Supporting Information for

An Interpretable and Transferrable Vision Transformer Model for Rapid Materials Spectra Classification

Zhenru Chen[1], Yunchao Xie[1*], Yuchao Wu[1], Yuyi Lin[1], Shigetaka Tomiya[2], and Jian Lin[1*]

[1]Department of Mechanical and Aerospace Engineering
University of Missouri, Columbia, MO 65201, United States

[2]Data Science Center, Graduate School of Advanced Science and Technology, Material Science Division, Nara Institute of Science and Technology (NAIST)
8916-5 Takayamacho, Ikoma City, Nara Prefecture 630-0192, Japan

*Email: linjian@missouri.edu; yxpx3@umsystem.edu.

Supplementary Notes

Supplementary Note S1: Data augmentation of theoretical XRD spectra

Data augmentation is typically employed to expand the size of the training data, thus bolstering the model generalization and avoid overfitting. It facilitates the acquisition of valuable features by exposing it to diverse perspectives and representations. Given that various synthesis methods often yield MOFs with varying sizes, morphologies, and crystalline structures, it becomes practical to replicate these phenomena. Simulation of these features includes (1) peak scaling[1] and (2) peak elimination,[2] which can be associated with preferred orientation and phase transitions, as well as (3) peak shifting[3] resulting from variations in lattice parameters, grain size, and instrumental factors. In this context, a physics-informed data augmentation technique was utilized to replicate these phenomena.[4] Before data augmentation, the XRD spectra were described as a discrete function f(2theta) = I → R^+, where I refers to a set of discrete angles and R^+ means the peak intensity. In detail, three sequential transformation functions, denoted as $f_1(S,c,n)$, $f_2(S,n)$, and $f_3(\varepsilon)$, were developed to augment the XRD spectra.

(1) Peak scaling is achieved by applying a scaling factor, denoted as c, to a randomly selected subsets of peaks corresponding to angles S_n. This accounts for variations in preferred orientations.

$$f_1(S,c,n) = f(S_n) \times c_n + f(I \backslash S_n) \quad (1)$$

(2) Peak elimination is carried out along the 2theta axis by nullifying the intensities at randomly selected angles S_n with a specific probability of 0.7.

$$f_2(S,n) = f(S_n) \times 0 + f(I \backslash S_n) \quad (2)$$

(3) Peak shifting is performed by introducing a small random shift ε along the 2theta axis, causing either a blue shift (to lower angles) or a red shift (to higher angles) across the entire XRD spectra. This accounts for variations in lattice parameters, grain size, and instrumental factors.

$$f_3(\varepsilon) = f(2\theta - \varepsilon) \quad (3)$$

In contrast to the approach in the literature,[4] where the length of S_n is fixed, in this context, a random length of S_n was employed. This randomization strategy yielded notably improved prediction accuracies, particularly for the Top-1 accuracy, which exhibited an increase of nearly 10%, as illustrated in Figure S2. This approach contributes to a more diverse and resilient collection of augmented data.

Supplementary Note S2: Data preprocessing of experimental XRD spectra

Two steps of the Savitzky-Golay smoothing and background subtraction were employed to preprocess these experimental XRD spectra.[4]

1. Savitzky-Golay smoothing

The Savitzky-Golay smoothing is a signal processing technique that preserves the underlying trends while reducing the impact of noise. It works by fitting a local polynomial regression model

to a small window of data and estimating the smoothed values for each point. Here, a Savitzky-Golay filter with a window size of 21 and a polynomial degree of 3 was employed to eliminate high-frequency noise.

2. Background subtraction

XRD spectra often contain background signals caused by various factors like solvent absorption, instrument noise, scattering, or fluorescence from impurities. Background subtraction is essential in materials spectroscopy analysis since it helps reveal the true spectral signatures of the materials being studied. Accurate background subtraction can improve the precision of measurements and aid in the identification and characterization of the materials. First, a peak-finding function (scipy.signal.find_peaks) was employed to identify a peak window in the smoothed spectra. Specifically, the prominence and the width of the peak window were set to be 0.5 and 1, respectively. For more accurate peak identification, spurious peaks caused by noise were sifting out. Then, the minimum values within this peak window were screened using the np.partition function. Finally, the values of the peak window were subtracted by this minimum number to accentuate the peaks.

Supplementary Note S3: Details of decision tree for scoring hyperparameters of ViT-XRD model

First, the ViT-XRD model was trained individually using combination of hyperparameters for at least 10 times. Then, the combination of the hyperparameters and the performance (Top-1, Top-3, and Top-5 accuracies) as the input and output were fed to a decision tree model. Figures S5, S6, and S7 provide a graphical representation of the decision tree models for Top-1, Top-3, and Top-5 accuracies, respectively. This approach offered a comprehensive understanding of the

performance landscape, aiding in not only identifying the optimal hyperparameter combinations but also in understanding the relative importance of each hyperparameter in the prediction accuracies.

Supplementary Note S4: t-SNE for dimension reduced visualization

As a dimension-reduction technique, t-SNE maps the high-dimensional data into 2D space while preserving the structure and relationship among data points. In our case, t-SNE was employed to reduce the dimensions of (1) the theoretical XRD spectra of 2000 MOFs, representations learned by (2) the CNN-XRD model, and (3) the ViT-XRD model. The representations learned by the CNN-XRD and ViT-XRD models were generated by outputting the last full-connected layer and the last layer of the Transformer encoder, respectively. Consequently, three 2D space maps were obtained, shown in Figure 4. For the ViT-XRD model, the features from the layer of the Transformer encoder were used. For the CNN-XRD model, we utilized features from the penultimate fully connected layer, as the last layer contains a softmax activation which is not ideal for t-SNE visualization.

Identification of 5 MOFs sharing the isolated distance. A k-nearest neighboring algorithm was developed to compute the Euclidean distance of each MOF to its neighbors in the dimension-reduced 2D space (Figure 4a). Then, the Euclidean distance was sorted, and the top five values were selected as the highest isolation distances.

