

# Topological Descriptors Help Predict Guest Adsorption in Nanoporous Materials

Aditi S. Krishnapriyan\*    Maciej Haranczyk<sup>†</sup>    Dmitriy Morozov<sup>‡</sup>

## Abstract

Machine learning has emerged as an attractive alternative to experiments and simulations for predicting material properties. Usually, such an approach relies on specific domain knowledge for feature design: each learning target requires careful selection of features that an expert recognizes as important for the specific task. The major drawback of this approach is that computation of only a few structural features has been implemented so far, and it is difficult to tell a priori which features are important for a particular application. The latter problem has been empirically observed for predictors of guest uptake in nanoporous materials: local and global porosity features become dominant descriptors at low and high pressures, respectively.

We investigate a feature representation of materials using tools from topological data analysis. Specifically, we use persistent homology to describe the geometry of nanoporous materials at various scales. We combine our topological descriptor with traditional structural features and investigate the relative importance of each to the prediction tasks. We demonstrate an application of this feature representation by predicting methane adsorption in zeolites, for pressures in the range of 1–200 bar. Our results not only show a considerable improvement compared to the baseline, but they also highlight that topological features capture information complementary to the structural features, which is especially important for the adsorption at low pressure, a task particularly difficult for the traditional features. Furthermore, by investigation of the importance of individual topological features in the adsorption model, we are able to pinpoint the location of the pores that correlate best to adsorption at different pressure, contributing to our atom-level understanding of structure-property relationships.

## 1 Introduction

Nanoporous materials, such as zeolites, are critical to many industrial sectors and are commonly used as membranes, adsorbents for separations, storage and delivery, and heterogeneous catalysts [1]. Along with zeolites, advanced porous materials (APMs) such as metal organic frameworks (MOFs), covalent organic frameworks (COFs), and related families of materials, have been investigated in the context of the above and other applications [2]. An important factor that determines the utility of any nanoporous material is that its pore morphology and the corresponding material structures are configured to maximize the performance in its respective application. Finding the materials that satisfy the latter is therefore necessary to develop more efficient, cleaner, and cheaper technologies. The nanoporous materials space is immense: there are more than 1400 zeolite materials already reported that belong to more than 200 unique topologies [3]. However, these experimentally known zeolites constitute only a very small fraction of more than 2.7 million of hypothesised structures [4], though the latter number may be orders of magnitude smaller than can be synthesized [5]. APMs have an even higher number of both already synthesized and predicted structures [6, 7], as their frameworks allow for greater flexibility in choosing diverse building blocks of various chemistries and topologies.

---

\*Stanford University, Stanford, California, United States of America

<sup>†</sup>IMDEA Materials Institute, C/Eric Kandel 2, 28906 - Getafe, Madrid, Spain

<sup>‡</sup>Lawrence Berkeley National Laboratory, Berkeley, California, United States of America

Over the last decade, there has been a growing effort to employ computational techniques to navigate the space of possible nanoporous materials and choose promising candidates for various applications. Molecular simulations allow the study of important guest molecule phenomena in zeolites and APMs, such as adsorption and diffusion [8]. Henry coefficients, heats of adsorption, adsorption isotherms, and diffusion coefficients of guest species can be predicted and there is solid evidence in the literature that these predictions are accurate [9, 10]. Initial attempts to screen for optimal materials have employed quite small collections of nanoporous materials, which could be characterized in their totality using molecular simulations (i.e. brute-force screening). The size of these studies were limited by the size of crystal structure databases, which at that time contained only experimentally known zeolites and early MOFs. The research landscape has changed dramatically with the introduction of databases of predicted nanoporous materials such as hypothetical zeolites [4], which posed a significant computational challenge for the brute-force approach.

As a result, an entire spectrum of high-throughput screening tools and approaches have been developed in the last decade, which can be grouped into two categories. The first consists of techniques that limit the number of structures to be investigated using expensive molecular simulations, examples of which include 1) tools enabling screening filters based on porosity descriptors [11], 2) substructure similarity screening and material diversity-based data reduction [12, 13], and 3) efficient searches using genetic algorithms or optimization approaches [14]. The second category consists of techniques accelerating prediction of properties of each material in the considered set, examples of which are 1) GPU-accelerated molecular simulations [15], 2) simplified simulations and models, and finally 3) machine learning prediction trained using the simulation results of a relatively small subset of materials [16, 17]. The latter approach is of particular interest as there is sufficient evidence to show that it can provide high accuracy, robustness, and the throughput necessary to tackle ever-increasing datasets.

Machine learning models aim to predict the material property of choice by using a function fitted on a set of known properties and operating on a material representation as its input. There are a number of frameworks for these fittings that include algorithms such as random forests, neural networks, and support-vector machines, all of which have examples of successful applications in materials, and are often interchangeable. A suitable material representation poses a greater challenge. Typically, domain experts need to select specific material features to serve as model input, used for making predictions about a particular property. Standard examples of such features include density of the material, cohesive energy per atom, and volume per atom [18].

In the context of nanoporous materials, one of the most desired properties to predict is a guest adsorption loading at given temperature–pressure conditions. Naturally, guest adsorption is determined by the morphology of the materials’ pores, where guest molecules will be located and interacting with the material structure. Porosity descriptors such as accessible surface area (ASA), largest cavity diameter (LCD), and others have been primary targets to be used to construct feature vectors. There are a growing number of models successfully trained and employed to predict adsorption of methane at conditions relevant to vehicular storage applications [19], CO<sub>2</sub> in carbon capture [20], and Xe in Xe separations from air [17]. These early studies have highlighted challenges related with the selection and implementation of relevant feature vectors.

