Persistent homology advances interpretable machine learning for nanoporous materials

Aditi S. Krishnapriyan1, 2, Joseph Montoya2, Jens Hummelshøj2, and Dmitriy Morozov1

1Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States of America
2Toyota Research Institute, Los Altos, California 94022, United States of America

Abstract

Machine learning for nanoporous materials design and discovery has emerged as a promising alternative to more time-consuming experiments and simulations. The challenge with this approach is the selection of features that enable universal and interpretable materials representations across multiple prediction tasks. We use persistent homology to construct holistic representations of the materials structure. We show that these representations can also be augmented with other generic features such as word embeddings from natural language processing to capture chemical information. We demonstrate our approach on multiple metal–organic framework datasets by predicting a variety of gas adsorption targets. Our results show considerable improvement in both accuracy and transferability across targets compared to models constructed from commonly used manually curated features. Persistent homology features allow us to locate the pores that correlate best to adsorption at different pressures, contributing to understanding atomic level structure-property relationships for materials design.

1 Introduction

Metal–organic frameworks (MOFs), and the family of nanoporous materials in general, have exceptional properties beneficial for a number of applications. They have played an important role in natural gas storage and delivery, as adsorbents for separation, and in catalysis applications [1, 2, 3]. The number of MOF structures is massive: there are many experimentally synthesized structures, but also many more hypothesized structures. Accordingly, there have been increasing efforts to employ computational approaches to navigate the vast space of MOFs and to choose promising candidates for various applications. These computational approaches have played an important role in the field by creating more guided directions for experiments and by providing some of the underlying theory to explain experimental results. Molecular simulations and ab initio calculations have allowed computation of key properties such as gas adsorption, Henry’s coefficients, and diffusion [4]. They belong to the wider toolbox of high-throughput screening tools aimed at accelerating materials property prediction and design.

More recently, machine learning techniques have entered this toolbox [5, 6]. An ML model learns a function fitted on a set of known properties that represent a material. This model is then used to accurately predict a target value for a new material that was not included in the learning process. The computational cost of machine learning algorithms is appealing: they are often several orders of magnitude faster than the conventional simulation approaches.

For nanoporous materials, one of the most desired properties to predict is the adsorption capacity of a gas at given temperature-pressure conditions. Different gas adsorptions have different applications: for example, adsorption of methane is relevant to novel on-board vehicular fuel storage technologies [7], while adsorption of carbon dioxide at low pressure is important for capture and storage of emissions from flue gases [8]. Often, multiple physical models are necessary to capture adsorption across the full range of pressures.
Creating a universal material representation, suitable for all different prediction tasks, is incredibly complicated. Typically, domain experts select specific material features as the model input: these features are usually context specific, used for making predictions about a particular property of interest.

The construction of descriptors that are able to achieve good prediction performance on different properties, as well as at different conditions, is a challenging task. For example, in the case of gas adsorption at high pressure, guest molecules tend to occupy the entire void space in a material. Thus, accessible surface area is commonly used in predictive models. In contrast, for gas adsorption at low pressures, the guest molecules aggregate in the strongly binding regions of the material’s pore — standard structural descriptors are not able to capture this information as well. Additionally, chemical interactions of the system are important in determining some gas adsorption properties; this information also needs to be encoded in material descriptors.

The current state of the field suggests a significant need and potential payoff for the development of robust generic descriptors that would support machine learning models across many targets. Much of the previous work in this area has focused on using basic porosity descriptors such as accessible surface area (ASA), largest cavity diameter (LCD), and pore limiting diameter (PLD) [10, 11, 12, 13, 14]. Although these features perform well for some target properties, there is a lot of room for improvement in capturing finer description of the pore structures. Moreover, it is hard to interpret these features in terms of the specific porous frameworks that make up nanoporous materials.

In this paper, we use an alternative to these standard structural descriptors. We apply topological data analysis, specifically, persistent homology [15], to compute signatures of the channels and voids in the pores of the material, recording their number and size. These topological signatures are then converted to vector representations, allowing supervised machine learning algorithms to use them as input for predicting target material properties. The topological features we compute are generic, not tailored to a specific application. Instead, the labels in the training data allow machine learning algorithms to decide which subset of the features is important to a particular target prediction. As we demonstrate, these topological descriptors beat the standard structural descriptors in predicting a variety of materials properties.

We previously introduced this approach with an application to learning zeolite methane adsorption across different pressures [16], where we showed that topological descriptors consistently outperformed conventional structural descriptors. In this paper, we extend this technique to MOFs and demonstrate its merit for more complicated structures with significantly more diverse elemental compositions.

There has been previous work applying topological data analysis to nanoporous materials [17]. This work used standard metrics for unsupervised clustering of crystal structures based on their topological signatures. We instead use topological signatures with supervised machine learning to predict a material property, allowing us to identify the parts of the signature (and thus, the material structure) important for the prediction.

