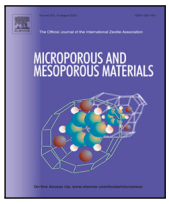
Microporous and Mesoporous Materials 398 (2025) 113796

Contents lists available at ScienceDirect

Microporous and Mesoporous Materials

journal homepage: www.elsevier.com/locate/micromeso



A physically motivated Machine Learning model for accurate gas adsorption predictions in nanoporous materials

Loukas Manitsas, George S. Fanourgakis *

Department of Chemistry, Aristotle University of Thessaloniki, Thessaloniki, GR-54124, Greece

ARTICLE INFO

Keywords:
Metal-Organic Frameworks
Gas adsorption
Machine Learning
Feature engineering

ABSTRACT

In this study we introduce a new set of descriptors to be employed by machine learning (ML) algorithms in order to accurately predict the gas adsorption capacities of Metal-Organic Frameworks (MOFs). The development of these descriptors is based on chemical intuition, specifically the realization that the ability of a porous material to adsorb gas depends not only on the total voids in the material's framework but also on the distribution of these voids. This distribution is efficiently calculated by computing the statistical moments of the helium void fraction. The new approach requires almost no additional computational cost beyond that needed for calculating the standard structural features of nanomaterials. For the development and evaluation of this approach, as well as for comparison with existing methods, previously published computational data are used for the uptake capacities of rare gases and small linear and branched hydrocarbons in Topologically Based Crystal Constructor (ToBaCCo) MOFs under various thermodynamic conditions. Extensive analysis of all results reveals that, despite its simplicity, the new approach provides reliable predictions with the same or even higher accuracy compared to previous methods. ML models developed using the same approaches and additional theoretical data for the adsorption of ethane gas by computation-ready, experimental (CoRE) MOFs demonstrate lower accuracy predictions. Our analysis, aimed at clarifying this point, leads to useful conclusions about the factors that determine the accuracy of each approach and the features needed in the training data to develop predictive models for a wide range of nanoporous materials.

1. Introduction

In the quest for sustainable solutions in energy and environment related problems, the discovery of new Metal–Organic Frameworks (MOFs) with high gas storage capacities has emerged as a pivotal area of research. MOFs represent a class of highly ordered, porous crystalline materials, constructed from metal ions or cluster nodes connected by organic linkers. Their highly tunable nature, vast surface areas, that can surpass $7000\,m^2\,g^{-1}$, high porosity which can exceed 90 % of their volume and the ability to modify their chemical functionality make MOFs exceptional candidates for a variety of applications, particularly in gas adsorption and separation processes. In energy related problems MOFs offer promising solutions for hydrogen storage, which can revolutionize fuel cell technology, and carbon capture [1,2]. Additionally, MOFs are explored for their catalytic properties to enhance the efficiency and sustainability of renewable energy technologies [3,4]. Environmentally, they play a vital role in reducing greenhouse gas emissions [5–7] and purifying air and water, thereby contributing to cleaner ecosystems. In industry, MOFs are revolutionizing processes with their exceptional ability to separate and store gases, leading to more efficient manufacturing and material usage [8].

Screening MOFs experimentally to identify optimal structures for specific applications is highly challenging due to the vast number of potential MOF structures that can be synthesized. In recent years, computer algorithms have been developed to combine MOF building blocks (metal nodes, organic linkers, and functional groups) within given topologies. These algorithms facilitate the design of hypothetical MOFs, which can potentially be synthesized in the lab. One of the first attempts at *in silico* design of MOFs was the database generated by Wilmer et al. [9] containing around 137 000 hypothetical MOFs. More recently, a database containing approximately 325 000 structures was generated using the same algorithm but by combining different building units [10]. This database was used for computational studies of CO_2 adsorption and CO_2/CH_4 separation [10,11]. Even larger databases have been reported; for example, by using a topology-based MOF constructor, several trillions of MOFs were generated by combining a relatively small number of building blocks in different topologies [12].

* Corresponding author.
E-mail addresses: manilouk@chem.auth.gr (L. Manitsas), fanourg@chem.auth.gr (G.S. Fanourgakis).

https://doi.org/10.1016/j.micromeso.2025.113796
Received 16 May 2025; Received in revised form 19 July 2025; Accepted 3 August 2025
Available online 11 August 2025

Each experimental synthesis and characterization process requires significant time, resources, and specialized equipment. Today, according to the Cambridge Structural Database (CSD), [13,14] over 100 000 MOFs have been synthesized and deposited in. However, only a portion of them has been experimentally characterized.

