Supplementary Information (SI) for Journal of Materials Chemistry A.

Comprehensive Overview of Machine Learning Application in MOFs: From Modeling Processes to Latest Applications and Design Classifications

Yutong Liu, Yawen Dong, Hua Wu*

Department of Chemistry, College of Sciences, Nanjing Agricultural University, Nanjing 210095, P. R. China

* corresponding authors

E-mail: wuhua@njau.edu.cn

13 pages

Contains 5 tables.

Contents

Table S1. Summary of data sources.

Table S2. Summary of commonly used material databases.

Table S3. Summary of common material descriptors.

Table S4. Three mainstream Cross Validation (CV) methods.

Table S5. The application of MOFs combined with ML in the adsorption and separation of various gases.

References

Table S1. Summary of data sources.

Sources	Descriptions	Advantages	Disadvantages	Future trend
Experiment	The scientific study of synthesizing specific MOFs by manually adjusting various experimental parameters.	Because reliable data cannot be obtained, synthetic experiments under certain conditions cannot be replaced by computational methods.	Experiment is costly and time-consuming, and the sensitivity of MOFs to the synthesis conditions will lead to the deviation of the results.	High-throughput experimental (HTE) + Machine Learning (ML)
Literature	Searchable written information that has been published in a journal or book.	A large amount of experimental information to be used is stored.	Provided as text, manual extraction of information is time-consuming, laborious and error-prone.	Natural Language Pprocessing (NLP) and Large Language Models (LLM)
Database	A publicly available or paid collection of crystal structures data.	A plenty of data and a variety of structural parameters are contained.	There are many sources, disordered relationships and poor universality.	Standardize data format, enhance its universality and systematicness.
Calculation	According to the existing data and combined with mathematics, physics and chemistry knowledge to expand the data.	Computational data sources are more readily available and apply to target attributes that lack sufficient data.	Some data are unreasonable and lack of experimental verification, which will interfere with the results of material development.	High-Throughput Computational Screening (HTCS) + Machine Learning (ML)

Table S2. Summary of commonly used material databases.

Database	Number	Year	Descriptions	URL
Inorganic Crystal Structure Database (ICSD)	over 240,000	1913	The largest database in the world for totally identified inorganic crystal structures, provided by FIZ Karlsruhe GmbH.[1-3]	https://psds.ac.uk/icsd
Cambridge Structure Database (CSD)	over 1,000,000	1965	A database of over 1,000,000 small-molecule organic and organometallic crystal structures, including 69,666 MOFs.[4, 5]	https://psds.ac.uk/csd
Crystallography Open Database (COD)	500,000	2003	An available collection of crystal structures of organic, inorganic, metal-organic compounds and minerals, excluding biopolymers.	http://crystallography.net/cod/browse.html
Materials Project	154,718	2011	A core program of the Materials Genome Initiative that reveals the properties of all known inorganic materials using high-throughput computing.[6]	https://next-gen.materialsproject.org/
Automatic Flow (AFLOW)	3,530,330	2012	A globally available database of 3,530,330 material compounds with over 734,308,640 calculated properties, and growing.[7]	http://www.aflowlib.org/
Hypothetical MOFs Database (hMOFs)	137,953	2012	A crystal structure database for screening hypothetical MOFs by large-scale assembly of metal clusters and organic ligands.[8]	http://hmofs.northwestern.edu
Computation-Ready, Experimental MOFs Database (CoRE)	5,109	2014	A collection of MOF structures derived from experimental data which can be immediately applied to molecular simulation.[9]	https://zenodo.org/records/7691378
	14,142	2019	A part of an update to the CoRE MOF 2014 Database including over 14 000 porous, three-dimensional MOF structures.[10]	
The Open Quantum	1,022,603	2015	An Open-access database of DFT calculated thermodynamic and structural properties for	https://www.oqmd.org/

Materials Database (OQMD)			more than 1 million materials, created in Chris Wolverton's group at Northwestern University.[11]	
Material Genome Engineering Databases (MGED)	710,050	2018	A database / application software integrated system platform based on material genetic engineering.	https://www.mgedata.cn/
Quantum MOF (QMOF)	14,482	2021	An online database of computed quantum-chemical properties for more than 14,000 experimentally synthesized MOFs.[12]	https://figshare.com/articles/dataset/QMOF_Database/13147324
MOFX-DB	over 160,000	2023	A publicly available Database of Computational Adsorption Data for Nanoporous Materials.[13]	https://mof.tech.northwestern.edu
ARC-MOF	over 280,000	2023	A diverse database of MOFs with DFT-derived partial atomic charges and descriptors.[14]	https://doi.org/10.5281/zenodo.6908727

Table S3. Summary of common material descriptors.