Besides geometry and topology, chemical information is key for predicting MOF properties. Chemistry is especially important for predicting adsorption capacities at low pressures, as well as chemical properties like the Henry’s coefficient, corresponding to the infinite dilution region. Previous approaches have constructed chemical descriptors by incorporating information from MOF building blocks, such as functional groups [14, 18]. These approaches have resulted in some improvements in predictive capabilities, however, as Fanourgakis et al. [12] noted, since the number of building blocks used to create a MOF is large, using these as descriptors for machine learning results in low transferability of the model and the need for a much larger training database. To combat this, Fanourgarkis et al. created chemical descriptors by identifying atom types using force fields. These atom types were created to account for different elements, as well as different hybridization and connectivities in different atoms. In this paper, we capture chemical information by using word embeddings constructed from a large corpus of abstracts with the word2vec algorithm [19]. The only input required is the elemental composition of the MOFs. By using word embeddings over element-property features we maintain the quasi-unsupervised nature of our machine learning pipeline. While the use of word embeddings to featurize composition do represent an implicit knowledge that the chemical elements are distinct, they use no explicit element-specific properties and are themselves derived from an unsupervised learning procedure on raw text. In addition, word embeddings are equally or more effective as other methods
of composition featurization for capturing the necessary information to accurately predict material properties in other contexts [19], further motivating our use.

We demonstrate the combined machine learning framework by predicting different targets across three datasets. The targets include methane and carbon dioxide adsorption at different pressures, as well as Henry’s coefficients. We show that the topological descriptors outperform standard porosity descriptors. Furthermore, we augment the topological features with word embeddings and show that this model improves prediction performance and compares favorably to prior work.

The key advantage of our approach is interpretability. It identifies channels and voids in a MOF that correlate best with adsorption at different pressures, further contributing to an atomic-level understanding of structure-property relationships.

2 Methods

2.1 Datasets

We demonstrate our approaches on three different datasets. The first dataset we consider is the hypothetical MOFs (hMOFs) database generated by Wilmer et al. [20] to predict gas uptakes in carbon dioxide at different pressures. This database consists of MOFs constructed by using a “Tinkertoy” algorithm to combine different MOF building blocks. We use the hMOF structures from MOFDB,¹ which has adsorption capacities (determined using grand canonical Monte Carlo simulations) for carbon dioxide at five different pressures ranging from 0.05 bar to 2.5 bar.

We also use the 2019 CoREMOF dataset [21] to predict methane and carbon dioxide adsorption capacities, and methane and carbon dioxide Henry’s coefficients (determined using grand canonical Monte Carlo and Widom insertion simulations). This dataset consists of a selection of experimentally synthesized MOFs. The other dataset we consider is the Boyd-Woo database [22] to predict the same targets as for the CoREMOF dataset (also determined using grand canonical Monte Carlo and Widom insertion simulations). These reflect gas adsorption in a low pressure, high pressure, and infinite dilution context. Similar to Moosavi et al. [13], we use a subset of approximately 20,000 structures from this database for our studies — we call this the BW dataset.

For each structure in our dataset, as in our previous work [16], we have determined the values of the following commonly used geometric descriptors. We call these structural descriptors, and use them as a baseline to compare against topological descriptors: (a) pore limiting diameter (PLD), in (Å), the diameter of the largest sphere to percolate through a material; (b) largest cavity diameter (LCD), in (Å), the diameter of the largest sphere than can fit inside the material’s pore system; (c) crystal density (\(\rho\)), in (kg/m\(^3\)); (d) accessible volume (AV), in (cm\(^3\)/g); (e) accessible surface area (ASA), in (m\(^2\)/cm\(^3\)). The values for these descriptors were computed using the Zeo++ software package [23].

2.2 Persistent homology

We describe the topological structure of the MOFs using persistent homology and follow the approach introduced in our earlier paper [16], where it was applied to zeolites. We review it briefly here.

To normalize the size of each MOF, expressed as (periodic) base cells of different sizes, we fill a (100 Å)\(^3\) cell with the atoms of the MOF. The size is chosen to be large enough to capture the statistics of the distribution of the topological features in every structure.

We represent a MOF as a union of balls centered on its atoms. We increase the radii of these balls and keep track of the changes in the topology of their union. The changes come in two types: a topological feature, such as a loop or a void, either appears or disappears. An important consequence of the algebraic formulation of this process is that these events can be paired uniquely, resulting in a set of birth–death pairs of radii, called a persistence diagram; see Figure 1. There are two persistence diagrams relevant to us: a diagram that tracks births and deaths of loops that we interpret as tunnels in the MOF (we call

¹http://hmofs.northwestern.edu
Figure 1: **Schematic outlining point cloud to persistence diagram.** (left) A point set (representing atomic centers) with balls of increasing radius around the points, (right) 1-dimensional persistence diagram of the point set. Representative cycles, corresponding to the points in the diagram, are highlighted with matching colors.

these 1-dimensional features), and a diagram that tracks voids that we think of as pockets in the MOF (2-dimensional features). The difference in birth–death values is called *persistence* of the pair. Pairs of larger persistence capture more prominent pores in the MOF. We compute persistence diagrams using Dionysus library.²

To translate persistence diagrams into vectors suitable as input for machine learning algorithms, we construct *persistence images* introduced by Adams et al. [24]. The birth–death pairs \((b, d)\) are transformed into birth–persistence pairs \((b, d - b)\). They are then convolved with Gaussians and discretized onto a grid of a fixed size, by integrating the resulting mixture of Gaussians in the cells of the grid. We use a modified version of the PersIm package³ to compute persistence images. We use the resolution of 50 × 50 and a Gaussian spread of \(\sigma = 0.15\).

