Machine learning prediction of open metal sites in metal-organic framework catalysts

Hui Xu an Guo1, Jiahao Ma2, Yueshan Zhong3, Yuhui Yin4, and Mingcan Chen5

1 School of Chemistry, Chemical Engineering and Biotechnology, Nanyang Technological University, Singapore 639798
2 Department of Marine Science and Technology, Harbin Institute of Technology, Weihai, Shandong Province, China 264200
3 Department of Biological Science, Xian Jiaotong Liverpool University, Suzhou, Jiangsu Province, China 215000
4 School of Chemical Engineering and Technology, State Key Laboratory of Chemical Engineering, Tianjin University, Tianjin, China 300072
5 Joint School of National University of Singapore and Tianjin University, International Campus of Tianjin University, Binhai New City, Fuzhou, China 350207

* Emails: 1 C210177@e.ntu.edu.sg; 2 2190790308@stu.hit.edu.cn; 3 Yueshan.Zhong20@student.xjtlu.edu.cn; 4 yinyuhui@tju.edu.cn; 5 mingcanchen7@163.com

Abstract

Metal–organic frameworks (MOFs) are porous materials assembled from inorganic metal clusters ("secondary building units") and organic building blocks, where these building units are

Introduction

Metal–organic frameworks (MOFs) are porous materials assembled from inorganic metal clusters ("secondary building units") and organic building blocks, where these building units are
arranged to form an extended three-dimensional (or two-dimensional) crystalline solid material. Their excellent porosity, satisfactory surface area and unprecedent structural diversity have inspired numerous research on their applicability in fields including gas adsorption, separation and catalysis, many of which have yielded valuable scientific insights. The combination of organic and inorganic chemistry not only contributes to the versatility of MOFs, but also shapes the promising future of MOF-related researches.¹

Many molecular modelers and researchers have been motivated to develop effective methods and models to assess the property of MOF according to a large number of possible structures of MOF and their potential applications. Some methods including machine learning models and large-scale highthroughput molecular simulations have been developed in this field. Despite these findings, the huge number of potential materials and conditions could still be a formidable challenge for model development which includes large datasets. Therefore, there is a clear need for algorithms that can efficiently explore the spaces while balancing simulation cost and prediction accuracy.

Open metal sites (OMS) are defined as unsaturated Lewis acid sites present on metal ions or clusters, which is used to be one of the most important performance criteria for MOFs.⁵ MOFs with high density OMS have recently been reported to have better performance in the separation work, such as hydrogen isotopes, carbon dioxide and Xe/Kr. Meanwhile, another significant function of MOFs, catalysis, also depends on the OMS. Many methods to improve the catalytic performance of MOFs are also based on OMS.⁶ It was even reported that MOFs with multiple OMS have better performance in cancer treatment.⁷

Often functioning as the most powerful binding sites, OMS are being actively studied with the aim of raising the performance of MOFs in various fields. For instance, various MOFs have been identified as excellent adsorbents for molecules such as H₂, CO₂, I₂, CO and acetylene, in which the binding of molecules with OMS plays a significant role. More detailed descriptions are available in the thorough review by Easun et al.²⁷

Despite their evident importance as mentioned, OMS detection remains a field at infancy. In situ measurements of OMS are usually conducted by infrared spectroscopy (IR), adsorption heat change, or Temperature-Programmed Desorption (TPD). Other techniques include indicators, titration, characteristic reactions, NMR, diffraction, etc. Their indirect nature and the resulting dependency on the crystalline structure of the MOF and the choice of probe molecules have led to their inevitable fragility to external distractions and noises, affecting the accuracy of their results. Research investigating the OMS sites on simulated MOF structures has been emerging as well with the surge of computational MOF analysis. Popular methods consist of grand canonical Monte Carlo (GCMC), electronic structure calculations and simulations (BSSP) and density functional theory (DFT) calculations. Nonetheless, the adsorbing ability of OMS was discovered to be underestimated by DFT, while the considerable computational cost of such first-principle methods and ensemble simulations still acts as a major hindrance against efficient OMS detection.

