User Manual

Gianmarco Terrones*

HJK Group, MIT

1 Overview

The MOFSimplify website, available at `mofsimplify.mit.edu`, serves as an interface through which one can interact with the Computation-Ready Experimental (CoRE) [1] Metal-Organic Framework (MOF) database data curated by the Kulik group and the models trained by the Kulik group on this data [2]. Specifically, a user can attain property predictions on new MOFs by using MOFSimplify. These predictions are generated by artificial neural networks (ANNs) trained on the curated MOF data. In addition to these predictions, the user can see which CoRE MOFs are most similar to new MOFs as determined by the ANNs. *Our hope is that these ANN models help MOF experimentalists to focus their efforts on stable MOFs.*

2 Selecting a MOF

Before the user can visualize a MOF, get a MOF's components, or run a prediction on a MOF's stability, the user must select a MOF. The user has three ways to select a MOF for analysis on the website. These options are available in the Main tab under **1) Select a MOF (metal-organic framework) for analysis**.

The first way is to click on the **Example MOF** button, which allows a user to load either an HKUST-1 cif file or any CoRE 2019 MOF cif file (as identified by refcode).

The second way is to click on the **Custom (upload cif file)** button, which will allow the user to select a file from their computer to upload. The file must be a `cif` file; the website will alert the user if she accidentally upload something else. The rest of the functionality on the website has a better probability of success if the MOF the user uploads is "clean." In other words, the cif file should be disorder free and should not have floating solvent.

The third way is to click on the **Building block assembly** button, which redirects the user to a tab where she can generate a MOF out of building blocks. Before clicking the button, however, a linker, SBU (secondary building unit), and MOF net must be selected. The code behind MOF building block construction comes from `https://github.com/tobacco-mofs/tobacco_3.0`. This repository is affiliated with recent work [3] which builds off of the Topologically Based Crystal Constructor (ToBaCCo) [4]. The linkers and SBUs are pieced together according to the chosen net. Not all combinations of linker, SBU, and net can generate a MOF, and if the user selects an invalid combination and attempts generation, an alert is triggered. An example of a valid combination that a user can try is 3B_4H_Ch as the linker, 6c_Al_1 as the SBU, and acs as the net. MOFSimplify does not have the capability to make ToBaCCo MOFs with more than one type of linker or SBU, but a user can download the ToBaCCo 3.0 code for this purpose if she wishes to do so.

Whenever a new MOF is selected in any of the three possible ways, all visualizations, status messages, and predictions are cleared. The name of the selected MOF is displayed as the first line in the **Status messages and MOF predictions** section of MOFSimplify.

*gterrone@mit.edu

3 MOF visualization

The user can visualize the unit cell of the selected MOF in the Visualization tab. The code behind the visualization comes from 3Dmol.js [5] and also `https://github.com/snurr-group/web-mofid`, which is affiliated with recent work [6]. If a MOF has too large a unit cell, visualization on MOFSimplify will not work. I would recommend trying a program like VESTA in this case.

4 MOF component identification

The user can determine the linkers and SBUs that comprise the selected MOF in the Component Analysis tab, using the **Get selected (1) MOF components** button. The (1) in the button text refers to the selection described in Selecting a MOF. An alert is displayed if component identification fails. The operation tends to fail, or at least take longer, if the selected MOF has a very large unit cell. Once the MOF components have been identified, the dropdowns "Linkers" and "SBUs" are populated. The user can select a linker (SBU) from the appropriate dropdown menu and click on the **Linker** (**SBU**) button to visualize it. The SMILES string of the visualized component will also appear next to the visualization. In addition, there is a button **Filter by connectivity** which eliminates duplicate components using molecular graph determinants. This button can be useful when many non-unique linkers or SBUs in a MOF are incorrectly found to be distinct by the molSimplify code. However, this button filters out isomers as well, so upon being clicked it is replaced with an **No filter** button which gives the user the option to undo the filtering in the dropdowns. The **Reset zoom** button under the component visualization pane resets the zoom level of the visualized component. The selected linker's (SBU's) xyz file can be downloaded by clicking the **Download linker** (**Download SBU**) button.

The code behind MOF component identification is available at `molSimplify.Informatics.MOF.MOF_descriptors`. molSimplify is available on GitHub at `https://github.com/hjkgrp/molSimplify` and is affiliated with recent work [7].