### 2.3 Representative cycles

The algorithm used to compute persistence [25] tracks cycles that represent the topological features summarized in the persistence diagram. The cycles are not unique, but they reveal the atomic structures responsible for particular birth–death pairs. In a crystal structure, representative cycles correspond to channels or voids in the material. We visualize the cycles to better understand the topological features that appear in the MOFs. We choose which cycle to visualize using the feature importances found by the machine learning algorithms. We compute the representative cycles using Dionysus.⁴

### 2.4 Word embeddings

To capture the MOF’s chemical information, we use word embeddings of the chemical elements to represent a given MOF’s stoichiometric formula. The chosen embeddings were constructed from a large corpus of abstracts with the word2vec algorithm [19]. We construct feature vectors based on the composition of each MOF structure that represent word embeddings for the different elements in the MOF using the

---

²[https://github.com/mrzv/dionysus](https://github.com/mrzv/dionysus)
³[persim.scikit-tda.org](https://persim.scikit-tda.org)
⁴[github.com/mrzv/dionysus](https://github.com/mrzv/dionysus)
“matscholar_el” preset ElementProperty featurizer in matminer [26]. The feature vectors correspond to 200 embedding dimensions, with the minimum, maximum, range, mean, and standard deviation for each dimension, for a total of 1000 values. We note that the different datasets have different numbers of unique elements, ranging from eight in the hMOF dataset to 72 in the CoREMOF dataset.

2.5 Machine learning

We use random forest [27] regression to predict various targets across the different datasets, including carbon dioxide and methane adsorption at different pressures, as well as the Henry’s coefficient for carbon dioxide and methane. One of our motivations for using the random forest is the ability to determine the feature importances in the model. The random forest algorithm builds an ensemble of decision trees and chooses a random subset of features for each one. The frequency with which a particular feature is chosen for a split is an estimate for the importance of the said feature.

We build trees for different groups of features: topological features, standard structural features, word embeddings, a combined model of topological features and word embeddings, a combined model of topological and structural features, and a combined model of topological features, structural features, and word embeddings. The topological features consist of both the 1D and 2D persistence images. We train the random forest on the specific target prediction of each material. Each of the forests consists of 500 trees, and the final prediction is the average of the prediction of all trees in the forest. After training the random forest on a training set, predictions are made on an unseen test set. For most of the predictions, we use an 80%/20% training-test split. The quality of the prediction is evaluated by comparing the predicted adsorption values and the correct adsorption values. We quantify our predictions by computing the root-mean-square deviation, $\sqrt{\sum (\hat{y}_i - y_i)^2/n}$, and the coefficient of determination ($R^2$), $1 - \sum(y_i - \hat{y}_i)^2/\sum(y_i - \bar{y})^2$.

3 Results

We evaluate the accuracy of the machine learning models by predicting a number of different targets across the different datasets. For each target, we calculate the root-mean-square deviation (RMSD) and coefficient of determination ($R^2$ score). For each target and each dataset, we include results from models trained on only the topological features, only the word embeddings, and both the topological features and the word embeddings (T + WE). We also include results from the structural descriptors, described in Section 2.1, as a baseline. Finally, we incorporate the standard structural descriptors by including models combining topological and structural descriptors (T + S), as well as topological descriptors, structural descriptors, and word embeddings (T + S + WE).

3.1 hMOF dataset

For the hMOF dataset, we predict carbon dioxide adsorption capacities at different pressures, as shown in Figure 2. The RMSD is low at lower pressures because the distribution of carbon dioxide adsorption capacity has low variance in this regime. While the topology-based model outperforms the word embeddings, the model combining the two performs even better. We also see that the topological features always outperform the structural features. The word embeddings do not perform as well here. This is likely due to the hMOF dataset lacking compositional diversity: the hMOF data set contains only eight unique elements; other datasets, like the CoREMOF dataset, have 72. Nevertheless, word embeddings help boost the overall model performance when combined with the topological features.

We achieve the best performance by combining all three features together, but the accuracy achieved by subsets of the features is revealing. Adding structural to topological features slightly improves the performance, but doesn’t match that of all three features combined. On the other hand, the T + WE model performs only slightly worse than the T + S + WE model, indicating that the topological features capture most of the information that the structural features provide.
Figure 2: **Model performances for hMOF dataset and CO$_2$ adsorption.** Comparison of root-mean-square deviation (left), coefficient of determination (right) in predicting gas uptakes in CO$_2$ for different features at different pressures for the hMOF dataset. The topological features consistently outperform the standard structural features at all pressures. The T + WE and T + S + WE models achieve the best performance in general.

| Descriptor     | 0.01 bar | 0.05 bar | 0.1 bar | 0.5 bar | 2.5 bar |
|----------------|----------|----------|---------|---------|---------|
| Structural     | 0.45     | 0.55     | 0.61    | 0.67    | 0.71    |
| Topological    | 0.57     | 0.64     | 0.68    | 0.75    | 0.80    |
| T + S          | 0.70     | 0.70     | 0.72    | 0.78    | 0.84    |
| T + WE         | 0.71     | 0.84     | 0.85    | 0.90    | 0.93    |
| T + S + WE     | 0.70     | 0.86     | 0.88    | 0.92    | 0.94    |
| Best model, Fanourgakis et al. [12] | –        | 0.65     | –       | 0.90    | 0.93    |

Table 1: **Summary of model performances for hMOF dataset and CO$_2$ adsorption.** Machine learning results for carbon dioxide adsorption predictions on the hMOF dataset at different pressures, represented by $R^2$ score. The best performing model for a given pressure is highlighted.

