Supporting Information

Deep transfer learning for predicting frontier orbital energies of organic materials using small data and its application to porphyrin photocatalysts

An Su[*], Xin Zhang, Chengwei Zhang, Debo Ding, Yun-Fang Yang, Keke Wang, Yuan-Bin She[*]

College of Chemical Engineering, Zhejiang University of Technology, Hangzhou 310014, P. R. China

* Corresponding authors

Prof. An Su

Email: ansu@zjut.edu.cn

Prof. Yuan-Bin She

Email: sheyb@zjut.edu.cn

College of Chemical Engineering, Zhejiang University of Technology, Hangzhou 310014, P. R. China

1. Supporting information for Metalloporphyrins and Porphyrins Database (MpPD)

We constructed the MpPD containing 377 porphyrin structures by DFT calculation and literature review. The general porphyrin structure is shown in Figure S1. In the MpPD dataset, there are 69 A_4-type porphyrin molecules that have been synthesized by our group[1] and 16 porphyrin structures that were retrieved from the literature[2]. We designed 292 A_2B_2-type porphyrins by combining the side chain groups of these porphyrins with some common electron-absorbing and donating groups, a total of 11 substituent groups, and several metals such as Fe, Ni, Cu, Mn, and Co in different combinations structures.
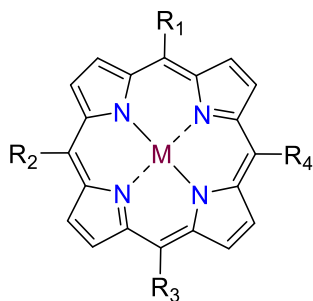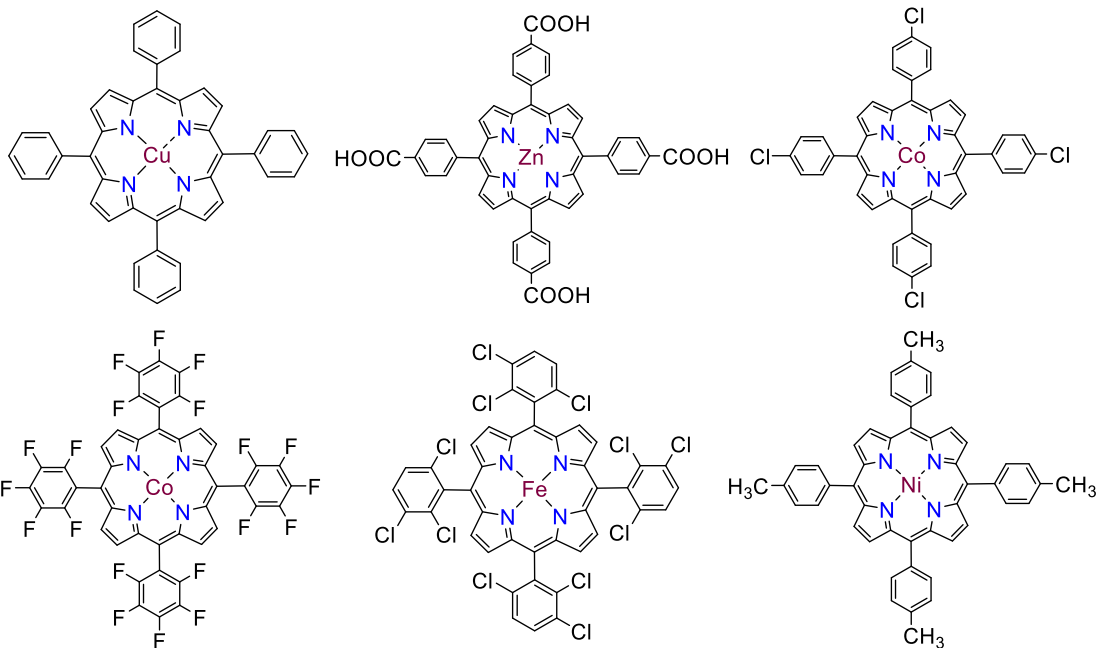