In the case of high pressure gas adsorption, guest molecules generally occupy the entire void space available in a material. Therefore, inclusion of a “global” material porosity descriptor such as ASA leads to predictive models with these features becoming the most important to the model. Empirically, this and similar correlations have been observed: for example, the Chachine rule in hydrogen storage application, which states that one weight percent of H<sub>2</sub> can be stored per 500 m<sup>2</sup>/g of a materials’ ASA [21]. However, in the case of low pressure adsorption, guest molecules tend to be localized in the strongly binding regions of the material’s pore [22], which are typically not captured well with the standard porosity descriptors. To account for this, specific descriptors can be developed: for example, Voronoi energy was introduced [17], but at substantial development effort and with unknown application transferability potential.

The current state of the field suggests a significant need and potential payoff for the development of robust generic descriptors that would capture porosity features across pore scales and allow for generalized models for prediction of adsorption loadings at various pressures. In this paper, we focus on multi-scale topological features that describe the connectivity of the pore space of the material. Specifically, we use persistent homology [23] to compute topological signatures of the channels and voids constituting the porosity

of a material, thereby capturing their number and size. We convert these signatures to vector representations, facilitating supervised machine learning algorithms to predict material properties. We note that persistent homology is very general and 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 problem.

While there has been previous work applying topological data analysis to nanoporous materials [24], it focused on using standard, global metrics on topological signatures to cluster the crystal structures. A more recent work [25] was also concerned with using topological signatures for supervised learning of nanoporous material properties. However, the authors turned topological descriptors into simplified feature vectors, which lost the ability to correlated the performance with specific topological features. In contrast, our goal is to use topological signatures with supervised machine learning to predict a material property, while simultaneously identifying the parts of the signature important for the prediction.

We demonstrate our approach in an application of prediction of methane adsorption in all silica zeolite structures, for which large sets of simulated data are available [26]. We show that our topological model outperforms existing porosity descriptors when it comes to predicting adsorption properties, using a number of metrics including root-mean-square error,  $R^2$  score, and the Spearman correlation coefficient. Furthermore, we show that unlike the handcrafted structural features, the topological features give accurate predictions across a wide range of pressures. We also show that the best model includes a combination of topological and structural features, meaning that the two capture complementary information about the material. Furthermore, our approach allows for identification of the structural features related with the important topological descriptors identified within the topological model, hence providing atom-level understanding of the origin of the property of interest.

## 2 Methods

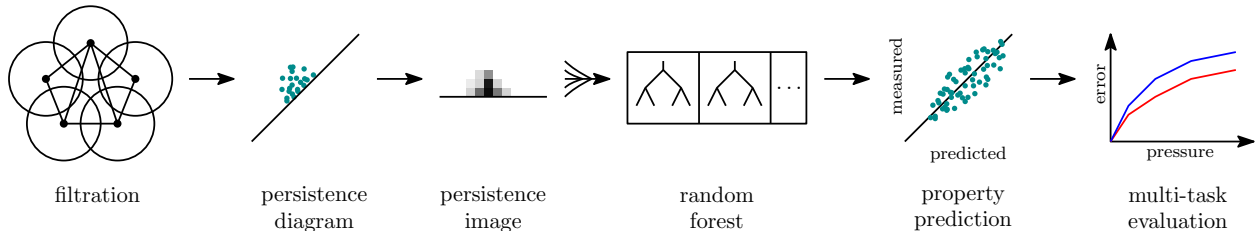


Figure 1: Computational pipeline.

A schematic view of the computational pipeline is depicted in Figure 1. In the initial step, material structures are analyzed to identify features present at various length scales. In particular, after normalizing the size of the structure representations, we compute their alpha shape filtrations, whose persistence diagrams we use as the topological signature of each material. Then, we vectorize these diagrams by turning them into so-called persistence images. Such vector representations are used together with the results of predicted guest uptakes to train random forest regression models for adsorption at different pressures. Finally, we compare different measures of the quality of the regression across choices of features and pressures. The rest of this section explains the methodology of the pipeline in more detail, as well as the material dataset used to demonstrate the approach.

**Zeolite morphology.** Zeolites are crystalline structures. They are typically represented by the positions of atoms with respect to the periodic cell, which can be implicitly replicated over an arbitrarily large volume. The cells themselves have vastly different sizes. To normalize the calculation, we prepare a  $100\text{\AA} \times 100\text{\AA} \times 100\text{\AA}$  supercell of each material. This size is chosen to be at least 2-3 times larger than the largest base cell. A similar approach of representing materials by their spherical supercell was employed earlier [27].

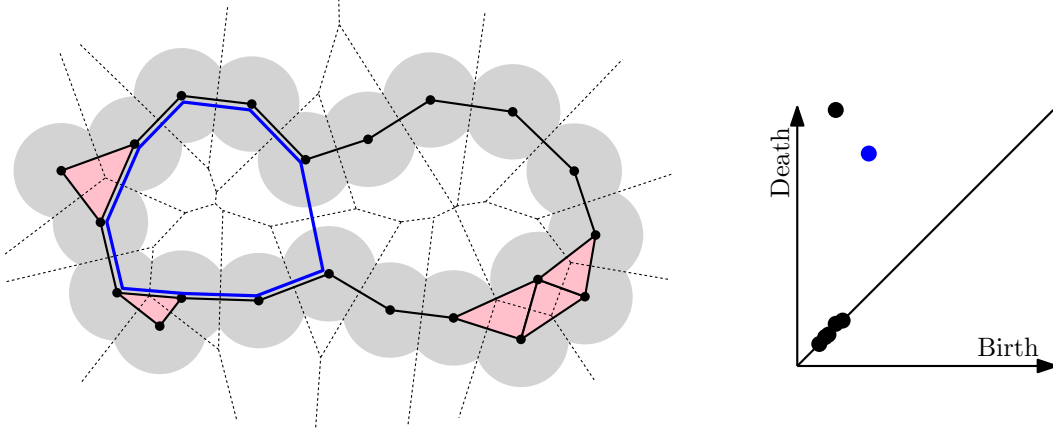


Figure 2: Left: a point set  $P$ ; a union of balls (gray) of fixed radius  $r$ , centered on the point set; a Voronoi diagram (dashed); an alpha shape  $\text{Alpha}_r(P)$  (points, solid edges, and pink triangles). Right: 1-dimensional persistence diagram of the alpha shape filtration of  $P$ ; the blue point represents the cycle highlighted in blue on the left.

