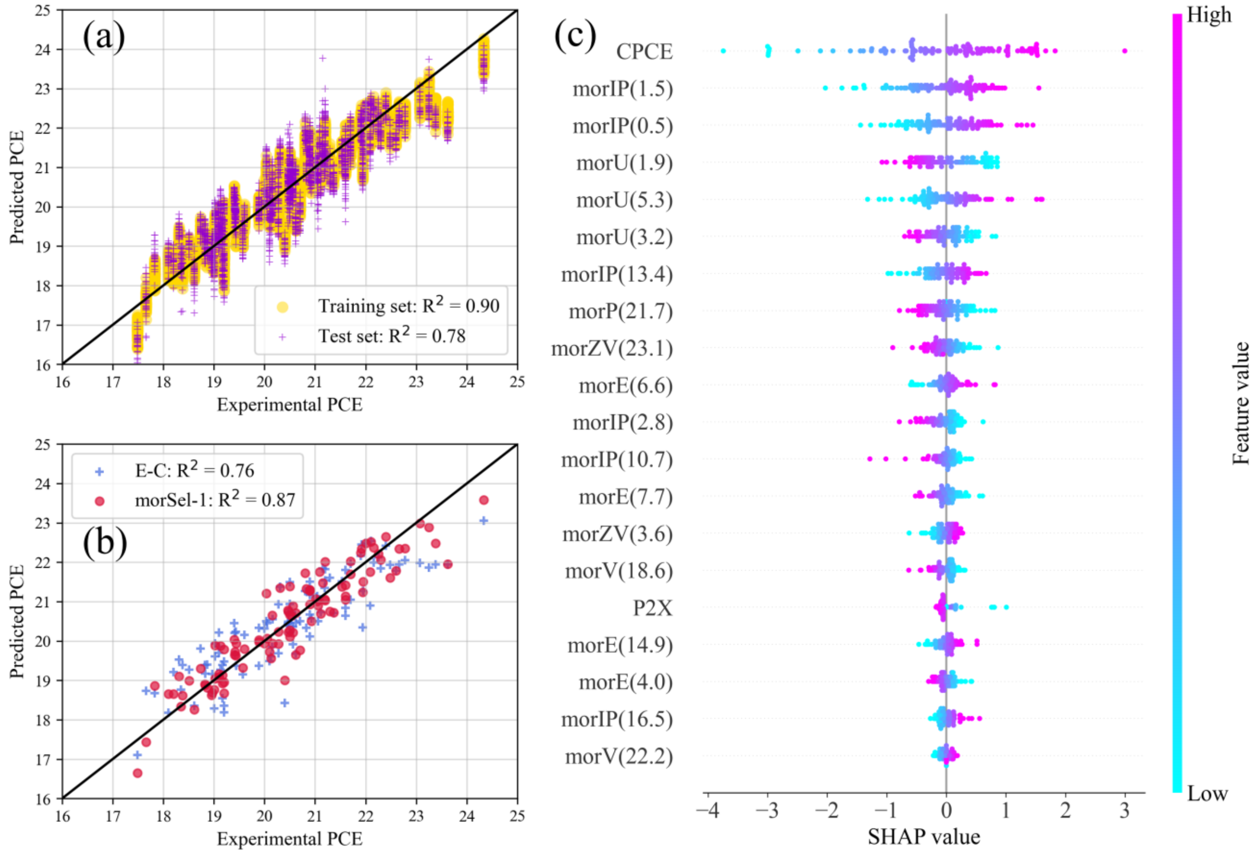

Fig. S18 The predicted PCE versus experimental PCE based on morSel-1 with (a) 30% samples as test set, and (b) LOO. (c) The SHAP summary plot; The value in the bracket is $s * s_L$.