Figure S1. General porphyrin structure in MpPD

A_4-type Porphyrin Structure. A_4-type porphyrins have all four side-chain substituents identical. In other words, the R1, R2, R3, and R4 groups are the same in Figure S1. 69 A_4-type porphyrin structures are designed or collected by combining 26 substituents with 6 central metals, and the representative structure is shown in Figure S2. The structure of the substituents is visualized in Figure S3, and the central metal is shown in Figure S4.
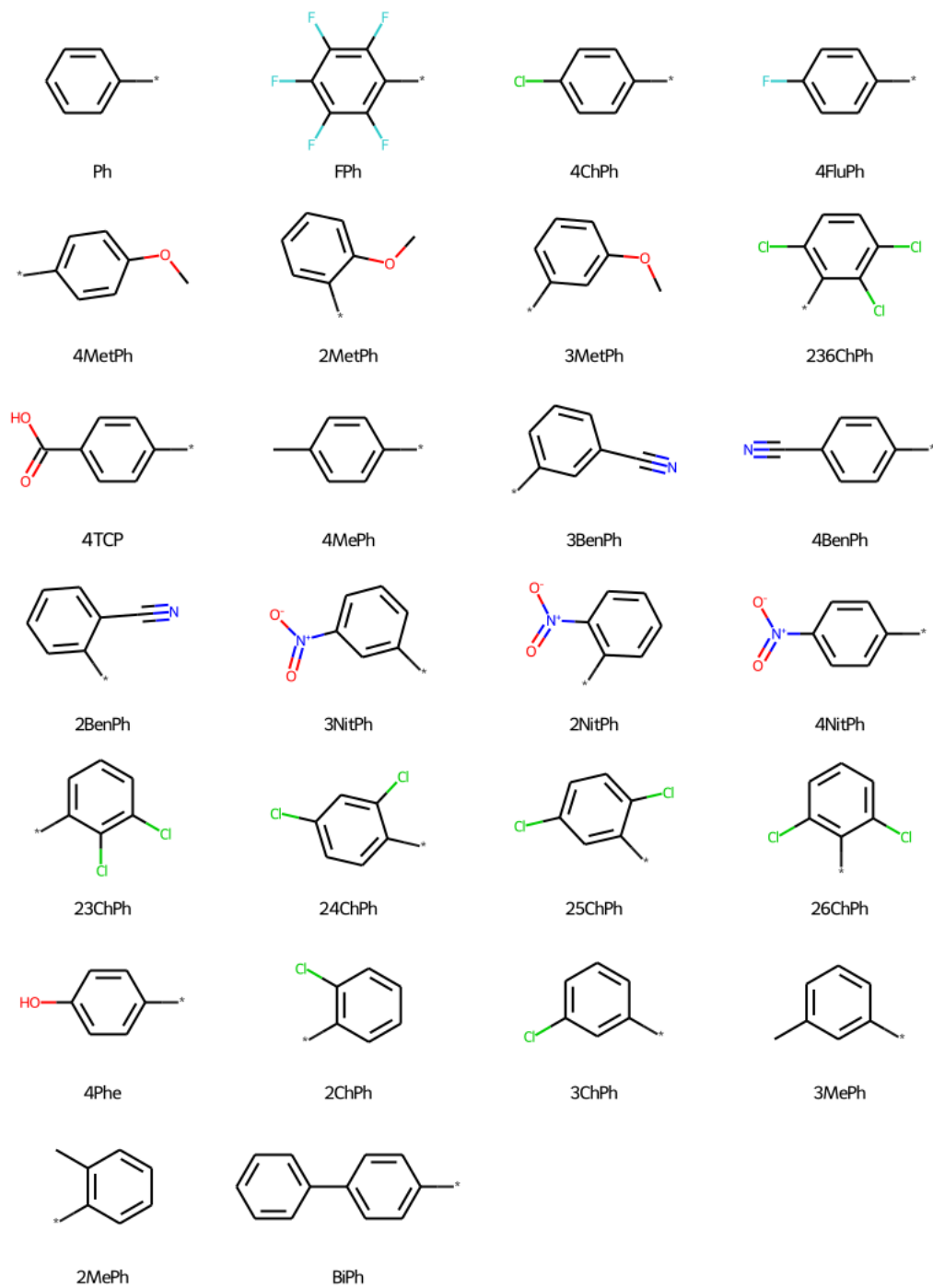
20

21   **Figure S2**. Representative structures of A$_4$-type porphyrins in MpPD

22

Ph   FPh   4ChPh   4FluPh

4MetPh   2MetPh   3MetPh   236ChPh

4TCP   4MePh   3BenPh   4BenPh

2BenPh   3NitPh   2NitPh   4NitPh

23ChPh   24ChPh   25ChPh   26ChPh

4Phe   2ChPh   3ChPh   3MePh

2MePh   BiPh

Figure S3. Structures of A_4-type porphyrin side-chain substituents in MpPD

Cu   Zn   Co   Fe   Mn   Ni

Figure S4. Central metals of A_4-type porphyrins in MpPD

28 **A2B2-type Porphyrin Structure.** Using the 16 porphyrin side chain groups retrieved

29 from the literature, a total of 11 substituents in combination with some common

30 electron-absorbing and donating groups and central metals including Fe, Ni, Cu, Mn, and

31 Co, 292 porphyrin structures of A2B2 type were designed. The A2B2 type porphyrin

32 structures have R1 and R3 side-chain substituents the same and R2 and R4 the same (