We compare our results to Fanourgakis et al. [12], who used standard structural features and a featurization based on atom types to predict CO$_2$ adsorption capacity in the hMOF dataset. Table 1 shows results for each of our models at different pressures, along with the best model from Fanourgakis et al. [12].

Our model does particularly well at low pressures, achieving an $R^2$ score of 0.86 at 0.05 bar, compared to 0.65 from [12]. Carbon dioxide adsorption at low pressure has an important application: carbon capture from flue gases. Thus, it is particularly promising to have a generalized framework for accurate prediction of these targets. In general, our model transfers well across different pressures, as demonstrated by consistently high performance.
3.2 BW and CoREMOF datasets

We evaluate the accuracy of the machine learning predictions on the BW and CoREMOF datasets. We predict six targets grouped into three categories: the Henry’s coefficient \((\log(K_H))\) for \(\text{CO}_2\) and \(\text{CH}_4\), the gas uptakes for \(\text{CO}_2\) at 0.15 and 16 bar, and the gas uptakes for \(\text{CH}_4\) at 5.8 and 65 bar. Since each category has different units, to make them easier to compare, we normalize each of the three categories with respect to the maximum RMSD in that category.

Figure 3 shows the results of these predictions for the BW dataset. As a general trend, topological features outperform the structural features. This is especially apparent for the Henry’s coefficient predictions and the \(\text{CO}_2\) and \(\text{CH}_4\) gas uptakes at low pressure. Although word embeddings are the least accurate in all categories, they boost performance when combined with the topological features. This observation reinforces the idea that adding chemical information can improve performance, particularly for targets like the Henry’s coefficient.

As with the hMOF dataset, we see that the model combining the three features performs best across all targets. The \(T + S\) model typically performs only slightly better than the topological features, and usually much better than the structural features. This corroborates that the topological features capture almost all the information in the structural features, and much more.

Figure 4 shows the results of the machine learning predictions for the CoREMOF dataset. For predictions of the Henry’s coefficient for both \(\text{CO}_2\) and \(\text{CH}_4\), as well as the \(\text{CO}_2\) adsorption capacity at low pressure, the structural features perform especially poorly in comparison to the other models. For these three targets, unlike in the BW dataset, the word embeddings outperform the topological features. This confirms that these properties depend more on chemical information. The reason why this is especially noticeable in the CoREMOF dataset is its higher composition diversity. The CoREMOF dataset has 72 unique elements, while our subset of the BW dataset has only 16. The one target where structural descriptors outperform the topological descriptors is the \(\text{CO}_2\) gas uptake at 16 bar. In general, structural (geometric) information is more important at higher pressures.

For low pressure adsorption targets like 0.15 bar \(\text{CO}_2\), the \(T + S\) model performs about the same as the topological features, but considerably better than the structural features. This suggests that there is a lot of relevant geometric information that the conventional structural descriptors don’t capture. For low pressure and infinite dilution gas adsorption predictions, to our knowledge, these topological descriptors are currently the best-performing descriptors that only take into account geometric information about the MOF.

Overall, across all three datasets, the model combining all three feature types \((T + S + WE)\) has the highest predictive capability.
Figure 3: **Model performances on BW dataset.** Comparison of root-mean-square deviation (left), coefficient of determination (right) in predicting the Henry's coefficient ($\log k_H$) for CO$_2$ and CH$_4$, gas uptakes for CO$_2$, and gas uptakes for CH$_4$, for different features for the BW20K dataset. For each target, the units are mol kg$^{-1}$ Pa$^{-1}$, mmol/g, and VSTP/V respectively. Due to the difference in units between targets, RMSD values are normalized with respect to the maximum value in each category. The black, dashed line defines the categories that share the same units. In the case of RMSD, these categories also share a normalization factor. The actual RMSD and $R^2$ values are shown above each bar.
Figure 4: Model performances on CoREMOF dataset. Comparison of root-mean-square deviation (left), coefficient of determination (right) in predicting the Henry's coefficient ($\log k_H$) for CO$_2$ and CH$_4$, gas uptakes for CO$_2$, and gas uptakes for CH$_4$, for different features for the CoREMOF dataset. For each target, the units are mol kg$^{-1}$ Pa$^{-1}$, mmol/g, and VSTP/V respectively. Due to the difference in units between targets, RMSD values are normalized with respect to the maximum value in each category. The black, dashed line defines the categories that share the same units. In the case of RMSD, these categories also share a normalization factor. The actual RMSD and $R^2$ values are shown above each bar.
3.3 Feature analysis

The random forest algorithm infers the importance of individual features by measuring how frequently they are used by the decision trees to make a prediction about a MOF. In our methodology, there are three distinct types of features: topological, structural, and word embeddings. Further, topological features come in two types, 1-dimensional features that capture the distribution of channels in the MOF and 2-dimensional features that describe the voids. Each of those consists of 2500 individual features (pixels in the persistence image), but we combine them to infer the aggregate importance of the different feature types. In this section, we analyze contributions from the topological and word embedding features, since the structural features contribute little extra information.

Figure 5 shows the relative importance of topological descriptors and word embeddings. For the BW dataset, 2D features are most important for the prediction, with word embeddings playing a larger role in the predictions of the Henry’s coefficient. For the CoREMOF dataset, word embeddings are more important, especially for the Henry’s coefficient and CO$_2$ adsorption at low pressure, where they account for 40-60% of the decisions, with topological features dominating the importance of predictions at higher pressure. For the hMOF dataset, 1D topological features are most important at low pressures, with 2D being more important at higher pressure, and word embeddings used in ~ 30% of the decisions.