Random Forest (RF) and Extreme Gradient Boosting (XGBoost) are two of the numerous algorithms widely implemented in MOF-related fields. They were proved to perform well in predicting adsorption and diffusion properties of various chemical species, electronic and force field parameters, and catalytic abilities including carbon fixing and photodegradation.

In this research, the aforementioned algorithms were trained on CoRE 2019, a dataset comprising of 12,000+ computation-ready experimental MOF structures using geometric-based, elemental and electronic descriptors following the MENA criteria. The accuracy of model prediction as well as feature importance was obtained and analyzed.

RF was found to perform better in the prediction, classifying 85% of the test data correctly with an f1 score of 0.891. XGBoost classified 80.6% of the test data correctly, with an f1 score of 0.865.
It was observed that fraction of electrons in d orbitals, average ionization energy, average metal electron affinity, average number of metal atoms and average boiling point were the most descriptive features in predicting the existence of OMS.

Methods

Dataset and descriptors

CoRE 2019 was chosen as the primary database of interest in this research, mainly in consideration of its reliable experimental and simulation-refined data source\(^\text{47,48}\). Compared to databases generated purely in silico, structures with proved existence from CoRE 2019 add to the credibility of our model, enhancing its application value.

According to the MENA (Meaningful, Efficient, small descriptor Number, Accurate) criteria proposed by Tawfik et al.\(^\text{49}\), the input descriptors were categorized into: (i) Geometry-based features, (ii) Elemental features, (iii) Electronic structure features. Detailed information of the descriptors were listed in Table 1. As the definitions of largest cavity diameter, pore limiting diameter and largest sphere along the free path from previous literatures vary\(^\text{50,51}\), a graphical representation of the definition applied in this research was included as Figure 1a. Figure 1b illustrated solvent accessible and non-accessible (excluded) areas. All features were normalized for better training.

Table 1. List of all descriptors. Corresponding explanations of geometry-based descriptors are attached, while definitions of elemental and electronic features follow the convention. Three versions of each elemental and electronic feature – the weighted average of all atoms, metal atoms only and non-metal atoms only respectively – were included in the model.

<table>
<thead>
<tr>
<th>Geometry-based descriptors</th>
<th>Elemental features</th>
<th>Electronic structure features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Largest cavity diameter (Å)</td>
<td>The diameter of the largest sphere that can fit in a pore</td>
<td>Electronegativity</td>
</tr>
<tr>
<td>Pore limiting diameter (Å)</td>
<td>The diameter of the largest rigid sphere capable of moving through a path</td>
<td>Atomic mass (amu)</td>
</tr>
<tr>
<td>Largest sphere along the free path (Å)</td>
<td>The largest spherical diameter along a path</td>
<td>Atomic radius (Å)</td>
</tr>
<tr>
<td>MOF density (g/cm(^3))</td>
<td>Mass density of a given MOF structure</td>
<td>Ionization energy (eV)</td>
</tr>
<tr>
<td>Accessible surface area (m(^2)/cm(^3))</td>
<td>Surface area accessible to solvent molecules</td>
<td>Electron affinity (eV)</td>
</tr>
<tr>
<td>Non-accessible surface area (m(^2)/cm(^3))</td>
<td>Solvent-excluded surface area</td>
<td>Melting point (K)</td>
</tr>
<tr>
<td>Void fraction</td>
<td>The fraction of the volume of void space over the total volume</td>
<td>Boiling point (K)</td>
</tr>
<tr>
<td>---------------</td>
<td>---------------------------------------------------------------</td>
<td>------------------</td>
</tr>
<tr>
<td>Accessible volume (cm³)</td>
<td>The volume accessible to solvent molecules</td>
<td></td>
</tr>
<tr>
<td>Non-accessible volume (cm³)</td>
<td>The volume not accessible to solvent molecules (isolated pockets)</td>
<td></td>
</tr>
<tr>
<td>Number of atoms</td>
<td>Total number of atoms in the simplest MOF building block</td>
<td></td>
</tr>
<tr>
<td>Volume</td>
<td>Total volume of one MOF building block</td>
<td></td>
</tr>
<tr>
<td>Chemical composition (represented by corresponding symbols of the element)</td>
<td>Number of the respective atom divided by the total number of atoms in the simplest MOF building block</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 1.** a) Definitions of pore limiting diameter, largest sphere along the free path and largest cavity diameter, represented in a MOF building unit. b) Illustration of solvent accessible surface and solvent excluded surface.