Categories	Descriptors
Geometrical descriptor	Pore size, dominant pore size, maximum pore size
	Available pore volume (V_a)
	Gravimetric surface
	Surface area (SA), Accessible surface area (ASA), volumetric surface area (VSA), gravity surface area (GSA)
	void fraction (VF)
	global cavity diameter (GCD)
	largest cavity diameter (LCD)
	pore limiting diameter (PLD)
	pore size distribution (PSD)
	pore volume (PV)
	density (ρ)
Topological descriptor	pore connectivity
	pore morphology
	porosity
	cavity size
	coordination numbers
	bond angles
	atom-specific persistent homology (ASPH)
Chemical descriptor	atomic type and number
	degree of unsaturation, total unsaturation
	electronegativity
	atomic composition
	electronic configurations
	metallic percentage
	oxygen to metal ratio (OMR)
	nitrogen to oxygen ratio (NOR)

Energy descriptor	crystal structure
	electrostatic potential-derived charge (ESPC)
	cohesive energies
	voronoi energies
	electronic band structure
	density of states
	heat of adsorption ($\Delta_{ads}H$)
	working capacity (ΔW)
	energy efficiency
	isosteric heat (Q_{st})
	Henry coefficient (K_H)
	effective point charge (EPoCh) [15]
	potential energy surface (PES)

Table S4. Three mainstream Cross Validation (CV) methods.

CV Type	Application situation	Advantages	Disadvantages
Hold-out CV	It is common in early tasks such as decision tree, naive Bayesian classifier, linear regression and logistic regression.	The dataset partition is simple and easy to operate.	Only part of the data is used in the model training, and the dataset is divided only once. So the result is accidental.[16]
LOOCV	It is suitable for small sample datasets.	All data points are utilized, so the bias is low.	Training is more complex and time-consuming.[17] And the validity of the test model changes greatly. Because testing for one data point, the estimated value of the model is greatly affected by the data point.
K-CV	It is suitable for large sample datasets.	The use of data is more efficient after multiple divisions, and the contingency of the results is greatly reduced, thereby improving the accuracy of the model.	Random and equal division of data is not suitable for datasets containing different categories.

Table S5. The application of MOFs combined with ML in the adsorption and separation of various gases.

Gas type	ML Algorithms	References
Methane (CH_4)	Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Poisson regression, Neural Network (NN), Unsupervised Transfer learning (TL), etc.	[18-27]
Carbon Dioxide (CO_2)	SVM, DT, RF, NN, Gradient Boosting Machines (GBM) Multiple Linear Regression (MLR), etc.	[28-36]
CO_2 / CH_4	DT, SVM.	[37]
Hydrogen (H_2)	DT, RF, Support Vector Regression (SVR), Linear Regression (LR), K-NearesNeighbor (KNN), Gradient Boosting Regression (GBR), etc.	[38-40]
CO_2 / H_2	Gradient Boosted Regression Tree (GBRT)	[41]
Nitrogen (N_2)	K-means clustering	[42]
O_2 / N_2	RF, GBRT, and Extreme Gradient Boosting (XGB).	[43]
Xenon / Krypton (Xe / Kr)	MOF-NET and Multi-Species Genetic Algorithm (MSGA) [44]	[45]
Ethane / Ethylene	RF, LR, DT, SVM, kNN, GBM, etc.	[46-48]
Propane / Propylene	RF, DT, etc.	[49-52]
Isobutene / Isobutane	LASSO, Elastic Net, SVM, XGBoost, Ridge Regression (RR), Bayes Regression (BR), and Artificial Neural Network (ANN).	[53]
Acetylene	DT, SVM, Gradient Boosting Decision Tree (GBDT), and Back Propagation Neural Network (BPNN).	[54]
High-sour natural gas	BPNN and the Partial Least-Square (PLS)	[55]
H_2S / CO_2 / CH_4	XGB, GBRT, Multi-layer perceptron (MLP), and the model obtained from the Tree-based Pipeline Optimisation Tool (TPOT).	[56]

References

[1] D. Zagorac, H. Mueller, S. Ruehl, J. Zagorac, S. Rehme, *J. Appl. Crystallogr.* **2019**, *52*, 918.

[2] G. Bergerhoff, R. Hundt, R. Sievers, I. D. Brown, *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 66.

[3] A. Belsky, M. Hellenbrandt, V. L. Karen, P. Luksch, *Acta Crystallogr B* **2002**, *58*, 364.