Identification of 5 MOFs sharing the closest distance. The density-based spatial clustering of Applications with Noise (DBSCAN) algorithm was developed to locate the most densely clustered MOFs. The five most densely populated clusters were identified based on the frequency of unique labels.
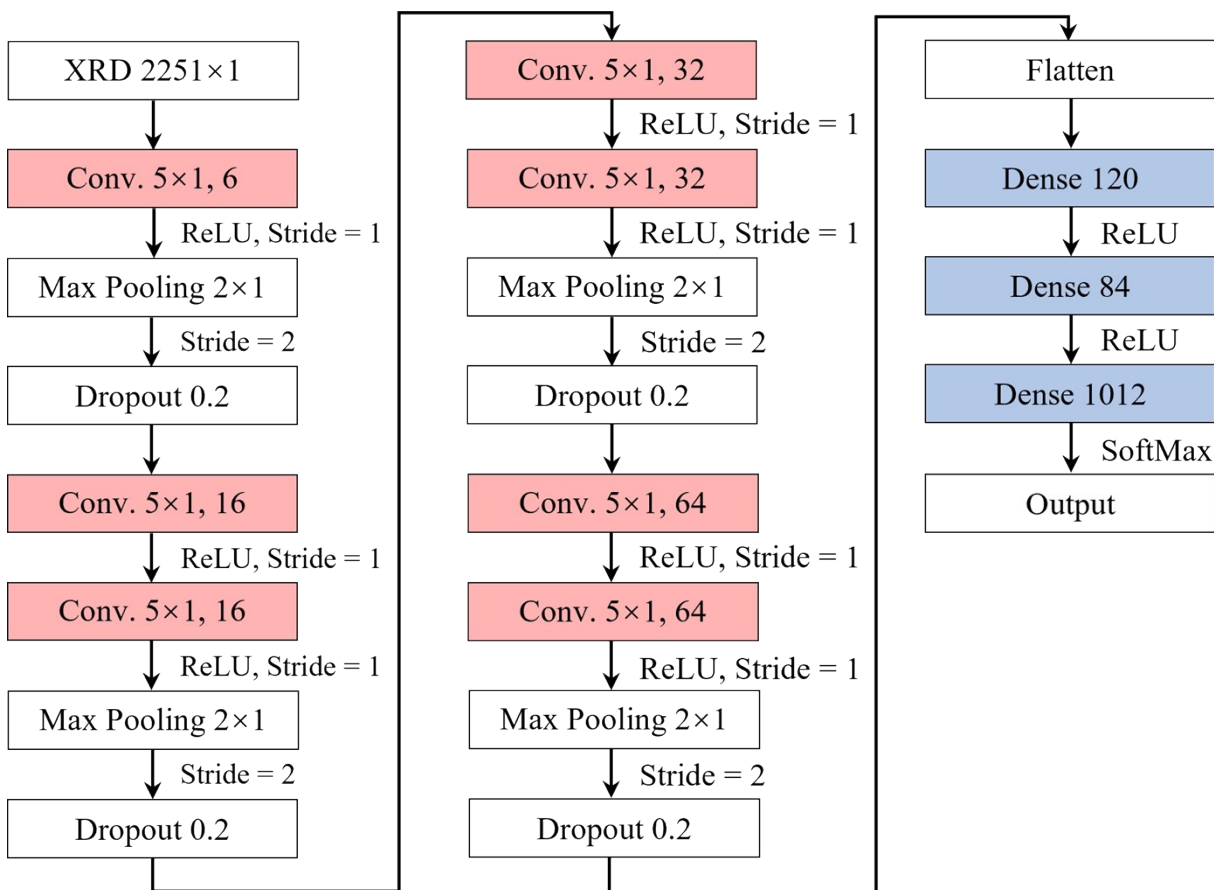
Supplementary Figures

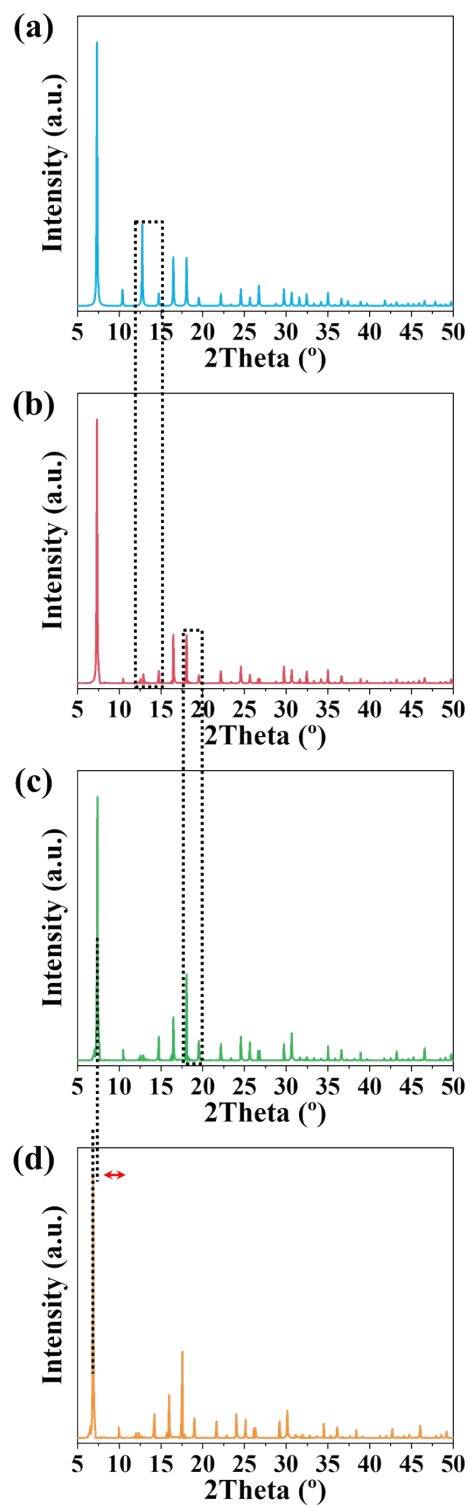**Figure S1.** Architecture of the CNN-XRD model.