Molecular simulations, have been pivotal in predicting the behavior of gases within MOFs. These simulations can save considerable time and resources by pre-screening MOF candidates for specific applications. In particular Monte Carlo simulation in the grand-canonical ensemble (GCMC) have been almost exclusively used for the study of gas adsorption and separation by nanoporous materials. Nevertheless, they are not without limitations. GCMC simulations can be computationally expensive, especially for large systems, and they rely heavily on the accuracy of the intermolecular potential models used, which may not always capture the complexity of real-world interactions. For example, simple pairwise additive potentials describing van der Waals (i.e. Lennard Jones potentials) and electrostatic interactions are almost exclusively used in this type of studies while important many-body effects, such as atom polarization, are not considered, with only few exceptions [15–17].

The advent of Machine Learning (ML) offers a promising avenue to overcome some of these challenges. ML models, trained on large datasets of MOF structures and their corresponding gas adsorption properties, have the potential to predict adsorption behavior quickly and accurately. ML has already proven to be a valuable alternative to experimental methods and molecular simulations, with several studies conducted, the majority of which have emerged over the last five years.

However, developing ML predictive models is not trivial. The challenges include the need for large and high-quality datasets, the ability to generalize across different MOFs and adsorbates, and the interpretation of the models' predictions in a chemically meaningful way. In several cases, the reference data from studies (usually results of GCMC simulations) are becoming publicly available [11,18–23], enabling the reproducibility of results and facilitating other related studies for these materials. Nevertheless, specifying the input parameters, known as descriptors or features, that will provide the ML algorithm with the necessary information about a material and enable the identification of accurate relationships with the desired outcome (in this case, gas adsorption capacity) remains a field of intense research. Several approaches proposed have been summarized and discussed in recent review articles [24–26].

In this article we introduce a novel, physically motivated set of ML descriptors that, when combined with standard structural features of MOF, such as surface area and void fraction, achieves comparable (if not superior) predictive accuracy to more complicated approaches. Given the small number of new descriptors and their negligible additional computational cost when calculated alongside standard structural features, large databases can be screened very quickly, requiring only small training set sizes. Part of the results are compared across two different databases of hypothetical and experimentally synthesized MOFs, leading to useful conclusions regarding the significance of MOF features in the development and evaluation of ML models.

2. Methodology

2.1. Datasets

In this work the dataset created by Shi et al. [27] was almost exclusively used for the development and evaluation of various ML models. The dataset contains information about structural features of MOFs from the ToBaCCo 1.0 MOF [28] database: helium void fraction (VF), volumetric and gravimetric surface areas (VSA and GSA, respectively), pore limiting diameter (PLD), and largest cavity diameter (LCD).

The adsorption of several important gases was studied by Shi et al. [27] by performing GCMC simulations: Krypton (Kr), Xenon (Xe), Ethane (Eth), Propane (Pro), *n*-Butane (But), *n*-Hexane (Hex) and 2,2-dimethylbutane (DMB). In the present work, for the specification of the systems, we follow the original notation adopted by Shi et al. namely the gas abbreviation is followed by two numbers that correspond to the pressure and temperature of simulations. For example, Pro-5-298, corresponds to the adsorption of Propane at P=5 bar and T=298 K. The adsorption of each gas was calculated for 2000 MOFs at each temperature and pressure (6000 for the DMB). All technical details of the simulations can be found in the original work. For the sake of convenience, the list of abbreviations along with the thermodynamic conditions that the adsorbates were studied is provided in Table S1. As in the original study, in the present one the adsorptions are always expressed in volumetric units (v_{STP}/v).

Apart from the previous results, in order to investigate further the effect of the chemical diversity of MOFs on the performance of the ML predictive models, we performed in this work GCMC simulations for the Eth-4-298, in 2000 randomly selected MOFs from the CoRE 2019 database [29].

2.2. ML descriptors

In what follows we will evaluate and compare the performance of two existing approaches as well as a newly introduced one, in the prediction of the adsorption by MOFs of a number of gases. To that end we summarize the main features of each approach below. All approaches are based on sets of ML descriptors that are using features of potential energy surfaces, namely the interaction of the material with probe atoms (either real or hypothetical). However, they significantly differ on the construction of the energy-based descriptors, as explained below.

