Electronic Supplementary Material (ESI) for RSC Advances.

Supporting Information

Predicting band gaps of MOFs on small data by deep transfer learning with data augmentation strategies

Zhihui Zhang[1, †], Chengwei Zhang[1, †], Yutao Zhang[1], Shengwei Deng[1], Yun-Fang Yang[1], An Su[1, *] and Yuan-Bin She[1, *]

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

† These authors contributed equally to this work.

*** 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

Table of Contents

Table S1 PMOF202 data set statistics

Item	**Eg** (eV)
mean	1.147434
std	0.627529
min	0.000603
25%	0.715559
50%	1.350389
75%	1.627507
max	2.004765

A statistical summary of the results of the DFT calculations is provided in **Table S1**, which includes the mean and standard deviation (std) of Eg, as well as the quartiles, maximum, and minimum values.

	Original	Rotation		Mirror Rotation		
Label	(1)	(2)	(3)	(4)	(5)	(6)
Composition	[x, y, z]	[y, z, x]	[z, x, y]	[x, z, y]	[z, y, x]	[y, x, z]

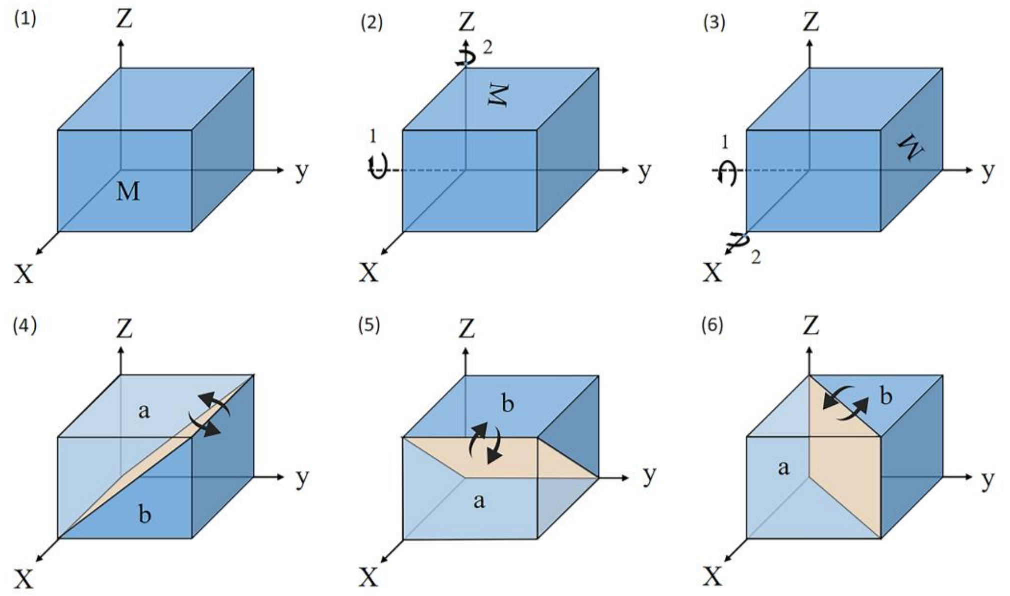

Figure S1 Schematic illustration of augmenting the PMOF168 dataset via rotation.

When using a three-dimensional matrix to represent a PMOF, the original encoded matrix (x, y, z) becomes (y, z, x), (z, x, y), (x, z, y), (z, y, x), and (y, x, z) by different ways of rotation, as shown respectively in (1 – 6)[1]

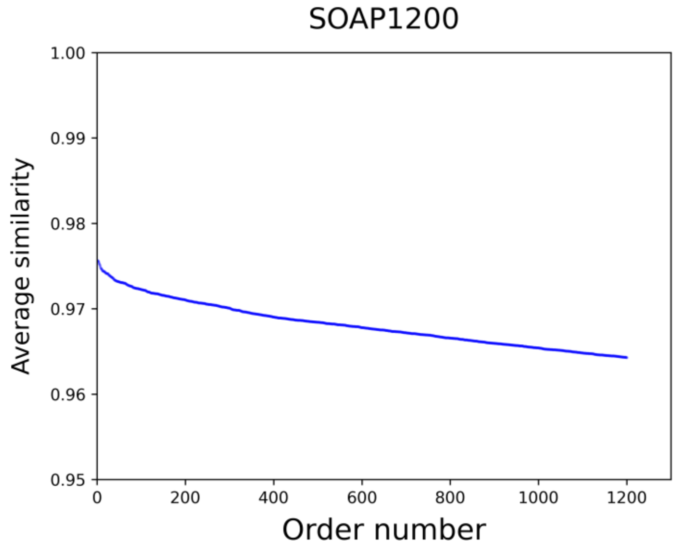

Figure S2 The top 1200 QMOFs with highest average similarity to PMOF202 dataset obtained by the Average SOAP kernel approach

The 1200 data sets are expanded, and the specific implementation is as follows: The Average SOAP kernel was used to calculate the global similarity between each MOF sample in the QMOF database and each porphyrin MOF material in the PMOF202 dataset. The global similarity between each MOF sample in the QMOF database and each porphyrin MOF material in the PMOF202 data set is summed and averaged to obtain the average similarity of each MOF sample in the QMOF database. For example, 1200 MOF samples in the QMOF database are selected in descending order of average similarity, denoted as SOAP1200. Combined with PMOF168, the SOAP1368 data set was obtained, which was used as a fine-tuning set to carry out in-depth migration learning for the four models respectively.

