

Preview

Preview of machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery

Sarah Callaghan^{1,*}

¹Cell Press, 50 Hampshire St., Cambridge, MA, USA

*Correspondence: s.callaghan@cell.com

<https://doi.org/10.1016/j.patter.2021.100239>

Metal–organic frameworks (MOFs) are a class of chemical compounds used for the storage of gases such as hydrogen and carbon dioxide. They also have potential applications in gas purification, catalysis and as supercapacitors. A database of quantum-chemical properties for over 14,000 MOF structures (the “QMOF database”) has been created and made available to the community along with code for machine learning and other related resources.

In the materials science field over recent years, significant work has been done on the design of novel metal–organic frameworks (MOFs), which are a class of materials composed of discrete inorganic nodes connected to one another via organic linkers. MOFs are currently being studied as materials that can store gases such as hydrogen (useful for fuel cells) and carbon dioxide (useful for carbon capture). They are also potentially of use in gas purification and separation, in catalysis, and as conduction solids and supercapacitors. MOFs have predictable and atomically defined structures with properties that are directly related to their underlying metal and organic building blocks, which allows them to be tailored for specific physical and chemical functions.

This ease of prediction has allowed tens of thousands of MOFs to be synthesized and a vast number to be proposed based on considering different combinations of their constituent building blocks. But due to the number of possible compositions, structures, and resulting properties, it can be hard to identify the highest performing MOFs for a given purpose. Researchers have therefore turned to high-throughput computational screening approaches based on classical simulations to efficiently explore the vast combinatorial space of MOF structures.

Unsurprisingly, large quantities of data have been generated during these computational screening studies, and to better engage with these, machine learning (ML) models have been developed to accelerate the discovery and

design processes for MOFs. These have provided excellent results in the areas of gas storage and separations, but that is only one potential use of MOFs. An untapped seam of investigation for MOFs is one where they are best described by quantum mechanical models, such as those based on the electronic, optical, magnetic, and/or catalytic properties.

Aside from the very large number of possible MOFs than can be built, the large number of atoms in MOF crystal structures makes it computationally demanding to carry out even moderate-scale quantum-chemical screening studies. Hence significant progress has been made in the development of ML models that can accelerate the quantum-chemical screening process for a wide range of compounds. High-throughput density functional theory (DFT) workflows are used to construct large-scale electronic structure property databases, and the combination of high-throughput DFT databases and ML has led to the discovery of a diverse range of materials with sought-after properties, including efficient organic light emitting diodes, superhard inorganic materials, and thermally conductive polymers.

In the May 5, 2021 issue of *Matter*, Rosen et al.¹ take advantage of a recently developed high-throughput periodic DFT workflow tailored for MOF structures to construct a large-scale database of MOF quantum mechanical properties. This publicly available dataset—the Quantum MOF (QMOF) database—contains computed properties for over

15,000 experimentally characterized MOFs after structure relaxation via DFT, including but not limited to optimized geometries, energies, band gaps, charge densities, density of states, partial charges, spin densities, and bond orders.

The authors anticipate that the QMOF database will serve two primary purposes: (1) materials discovery using the as-deposited data and (2) the evaluation and development of novel ML algorithms to reduce or remove the need for otherwise expensive DFT calculations.

The work done by Rosen et al.¹ ties in well with another paper published in *Patterns* by Balzer et al.²: “Wiz: A web-based tool for interactive visualization of big data.” In the Balzer et al.² descriptor article, the authors presented Wiz, a freely accessible web app that anyone with a browser can use to interactively visualize their data in multi-dimensions. Visual analytical methods used on cleaned and structured data can provide useful insights to help us better understand data. In the article, the authors used MOFs as a use case and developed an interactive multi-dimensional visualization tool for producing thousands of unique structure-property plots to convey the full information obtained on oxygen storage for all the structures under study. The Wiz visualization tool can be used for any MOFs data to get new insight from a huge amount of data that is difficult or impossible to be meaningful at first glance.

The QMOF database was used to develop several ML models for the prediction of MOF band gaps from nothing more



than an encoding of the experimental (i.e., unrelaxed) crystal structures. This drastically reduces the number of computationally demanding quantum mechanical simulations that need to be carried out in future screening studies. As most MOFs are known to be electronically insulating, which limits their potential use in electrocatalysis, sensing, energy storage, and other applications where some degree of electrical conductivity is necessary, a ML model that can predict MOF band gaps is particularly desirable. In the article, the authors identify a top-performing band gap regression model based on a crystal graph convolutional neural network and show how dimensionality reduction techniques can be used to discover overarching structure-property relationships for the identification of MOFs with specific electronic structure properties.

The QMOF database now makes it possible to investigate research areas in the field that are reliant on a large database of quantum-chemical properties for MOFs. For instance, with the success of

transfer and multi-task learning, the QMOF database can help to increase the accuracy and reduce the required training set size for ML models predicting new MOF properties not already present in the QMOF database. Even outside the areas of high-throughput DFT screening, data mining, and ML, there are many possible domain use-cases for the QMOF database.

The QMOF database is a living resource with updates to the QMOF database planned. The authors welcome the development of subsets, modifications, and supplements to the database that suit the diverse needs of the MOF community. For the data scientist, the QMOF database forms a well-defined, well-described database with which to develop and test new ML algorithms and techniques.

Data and code availability for the QMOF database

The landing page for the QMOF database can be found at the following GitHub repository: <https://github.com/arsen93/>

QMOF. Data associated with the QMOF database are hosted via Figshare and have the following permanent DOI: [10.6084/m9.figshare.13147324](https://doi.org/10.6084/m9.figshare.13147324). All data associated with the original research article are made publicly available, including results from the DFT calculations, Python scripts to reproduce the machine learning analyses, code to reproduce the automated DFT screening process, and other related resources.

DECLARATION OF INTERESTS

S.C. is the editor-in-chief of *Patterns* but otherwise has no competing interests to declare.

REFERENCES

1. Rosen, A.S., Iyer, S.M., Ray, D., Yao, Z., Aspuru-Guzik, A., Gagliardi, L., Notestein, J.M., and Snurr, R.Q. (2021). Machine learning the quantum-chemical properties of metal-organic frameworks for accelerated materials discovery. *Matter* 4, 1–20.
2. Balzer, C., Oktavian, R., Zandi, M., Fairen-Jimenez, D., and Moghadam, P.Z. (2020). Wiz: A web-based tool for interactive visualization of big data. *Patterns* 1, 100107.