33 Figure S1). A representative A2B2-type porphyrin structure is shown in Figure S5, and

34 the substituent structure is shown in Figure S6. The central metal is shown in Figure S7.
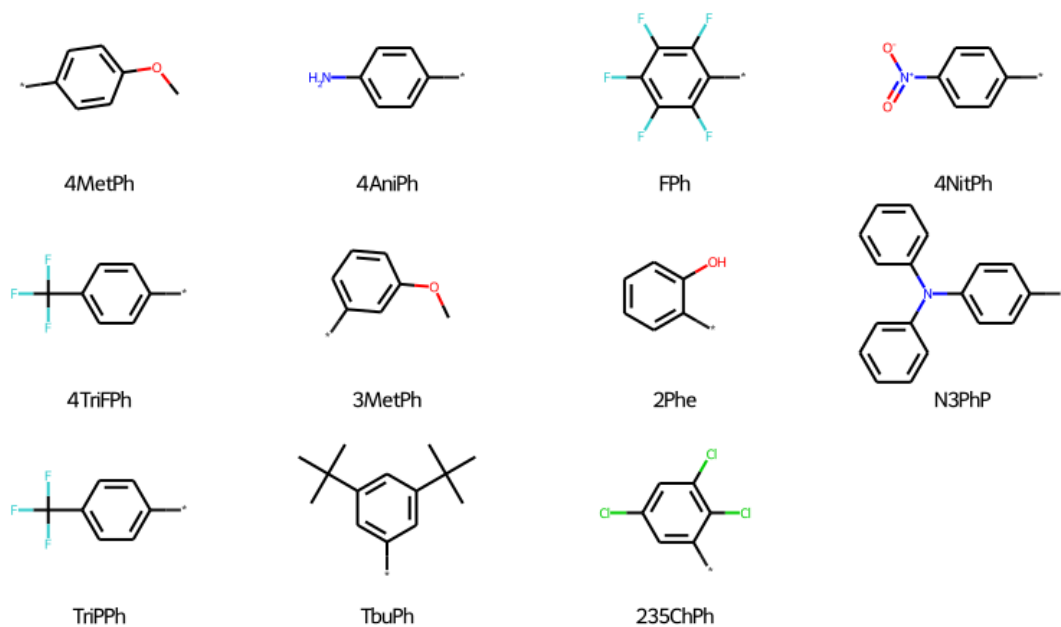
35
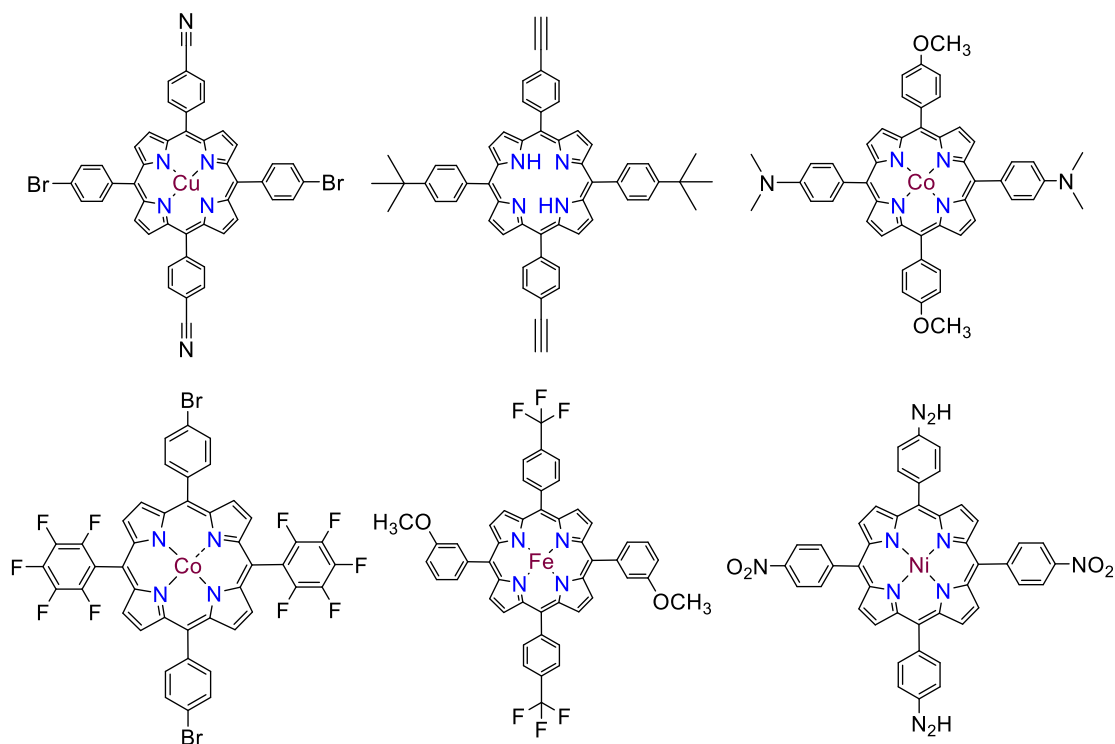
36

37 **Figure S5**. Representative structures of A2B2-type porphyrins in MpPD

38

4MetPh   4AniPh   FPh   4NitPh

4TriFPh   3MetPh   2Phe   N3PhP

TriPPh   TbuPh   235ChPh

Figure S6. Side-chain substituents of A_2B_2-type porphyrins in MpPD

Cu   Co   Fe   Mn   Ni

Figure S7. Central metals of A_2B_2 porphyrins in MpPD
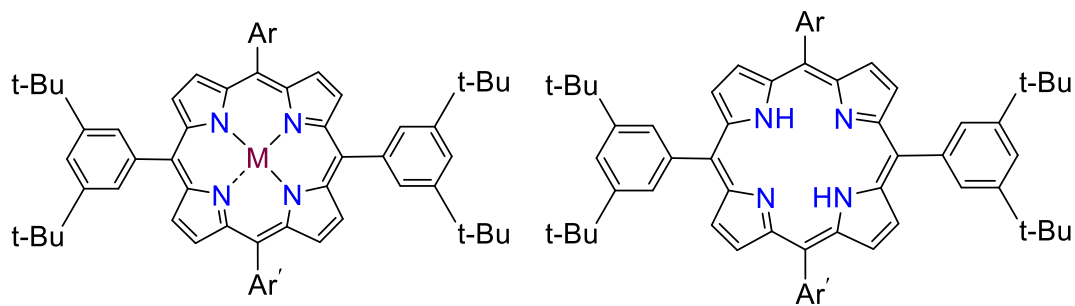
51   **Porphyrin Structures from literature.** The 16 porphyrin structures retrieved from the