We model the pore structure of the materials across scales. Specifically, we take the union of balls around the atoms and vary their radii. We track how the topology of this union (and, dually, of its complement — the pore space) changes. To represent this union combinatorially, we use alpha shapes. Briefly, given a point set  $P$ , a *Voronoi cell* of a point,  $p$ , consists of all points of the ambient space that are closer to  $p$  than to any other point in  $P$ :

$$\text{Vor}(p) = \{x \in \mathbb{R}^3 \mid \|x - p\| \leq \|x - q\| \forall q \in P\}.$$

An *alpha shape* of  $P$  for radius  $r$  consists of all the subsets  $\sigma$  of  $P$ , where the intersection of the Voronoi cells and the balls of radius  $r$  around the points  $p \in \sigma$  is not empty:

$$\text{Alpha}_r(P) = \left\{ \sigma \subset P \mid \bigcap_{p \in \sigma} B_r(p) \cap \text{Vor}(p) \neq \emptyset \right\}.$$

Figure 2 illustrates a point set and its alpha shape for a fixed radius. Geometrically, an alpha shape consists of points, edges, triangles and tetrahedra that are a subset of the Delaunay triangulation. More importantly for us, an alpha shape for radius  $r$  has the same topology as the union of balls of the same radius.

By construction, alpha shapes nest as we increase their radius:  $\text{Alpha}_r(P) \subseteq \text{Alpha}_{r'}(P)$  for  $r < r'$ . Their nesting across all radii forms an *alpha shape filtration*. We compute the alpha shapes using the software package DioDe,<sup>1</sup> which is a Python binding to the relevant functionality in the computational geometry library CGAL.<sup>2</sup>

**Persistent homology.** Persistent homology tracks the changes of homology groups in a filtration. We do not define *homology* formally — there are thorough surveys [23, 28] of all the relevant concepts — but note that this algebraic topological invariant records the number of connected components, loops, and voids in a space. Given an alpha shape filtration of a point set, the homology groups of the nested spaces are connected by linear maps:

$$H(\text{Alpha}_{r_1}(P)) \rightarrow H(\text{Alpha}_{r_2}(P)) \rightarrow \dots \rightarrow H(\text{Alpha}_{r_n}(P)).$$

By following the changes in the sequence, one can identify pairs of radii, when a homology class (e.g., a loop or a void) appears and subsequently disappears: i.e., when a tunnel or a pocket appears and then closes up

<sup>1</sup>[github.com/mrzv/diode](https://github.com/mrzv/diode)

<sup>2</sup>[cgal.org](https://cgal.org)

in the pore space. Such pairs of radii are called *birth–death pairs*,  $(b, d)$ . The bigger the difference,  $d - b$ , between death and birth, the more prominent, or *persistent*, is the class, and the more pronounced is the pore.

The full collection of birth–death pairs for a filtration is called a persistence diagram,  $\text{Dgm}_k(P)$ ; see Figure 2. There are three non-trivial diagrams for the alpha shapes in three dimensions. The 0-th diagram records the evolution of connected components in the filtration; the 1-st diagram, loops; the 2-nd diagram, voids. In the remainder, we use only the 1-st and 2-nd diagrams — the 0-th diagram, the structure of connected components, does not have much information beyond the distances between nearest neighbor atoms. We compute the persistence diagrams from alpha shapes using the software package Dionysus.<sup>3</sup>

Persistent homology has many attractive properties. A crucial one has to do with its resilience to noise: persistence is stable. If we perturb the point set a little, the persistence diagram changes only a little. To be precise, let  $d_H$  denote the Hausdorff distance between two points sets:

$$d_H(P, Q) = \max \left\{ \max_{p \in P} \min_{q \in Q} \|p - q\|, \max_{q \in Q} \min_{p \in P} \|p - q\| \right\}.$$

Let  $d_B$  denote the bottleneck distance between persistence diagrams:

$$d_B(\text{Dgm}_k(P), \text{Dgm}_k(Q)) = \inf_{\gamma} \sup_{x \in \text{Dgm}_k(P)} \|x - \gamma(x)\|_{\infty},$$

where  $\gamma : \text{Dgm}_k(P) \rightarrow \text{Dgm}_k(Q)$  is a bijection between the persistence diagrams.<sup>4</sup> Stability means that the former upper bounds the latter:

$$d_B(\text{Dgm}_k(P), \text{Dgm}_k(Q)) \leq d_H(P, Q).$$

**Persistence images.** Persistence diagrams are multi-sets of points in the plane. Most machine learning algorithms operate on points in a fixed-dimensional Euclidean space. To translate the former to the latter, Adams et al. [29] introduced *persistence images*, which, after transforming the birth–death pairs  $(b, d)$  into birth–persistence pairs  $(b, d - b)$ , replace each one by a Gaussian (to spread its influence across the neighborhood, since nearby points represent features of similar size), and weigh each point in the plane:

$$f(x, y) = w(y) \sum_{(b, d) \in \text{Dgm}(P)} \mathcal{N}((b, d - b), \sigma).$$

Persistence images result from discretizing  $f(x, y)$  by integrating its value within each grid cell of fixed dimensions. Adams et al. choose the weighting so that  $w(y) \rightarrow 0$  as  $y \rightarrow 0$ , for example, the linear weighting,  $w(y) = y$ . This choice places more weight on persistent points and ensures the stability of the persistence images. Since we do not know a priori which pore size is significant for our application, and because we are not as concerned with stability — zeolite structures are derived within crystal symmetry constraints and thus suffer at most from the floating point error in the atomic coordinates — we also experiment with using no weighting,  $w(y) = 1$ .