2.2.1. Two dimensional energy histogram descriptors (2D–EH)

Details for the two-dimensional energy histograms (2D–EH) approach can be found in the original work [27]. For the sake of completeness, we briefly summarize it here. Two-dimensional histograms are computed for each MOF describing in one dimension the interaction energy of the framework and a probe spherical atom and in the second one the corresponding magnitude of the potential energy gradient. The potential energy and the magnitude of its gradient are computed at the grid points of a discretized unit cell. As probe atoms the Kr and Xe atoms were used, while for the case of alkanes, the methyl group, which is treated as a spherical particle, is also considered. In all cases the interactions between the probe atoms and the framework are described by Lennard-Jones (LJ) functions. All elements of the normalized $n \times m$ histogram matrix were used as descriptors by the ML algorithms. When methyl groups were considered the total number of features was 231. A graphical illustration of the 2D–EH descriptors for a specific MOF is presented in Fig. 1 of the original work [27].

The 2D–EH approach is an extension of a previous approach [30] in which one-dimensional histograms of the probe-framework potential energy were considered (1D–EH). The idea of the 2D–EH approach is that the gradients of the energy also contain spatial information that is missing from the 1D–EH. For example, the 1D–EH descriptors cannot distinguish if a grid point is close or not to framework's atoms. The thorough evaluation of the two approaches [27] for a variety of gases at different thermodynamic conditions revealed that in most cases the 2D–EH is significantly more accurate than the 1D–EH [22].

2.2.2. Structural and probe atoms descriptors (str+Vprb)

The idea of probe atoms is to create a reduced representation of the potential energy surface of the nanomaterial. For that, a hypothetical atom (probe) is inserted in a position r_i of the framework and its interaction energy with the material V is computed. Each probe is characterized by its (Van der Waals) diameter σ and its energy parameter ε that is used to describe the strength of its interaction with the nanomaterial. The value of a probe atom is computed as the average

Boltzmann factor after a large number N of insertions of the probe in different positions of the nanomaterial:

$$\text{probe}(\sigma,\varepsilon) = \frac{1}{N}\sum_{i=1}^{N}\exp\left(-\beta V(\mathbf{r}_i;\sigma,\varepsilon)\right), \tag{1}$$

Since a probe is given as an average, its value may be similar for materials of very different properties. In order to create a unique representation of each material, a number of different probes is considered by choosing different sets of (σ,ε) parameters.

In practice, since all gases examined in this work interact with the frameworks by weak van der Waals interactions, we can consider that the interaction energy V is described by a simple LJ potential. For a probe atom with LJ parameters (σ,ε) at the position $\mathbf{r}_i$ of the framework, the energy is computed as

$$V(\mathbf{r}_i;\sigma,\varepsilon) = \sum_{j=1}^{\text{Nat}} 4\varepsilon_j^0\left[\left(\frac{\sigma_j^0}{r_{ij}}\right)^{12} - \left(\frac{\sigma_j^0}{r_{ij}}\right)^{6}\right] \tag{2}$$

where Nat is the number of atoms in the unit cell of MOF, $r_{ij} \equiv |\mathbf{r}_{ij}| = |\mathbf{r}_i - \mathbf{r}_j|$ is the distance of the probe at the position i with the jth atom of the MOF. Finally, $\sigma_j^0 = (\sigma+\sigma_j)/2$ and $\varepsilon_j^0 = \sqrt{\varepsilon\varepsilon_j}$ (Lorentz–Berthelot combining rules [31,32]) with (σ_j,ε_j) the LJ parameters of the jth atom of the MOF.

We notice that for polar molecules such as CO_2, H_2S and H_2 electrostatic interactions between the guest molecules and the framework should be taken into account, to that end a different type of probe atoms is used in addition. These probes carry a small permanent dipole moment [33]. In order to distinguish the two types of probes, the probes used in the present study are called Vprobes, while the probes with the permanent dipole moment Dprobes.

In the original development of the approach [34] in which the adsorption of methane by 4932 CoRE MOFs [35] was investigated, it was concluded that using 4 probes of sizes 2.5, 3.0, 3.5 and 4.0 Å together with the structural features of MOFs, were sufficient for obtaining converged results. This set of descriptors will be call str+Vprb, hereafter. The addition of smaller or larger probe sizes did not lead to improved predictions. It was found that compared to the case that the ML model is constructed using solely structural features, the addition of Vprobes leads to large improvements in its predictive accuracy. The molecule of methane was described as a spherical particle which interacts with other atoms or molecules with LJ functions. In the present study, in which additional larger, non-spherical and non-symmetrical molecules were adsorbed by the nanomaterials, the previous conclusions as well as the parameters and the optimum number of probe atoms were revisited.