These results reveal the importance of different properties for different tasks. They support the claim that chemical information is more important for infinite dilution and low-pressure CO$_2$ adsorption. In these conditions, the electrostatic interactions between the gas and the MOF framework play an important role in adsorption capacity — the word embeddings capture this non-structural information. On the other hand, methane adsorption at higher pressure is mostly described by 2D topology, a trend that we also observed in zeolites [16].

Our results also suggest why the conventional structural descriptors perform especially poorly when predicting CO$_2$ adsorption in hMOFs at low pressure. The standard structural features describe the pore geometry by the largest sphere to percolate through the materials and the largest sphere that can fit inside its pore system. In contrast, topological features record the widths of the channels that criss-cross the MOF as well as the sizes of different cavities. They also distinguishing between the distribution of channels and voids, by separating 1D and 2D topological features, and record other finer information about their shape.

As Figure 5 shows, topological features play a major role in predicting gas adsorption, with the 1-dimensional channels being especially important for adsorption at low pressures in the CoREMOF and hMOF datasets, and 2-dimensions voids being important for the predictions with the BW dataset. The differences in feature importances can also be linked back to the data: for example, the CoREMOF MOFs tend to have smaller pores than the BW MOFs. Ultimately, there is at least some complementary information being captured by the structural and topological features, as the best models across all the datasets include both.

Our results also suggest guiding principles for developing MOFs with high adsorption capacity. Channels in the material are important for adsorption at lower pressures, while adsorption at higher pressures is more influenced by voids in the material. For a MOF to have high adsorption capacity across pressures, a structure with diverse pore geometry, with both channels and voids, may be important.
Figure 5: **Feature analysis of machine learning models.** Summary of relative feature importance across different targets for the 1D, 2D topological features, and word embeddings. Results are from the restricted persistence images. The BW, CoREMOF, and hMOF datasets are shown here.

## 4 Interpretability

### 4.1 Topological features and representative cycles

Nanoporous materials, and especially MOFs, are known for how tunable they are: experimentalists can synthesize materials with precisely sized pores. Understanding how structure influences a particular material property helps guide this process. Persistent homology is especially helpful here. The features it defines have an immediate interpretation that can be incorporated into synthesis.

The points in a persistence diagram correspond to voids and channels of specific sizes. A point \((b, d)\) in a
2-dimensional diagram is generated by a cavity that can fit the largest sphere of radius $d$; the largest sphere that can escape from the cavity (and move to a larger cavity) has radius $b$. A point $(b, d)$ in a 1-dimensional diagram is produced by a channel in the material, specifically, by its narrowest “bottleneck.” The death value, $d$, records the radius of the largest sphere that can pass through this bottleneck. The birth value, $b$, records how close the atoms of the bottleneck are to each other.

For each dataset and each target property, the most important 1D and 2D birth-death points, as identified by the random forest algorithm, are listed in Tables 2, 3, and 4. We note a few patterns. In the case of methane adsorption in all three regimes (infinite dilution, low pressure, and high pressure), the 2D birth and death values are similar for both the BW and CoREMOF datasets — in fact, almost identical for the infinite dilution and high pressure cases. Specifically, birth values are around 2.3 – 2.4 Å for high pressure methane adsorption, and 3.4 – 3.8 Å for low pressure and infinite dilution methane adsorption. Death values are 3.2 Å for high pressure methane adsorption, and 4 – 4.6 Å for low pressure and infinite dilution methane adsorption. Another pattern to note in the hMOF dataset is that 1D death values get larger as CO$_2$ adsorption increases, meaning the size of the largest sphere able to pass through the channel increases.

| Target property | 1D birth | 1D death | 2D birth | 2D death |
|-----------------|----------|----------|----------|----------|
| log($K_H$) CO$_2$ | 1 | 4 | 3.3 | 4.1 |
| log($K_H$) CH$_4$ | 1.6 | 2 | 3.6 | 4.4 |
| 0.15 bar CO$_2$ | 3.5 | 3.6 | 3.4 | 4 |
| 16 bar CO$_2$ | 1.7 | 2 | 3.1 | 3.9 |
| 5.8 bar CH$_4$ | 1.4 | 3 | 3.8 | 4.6 |
| 65 bar CH$_4$ | 3.6 | 4.3 | 2.3 | 3.2 |

Table 2: Most important 1D/2D birth–death points for the BW dataset.

| Target property | 1D birth | 1D death | 2D birth | 2D death |
|-----------------|----------|----------|----------|----------|
| log($K_H$) CO$_2$ | 0.1 | 0.6 | 2.4 | 3.2 |
| log($K_H$) CH$_4$ | 0.3 | 1 | 3.6 | 4.4 |
| 0.15 bar CO$_2$ | 0.3 | 1.3 | 3.7 | 3.8 |
| 16 bar CO$_2$ | 1.5 | 2.4 | 2.2 | 2.9 |
| 5.8 bar CH$_4$ | 0.3 | 1 | 3.4 | 4 |
| 65 bar CH$_4$ | 1.1 | 3.1 | 2.4 | 3.2 |

Table 3: Most important 1D/2D birth–death points for the CoREMOF dataset.

| Target property | 1D birth | 1D death | 2D birth | 2D death |
|-----------------|----------|----------|----------|----------|
| 0.01 bar CO$_2$ | 0.02 | 0.7 | 3.2 | 3.5 |
| 0.05 bar CO$_2$ | 1.1 | 1.6 | 1.6 | 2.1 |
| 0.1 bar CO$_2$ | 1.1 | 2.7 | 4.4 | 5.5 |
| 0.5 bar CO$_2$ | 1.3 | 3.5 | 4.7 | 5.8 |
| 2.5 bar CO$_2$ | 1 | 3.7 | 4 | 5.1 |

Table 4: Most important 1D/2D birth–death points for the hMOF dataset.