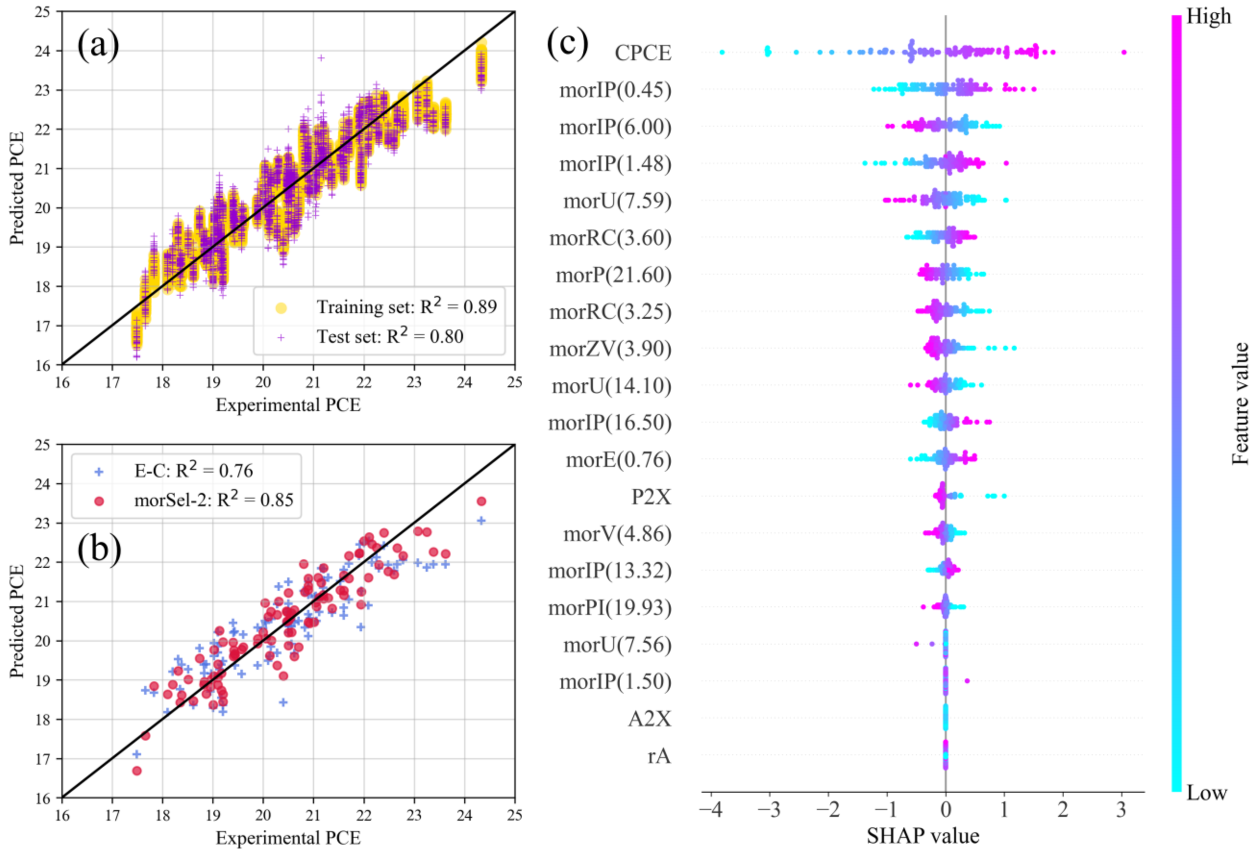

Fig. S19 The predicted PCE versus experimental PCE based on morSel-2 with (a) 30% samples as test set, and (b) LOO. (c) The SHAP summary plot; The value in the bracket is $s * s_L$.

Attempt of GNN Model

As a promising direction, graph neural networks (GNN) have been widely applied in molecular information extraction and property prediction.[9-11] GNN extracts molecular features based on molecular graphs containing nodes and edges as input. Nodes represent atoms, and edges represent connectivity between different atoms. For comparison and exploration of the potential application of optimized 3D-MoRSE, we set two examples that atomic weights and atomic 3D-MoRSE values were set as node features, respectively, shown in Fig. S20. Torch and torch-geometric packages were utilized[12]. GraphSAGE (Sample and aggregate) model[13] was used in graph convolution layers, and global mean pooling was used in the global pooling layer. The model was evaluated 20 times with different data splits each time (test set: 30%) and the results were averaged. For example A, thirteen atomic weights as shown in Table S1 were set as node features. Two graph convolution layers were added. The R^2 reached 0.65 and RMSE reached 0.83. For example B, the node features were defined as $\sum_{j=1}^{N} A_i A_j f(s, r_{ij})$ for ith atom, where $f(s, r_{ij}) = \sin(s * s_L * r_{ij})/(s * s_L * r_{ij})$ when $i \neq j$, and $f(s, r_{ij}) = 0$ when $i = j$. One graph convolution layer was added. The R^2 reached 0.63 and RMSE reached 0.86. As our dataset is too small, applying GNN is relatively inappropriate and did not perform well. However, a potential application for optimized 3D-MoRSE was provided.