2.2.3. Structural and void fraction moments descriptors (str+VFn)

Experimentally, pore volume v_{pore} (the free volume of the nanomaterial) is a very important property usually determined by measuring at low temperature the amount of nitrogen gas adsorbed in the pores of the nanomaterial. It is directly related to the VF that expresses the ratio of the available volume over the total volume of the nanomaterial via the relation $\text{VF} = v_{\text{pore}}\rho_{\text{crystal}}$. Theoretically, there are several approaches to compute the VF, although discrepancies among the theoretical results as well as between the theoretical and the measurable results are often observed. It is beyond the scope of the present work to investigate the reasons behind these discrepancies, which have been extensively discussed in previous works [36]. Instead, the most commonly used approach, the helium VF will be used here. It is computed as the average Boltzmann factor of the interaction energy of a helium atom with the materials' framework

$$\text{VF} = \frac{1}{N}\sum_{i=1}^{N}\exp\left(-\beta V^{\text{He}}(\mathbf{r}_i)\right). \tag{3}$$

The previous equation is identical to the Eq. (1), but instead of a probe atom the function V describes the interaction energy of a helium atom with the material. The LJ potential (Eq. (2)), with the proper parameters of the helium atom is used again for the calculation of V^{He}. In the following text, any mention of VF refers to the helium void fraction.

One of the claims of this work is that apart from the average value of the Boltzmann factor, its distribution provides also valuable information about the gas–adsorption capacity of a material and as such, can be used as a descriptor by ML algorithms. Standard deviation, skewness and kurtosis are common measures used to describe the shape of a distribution. For example, standard deviation is a measure of the amount of the variation of the values of the variable about its mean, skewness quantifies the extent to which the data deviates from symmetry, while kurtosis indicates how much data resides in the tails.

The previous quantities can be computed using the non-central moments of the distribution. For the case of the VF, these moments can be written as

$$\text{VF}^{(n)} = \frac{1}{N}\sum_{i=1}^{N}\exp\left(-n\beta V^{\text{He}}(\mathbf{r}_i)\right), \tag{4}$$

where $n = 2,3,4,\ldots$ corresponds to the second, third, fourth, and higher non-central moments of the VF, while for $n=1$, $\text{VF}^{(1)}$ is equivalent to the VF of Eq. (3). Based on VF and $\text{VF}^{(2)}$, the variance can be computed as $\text{Var(VF)} = \text{VF}^{(2)} - \text{VF}^2$, while standard expressions for skewness and kurtosis can be found in statistical textbooks [37].

In this work, the non-central moments $\text{VF}^{(2)}$, $\text{VF}^{(3)}$, and $\text{VF}^{(4)}$ from Eq. (4) (collectively referred to as VFn) were used as descriptors alongside the structural features of the MOFs. This combined set is denoted as str+VFn hereafter. Although $\text{VF}^{(3)}$ and $\text{VF}^{(4)}$ exhibit strong correlations with $\text{VF}^{(2)}$ in the ToBaCCo dataset, they are retained due to their negligible computational cost and potential relevance in capturing shape-dependent features across other material classes. VF (i.e., $\text{VF}^{(1)}$) is consistently treated as a structural descriptor throughout.

From a technical standpoint, the VF and $\text{VF}^{(n)}$ descriptors can be efficiently computed alongside the Vprobe descriptors in a single-step calculation. The most computationally demanding task in evaluating Eqs. (1), (3), and (4) is the calculation of the interaction energy between helium atoms (or probes) and all framework atoms located within a real-space cutoff radius R_c. Because all interactions follow the same functional form (Eq. (2)), distances and their powers (r^{-6}, r^{-12}) need only be computed once, yielding substantial efficiency gains. Implementation is straightforward: the necessary modifications to the Poreblazer package [38] are minimal and detailed in the Supporting Information. A version of the package that includes these modifications is freely available at https://github.com/fanourgg/Poreblazer_VFn.