Figure S2. Flowchart of data augmentation procedure. (a) Original data. (b) Peak elimination.
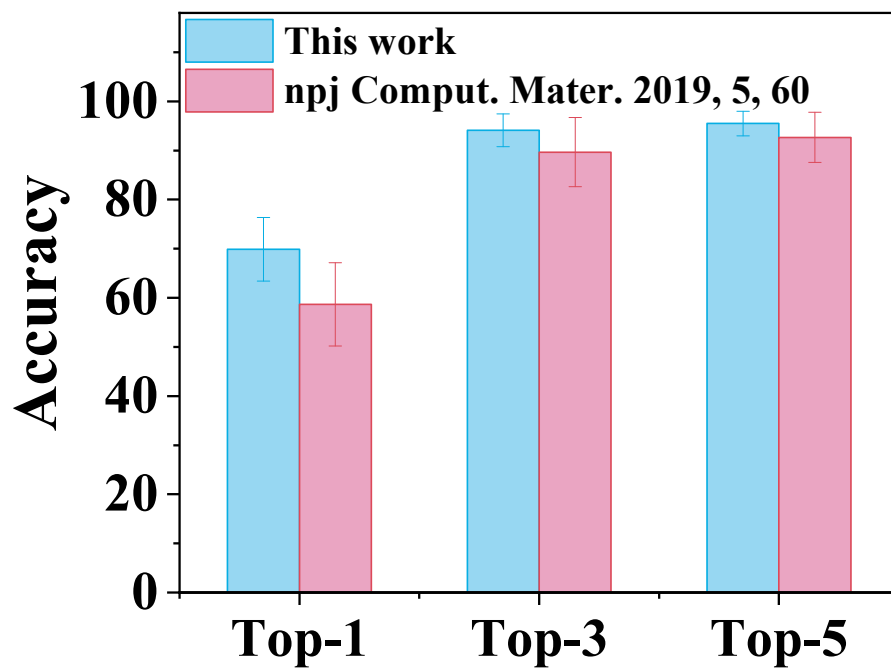
(c) Peak scaling. (d) Peak shift.