52   literature are shown in Figure S8.[2]

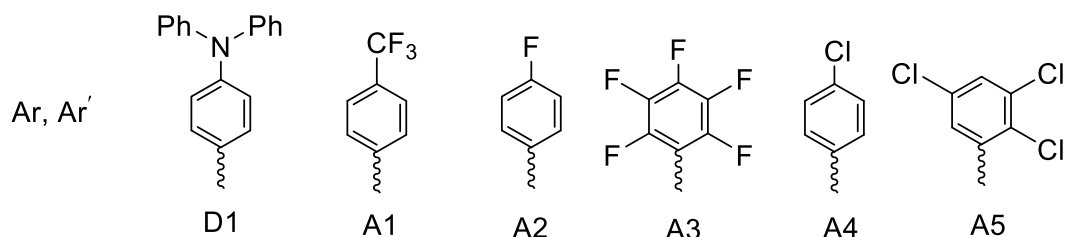53

PP: **DD**   (Ar=Ar′=D1)   **PP-M**

         AA1-AA5 (Ar=Ar′=A1-A5)   (M=Ni, Co, Zn)

         DA1   (Ar=D1, Ar′=A1)

54

55   **Figure S8**. Structures of porphyrins retrieves from literature[2]

56

57

58

59

60

61   **Statistical summary of DFT calculations of MpPD.** A statistical summary of the results

62   of the DFT calculations is provided in Table S2, which includes the mean and standard

63   deviation (std) of HOMO, LUMO and E_gap, as well as the quartiles, maximum, and

64   minimum values.

65   **Table S2.** MpPD data set statistics

	HOMO (eV)	LUMO (eV)	E_gap (eV)
count	377	377	377
mean	-5.5049	-2.8510	2.6544
std	0.3630	0.3602	0.2379
min	-6.2695	-3.7337	1.7799
25%	-5.7315	-3.1372	2.5176
50%	-5.4817	-2.8450	2.7176
75%	-5.2500	-2.5731	2.8345
max	-4.6831	-2.0248	2.9982

66

67

68

69

70

71
72
73

2. Supporting Information for Porphyrin-based Dyes Database (PBDD)

For the model pre-training data, we used the Porphyrin-based Dyes Database (PBDD) from the Computational Materials Repository (CMR)[3]. Currently, 12,096 porphyrin molecular structures and information such as HOMO, LUMO, and energy gap information are stored (Table S1).

Table S1. Molecular information in PBDD

key	description	unit
A	Anchor group	
E_HOMO	HOMO location calculated as ionization potential	eV
E_LUMO	LUMO location calculated as electron affinity	eV
E_gap	Electronic gap calculated as E_LUMO – E_HOMO	eV
ID	The unique number of the molecule in the database	
M	Metal center	
Mass	Relative molecular weight	
R1	First side group	
R2	Second side group	
R3	Third side group	

3. Hyperparameters optimization of PorphyBERT and PorphyDMPNN and learning curves of optimized models

We tuned the hyperparameters of the PorphyBERT and PorphyDMPNN models, mainly on the epoch and dropout. The optimization results are listed in Table S3 and Table S4. With the optimized hyperparameters, learning curves were plotted in Table S5 and Table S6.

Table S3. Parameters tuning of PorphyBERT and the performance on training, validation and test sets

Epoch	Dropout	Train RMSE (eV)	Train MAE (eV)	Validation RMSE (eV)	Validation MAE (eV)	Test RMSE (eV)	Test MAE (eV)
E_gap							
30	0.4	0.0598	0.0466	0.1043	0.0646	0.0933	0.0589
38	0.4	0.0426	00269	0.0901	0.0412	0.0378	0.0261
50	0.4	0.0904	0.0700	0.1177	0.0884	0.1005	0.0853
HOMO							
30	0.4	0.0498	0.0288	0.0871	0.0458	0.0431	0.0314
50	0.4	0.1419	0.1170	0.1465	0.1202	0.1820	0.1507
LUMO							
30	0.4	0.0867	0.0632	0.1520	0.0999	0.1104	0.0793
39	0.2	0.1388	0.1182	0.1757	0.1378	0.1771	0.1530
43	0.5	0.0685	0.0522	0.1268	0.0809	0.0975	0.0726

98   **Table S4**. Parameter tuning of PorphyDMPNN and its performance on training, validation and
99   test sets