[4] P. Z. Moghadam, A. Li, S. B. Wiggin, A. Tao, A. G. P. Maloney, P. A. Wood, S. C. Ward, D. Fairen-Jimenez, *Chem. Mater.* **2017**, *29*, 2618.

[5] C. R. Groom, F. H. Allen, *Angew Chem Int Edit* **2014**, *53*, 662.

[6] A. Jain, O. Shyue Ping, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, *Apl Mater.* **2013**, *1*, 011002.

[7] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. W. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, O. Levy, *Comput. Mater. Sci.* **2012**, *58*, 227.

[8] C. E. Wilmer, M. Leaf, C. Y. Lee, O. K. Farha, B. G. Hauser, J. T. Hupp, R. Q. Snurr, *Nat. Chem.* **2012**, *4*, 83.

[9] Y. G. Chung, J. Camp, M. Haranczyk, B. J. Sikora, W. Bury, V. Krungleviciute, T. Yildirim, O. K. Farha, D. S. Sholl, R. Q. Snurr, *Chem. Mater.* **2014**, *26*, 6185.

[10] Y. G. Chung, E. Haldoupis, B. J. Bucior, M. Haranczyk, S. Lee, H. Zhang, K. D. Vogiatzis, M. Milisavljevic, S. Ling, J. S. Camp, B. Slater, J. I. Siepmann, D. S. Sholl, R. Q. Snurr, *J. Chem. Eng. Data* **2019**, *64*, 5985.

[11] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Ruehl, C. Wolverton, *npj Comput. Mater.* **2015**, *1*, 15010.

[12] A. S. Rosen, S. M. Iyer, D. Ray, Z. Yao, A. Aspuru-Guzik, L. Gagliardi, J. M. Notestein, R.

Q. Snurr, *Matter* **2021**, *4*, 1578.

[13] N. S. Bobbitt, K. Shi, B. J. Bucior, H. Chen, N. Tracy-Amoroso, Z. Li, Y. Sun, J. H. Merlin, J. I. Siepmann, D. W. Siderius, R. Q. Snurr, *J. Chem. Eng. Data* **2023**, *68*, 483.

[14] J. Burner, J. Luo, A. White, A. Mirmiran, O. Kwon, P. G. Boyd, S. Maley, M. Gibaldi, S. Simrod, V. Ogden, T. K. Woo, *Chem. Mater.* **2023**, *35*, 900.

[15] I. B. Orhan, T. C. Le, R. Babarao, A. W. Thornton, *Commun. Chem.* **2023**, *6*, 214.

[16] Q. Li, Int. J. Comput. Ass. Rad. **2006**, *1*, 347.

[17] J. Zhang, S. Wang, *Neural Comput. Appl.* **2016**, *27*, 1717.

[18] R. Wang, Y. Zhong, L. Bi, M. Yang, D. Xu, *Acs Appl. Mater. Interfaces* **2020**, *12*, 52797.

[19] R. Gurnani, Z. Yu, C. Kim, D. S. Sholl, R. Ramprasad, *Chem. Mater.* **2021**, *33*, 3543.

[20] X. Wei, Z. Lu, Y. Ai, L. Shen, M. Wei, X. Wang, *Sep. Purif. Technol.* **2024**, *330*, 125291.

[21] X. Wu, Z. Cao, X. Lu, W. Cai, *Chem. Eng. J.* **2023**, *459*, 141612.

[22] R. Wang, Y. Zou, C. Zhang, X. Wang, M. Yang, D. Xu, *Microporous Mesoporous Mater.* **2022**, *331*, 111666.

[23] M. Suyetin, Faraday Discuss. **2021**, *231*, 224.

[24] S. Y. Kim, S. I. Kim, Y. S. Bae, *J. Phys. Chem. C* **2020**, *124*, 19538.

[25] Z. Gulsoy, K. B. Sezginel, A. Uzun, S. Keskin, R. Yildirim, *Acs Comb. Sci.* **2019**, *21*, 257.

[26] G. S. Fanourgakis, K. Gkagkas, E. Tylianakis, E. Klontzas, G. Froudakis, *J. Phys. Chem. A* **2019**, *123*, 6080.

[27] M. Pardakhti, E. Moharreri, D. Wanik, S. L. Suib, R. Srivastava, *Acs Comb. Sci.* **2017**, *19*, 640.

[28] B. Zheng, F. L. Oliveira, R. N. B. Ferreira, M. Steiner, H. Hamann, G. X. Gu, B. Luan, *Acs Nano* **2023**, *17*, 5579.