We can dissect topological representations further and extract representative cycles for each point. Although these cycles are not unique — we are at the mercy of certain choices persistent homology calculation makes — they are helpful in visualizing the cavities and channel bottlenecks represented by the points in the persistence diagram.

Since we train our machine learning algorithm on persistence images, we have to take an extra step to identify the points in a persistence diagrams with relevant representative cycles. We illustrate our steps on the following example. Figure 6(b) shows the feature importances for predicting CO$_2$ adsorption at low pressure in the BW dataset. We identify the birth and persistence values of the most important pixel in the
persistence image and highlight its center in orange in Figure 6(c), which also shows in blue the persistence diagram for str-m3-o10-o15-pcu-sym.49, a MOF with high CO\textsubscript{2} adsorption at low pressures. Figure 6(d) shows a representative cycle for the point in the persistence diagram that falls within the Gaussian spread factor used to construct the persistence image.

![Diagram](image)

Figure 6: **Schematic outlining identification of a representative cycle in a crystal structure.** (a) the original crystal structure for MOF str-m3-o10-o15-pcu-sym.49 from the BW dataset. (b) The feature importance, shown as an image, for the 2D topology features for CO\textsubscript{2} adsorption at 0.15 bar. The color bar for this figure is shown at the bottom. (c) The 2D birth vs. persistence plot for str-m3-o10-o15-pcu-sym.49, with the (birth, persistence) point with the highest feature importance (as determined by the machine learning algorithm) in orange. The dashed lines around the orange point show the Gaussian spread factor. (d) The representative cycle for the closest point in the birth vs. persistence diagram to the orange point in (c).
We extracted the representative cycles from the high gas adsorption MOFs from different databases. Four examples, including both 1D topology (channels) and 2D topology (voids), appear in Figure 7. One notable trend is that the loop in Figure 7(c) is present in many of the materials in the hMOF dataset that have high CO$_2$ adsorption at low pressure. Similarly, the void size seen in Figure 7(a) is present in many of the MOFs with high CO$_2$ adsorption.

4.2 Word embeddings and material properties

We explore the interpretability of the word embeddings by relating their importances in predicting MOF properties and in predicting chemical properties of individual elements. The former we obtain from the random forests just as the importances of the topological features. To calculate the importances for individual elements, we retrieve word embeddings for all the elements in the matscholar database [19] and use these as features to train models to predict various chemical properties — electronegativity, atomic radius, electrical resistivity, melting point, etc. — of the pure elements contained in pymatgen’s ‘periodic_table’ module [28]. We extract the feature importances for each of these models. Because each MOF has 1000 features, summarizing the distribution of 200 features over its elements, as described in Section 2.4, we sum up the MOF feature importances corresponding to the same elemental feature.

We take the subset of feature importances that account for 90% of the random forest decisions. By definition, these features describe the subspace of our input where most of the decisions are made to make a prediction about the given target property. Given a MOF target property and a chemical target property,
we compute the Jaccard similarity between the two subsets of features. This metric measures the relative size of the subspace, important for the random forest decisions for both targets.

Table 5 lists the top three materials properties by similarity to each MOF target property; all of them have a Jaccard similarity greater than 0.4. Following this procedure, we identify the chemical property with the strongest informational relevance to a given MOF target property. We focus on interpreting the

| Target property          | 1                  | 2                  | 3                              |
|--------------------------|--------------------|--------------------|--------------------------------|
| log(K_H) CO_2            | electronegativity  | Poisson’s ratio    | Mendeleev’s number             |
| log(K_H) CH_4            | electronegativity  | Poisson’s ratio    | thermal conductivity           |
| 0.15 bar CO_2            | atomic radius      | electronegativity  | Poisson’s ratio                |
| 16 bar CO_2              | thermal conductivity| Brinell’s hardness | electron negativity            |
| 5.8 bar CH_4             | thermal conductivity| Poisson’s ratio    | Brinell’s hardness             |
| 65 bar CH_4              | thermal conductivity| electron negativity| melting point                  |

Table 5: **Material properties sharing overlap with word embedding feature importances.** Machine learning models trained with elemental word embeddings and materials properties are compared to the models trained with MOF composition word embeddings and MOF target properties for the CoRE2019 dataset. The feature importances of each model are analyzed, and compared by Jaccard similarity. The top three materials properties most similar to the model trained to MOF target properties are listed.

results from a MOF design perspective. The word-embedding features play a bigger role than topology in predicting 0.15 bar CO_2 adsorption, log(K_H) CO_2, and log(K_H) CH_4 (Figure 4). For these three targets, the machine learning model trained on electronegativity was the most or second most similar to the model trained on the word embeddings for each MOF. This suggests that local electrostatic interactions are more significant in carbon dioxide adsorption at low pressure, which is consistent with qualitative descriptions of low pressure or dilute-limit profiles of absorptivity in porous materials from literature [13]. Thermal conductivity also appears multiple times, and is the most relevant elemental property for high pressure CO_2 and CH_4 adsorption.