We use a modified version of the PersIm package<sup>5</sup> to compute persistence images. We choose a resolution of  $50 \times 50$  and a spread  $\sigma = 0.15$ . We also tried spreads of 0.05, 0.10, 0.15, and 0.20, but the results had similar accuracy in the range of 0.10–0.20. Similarly, we also tried different resolutions ranging from  $10 \times 10$  to  $100 \times 100$  and found that the middle range of  $50 \times 50$  had the best accuracy. The images were scaled on the  $x$ -axis by the maximum birth value across all diagrams and on the  $y$ -axis by the maximum persistence value across all diagrams.

<sup>3</sup><https://github.com/mrzv/dionysus>

<sup>4</sup>For technical reasons, including to make the bijection  $\gamma$  well-defined, persistence diagrams include all the points on the diagonal, taken with infinite multiplicity. We do not dwell on this detail, but refer the interested reader to the surveys [23, 28].

<sup>5</sup>[persim.scikit-tda.org](https://persim.scikit-tda.org)

**Data sets.** To demonstrate an application of our approach, we have employed a dataset constituting of crystal structures of all silica zeolites, for which methane adsorption isotherms were determined previously [26]. Specifically, this set contains 55,039 structures composed on a subset of the PCOD hypothetical zeolite database [4] and a subset of the IZA-recognized zeolites [3]. Adsorption isotherms were determined using grand-canonical Monte Carlo simulations involving classical Lennard-Jones potential models the framework-guest and guest-guest interactions [26]. Isotherms were sampled at 14 pressures in the 1–200 bar range.

For each structure in our set, we have determined the values of the following commonly used descriptors which we refer to as baseline descriptors:

- Accessible volume (AV), in ( $\text{cm}^3/\text{g}$ ), calculated as the total AV resulting from a summation of contributions from methane-accessible channels and non-accessible pockets;
- Accessible surface area (ASA), ( $\text{cm}^2/\text{cm}^3$ ), also calculated with the inclusion of inaccessible pockets;
- Crystal density ( $\rho$ ), in ( $\text{kg}/\text{m}^3$ );
- Pore Limiting Diameter (PLD), in ( $\text{Å}$ ), the diameter of the largest sphere to percolate through material;
- Largest Cavity Diameter (LCD), in ( $\text{Å}$ ), the diameter of the largest sphere than can fit inside the material’s pore system.

The values for the above descriptors were computed using the Zeo++ software package [30].

**Random forests.** We use random forest [31] regression to predict the methane adsorption at 14 different pressures, from 1 to 200 bar. Random forests are ensembles of decision trees. Each decision tree is built recursively by finding a feature, which allows for a cut that maximizes the information gain about the target value. Each tree in the ensemble is built on features chosen at random. The frequency with which each feature is chosen for making a cut serves as an estimate of its importance for the target problem. Such an ability to assign importance to features and to combine features across different modalities, independently of their scale, is the main reason for our choice of random forests.

We build the trees for different combinations of the features to evaluate their relative importance. We use 1- and 2-dimensional persistence images separately and together. Additionally, we employ baseline structural features, described in the previous paragraph, as well as a combination of the two types of features, topological and structural.

We train the random forest on the specific methane adsorption of each material at a given pressure. Each of our forests consists of 500 trees, and the final methane prediction is the average of the prediction of all the trees in the forest. After training the random forest on the training set, predictions are made on an unseen test set. We evaluate the quality of the prediction by comparing the predicted and the correct adsorption values. We compute the root-mean-square deviation,  $\sqrt{\sum(\hat{y}_i - y_i)^2/n}$ , the coefficient of determination ( $R^2$ ),  $1 - \sum(y_i - \hat{y}_i)^2 / \sum(y_i - \bar{y})^2$ , and the Spearman rank coefficient, which computes how well the prediction captures the ordered ranking of the materials by computing the Pearson correlation between the ranks of the predicted and true methane adsorption values, which represents a typical material screening application.

### 3 Results

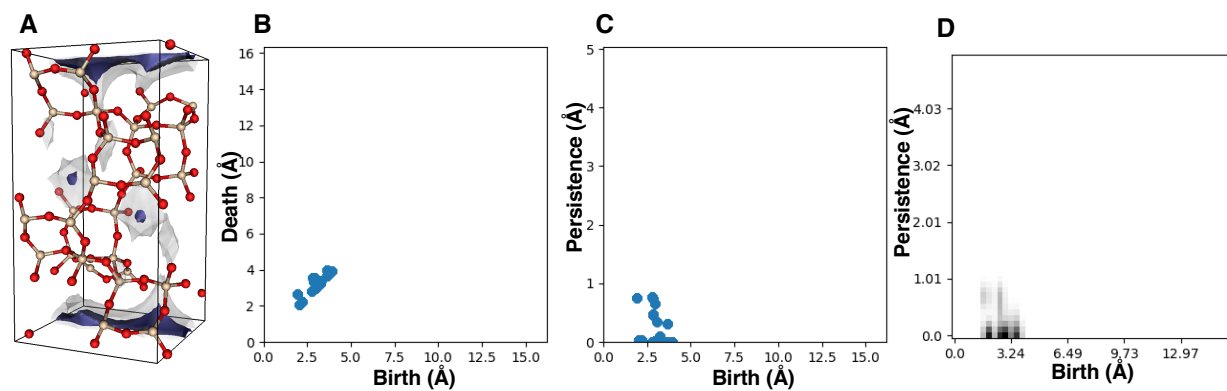


Figure 3: Schematic outlining the (a) zeolite structure, (b) persistence diagram plotted as birth vs. death, which is then (c) linearly transformed to birth vs. persistence, and then expressed as a (d) persistence image. All the diagrams are scaled based on the maximum birth, death, and persistence of the whole data set.