Learning Curves

Please note that the ranges of the y axes are different for each figure.

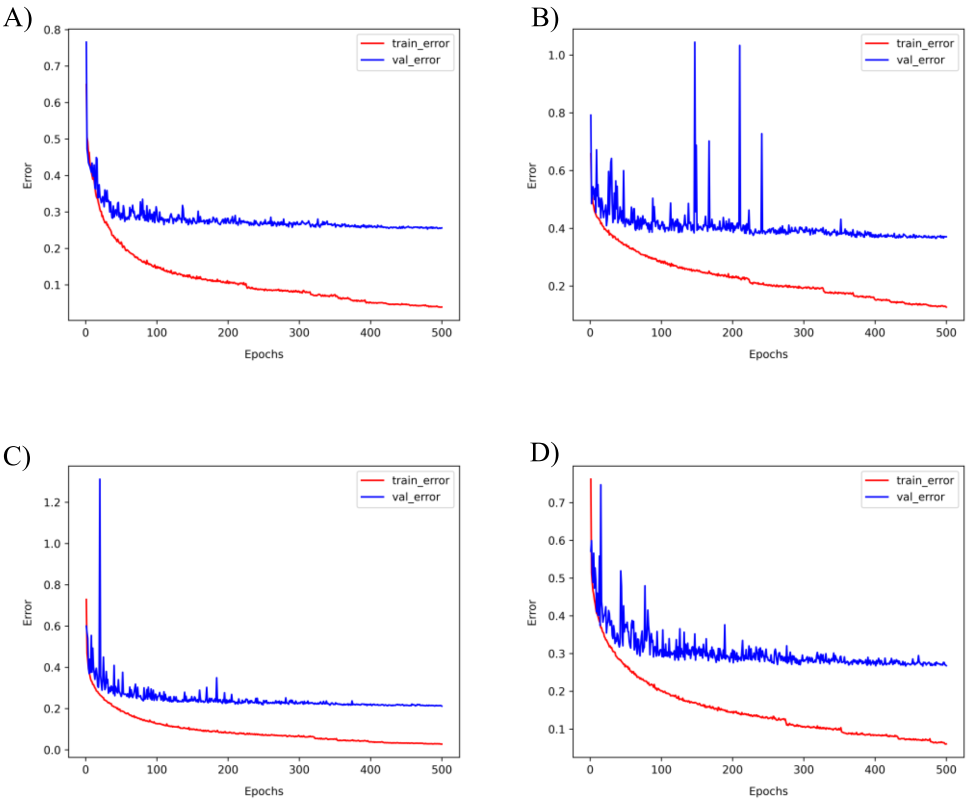

Figure S3. The learning curves of four models pretrained by QMOF. A) CGCNN; B) GCN; C) MEGNet; D) SchNet; epochs=500