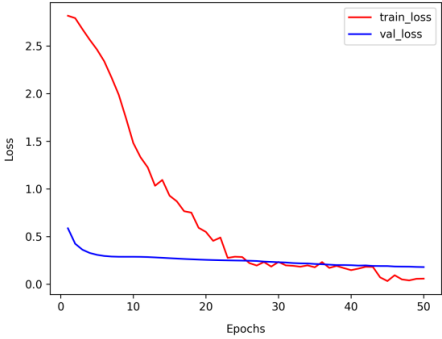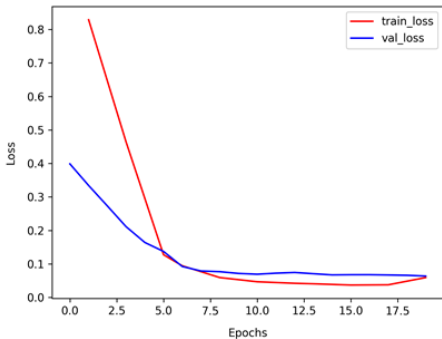| Epoch	Dropout	Train		Validation		Test	
		RMSE (eV)	MAE (eV)	RMSE (eV)	MAE (eV)	RMSE (eV)	MAE (eV)
E_gap							
100	0.4	0.0709	0.0498	0.0947	0.0682	0.0690	0.0527
110	0.4	0.0638	0.0425	0.0872	0.0598	0.0564	0.0427
150	0.4	0.0535	0.0326	0.0811	0.0467	0.0417	0.0315
HOMO							
15	0.4	0.0648	0.0479	0.0972	0.0702	0.0673	0.0494
LUMO							
30	0.4	0.1419	0.1063	0.1388	0.1048	0.1693	0.1334
40	0.4	0.1365	0.1030	0.1234	0.0982	0.1320	0.0973
60	0.4	0.1149	0.0860	0.1179	0.0913	0.1132	0.0864

100

101

102
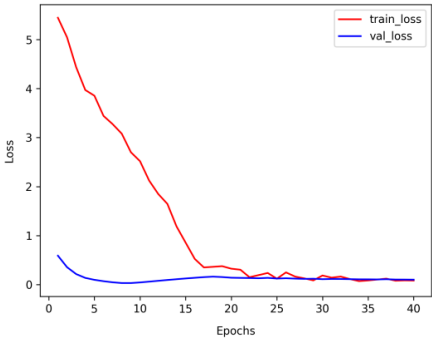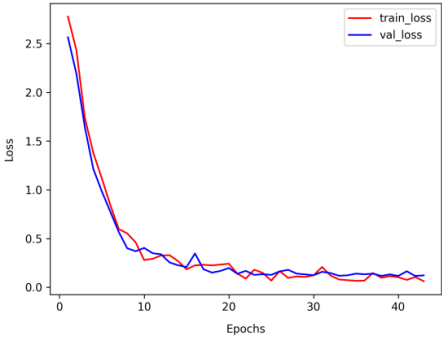
103

104

105
106
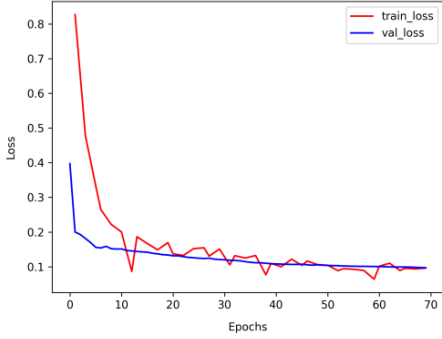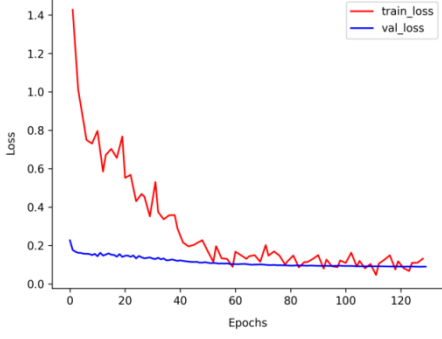
Table S5. The learning curves of PorphyBERT on HOMO, LUMO and E_gap

Type	Epoch	Learning Curve
HOMO	40	
LUMO	40	
E_gap	50	

108

109

110

111

112   **Table S6**.   The learning curves of PorphyDMPNN on HOMO, LUMO and E_gap

Type	Epoch	Learning Curve
HOMO	20	
LUMO	70	
E_gap	120	

113

114

4. PorphyDMPNN fine-tuned with the feature-based approach

This section shows the performance of PorphyDMPNN fine-tuned using the feature-based approach rather than the fine-tuning approach shown in the main text. The performance on the training and validation sets is shown in Table S7 and performance on the test set is shown in Figure S9.

Table S7. The prediction performance of PorphyDMPNN fined by tge feature-based approach on training and validation sets