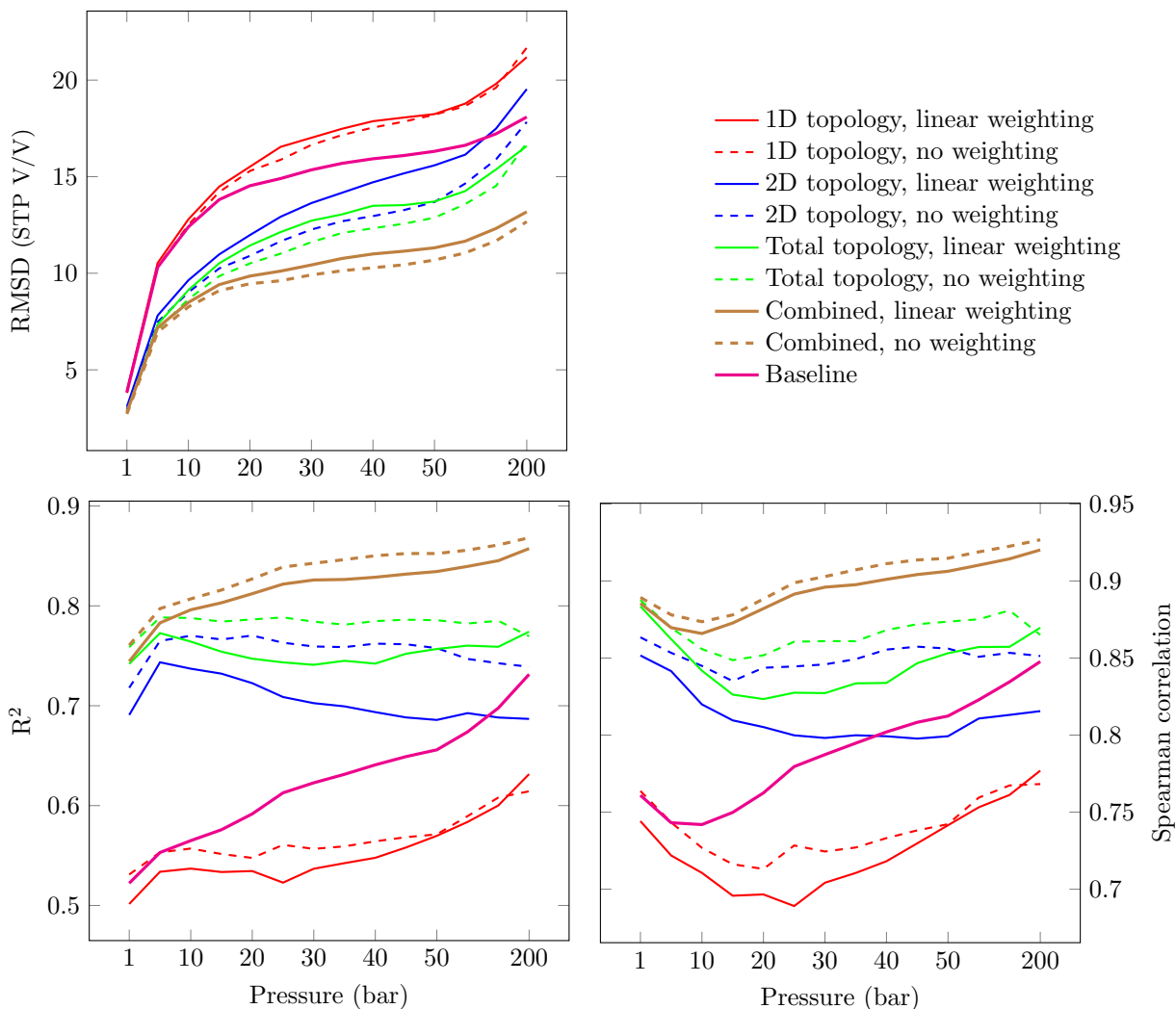


Figure 4: Comparison of root-mean-square deviation (top-left), coefficient of determination (bottom-left), and Spearman rank coefficient (bottom-right) in predicting methane adsorption for different features at different pressures. Top-right: plot legend.

Figure 4 shows the performance of different sets of features in predicting methane adsorption in the 1-200 bar pressure range, measured using the root-mean-square deviation (RMSD), the  $R^2$  score, and the Spearman rank coefficient. The RMSD is low at lower pressures because the distribution of methane adsorption has low variance in that regime. The overall trend of all three metrics is the same: the baseline features perform better than 1-dimensional topological features alone; they also perform better than 2-dimensional features with linear weights at higher pressures. 2-dimensional topological features perform better than baseline (except with linear weights at higher pressures), and combined 1- and 2-dimensional topological features always perform better than the baseline. However, the best results are derived from combining topological and baseline structural features. This clearly shows that topology contributes complimentary information to the baseline and significantly improves performance of our regression model (with  $R^2$  increasing from .52 to .76 at 1 bar and from .73 to .87 at 200 bar).

A particularly interesting revelation of our experiments is the role the weights, used to construct the persistence images, play in the performance of our models. The linear weights, which guarantee stability of the persistence images, perform worse for all but the highest pressure than no weights, which make no



such guarantees. Our best explanation of this behavior is the following. Linear weights make an implicit judgement about the significance of the width of the channels and the diameter of the voids, determining the narrower and smaller topological features (those with lower persistence) to be less important. This scheme is reasonable for many uses of persistence, and indeed ensures the stability of persistence images by limiting the effect of the smaller topological features that result from noisy perturbation of the data. However, with zeolite structures, there is no such error. The atomic coordinates lie on a crystal lattice and the only error we expect is the floating point error. On the other hand, small topological features are related to the roughness of the pore surface, making them important to our application. These observations highlight a major strength of our approach: instead of relying on the pre-determined weights that assign significance to the different regions of the persistence diagrams and images, we let the supervision involved in the learning process to assign the appropriate weights to the different parts of the images. As Figure 5 illustrates, the weights the algorithm chooses do not decay with persistence.