**Machine learning methods**

As reviewed in the previous section, XGBoost and Random Forest (RF) were chosen to analyze data in this research. XGBoost provides parallel tree boosting which plays an important role in leading the machine learning library for ranking problems, classification, and regression. For the mathematical explanation of XGBoost, it could be concluded with this equation:

\[
L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \sum_k (f_k)
\]

where the first part of the equation stands for the loss function term and the second part is the regularization term. The residual is the difference between the real value and the calculated value.
Additionally, RF is a widely used machine learning tool that combines the output of multiple decision trees to reach a single result, and the mathematical theory of RF could be described with a function:

\[
L(x) = \frac{1}{n} \sum_{i=1}^{n} \alpha_i T_i(x)
\]  

(2)

**Data pre-processing**

In this study, the data processing library pandas is used for data preprocessing, including one hot coding for chemical components and Min Max standardization for all data.

\[
MaxMinNormalisation(x) = \frac{x - \min(x)}{\max(x) - \min(x)}
\]  

(3)

**Feature importance score**

Feature importance can be calculated by impurity-based methods or permutation methods. In the research, the permutation feature importance for analyzing data was employed. Permutation feature importance is based on the reduction in a model score with randomly shuffled single feature values. Particularly, this model inspection technique shuffles features n times and refits to estimate the importance. This method is mainly focused on breaking the association between the target and the feature, which makes it possible to measure how the model depends on the feature with an indicative drop in the model score. This is especially helpful for avoiding bias toward high-cardinality features and non-linear or opaque estimators.

On the other hand, the XGBoost algorithm calculated importance score based on gain, which refers to the improvement of classification accuracy when a feature is included in the prediction.

**Results and discussion**

**Model accuracy**

**Table 2.** The tuning parameters for RF. ‘Estimator’ refers to the maximum number of iterations of the weak learners. ‘Depth’ refers to the maximum depth of decision tree.

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Estimator</th>
<th>Depth</th>
<th>Min. no. of leaf nodes</th>
<th>Min. no. of samples required for partition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal Value</td>
<td>174</td>
<td>22</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

174 estimators were used in the model. Generally speaking, `n_estimators` is too small, easy to underfit, `n_estimators` is too large, and easy to overfit, and generally choose a moderate value. The maximum depth of the decision tree was 22. The default is "None", and the decision tree does not limit the depth of the subtree, so that each leaf node has only one category, or that it reaches the `min_samples_split`. In general, this value can be ignored when there is little data or few features. If the model has a large sample size and many features, the recommendation limits this maximum depth, and the specific value depends on the distribution of the data. Common values can be between 10-100.
The minimum number of leaf nodes was 1. This value limits the number of samples with the least leaf nodes, which are pruned together with the siblings if the number of certain leaf nodes is less than the number of samples. The default is 1, which can enter an integer of the minimum number of samples, or a percentage of the total number of samples. If the sample size is not large, this value is not needed.

The minimum number of samples required for internal node partition in the model was 2. This value limits the condition under which the subtrees continue to divide with a smaller number of samples than \textit{min\_samples\_split}. The default is 2, and the value is not required if the sample size is not large. The sample size is recommended if it is very large.