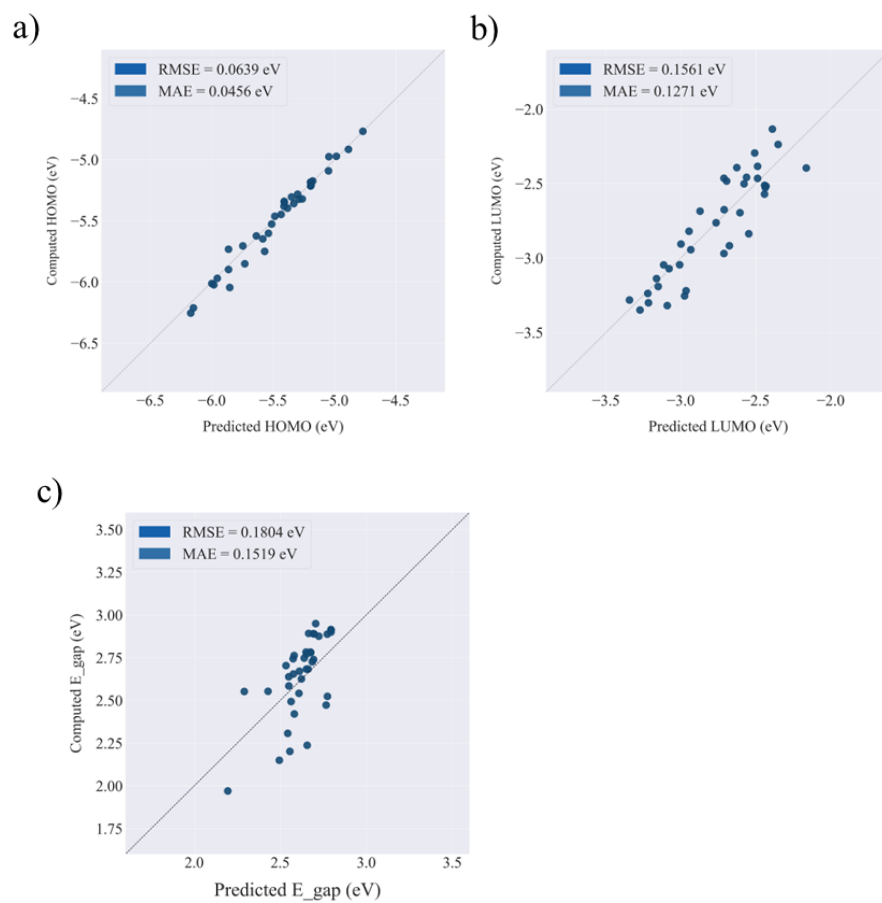| Energy type	Train		Validation	
	RMSE (eV)	MAE (eV)	RMSE (eV)	MAE (eV)
HOMO	0.0511	0.0381	0.0728	0.0517
LUMO	0.1500	0.1167	0.1730	0.1423
E_gap	0.1655	0.1329	0.1607	0.1288

Figure S9. Performance of PorphyDMPNN fine-tuned with the feature-based approach in predicting HOMO, LUMO levels and E_gap. a) HOMO level, b) LUMO level, c) Energy gap (E_gap).

5. Performance of baseline models trained with PBDD or MpPD only

A BERT model was trained only with PBDD or MpPD to directly predict its HOMO and LUMO energy levels and E_gap. The performance of the model is shown in Table S8 (training set and validation set) and Figure S10 (test set).
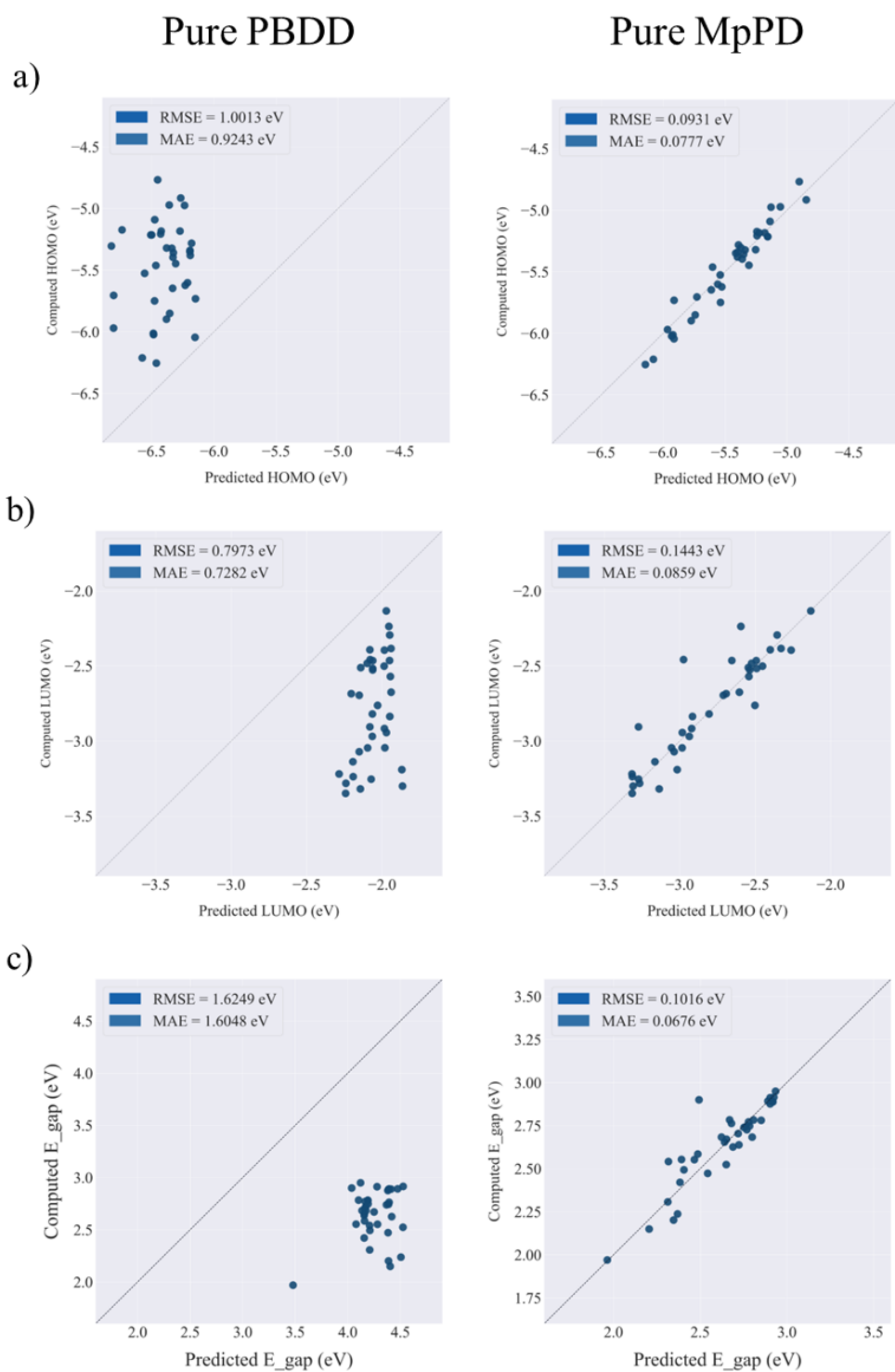
Table S8. Performance of BERT model trained only on PBDD or MpPD in predicting HOMO, LUMO level and E_gap (training and validation)