Figure S3. Prediction accuracies of Top-1, Top-3, and Top-5 over augmentation methods of ours and literature.[4]
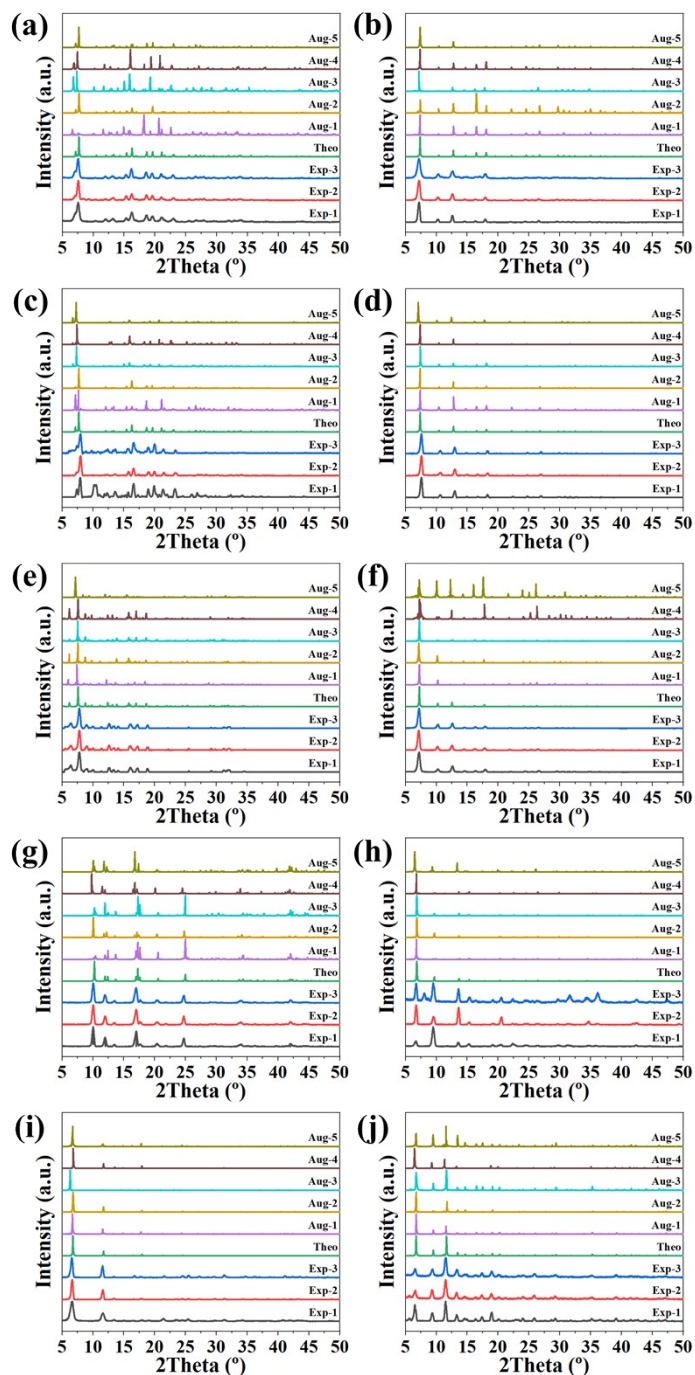
Figure S4. Representative XRD spectra of the ten MOFs. (a) ZIF-7, (b) ZIF-8, (c) ZIF-9, (d) ZIF-67, (e) ZIF-71, (f) ZIF-90, (g) MOF-2, (h) MOF-5, (i) MOF-74, and (j) MOF-199. Here, 'Theo', 'Exp', and 'Aug' refer to theoretical, experimental and augmented data.
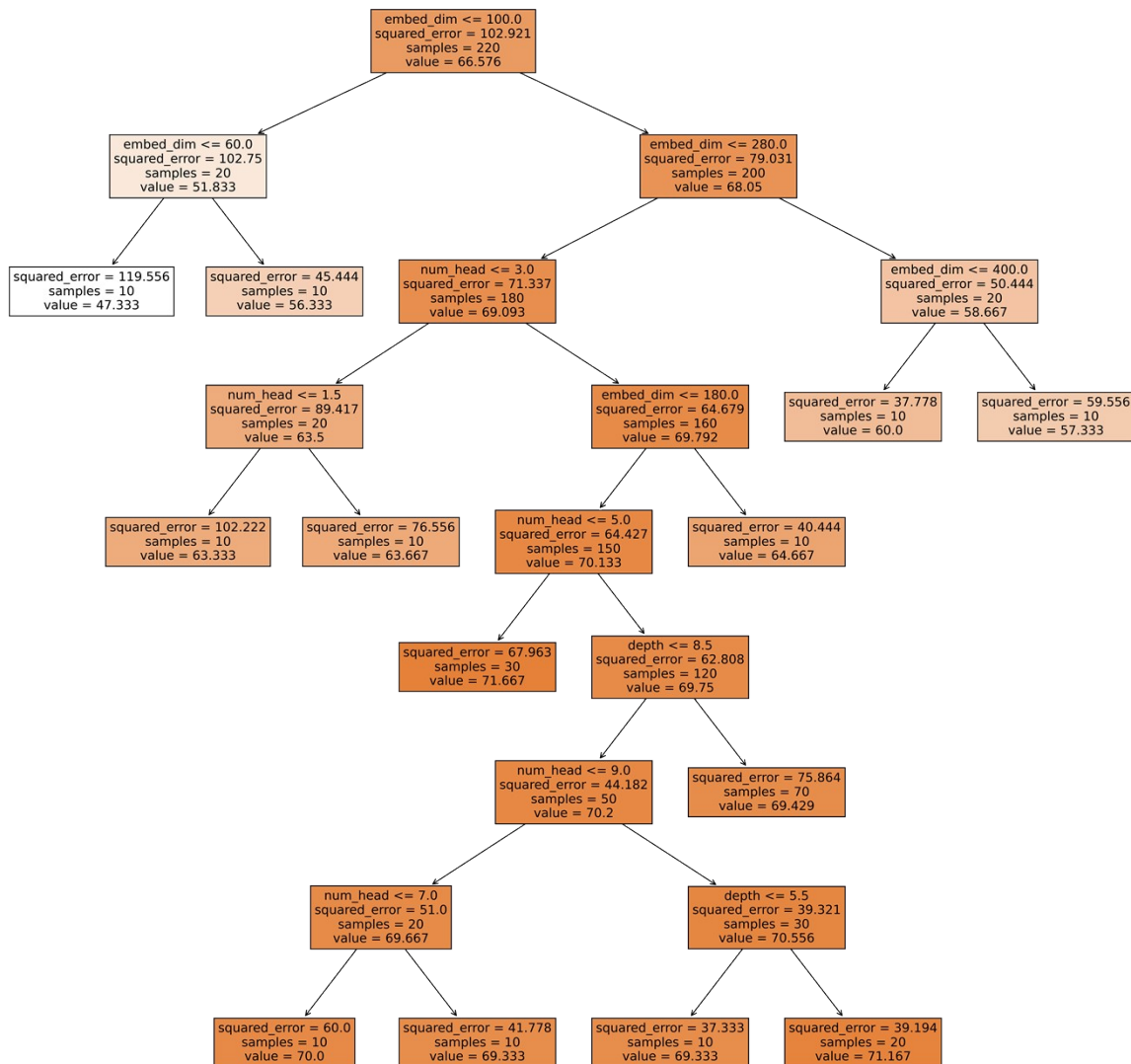