5 MOF property prediction

The user can predict the selected MOF's stability using the **Stability upon solvent removal** or the **Thermal stability** buttons in the Main tab. Both of these operations generate features by invoking code from `molSimplify.Informatics.MOF.MOF_descriptors` to get chemical information in the form of Revised Autocorrelations (RACs) [8] and by making Zeo++ [9] calls to attain geometric information. A prediction is then made using an ANN trained by our group. There is one ANN trained for solvent removal stability predictions, and there is one ANN trained for thermal stability predictions. Data was compiled using a widescale search and natural language processing [2, 10] .

For solvent removal stability predictions, the model is performing a classification task but quantifies its certainty in its prediction by how close the output is to zero (unstable) or one (stable).

For the thermal stability prediction, in addition to a temperature prediction, a percentile for the current MOF's predicted thermal breakdown temperature relative to the training data experimental breakdown temperatures is displayed. Furthermore, a plot is displayed showing the current MOF's predicted breakdown temperature relative to the experimental breakdown temperatures of the CoRE MOF database MOFs used to train the thermal stability ANN.

If the user happens to upload a MOF that is in the training data of the solvent (thermal) ANN, and the **Stability upon solvent removal** (**Thermal stability**) button is pressed, MOFSimplify will return the ground experimental truth for that MOF instead of making an ANN prediction. As an aside, the presence of a MOF in the training data is determined by examination of RAC and Zeo++ features. These features/descriptors are the numbers used to represent a MOF to an ANN as an array of numbers. When a prediction is made on MOFSimplify, the descriptors generated for the selected MOF can be downloaded as a `csv` file by clicking the **Download descriptors** button in the Main tab.

Predictions (or ground truths) are color coded. For the solvent removal case, green corresponds to the model being confident that the MOF is stable upon solvent removal (or it represents ground truth stability). Red corresponds to the model being confident the MOF is unstable upon solvent removal (or it represents ground truth instability). Yellow corresponds to the model being less sure of its prediction. For the thermal case, green indicates the current MOF's predicted or ground truth thermal breakdown temperature is high relative to the training data. Red indicates the current

MOF's predicted or ground truth thermal breakdown temperature is low relative to the training data. Yellow indicates an intermediate thermal breakdown temperature.

If a prediction fails at any point, an alert is presented to the user. The models may take a long time to run or even fail to complete if the selected MOF has a very large unit cell. A prediction operation can take around 12 seconds for some MOFs. The code is probably running after the user has clicked a button; please be patient! The user should not click on another button if she just clicked on a button and its operation has not yet completed. Predictions are allowed a maximum of 1 minute to complete. If a prediction is not completed in this amount of time, the website aborts the operation.

5.1 ANN nearest neighbors

Recent work has evaluated ANN latent space distance as a method for quantifying uncertainty in an ANN prediction [11]. In brief, latent space distance quantifies the similarity between two model inputs after they have passed through most of the neural network. In our case, a model input is a MOF which is translated into a bunch of numbers that describe it. Two MOFs which have a small ANN latent space distance can be expected to have a similar property, regarding the property that that ANN predicts for. We call MOFs with the smallest latent space distance between them to be nearest neighbors.

The ANN nearest neighbor section of the website appears at the bottom of the Main tab after a prediction is run. When a solvent removal stability or thermal stability prediction is made in MOFSimplify, the corresponding ANN nearest neighbors dropdown is populated with five MOFs. These nearest neighbors are the CoRE MOFs that have the smallest latent space to the selected MOF.

Information about a neighbor can be displayed by clicking the **Show** button in the solvent removal or thermal latent space nearest neighbor section. The information displayed includes the neighbor's experimental ground truth (data mined by us) and the DOI of the paper associated with the neighbor. By clicking the **Download** button, the user can download a `txt` file containing this information, as well as the `cif` file for the neighbor.

The **Visualize** buttons are analogs of the **Visualize selected (1) MOF** button in the Visualization tab. The difference is that they do not visualize the selected MOF, but rather the neighbor currently selected in the solvent or thermal neighbor dropdown. Once the user clicks on the **Visualize** button, she should wait for a few seconds and not click anything else (excluding the alert popup). She will be redirected to the Visualization tab where a visualization will appear.

Likewise, the **Get components** buttons are analogs of the **Get selected (1) MOF components** button in the Component Analysis tab. There is a line above the Linkers dropdown in the Component Analysis tab which indicates the name of the MOF for which components are listed. Once the user clicks on the **Get components** button, she should wait for a few seconds and not click anything else. She will be redirected to the Component Analysis tab where components will be presented.