Training datasets	Train		Validation	
	RMSE(eV)	MAE(eV)	RMSE(eV)	MAE(eV)
HOMO				
Pure PBDD	0.0394	0.0291	0.0550	0.0409
Pure MpPD	0.0858	0.0649	0.1011	0.0789
LUMO				
Pure PBDD	0.0667	0.0558	0.0708	0.0599
Pure MpPD	0.0824	0.0610	0.1476	0.1025
E_gap				
Pure PBDD	0.0887	0.0743	0.1278	0.1105
Pure MpPD	0.0551	0.0355	0.0714	0.0561

Figure S10. BERT model trained only on PBDD (left) or MpPD (right) in predicting a) HOMO energy, b) LUMO energy, and c) energy gap (E_gap).

157 Similarly, a DMPNN model was trained only on PBDD or MpPD and directly predict its

158 HOMO and LUMO energy levels and E_gap. The performance of the model is shown in

159 Table S9 (training set and validation set) and Figure S11 (test set).

160

161 **Table S9**. Performance of DMPNN model trained only on PBDD or MpPD in predicting
162 HOMO, LUMO level and E_gap

| Training datasets	Train		Validation	
	RMSE (eV)	MAE (eV)	RMSE (eV)	MAE (eV)
HOMO				
Pure PBDD	0.0468	0.0355	0.0491	0.0370
Pure MpPD	0.0307	0.0200	0.0401	0.0254
LUMO				
Pure PBDD	0.0627	0.0537	0.0609	0.0532
Pure MpPD	0.0627	0.0457	0.0968	0.0592
E_gap				
Pure PBDD	0.0966	0.0833	0.1002	0.0874
Pure MpPD	0.0756	0.0462	0.0667	0.0490

163
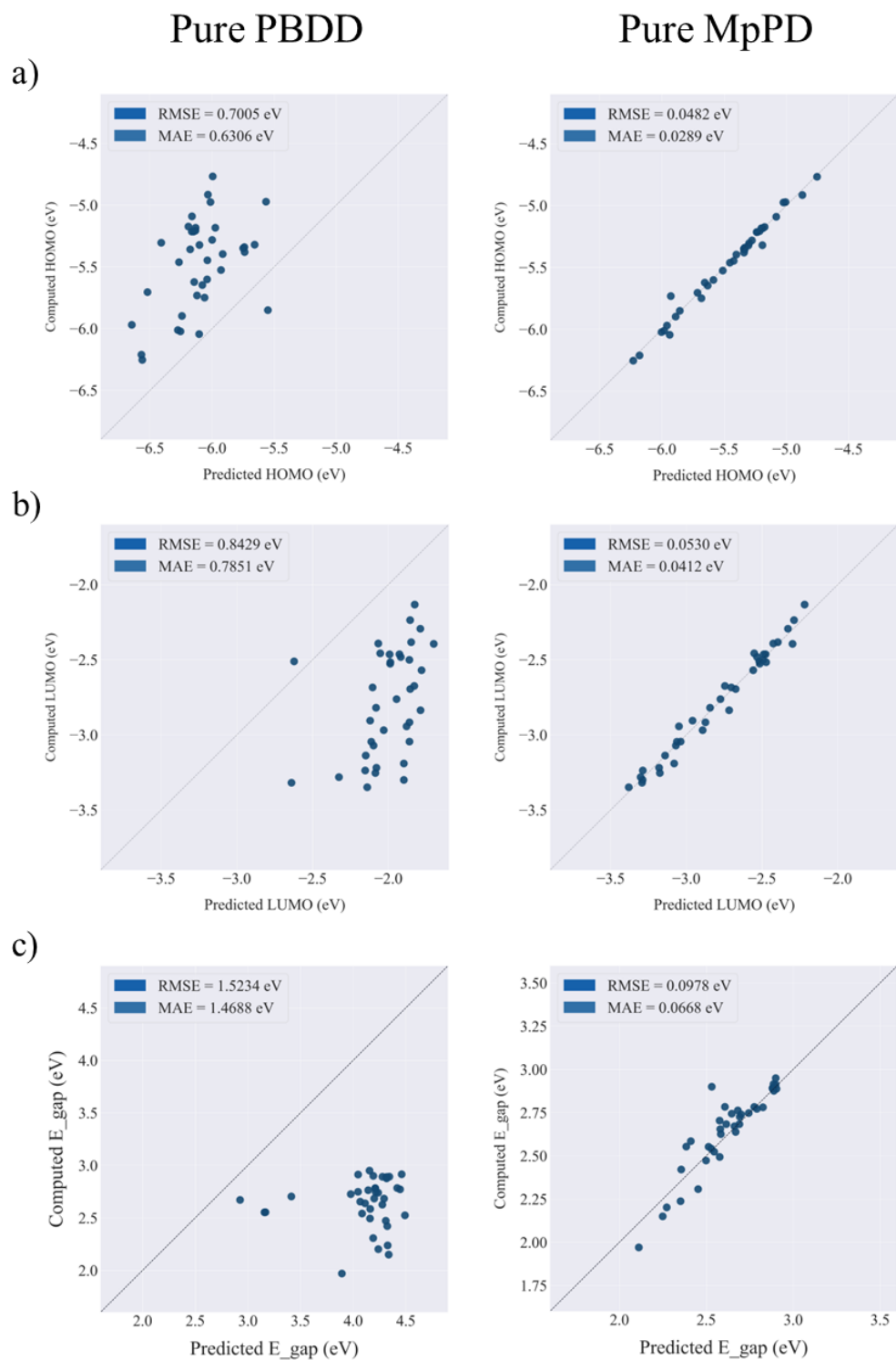
Figure S11. Performance of DMPNN model trained only on PBDD (left) or MpPD (right) in predicting a) HOMO energy, b) LUMO energy, and c) energy gap (E_gap).