Figure S5. Visualization of a decision tree for assessing hyperparameter importance in determining the Top-1 prediction accuracy.
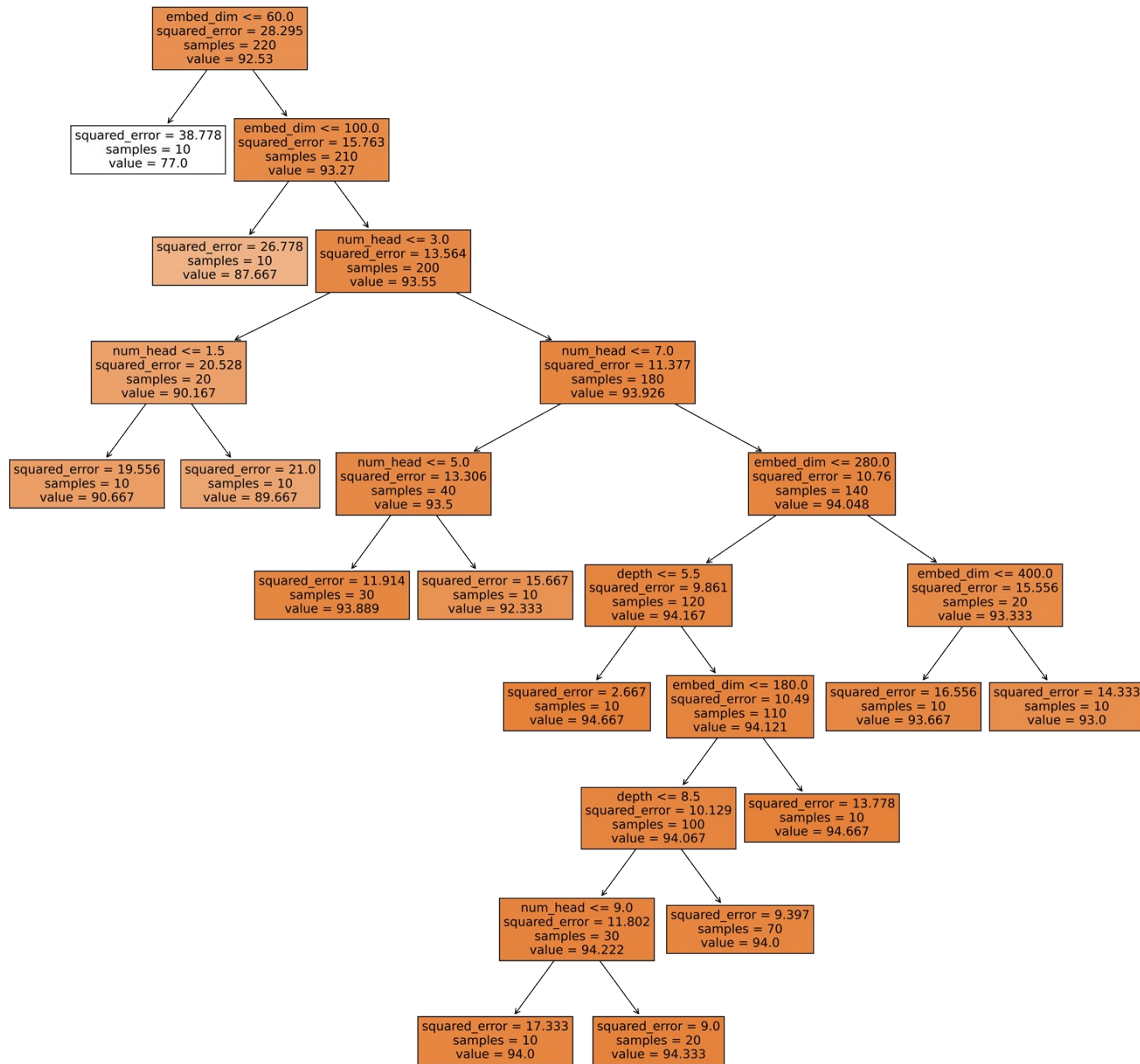
Figure S6. Visualization of a decision tree for assessing hyperparameter importance in determining the Top-3 prediction accuracy.
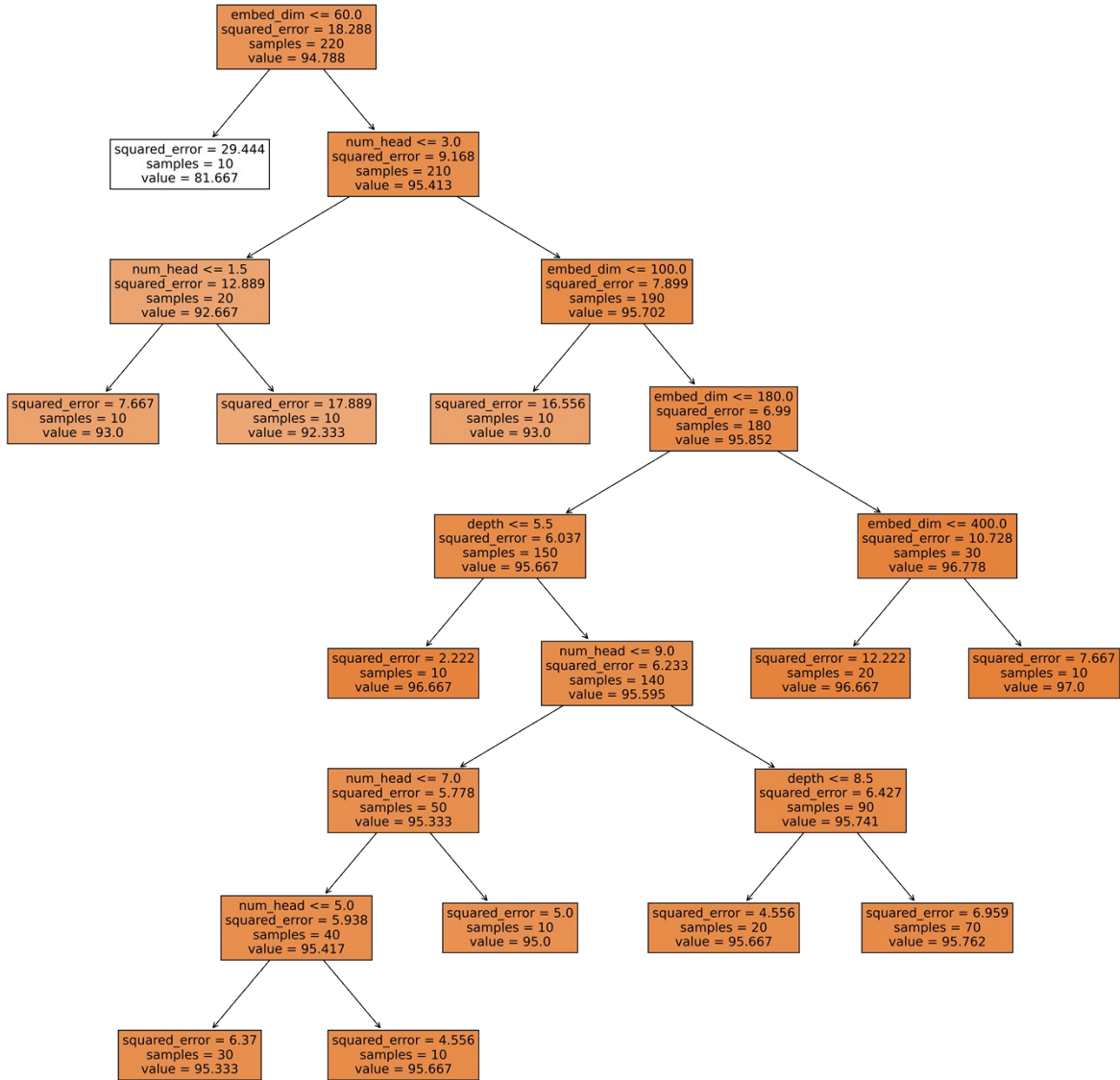
Figure S7. Visualization of a decision tree for assessing hyperparameter importance in determining the Top-5 prediction accuracy.