[29] S. Kancharlapalli, R. Q. Snurr, *Acs Appl. Mater. Interfaces* **2023**, *15*, 28084.

[30] Z. Zhang, X. Cao, C. Geng, Y. Sun, Y. He, Z. Qiao, C. Zhong, *J. Membr. Sci.* **2022**, *650*, 120399.

[31] Z. Sun, Y. Liao, S. Zhao, X. Zhang, Q. Liu, X. Shi, *J. Mater. Chem. A* **2022**, *10*, 5174.

[32] K. Choudhary, T. Yildirim, D. W. Siderius, A. G. Kusne, A. Mcdannald, D. L. Ortiz-Montalvo, *Comput. Mater. Sci.* **2022**, *210*, 111388.

[33] M. N. Amar, H. Ouaer, M. A. Ghriga, *Fuel* **2022**, *311*, 122545.

[34] X. Deng, W. Yang, S. Li, H. Liang, Z. Shi, Z. Qiao, *Appl. Sci.-Basel* **2020**, *10*, 569.

[35] T. D. Burns, K. N. Pai, S. G. Subraveti, S. P. Collins, M. Krykunov, A. Rajendran, T. K. Woo, *Environ. Sci. Technol.* **2020**, *54*, 4536.

[36] R. Anderson, J. Rodgers, E. Argueta, A. Biong, D. A. Gomez-Gualdron, *Chem. Mater.* **2018**, *30*, 6325.

[37] M. Z. Aghaji, M. Fernandez, P. G. Boyd, T. D. Daff, T. K. Woo, *Eur. J. Inorg. Chem.* **2016**, 4505.

[38] J. Park, Y. Lim, S. Lee, J. Kim, *Chem. Mater.* **2023**, *35*, 9.

[39] K. Salehi, M. Rahmani, S. Atashrouz, *Int. J. Hydrogen Energy* **2023**, *48*, 33260.

[40] X. Zhang, Q. R. Zheng, H. Z. He, *J. Taiwan Inst. Chem. Eng.* **2022**, *138*, 104479.

[41] H. Dureckova, M. Krykunov, M. Z. Aghaji, T. K. Woo, *J. Phys. Chem. C* **2019**, *123*, 4133.

[42] M. Fernandez, A. S. Barnard, *Acs Comb. Sci.* **2016**, *18*, 243.

[43] Y. Yan, Z. Shi, H. Li, L. Li, X. Yang, S. Li, H. Liang, Z. Qiao, *Chem. Eng. J.* **2022**, *427*, 131604.

[44] S. Lee, B. Kim, H. Cho, H. Lee, S. Y. Lee, E. S. Cho, J. Kim, *Acs Appl. Mater. Interfaces* **2021**, *13*, 23647.

[45] Y. Lim, J. Park, S. Lee, J. Kim, *J. Mater. Chem. A* **2021**, *9*, 21175.

[46] P. Halder, J. K. Singh, *Energy Fuels* **2020**, *34*, 14591.

[47] Z. Wang, T. Zhou, K. Sundmacher, Chem. Eng. J. **2022**, *444*, 136651.

[48] Y. Wu, H. Duan, H. Xi, Chem. Mater. **2020**, *32*, 2986.

[49] V. Daoo, J. K. Singh, Acs Appl. Mater. Interfaces **2024**, *16*, 6971.

[50] H. Tang, Q. Xu, M. Wang, J. Jiang, *Acs Appl. Mater. Interfaces* **2021**, *13*, 53454.

[51] Y. Wang, Z.-J. Jiang, D.-R. Wang, W. Lu, D. Li, *J. Am. Chem. Soc.* **2024**, *146*, 6955.

[52] X. Xue, M. Cheng, S. Wang, S. Chen, L. Zhou, C. Liu, X. Ji, *Ind. Eng. Chem. Res.* **2023**, *62*, 1073.

[53] X. Sun, W. Lin, K. Jiang, H. Liang, G. Chen, *Phys. Chem. Chem. Phys.* **2023**, *25*, 8608.

[54] P. Yang, G. Lu, Q. Yang, L. Liu, X. Lai, D. Yu, *Green Energy Environ.* **2022**, *7*, 1062.

[55] H. Liang, W. Yang, F. Peng, Z. Liu, J. Liu, Z. Qiao, *Apl Mater.* **2019**, *7*, 091101.

[56] W. Gao, W. Zheng, K. Yan, W. Sun, L. Zhao, *Fuel* **2023**, *350*, 128757.