19

6. Random Forest as benchmark model
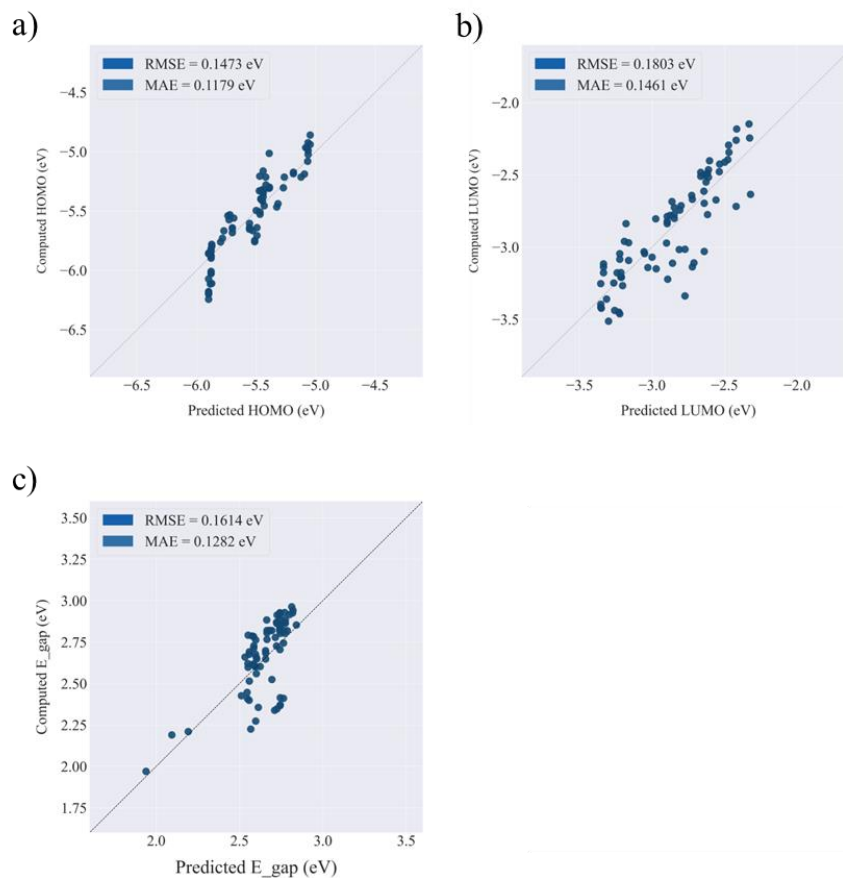
Figure S12. Performance of Random Forest model trained on MpPD in predicting a) HOMO energy, b) LUMO energy, and c) energy gap (E_gap).

References

177   **References**

178   1. Shen, H.-M.; Hu, M.-Y.; Liu, L.; Qi, B.; Ye, H.-L.; She, Y.-B., Efficient and

179   selective oxidation of tertiary benzylic C-H bonds with O_2 catalyzed by

180   metalloporphyrins under mild and solvent-free conditions. *Applied Catalysis A: General*

181   **2020,** *599*.

182   2. Chang, I. J.; Jeon, Y. S.; Hwang, K. J., Synthesis and Band Gap Analysis of

183   Designed Porphyrin Derivatives Containing Electron Donating and Accepting Group.

184   *Bulletin of the Korean Chemical Society* **2019,** *40* (2), 173-179.

185   3. Porphyrin based dyes[DB/OL]. https://cmr.fysik.dtu.dk/dssc/dssc.html.