The ability to determine the weights — more accurately, the feature importance — in our model is one of our main motivations for using the random forest. This algorithm builds an ensemble of decision trees, choosing a random subset of features for each one. The frequency with which a particular feature is chosen for a split, across the ensemble, is its importance. Figure 5 shows the importance of the features for the 1-, 2-dimensional topological and the baseline features at three different pressures (low, medium, and high). The color map differs for each plot (even within a single pressure), to make the low values more visible.

From these plots, it is possible to pinpoint the birth and persistence values that correlate most strongly to predicting methane adsorption. As these values describe the voids in a crystal structure, they directly correspond to the structure of the zeolite and thus make it possible to elucidate the porous structure of materials with high methane adsorption. Figure 6 shows an example of the representative cycle for the most important 2-dimensional topology features for a zeolite (ACO) with high methane adsorption, i.e. the porous structure corresponding to the birth and persistence values in Figure 5. A notable property of nanoporous materials is their tunability, i.e. the ability to produce a material with precisely sized pores. Thus, understanding the porous structure correlating to a material property of interest provides insight into synthesizing such a material.

Moreover, the feature that we identify in Figure 6 includes fragments, which are rings containing eight silica atoms. Similar features were identified in the top adsorbing zeolites identified by brute-force screening in an earlier study [26].

The relative importance of the three different types of features is shown in Figure 7. The 1- and 2-dimensional topology plots show the regions of birth and persistence values, which are the best predictors for methane adsorption. Although they change across pressures, these changes are small. In contrast, LCD is the most important feature at low pressures, but ASA becomes the most important feature at higher pressures. It numerically illustrates the empirical observations from earlier studies that local and global features are less transferable across different pressures in comparison to the topology-based features.

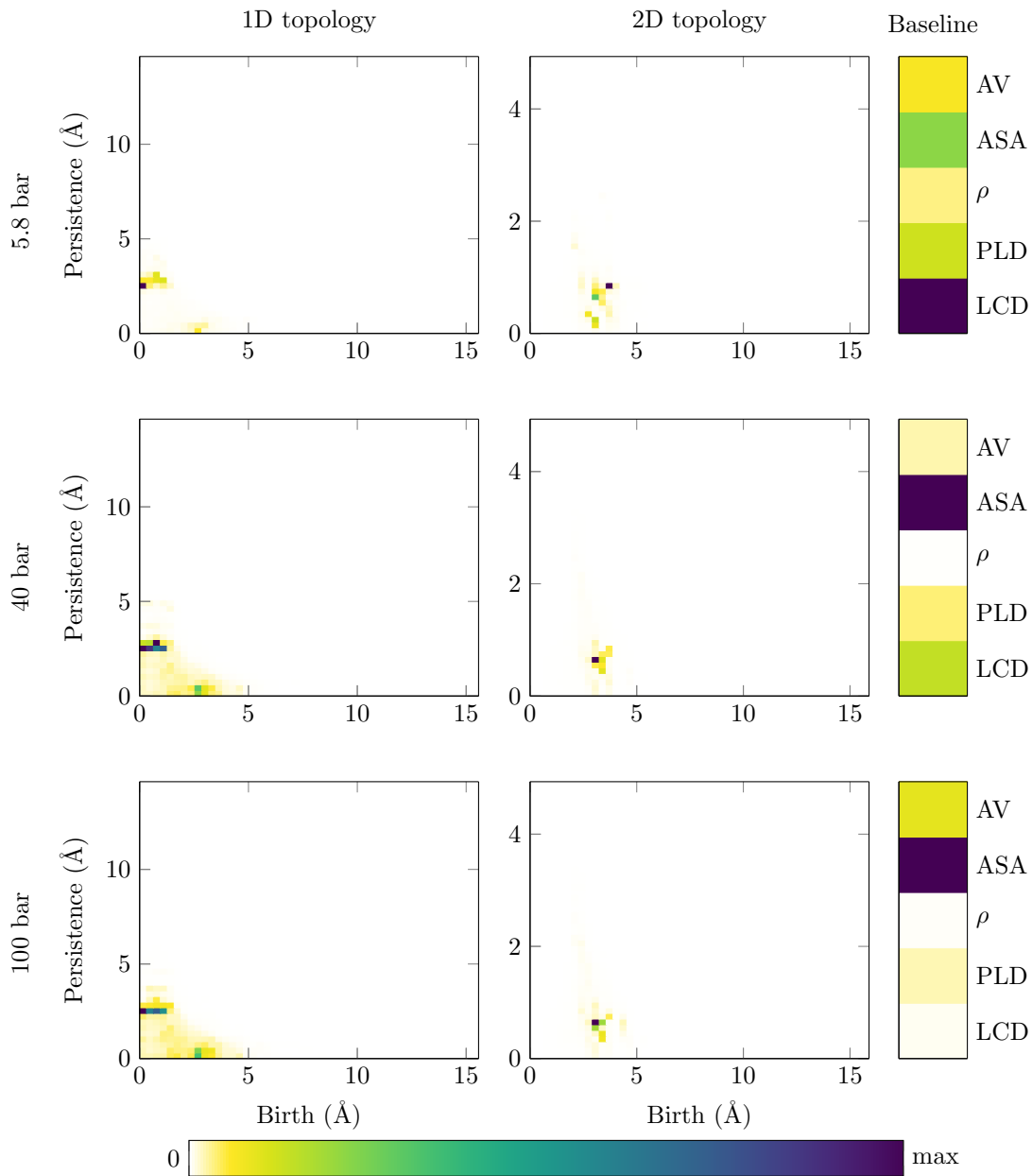


Figure 5: Feature importance of the 1D topology, 2D topology, and baseline features at low (5.8 bar), medium (40 bar), and high (100 bar) pressures. The feature importance colormap is scaled relative to each individual plot. The relative contribution of 1D, 2D topology and baseline features appears in Figure 7.