**Figure 1.** a) Confusion matrix based on test data: Random forests perform better in positive predictions. b) As the proportion of training data increases, the recall rate and accuracy rate change. c) The AUC (Area under the ROC Curve) of random forest is larger, with reliable accuracy. d) XGBoost recall rate drops faster, indicating a lower precision.
As observed in Fig. 1a, RF correctly predicted a total of 3030 out of 3560 (85.1%) test data points, with an f1 score of 0.891. XGBoost correctly predicted a total of 2910 out of 3610 (80.6%) test data points, with an f1 score of 0.865 and a mean squared error (MSE) of 0.183. Fig. 1b plotted the relationship of classification threshold versus precision and recall respectively. While the precision line of the two models behaved similarly, the recall rate in XGBoost dropped slightly more rapidly than RF as the classification threshold increased. Fig. 1c showed the ROC curves of both models. The AUC of RF model was larger than that of XGBoost, indicating a better prediction performance. Fig. 1d indicated the relationship between precision and recall. The data from both RF and XGBoost exhibited a descending trend, but the rate of descendence of XGBoost results was higher. Therefore, the average precision (AP) of the XGBoost model was lower.

Feature importance

Feature importance data were extracted from models trained with both XGBoost and RF. The top 10 most determinative features were plotted with corresponding importance score in a descending manner, as illustrated in Fig. 2.

Figure 2. a) All of the top 3 features describe electronic or elemental properties of atoms in MOF. b) Fraction of electrons in d orbitals was rated as the top 3 most important features in both algorithms. Despite slight difference in order, the list of most important features provided by the two algorithms are similar.

As illustrated in Fig. 2a, 5 out of 10 features with highest importance belonged to elemental features, 2 were electronic features and 3 were classified as geometry-based features. The strongest correlations were found between the existence of OMS and ionization energy, fraction of electrons d orbitals and the number of metal atoms, with importance scores of 0.011524, 0.007724 and 0.007518 respectively. A relatively large importance gap between the top feature and others were observed.

The number of elemental, electronic and geometry-based features in Fig. 2b were 6, 2, and 2 respectively. The top three most important features were average metal electron affinity, average boiling point and fraction of electrons in d orbitals, with importance scores of 0.078510, 0.060384 and 0.049473 respectively.

Among the features, fraction of electrons in d orbitals was rated by both algorithms as one of the top 3 most important features. The remaining 2 of the top 3 features assessed by one algorithm was both present in the top 10 most important features suggested by the other algorithm. The difference in order might result from the algorithms’ different mechanisms of decision tree selection.
Analysis of the top three features from both RF and XGBoost algorithms with respect to MOFs with and without OMS respectively was performed. The results and possible rationale are as follows.

**Fraction of electron in d orbitals**

![Fraction of electron in d orbitals](image)

**Figure 3.** MOFs with OMS have a slightly higher average number of electrons in d orbitals. The blue vertical line indicates the weighted average of the value among MOFs without OMS, while the orange vertical line indicates that of MOFs with OMS.

Fig. 3 indicated that the majority of MOF building units have around 1 electron in d orbitals on average. MOFs with OMS have slightly more electrons in d orbitals than those without OMS. Since elements of short periods such as H, N and O with no d electrons were also included in the calculation of weighted average, it is reasonable to infer that more MOFs with OMS contain metals with multiple d electrons such as Cr, Zr and Fe, which are relatively easy to achieve high valence states.

The dissociation energy of coordinative bonds is often lower than covalent bonds, making metal-ligand complexes liable to thermal degradation. According to coordination chemistry principles, high valence metal ions increase the electrostatic interaction between metal ions and ligands, strengthening their thermal stability. Therefore, the MOF structure constructed by high valence metal ions are more likely to remain integrated when excessive ligands are removed by heating to form OMS.