The relevance of thermal conductivity at higher pressures is more difficult to interpret, given that thermal conductivity contains an electronic and vibrational component. However, a relationship between thermal conductivity and MOF geometry has been suggested previously. Specifically, thermal conductivity correlates with pore size and porosity [29, 30], which in turn affects adsorption. Thus, when designing a MOF, choosing metal elements that decrease the thermal conductivity of a MOF could be one way to increase adsorption.

Another materials property that appeared multiple times for multiple MOF targets was the Poisson’s ratio, which reflects elasticity of a material. This is another property that fits in the existing paradigm of MOF design: namely, flexibility. MOFs with flexible frameworks often are better adsorbents [31], since they can accommodate a larger space to fit a gas molecule with less stress.

In summary, the latent information contained in the word embeddings overlaps with known descriptors for MOF gas adsorption, pointing to important chemical features for designing high adsorption MOFs.

### 5 Conclusions

We have developed a machine learning framework for MOFs, and nanoporous materials in general, using persistent homology and word embeddings. Our approach builds a holistic representation of the materials, requiring less handcrafting and domain expert guidance than the currently widely used porosity and chemical descriptors. Our topological representation is a vectorized persistence diagram, obtained from the atomic coordinates of the normalized supercell representation of a materials’ crystal structure. It can be used in any machine learning algorithm. We augment the topological information with element embeddings, constructed from a large set of scientific abstracts via the word2vec algorithm [19]. They provide a generalized representation of the MOF composition. We have tested this approach on three different datasets, predicting several gas adsorption targets at various pressures. These experiments show a significantly improved
performance compared to standard structural descriptors. The topological features we compute are generic and transferable across different property targets. As the topological descriptors consistently outperform standard structural descriptors, they provide a simple way to boost the performance of any machine learning algorithm. Additionally, to our knowledge, these descriptors are the best purely geometric descriptors for predicting gas adsorption at low pressures and in the infinite dilution regime. Moreover, these descriptors are interpretable: their components can be traced to specific channels and voids in the crystal structure, which contributes to a greater understanding of structure-property relationships in MOFs.

We conclude by highlighting the key strengths of our approach. 1) It provides a way to quickly screen any dataset to find the top MOFs for a particular task without the need to handcraft specific features. 2) It augments existing features for accurate performance predictions. 3) It helps guide materials design by directly connecting property predictions to the crystal structure.

6 Acknowledgements

This work was supported by the U.S. Department of Energy under Contract Number DE-AC02-05CH11231 at Lawrence Berkeley National Laboratory. A.S.K. is an Alvarez Fellow in the Computational Research Division at LBNL. A.S.K. also acknowledges support from the TRI-AMDD internship program and further support from TRI-AMDD for cloud computing resources and initial project planning. The authors acknowledge helpful discussions with Muratahan Aykol, John Dagdelen, Aayush Singh, Santosh Suram, and Ram Seshadri.

7 Competing interests

The authors declare no competing interests.

References

[1] Jesse L. C. Rowsell, Elinor C. Spencer, Juergen Eckert, Judith A. K. Howard, and Omar M. Yaghi. Gas adsorption sites in a large-pore metal–organic framework. Science, 309(5739):1350–1354, 2005.

[2] Jian-Rong Li, Julian Sculley, and Hong-Cai Zhou. metal–organic frameworks for separations. Chemical Reviews, 112(2):869–932, 2012. PMID: 21978134.

[3] Dong Yang and Bruce C. Gates. Catalysis by metal organic frameworks: Perspective and suggestions for future research. ACS Catalysis, 9(3):1779–1798, 2019.

[4] Samuel O. Odoh, Christopher J. Cramer, Donald G. Truhlar, and Laura Gagliardi. Quantum-chemical characterization of the properties and reactivities of metal–organic frameworks. Chemical Reviews, 115(12):6051–6111, 2015. PMID: 25872949.

[5] Kevin Maik Jablonka, Daniele Ongari, Seyed Mohamad Moosavi, and Berend Smit. Big-data science in porous materials: Materials genomics and machine learning. Chemical Reviews, 120(16):8066–8129, 2020. PMID: 32520531.

[6] Sanggyu Chong, Sangwon Lee, Backjun Kim, and Jihan Kim. Applications of machine learning in metal–organic frameworks. Coordination Chemistry Reviews, 423:213487, 2020.

[7] Yabing He, Wei Zhou, Guodong Qian, and Banglin Chen. Methane storage in metal–organic frameworks. Chem. Soc. Rev., 43:5657–5678, 2014.

[8] Kenji Sumida, David L. Rogow, Jarad A. Mason, Thomas M. McDonald, Eric D. Bloch, Zoey R. Herm, Tae-Hyun Bae, and Jeffrey R. Long. Carbon dioxide capture in metal–organic frameworks. Chemical Reviews, 112(2):724–781, 2012. PMID: 22204561.
[9] Shuji Himeno, Toshiya Komatsu, and Shoichi Fujita. High-pressure adsorption equilibria of methane and carbon dioxide on several activated carbons. *Journal of Chemical & Engineering Data*, 50(2):369–376, 2005.