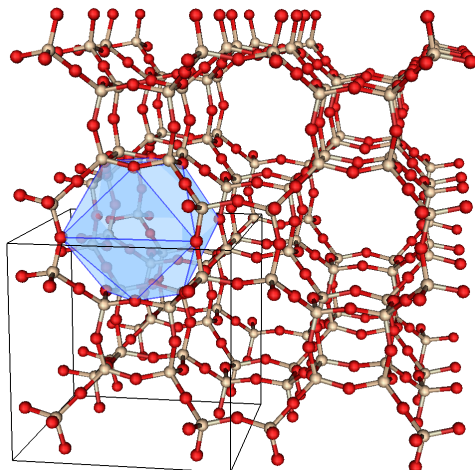


Figure 6: Representative cycle (as highlighted by the points enclosing the blue structure) of the most important 2-dimensional topology features of a high methane adsorption zeolite, ACO. The porous structure corresponds to the birth and persistence values in Figure 5 that are most predictive of methane adsorption.

Figure 7 shows how the relative importance of the three different types of features changes with pressure. Each bar is the sum of feature importance in the combined model, using all three types of features. The plot confirms what we have already observed in Figure 4. At low pressures, topological features perform significantly better than the baseline features and the model places most of its weight on them. As the pressure increases, more weight is assigned to the baseline features, although even at the highest pressure, the combined weight on the topological features exceeds 34% (with most weight assigned to the 2D topological features).

A more recent work [25] is also concerned with using topological signatures for supervised learning of nanoporous material properties. However, there is a key difference from our work: the way it turns topological descriptors into feature vectors relies on constructing ad hoc handcrafted summaries, making it impossible to localize its important subsets. Accordingly, as the authors show [25], the most important of these (global) features closely correlate with the standard structural features, namely, ASA and LCD in our terminology above. In contrast, we input topological descriptors into machine learning algorithms with minimal processing, allowing us both to extract finer information about their important subsets and to show that they capture information complementary to the traditional structural features. Although Zhang et al. [25] consider a slightly different learning task than us, it is sufficiently close for a reasonable comparison: we report higher  $R^2$  scores.

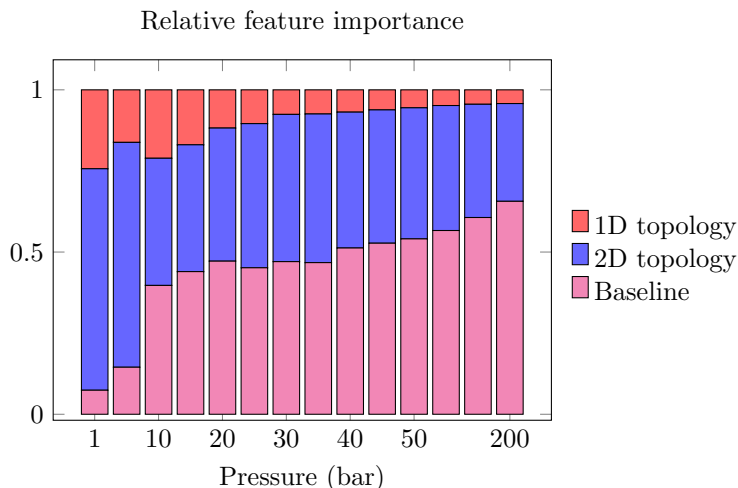


Figure 7: Summary of relative feature importance across different pressures for the 1D, 2D topological features, and baseline features. Combined topological features dominate for low pressures, but as pressure increases, the baseline features become more important. 2D topology features remain significant even at higher pressures.

## 4 Conclusions

We have developed an approach using topology to create a more holistic feature representation for nanoporous materials than the currently widely used porosity descriptors. Specifically, this representation is a vectorized persistence diagram obtained from analysis of the normalized supercell representation of the materials’ crystal structure, that can then be utilized in any machine learning algorithm.

We have tested this approach using a set of roughly 55 thousand zeolite structures and their methane adsorption isotherms obtained using molecular simulations, which enable us to train random forest regressors operating on our representations. We have demonstrated that this approach performs better than current scientific structural features and can accurately predict both adsorption loadings and structures rankings, highlighting the possibility of its applications in various screening studies.

We also note that the highest performance of statistical models was achieved by combining the new topology representations with the commonly used porosity descriptors.

## Acknowledgements

This work was partially supported by Laboratory Directed Research and Development (LDRD) funding from Berkeley Lab, provided by the Director, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231. M.H. acknowledges support from the Spanish Ministry of Economy and Competitiveness (RYC-2013-13949). A.S.K. was supported by the Department of Energy Computational Science Graduate Fellowship Program of the Office of Science and National Nuclear Security Administration in the Department of Energy under contract DE-FG02-97ER25308.