If a ground truth is returned instead of a prediction (see MOF property prediction), no ANN nearest neighbors are gathered.

Clicking on the **Show** button in the thermal nearest neighbor section displays, in addition to the things mentioned previously, a simplified thermogravimetric analysis (TGA) plot of the neighbor. The dot on the plot indicates the thermal breakdown temperature of the thermal neighbor MOF. The simplified TGA plots were extracted from TGA plots in the papers documenting the neighbor MOFs. For copyright reasons, the original TGA plots could not be displayed. Simplified plots were generated by choosing four points total, two on each segment of different slope on the TGA plot. A rapid decrease in mass indicates breakdown. The TGA plot is generated with Bokeh and can be downloaded (click the floppy disk icon).

6 User feedback

Help us improve our models! There are five feedback/community engagement forms on MOFSimplify. The user can give feedback on solvent removal or thermal stability predictions (or ground truths) after they are displayed. The user can also give feedback on stability ground truths mined from latent space nearest neighbor manuscripts, after the **Show** button is clicked for a given nearest neighbor. Lastly, the user can upload information on new MOFs using the Data Upload tab. We ask for emails in order to discourage spamming, and so that we can reach out to users if we have any questions. We ask for TGA traces as support for any stability claims. Finally, for new MOFs

sent to us in the Data Upload tab we ask for a cif file because we do not yet have a structure for the MOF the user is referring to (unlike in the four other feedback locations).

References

[1] Y. G. Chung, E. Haldoupis, B. J. Bucior, M. Haranczyk, S. Lee, H. Zhang, K. D. Vogiatzis, M. Milisavljevic, S. Ling, J. S. Camp, *et al.*, "Advances, updates, and analytics for the computation-ready, experimental metal–organic framework database: CoRE MOF 2019," *Journal of Chemical & Engineering Data*, vol. 64, no. 12, pp. 5985–5998, 2019.

[2] A. Nandy, C. Duan, and H. J. Kulik, "Using machine learning and data mining to leverage community knowledge for the engineering of stable metal–organic frameworks," *Journal of the American Chemical Society*, vol. 143, no. 42, pp. 17535–17547, 2021.

[3] R. Anderson and D. A. Gómez-Gualdrón, "Increasing topological diversity during computational "synthesis" of porous crystals: how and why," *CrystEngComm*, vol. 21, no. 10, pp. 1653–1665, 2019.

[4] Y. J. Colón, D. A. Gómez-Gualdrón, and R. Q. Snurr, "Topologically guided, automated construction of metal–organic frameworks and their evaluation for energy-related applications," *Crystal Growth & Design*, vol. 17, no. 11, pp. 5801–5810, 2017.

[5] N. Rego and D. Koes, "3dmol.js: molecular visualization with WebGL," *Bioinformatics*, vol. 31, no. 8, pp. 1322–1324, 2015.

[6] B. J. Bucior, A. S. Rosen, M. Haranczyk, Z. Yao, M. E. Ziebel, O. K. Farha, J. T. Hupp, J. I. Siepmann, A. Aspuru-Guzik, and R. Q. Snurr, "Identification schemes for metal–organic frameworks to enable rapid search and cheminformatics analysis," *Crystal Growth & Design*, vol. 19, no. 11, pp. 6682–6697, 2019.

[7] E. I. Ioannidis, T. Z. Gani, and H. J. Kulik, "molSimplify: A toolkit for automating discovery in inorganic chemistry," 2016.

[8] J. P. Janet and H. J. Kulik, "Resolving transition metal chemical space: Feature selection for machine learning and structure–property relationships," *The Journal of Physical Chemistry A*, vol. 121, no. 46, pp. 8939–8954, 2017.

[9] T. F. Willems, C. H. Rycroft, M. Kazi, J. C. Meza, and M. Haranczyk, "Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials," *Microporous and Mesoporous Materials*, vol. 149, no. 1, pp. 134–141, 2012.

[10] A. Nandy, G. Terrones, N. Arunachalam, C. Duan, D. W. Kastner, and H. J. Kulik, "MOFSimplify: Machine learning models with extracted stability data of three thousand metal-organic frameworks," *arXiv preprint arXiv:2109.08098*, 2021.

[11] J. P. Janet, C. Duan, T. Yang, A. Nandy, and H. J. Kulik, "A quantitative uncertainty metric controls error in neural network-driven chemical discovery," *Chemical science*, vol. 10, no. 34, pp. 7913–7922, 2019.