[10] Michael Fernandez and Amanda S. Barnard. Geometrical properties can predict CO2 and N2 adsorption performance of metal–organic frameworks (MOFs) at low pressure. *ACS Combinatorial Science*, 18(5):243–252, 2016. PMID: 27022760.

[11] Maryam Pardakhti, Elshan Moharrerii, David Wanik, Steven L. Suib, and Ranjan Srivastava. Machine learning using combined structural and chemical descriptors for prediction of methane adsorption performance of metal organic frameworks (MOFs). *ACS Combinatorial Science*, 19(10):640–645, 2017. PMID: 28800219.

[12] George S. Fanourgakis, Konstantinos Gkagkas, Emmanuel Tylianakis, and George E. Froudakis. A universal machine learning algorithm for large-scale screening of MOFs. *Journal of the American Chemical Society*, 142(8):3814–3822, 2020. PMID: 32017547.

[13] Seyed Mohamad Moosavi, Aditya Nandy, Kevin Maik Jablonka, Daniele Ongari, Jon Paul Janet, Peter G. Boyd, Yongjin Lee, Berend Smit, and Heather J. Kulik. Understanding the diversity of the metal–organic framework ecosystem. *Nature Communications*, 11(1):4068, 2020.

[14] Ryther Anderson, Jacob Rodgers, Edwin Argueta, Achay Biong, and Diego A. Gómez-Gualdrón. Role of pore chemistry and topology in the CO2 capture capabilities of MOFs: From molecular simulation to machine learning. *Chemistry of Materials*, 30(18):6325–6337, 2018.

[15] Herbert Edelsbrunner and John Harer. Persistent homology – a survey. *Contemp. Math.*, 453:257–282, 2007.

[16] Aditi S. Krishnapriyan, Maciej Haranczyk, and Dmitriy Morozov. Topological descriptors help predict guest adsorption in nanoporous materials. *The Journal of Physical Chemistry C*, 124(17):9360–9368, 2020.

[17] Yongjin Lee, Senja D. Barthel, Paweł Dłotko, S. Mohamad Moosavi, Kathryn Hess, and Berend Smit. Quantifying similarity of pore-geometry in nanoporous materials. *Nature Communications*, 8(1):15396, 2017.

[18] Giorgos Borboudakis, Taxiarchis Stergiannakos, Maria Frytsali, Emmanuel Klontzas, Ioannis Tsamardinos, and George E. Froudakis. Chemically intuited, large-scale screening of MOFs by machine learning techniques. *npj Computational Materials*, 3(1):40, October 2017.

[19] Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98, 2019.

[20] Christopher E. Wilmer, Michael Leaf, Chang Yeon Lee, Omar K. Farha, Brad G. Hauser, Joseph T. Hupp, and Randall Q. Snurr. Large-scale screening of hypothetical metal–organic frameworks. *Nature Chemistry*, 4(2):83–89, February 2012.

[21] Yongchul G. Chung, Emmanuel Haldoupis, Benjamin J. Bucior, Maciej Haranczyk, Seulchan Lee, Hongda Zhang, Konstantinos D. Vogiatzis, Marija Milisavljevic, Sanliang Ling, Jeffrey S. Camp, Ben Slater, J. Ilja Siepmann, David S. Sholl, and Randall Q. Snurr. Advances, updates, and analytics for the computation-ready, experimental metal–organic framework database: Core mof 2019. *Journal of Chemical & Engineering Data*, 64(12):5985–5998, 2019.

[22] Peter G. Boyd and Tom K. Woo. A generalized method for constructing hypothetical nanoporous materials of any net topology from graph theory. *CrystEngComm*, 18(21):3777–3792, 2016. Publisher: The Royal Society of Chemistry.
[23] 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*, 149:134–141, 2012.

[24] Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. Persistence images: A stable vector representation of persistent homology. *J. Mach. Learn. Res.*, 18(8):1–35, 2017.

[25] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete Comput. Geom.*, 28:511–533, 2002.

[26] Logan Ward, Alexander Dunn, Alireza Faghaninia, Nils E. R. Zimmermann, Saurabh Bajaj, Qi Wang, Joseph Montoya, Jiming Chen, Kyle Bystrom, Maxwell Dylla, Kyle Chard, Mark Asta, Kristin A. Persson, G. Jeffrey Snyder, Ian Foster, and Anubhav Jain. Matminer: An open source toolkit for materials data mining. *Computational Materials Science*, 152:60 – 69, 2018.

[27] L. Breiman. Random forests. *Int. J. Mach. Learn. Cybern.*, 20:273–297, 1995.

[28] Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L. Chevrier, Kristin A. Persson, and Gerbrand Ceder. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314 – 319, 2013.

[29] Iwan Sumirat, Y. Ando, and S. Shimamura. Theoretical consideration of the effect of porosity on thermal conductivity of porous materials. *Journal of Porous Materials*, 13(3):439–443, August 2006.

[30] Hasan Babaei, Alan J. H. McGaughey, and Christopher E. Wilmer. Effect of pore size and shape on the thermal conductivity of metal–organic frameworks. *Chem. Sci.*, 8:583–589, 2017.

[31] Francois-Xavier Coudert. Responsive metal–organic frameworks and framework materials: Under pressure, taking the heat, in the spotlight, with friends. *Chemistry of Materials*, 27(6):1905–1916, 2015.