## References

- [1] A. Phan, C.J. Doonan, F.J. Uribe-Romo, M. O’Keeffe, and O. Yaghi. Synthesis, structure, and carbon dioxide capture properties of zeolitic imidazolate frameworks. *Acc. Chem. Res.*, 43:58, 2010.
- [2] Hong-Cai Zhou, Jeffrey R. Long, and Omar M. Yaghi. Introduction to metal–organic frameworks. *Chem. Soc. Rev.*, 112:637, 2012.
- [3] Ch. Baerlocher and L.B. McCusker. Database of zeolite structures. <http://www.iza-structure.org/databases/>, 2017. [Online; accessed 15-Sep-2019].
- [4] Ramdas Pophale, Phillip A. Cheeseman, and Michael W. Deem. A database of new zeolite-like materials. *Phys. Chem. Chem. Phys.*, 13:12407–12412, 2011.
- [5] Jose L. Salcedo Perez, Maciej Haranczyk, and Nils E. R. Zimmermann. High-throughput assessment of hypothetical zeolite materials for their synthesizability and industrial deployability. *Zeitschrift für Kristallographie - Crystalline Materials*, 234:1–15, 2019.
- [6] H. Furukawa, K.E. Cordova, M. O’Keeffe, and O.M. Yaghi. The chemistry and applications of metal–organic frameworks. *Science*, 341:974, 2013.
- [7] 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*, 4:83–89, 2012.
- [8] Tina Duren, Youn-Sang Bae, and Randall Snurr. Using molecular simulation to characterise metal–organic frameworks for adsorption applications. *Chem. Soc. Rev.*, 38:1237, 2009.
- [9] Berend Smit and T.L.M. Maesen. Molecular simulations of zeolites: adsorption, diffusion, and shape selectivity. *Chem. Rev.*, 108:4125, 2008.
- [10] R. Krishna, J.M. van Baten, E. Garcia-Perez, and Sofia Calero. Diffusion of ch<sub>4</sub> and co<sub>2</sub> in mfi, cha and ddr zeolites. *Chem. Phys. Lett.*, 429:219, 2006.
- [11] Emmanuel Haldoupis, Sankar Nair, and David S. Sholl. Efficient calculation of diffusion limitations in metal organic framework materials: A tool for identifying materials for kinetic separations. *J. Am. Chem. Soc.*, 132:7528, 2010.
- [12] Richard L. Martin, Berend Smit, and Maciej Haranczyk. Addressing challenges of identifying geometrically diverse sets of crystalline porous materials. *J. Chem. Inf. Model.*, 52(2):308–318, 2012.
- [13] R.L. Martin, T.F. Willems, L.-C. Lin, J. Kim, J. Swisher, B. Smit, and M. Haranczyk. Similarity-driven discovery of zeolite materials for adsorption-based separations. *ChemPhysChem*, 13:3595–3597, 2012.
- [14] Y. G. Chung et al. In silico discovery of metal-organic frameworks for precombustion co<sub>2</sub> capture using a genetic algorithm. *Sci. Adv.*, 2:e1600909, 2016.
- [15] J. Kim, R.L. Martin, O. Ruebel, M. Haranczyk, and B. Smit. High-throughput characterization of porous materials using graphics processing units. *Journal of Chemical Theory Computation*, 8:1684–1693, 2012.
- [16] Benjamin J. Bucior, N. Scott Bobbitt, Timur Islamoglu, Subhadip Goswami, Arun Gopalan, Taner Yildirim, Omar K. Farha, Neda Bagheri, and Randall Q. Snurr. Energy-based descriptors to rapidly predict hydrogen storage in metal–organic frameworks. *Mol. Syst. Des. Eng.*, 4:162–174, 2019.
- [17] Cory M. Simon, Rocio Mercado, Sondre K. Schnell, Berend Smit, and Maciej Haranczyk. What are the best materials to separate a xenon/krypton mixture? *Chem. Mater.*, 27:4459–4475, 2015.

- [18] Logan Ward, Alexander Dunn, Alireza Faghaniniac, 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. *Comput. Mater. Sci.*, 152:60–69, 2018.
- [19] Maryam Pardakhti, Ehsan Moharreri, 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 Comb. Sci.*, 19:640–645, 2017.
- [20] Michael Fernandez, Peter G. Boyd, Thomas D. Daff, Mohammad Zein Aghaji, and Tom K. Woo. Rapid and accurate machine learning recognition of high performing metal organic frameworks for co2 capture. *J. Phys. Chem. Lett.*, 5:3056–3060, 2014.
- [21] R. Chahine and T. Bose. Characterization and optimization of adsorbents for hydrogen storage. *Hydrogen Energy Prog.*, 2:1259–1264, 1996.
- [22] L-C. Lin, A. Berger, R.L. Martin, J. Kim, J. Swisher, K. Jariwala, C.H. Rycroft, A. Bhowan, M.W. Deem, M. Haranczyk, and B. Smit. In silico screening for carbon capture materials. *Nature Materials*, 11:633–641, 2012.
- [23] Herbert Edelsbrunner and John Harer. Persistent homology – a survey. *Contemp. Math.*, 453:257–282, 2007.
- [24] Yongjin Lee, Senja D. Barthel, Paweł Dłotko, Seyed Mohamad Moosavi, Kathryn Hess, and Berend Smit. High-throughput screening approach for nanoporous materials genome using topological data analysis: Application to zeolites. *J. Chem. Theory Comput.*, 14(8):4427–4437, 2018.
- [25] Xiangyu Zhang, Jing Cui, Kexin Zhang, Jiasheng Wu, and Yongjin Lee. Machine learning prediction on properties of nanoporous materials utilizing pore geometry barcodes. *Journal of chemical information and modeling*, 59(11):4636–4644, November 2019.
- [26] Cory M. Simon, Jihan Kim, Li-Chiang Lin, Richard L. Martin, Maciej Haranczyk, and Berend Smit. Optimizing nanoporous materials for gas storage. *Phys. Chem. Chem. Phys.*, 16:5499–5513, 2014.
- [27] Shujiang Yang, Mohammed Lach-hab, Iosif I. Vaisman, and Estela Blaisten-Barojas. Identifying zeolite frameworks with a machine learning approach. *J. Phys. Chem. C*, 113:21721–21725, 2009.
- [28] Herbert Edelsbrunner and Dmitriy Morozov. Persistent homology. In Jacob E Goodman, Joseph O’Rourke, and Csaba D Tóth, editors, *Handbook of Discrete and Computational Geometry*. CRC Press, 2017.
- [29] 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.
- [30] 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.
- [31] L. Breiman. Random forests. *Int. J. Mach. Learn. Cybern.*, 20:273–297, 1995.