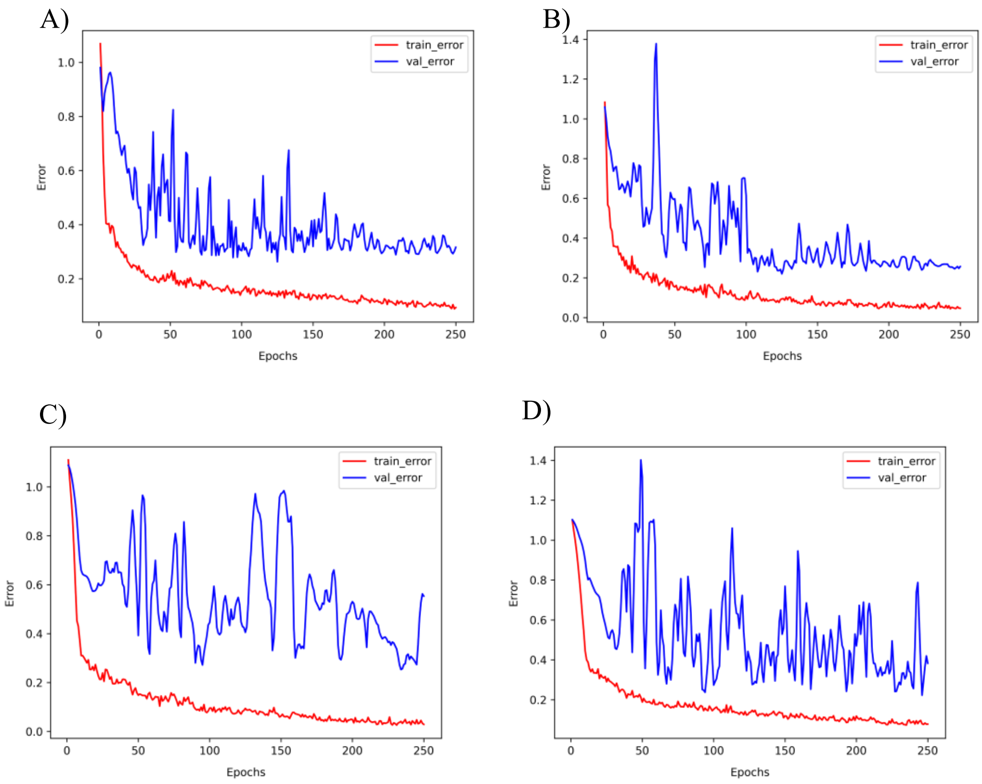

Figure S4 The learning curves of four models fine-tuned by PMOF168. A) CGCNN; B) GCN; C) MEGNet; D) SchNet; epochs=250

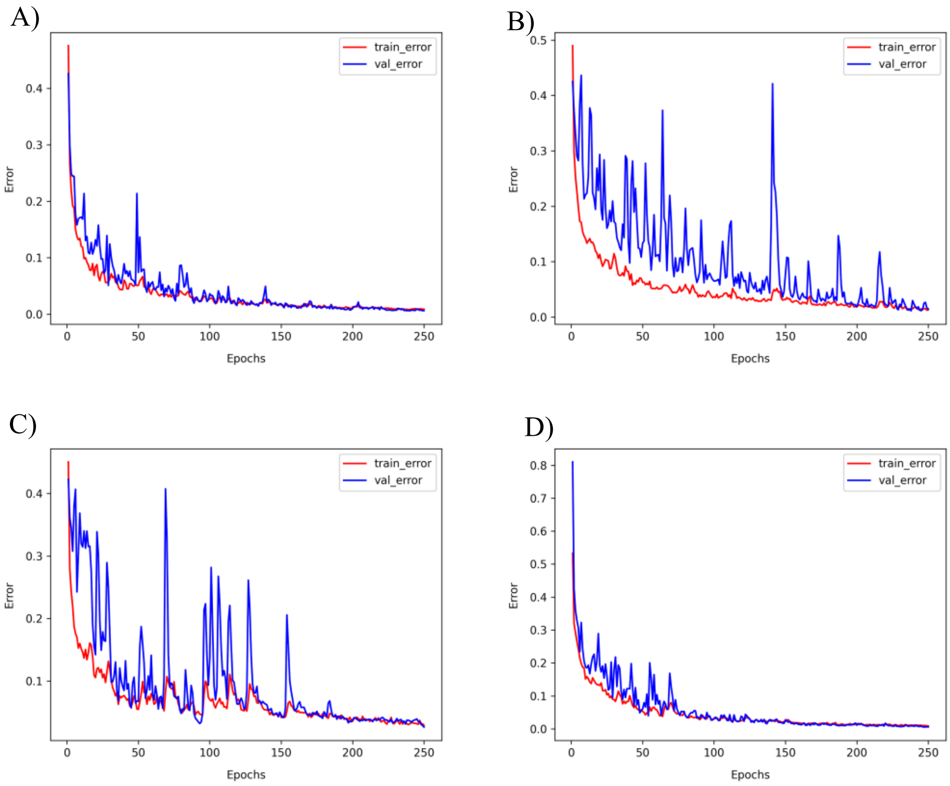

Figure S5 The learning curves of four models fine-tuned by DA1008. A) CGCNN; B) GCN; C) MEGNet; D) SchNet; epochs=250