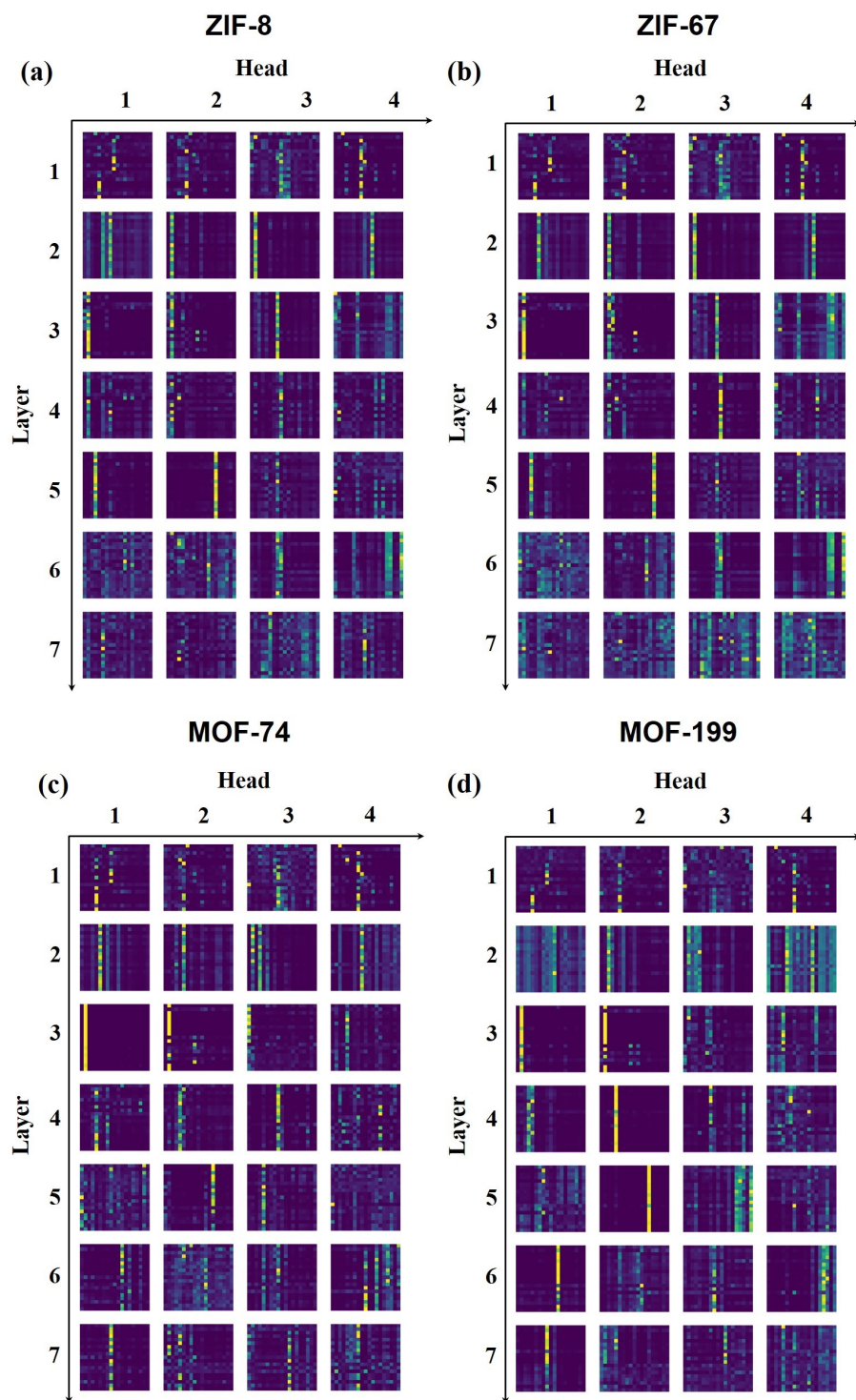
Figure S8. Heat maps of the learned attention weights from the ViT-XRD model across four heads across all encoder layers over the XRD spectrum of (a) ZIF-8, (b) ZIF-67, (c) MOF-74, and (d) MOF-199.
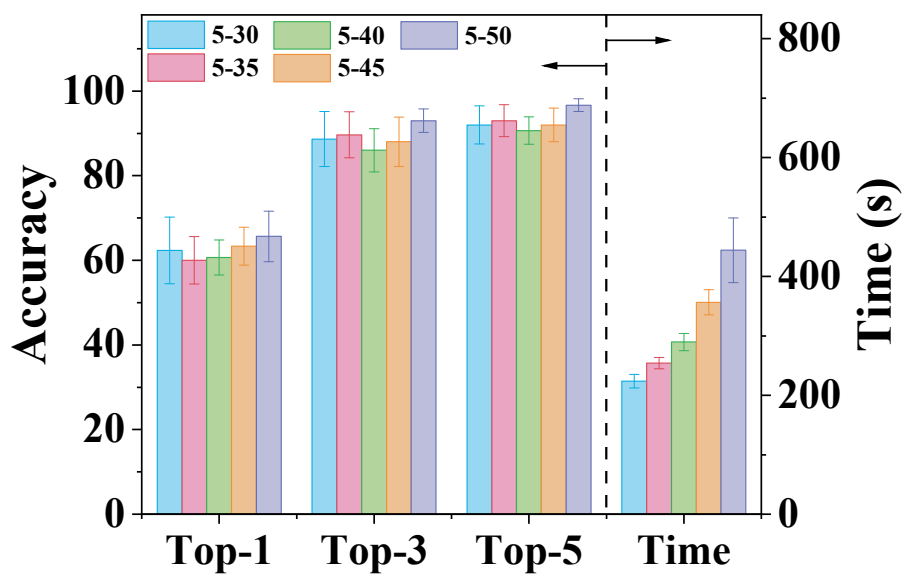
Figure S9. Prediction accuracies of Top-1, Top-3, and Top-5 and training time of ViT model over different 2theta ranges of XRD spectra.
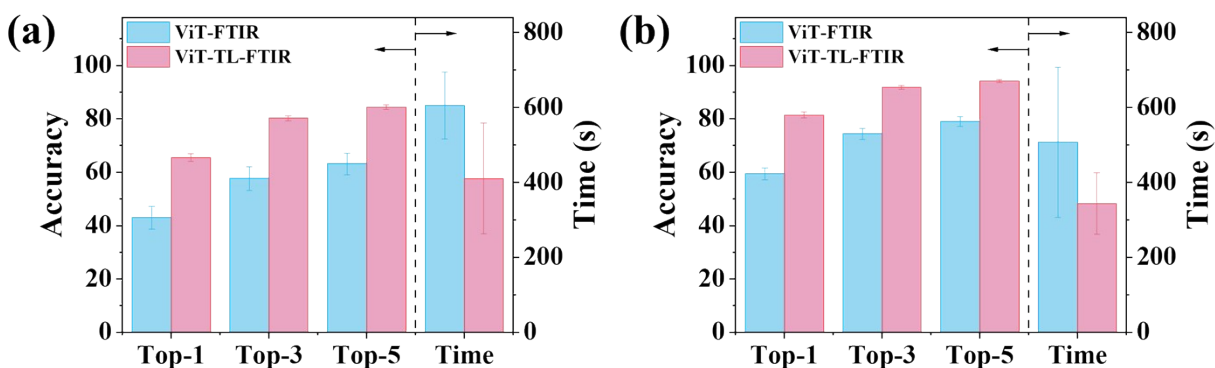