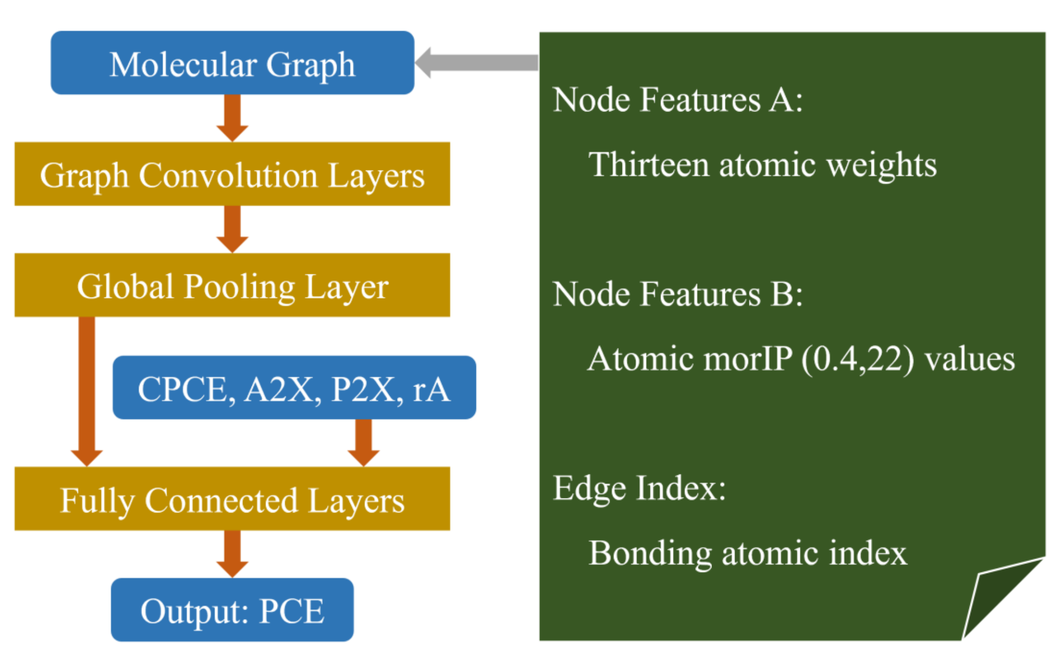

Fig. S20 Schematic diagram of the GNN models.

Reference:

1. O. Devinyak, D. Havrylyuk and R. Lesyk, *Journal of Molecular Graphics and Modelling*, 2014, **54**, 194-203.
2. W. Liu, Y. Lu, D. Wei, X. Huo, X. Huang, Y. Li, J. Meng, S. Zhao, B. Qiao, Z. Liang, Z. Xu and D. Song, *Journal of Materials Chemistry A*, 2022, **10**, 17782-17789.
3. L. H. Hall, B. Mohney and L. B. Kier, *Journal of Chemical Information and Computer Sciences*, 1991, **31**, 76-82.
4. C.-G. Zhan, J. A. Nichols and D. A. Dixon, *The Journal of Physical Chemistry A*, 2003, **107**, 4184-4195.
5. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman and D. J. Fox, *Journal*, 2016.
6. W. Zhang, Q.-S. Li and Z.-S. Li, *Applied Surface Science*, 2021, **563**, 150267.
7. X. Ye, W. Ou, B. Ai and Y. Zhou, *Physical Chemistry Chemical Physics*, 2023, **25**, 32250-32260.
8. J. Ma, Z. Du, Z. Lei, L. Wang, Y. Yu, X. Ye, W. Ou, X. Wei, B. Ai and Y. Zhou, *Journal of Chemical Information and Modeling*, 2023, **63**, 5089-5096.
9. X. Fang, L. Liu, J. Lei, D. He, S. Zhang, J. Zhou, F. Wang, H. Wu and H. Wang, *Nature Machine Intelligence*, 2022, **4**, 127-134.
10. Y. Wang, J. Wang, Z. Cao and A. Barati Farimani, *Nature Machine Intelligence*, 2022, **4**, 279-287.
11. C. W. Coley, R. Barzilay, W. H. Green, T. S. Jaakkola and K. F. Jensen, *J Chem Inf Model*, 2017, **57**, 1757-1772.
12. M. Fey and J. E. Lenssen, *Journal*, 2019, DOI: 10.48550/arXiv.1903.02428, arXiv:1903.02428.
13. W. L. Hamilton, R. Ying and J. Leskovec, *Journal*, 2017, DOI: 10.48550/arXiv.1706.02216, arXiv:1706.02216.