**Average ionization energy & average metal electron affinity**

![Average ionization energy & average metal electron affinity](image)

**Figure 4.** a) The majority of MOFs with OMS have lower average ionization energy compared to MOFs without OMS. b) Significantly more MOFs with OMS have high average metal electron affinity. The
blue vertical line indicates the weighted average of the value among MOFs without OMS, while the orange vertical line indicates that of MOFs with OMS.

As illustrated in Fig. 4a-b, the average ionization energy of MOFs with OMS was slightly lower than those without OMS, while the average metal electron affinity was higher. The activation of open metal sites in MOFs is often seen as a key in applications. The OMSs in MOFs often represent the strongest binding sites, which leads to an increased interaction with different sorbate molecules, compared to MOFs where the metal sites are fully occupied. Both lower ionization energy and higher electron affinity indicate easier ability to gain a charge, leading to more stable metal-ligand bonds that contributes to the formation of OMS by preventing thermal degradation and structural collapse.

Average number of metal atoms

![Average number of metal atoms](image)

**Figure 5.** The mean fraction of metal atoms in MOFs with OMS is higher than those without OMS. The blue vertical line indicates the weighted average of the value among MOFs without OMS, while the orange vertical line indicates that of MOFs with OMS.

**Fig. 5** illustrated that the mean fraction of metal atoms in MOFs with OMS is higher than those without OMS. As organic linkers are dissociated from the metal clusters to form OMS, the composition of non-metallic atoms in a MOF building unit drops, leading to an increase in fraction of metal atoms\(^{30}\). In addition, a high fraction of metal atoms usually indicates the presence of multi-atom metal clusters that increase the connectivity of MOF structures, making the building units structurally stable\(^ {32}\). Therefore, they are more capable of tolerating structural defects such as OMS.

Average boiling point

![Average boiling point](image)

**Figure 6.** The average boiling points are higher in MOFs with OMS. The blue vertical line indicates the weighted average of the value among MOFs without OMS, while the orange vertical line indicates that of MOFs with OMS.
Generally, the boiling point of nitrogen, oxygen and hydrogen have relative lower boiling points (N: -196 °C; O: -183 °C; H: -253 °C) compared with metal ions, and some common metal ions shown in the open metal sites are Cr3+, Fe3+, and Zr4+, these elements have relative high values in boiling points (Cr: 2672 °C; Fe: 2750 °C, Zr: 4377 °C), which may result boiling point an effective predictor for the existence open metal site.

Conclusions

In this study, XGBoost and Random Forest algorithms were employed to predict the existence of OMS in MOF structures in an accurate and cost-efficient manner. The determinative features as reported by the models were further analyzed and proper relationships between physically meaningful descriptors and OMS were established. By offering simple guidance to the synthetic exploration of MOFs with OMS, the study aims at accelerating the rational design and discovery of open-framework MOF structures with high value in adsorption, separation, and catalysis. The authors hope that the research can inspire further exploration of more suitable features, especially in the electronic structure category, as well as more accurate interpretation of the important features.

Acknowledgments

This work was supported by Touch Education Technology Inc. We acknowledge scientific & editorial support from Dr. J. S. Lim of University of California Los Angeles; and administrative support from C. Ding of Touch Education Technology Inc.

Author contributions

• Y.Z: Conceptualization, investigation, formal analysis, writing – reviewing and editing, visualization
• H.G: Conceptualization, implementation of code and algorithms, formal analysis, literature review, data curation, writing – original draft & editing, visualization
• Y.Y: Dataset generation, dataset processing, implementation of code & algorithms, tuning hyperparameters, data curation
• J.M.: Literature review of OMS application, Data normalization and visualization, Machine learning method and comparison, optimization and determination of some parameters and corresponding manuscript writing
• C.M.: Conceptualization, literature review, explanation of some concepts, writing and editing

Competing financial interests

The authors declare no competing financial interests.

References


Chung, Y. G. et al. (Zenodo, 2020).