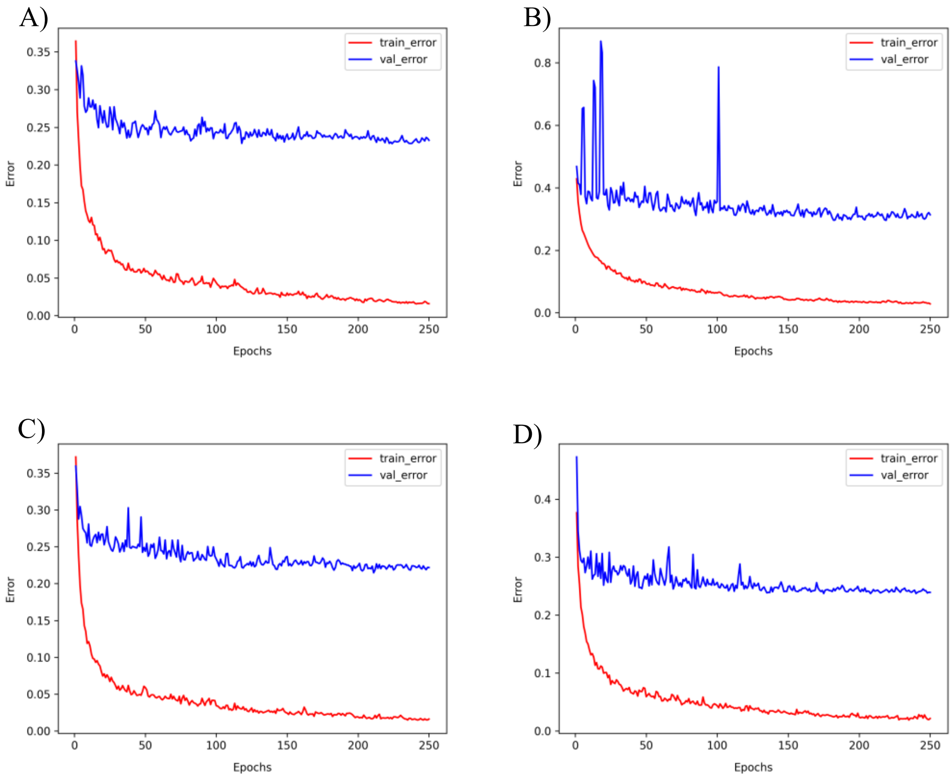

Figure S6 The learning curves of four models fine-tuned by SOAP1200+PMOF168**.** A) CGCNN; B) GCN; C) MEGNet; D) SchNet; epochs=250

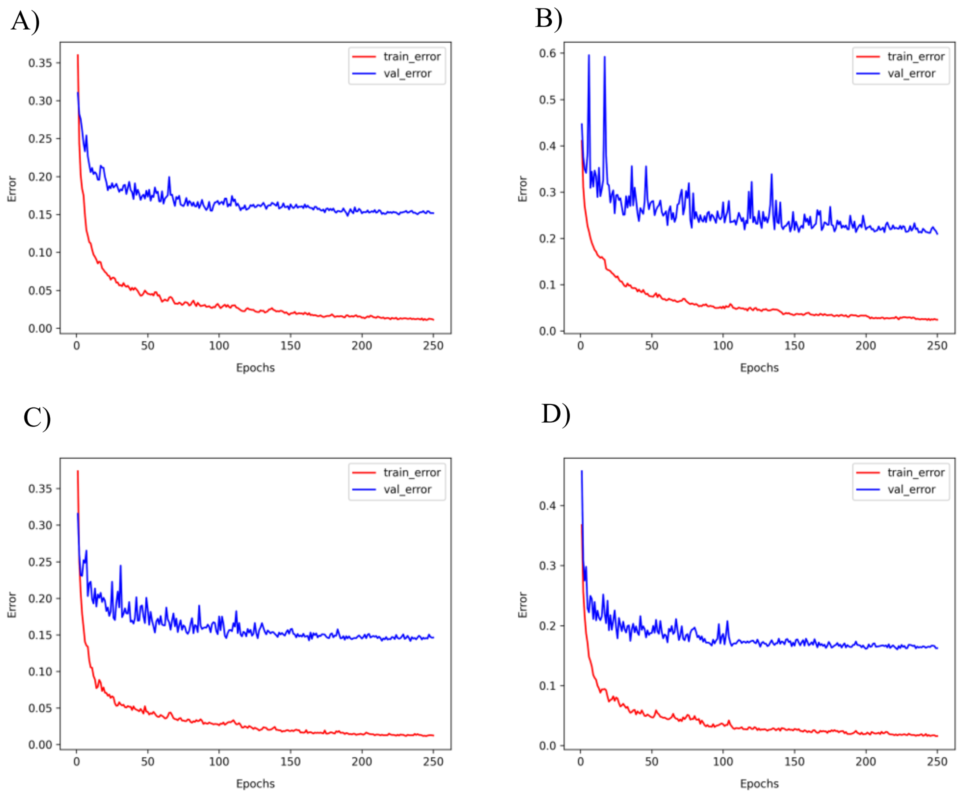

Figure S7 The learning curves of four models fine-tuned by SOAP1200+DA1008. A) CGCNN; B) GCN; C) MEGNet; D) SchNet; epochs=250

References

1. Hung, T.-H.; Xu, Z.-X.; Kang, D.-Y.; Lin, L.-C., Chemistry-Encoded Convolutional Neural Networks for Predicting Gaseous Adsorption in Porous Materials. *The Journal of Physical Chemistry C* **2022,** *126* (5), 2813-2822.