Figure S10. Prediction accuracies of Top-1, Top-3, and Top-5 as well as training time of transfer learning and non-transfer learning on FTIR data under different Embed_dim. (a) 120 and (b) 240.
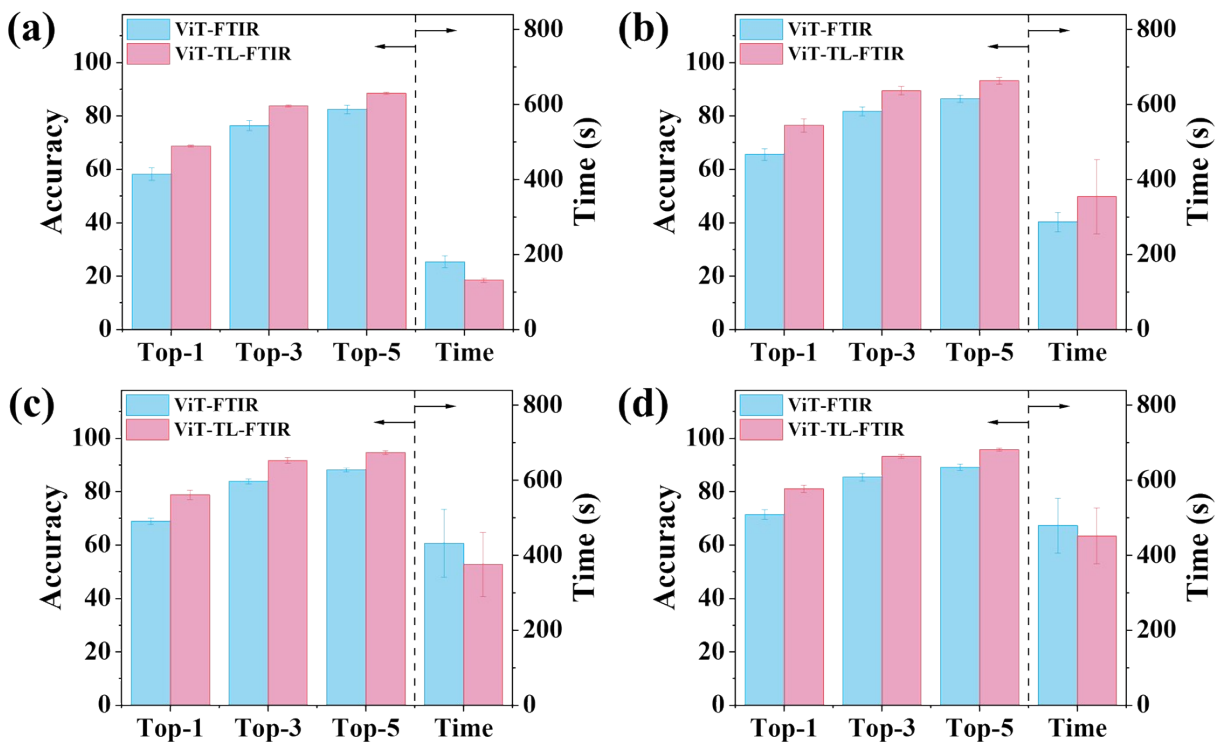
Figure S11. Prediction accuracies of Top-1, Top-3, and Top-5 as well as training time of transfer learning model over different augmentation times. (a) 10 times, (b) 20 times, (c) 30 times, and (d) 40 times.
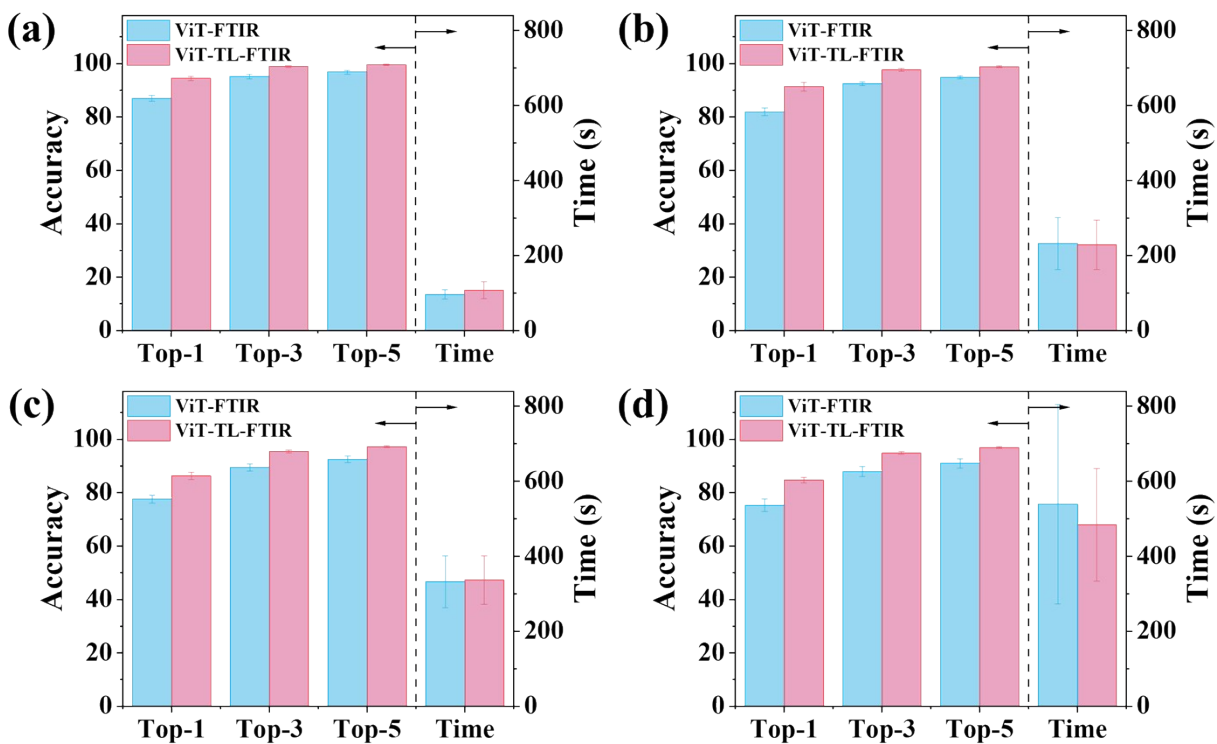
Figure S12. Prediction accuracies of Top-1, Top-3, and Top-5 as well as training time of transfer learning model over different number of organic molecules. (a) 500, (b) 1000, (c) 2000, and (d) 3000.

Supplementary Tables

Table S1. Comparison of prediction accuracies and training times over various models.

Model	Accuracies (%)			Time (s)
	Top-1	Top-3	Top-5	
NB	10	16.7	16.7	3.4±0.04
KNN	36.7	63.3	66.7	1.8±0.05
LR	63.3	93.3	100	4113±38.23
RF	NA	NA	NA	NA
XGB	NA	NA	NA	NA
CNN	60.0±5.58	87.6±2.23	89.5±2.90	378.5±69.7
ViT	70.0±7.41	93.2±3.68	94.9±3.01	268.9±14.14

Table S2. The CCDC numbers and full names of ten extra MOFs that are displayed in **Figure 4**.

CCDC numbers	Full name
RECHOX	catena-(bis(μ2-bis(1-Methyl-2-thioxo-4-imidazolin-3-yl)methane)-silver(i) tetrafluoroborate)
REFLAQ	catena-(pentakis(μ2-2,2'-Bi-imidazole)-bis(2,2'-bi-imidazole)-hexa-silver tetraperchlorate bis(acetate) tetrahydrate)
COLMUM	catena-((μ2-3,5-dihydroxybenzoicacid-O,O)-(μ2-aqua)-diaqua-sodium 3,5-dihydroxybenzoate dihydrate)
REFFIT	catena-(bis(1-n-Butyl-2,3-dimethylimidazolium) tetrakis(μ4-4,4'-oxybis(benzoato)-O,O,O',O'',O''')-bis(μ4-4,4'-oxybis(benzoato)-O,O',O'',O''')-bis(μ3-hydrogen 4,4'-oxybis(benzoato)-O,O',O'')-hexa-magnesium)
DIWXUD01	catena-(octakis(μ2-Oxo)-bis(1-(4-pyridyl)-2-(4-pyridinio)ethylene)-tetraoxo-cobalt(ii)-tetra-vanadium)
ZIDXAN	catena-[bis(μ5-Adamantane-1,3-dicarboxylato)-(μ2-1,1'-butane-1,4-diylbis(2-methyl-1H-imidazole))-tetra-silver dihydrate]
LOGXIQ	catena-[tris(2,2'-bipyridine)-ruthenium tetrakis(μ-selenido)-tetracosakis(μ-benzenethiolate)-tetrabromo-heptadeca-cadmium]
RAXCUQ	catena-((μ4-Azobenzene-3,3',5,5'-tetracarboxylato)-bis(μ2-1,3-bis(imidazolyl)propane)-di-zinc dihydrate)
NEFQAR	catena-((μ2-isophthalato-O,O')-aqua-tris(pyridine)-nickel(ii))
REFMAS	catena-((μ2-Isophthalato)-(2,2'-ethane-1,2-diylbis(1H-benzimidazole))-cobalt)

References

(1) L. Ye, J. Liu, L. Tian, T. Peng, L. Zan. The replacement of {101} by {010} facets inhibits the photocatalytic activity of anatase TiO_2. *Applied Catalysis B: Environmental* 2013, 134-135, 60-65.

(2) L. Ye, J. Liu, Z. Jiang, T. Peng, L. Zan. Facets coupling of BiOBr-g-C_3N_4 composite photocatalyst for enhanced visible-light-driven photocatalytic activity. *Applied Catalysis B: Environmental* 2013, 142, 1-7.

(3) T. Sawabe, M. Akiyoshi, K. Yoshida, T. Yano. Estimation of neutron-irradiation-induced defect in 3C–SiC from change in XRD peak shift and DFT study. *Journal of Nuclear Materials* 2011, 417, 430-434.

(4) F. Oviedo, Z. Ren, S. Sun, C. Settens, Z. Liu, N. T. P. Hartono, S. Ramasamy, B. L. DeCost, S. I. P. Tian, G. Romano, A. Gilad Kusne, T. Buonassisi. Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks. *npj Computational Materials* 2019, 5, 60.