In the reported results the modified version of Poreblazer was used for the calculation of VF, VFn and Vprobes. The materials' framework was discretized in small cubelets with linear dimension equal to 0.5 Å and the interaction energy of the helium atom (or probes) located in a cubelet with the framework was calculated. The Boltzmann factor was computed in all cubelets leading to its final average value. The procedure was repeated for all MOFs. The same descriptors were computed with the Poreblazer and a cubelet size equal to 1 Å. In an alternative approach, instead of a discretized grid, the helium atom (or probes) was randomly placed in the materials' framework. These calculations were performed using homemade codes while the number of random placements was 10^4, 10^5 and 10^6. The final ML results obtained with the VF, VFn and Vprobes computed using the previous two different approaches and the different levels of accuracy were compared, and no significant differences were noted. Therefore, no comparative results will be reported here.

It should be noted here that the LJ parameters of the helium atom used for the calculation of the VF and VFn were σ_{He} =2.58 Å and $\varepsilon_{\text{He}}/k_B$ =10.22 K. These values are slightly different from the helium parameters used by Shi et al. [27] (σ_{He} =2.64 Å and $\varepsilon_{\text{He}}/k_B$ =10.9 K). However, as will be seen later on, this has no noticeable effect on the final results.

It is important to note that the VF and $VF^{(n)}$ descriptors encode purely spatial characteristics of the pore environment and are derived from static, equilibrium configurations of the material. These features reflect geometric and energetic distributions throughout the MOF framework, and do not carry any temporal or sequential information. Accordingly, they are well-suited for use in conventional machine learning algorithms that operate on vectorized inputs, such as decision trees and feedforward networks.

2.3. ML algorithms, training and evaluation

In this work, in order to develop ML predictive models we employed the random forest (RF) [39] and the extremely randomized trees (ERT) [40] algorithms implemented in the scikit-learn module (1.2.2) [41] of python. In addition, we also used the XGBoost implementation of the gradient-boosting decision trees algorithm (XGB) [42].

The results of the 2D–EH approach, as reported by Shi et al. [27], were obtained using a fixed train/test split, with 1000 MOFs assigned to both the training and test sets. For the But-1.2–298 case, the split consisted of 4800 training and 1200 test structures. To enable direct comparison with those results, we adopted the same fixed single-split strategy in these cases.

For independent evaluation of our descriptors, however, we employed a Monte Carlo cross-validation (MCCV) scheme [43] to more rigorously assess model robustness. Specifically, maintaining the same training set sizes, we performed 20 independent train/test runs using randomized splits and report the average values of the following statistical metrics:

$$R^2 = 1 - \frac{\sum_i^n (y_i - \hat{y}_i)^2}{\sum_i^n (y_i - \bar{y})^2} \tag{5}$$

$$MAE = \frac{1}{n} \sum_i^n |y_i - \hat{y}_i| \tag{6}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_i^n (y_i - \hat{y}_i)^2} \tag{7}$$

$$WAPE = \frac{\sum_i^n |y_i - \hat{y}_i|}{\sum_i^n y_i} \times 100\% \tag{8}$$

where y_i and $\hat{y}_i$ are the reference (obtained from GCMC simulations) and the ML predicted value of gas adsorption by the material i, respectively. The average of the n reference values y_i is denoted by $\bar{y}$. We note that the coefficient of determination (or r-squared) R^2 and the weighted average percentage error (WAPE) are dimensionless quantities, while the mean absolute error (MAE) and the root-mean-square error (RMSE) have the units of gas adsorption. R^2 takes its maximum value (=1) when perfect agreement between all y_i and $\hat{y}_i$ is observed. On the other hand the better the predictive model the lower the values of the WAPE, MAE and RMSE are.

In addition to performance evaluation, we employed the SHapley Additive exPlanations (SHAP) method [44] to quantify feature attribution in the trained ML models. SHAP provides a unified framework for interpreting the contribution of each descriptor to model predictions, grounded in cooperative game theory. In our context, it is used to rank descriptors by their average impact on the predicted adsorption values. SHAP values were computed for the tree-based models using the TreeExplainer implementation provided in the shap package.

3. Results & discussion

3.1. Evaluation of str+Vprb descriptors

One of the motivations of the present work is to use the database created by Shi et al. [27] in order to investigate the performance of the previously developed set of descriptors, the Vprobes, for a diverse set of adsorbates and for a variety of thermodynamic conditions, and to determine their optimal parameters. So far, this set of descriptors has been evaluated only during its initial development, [34] for the case of the CH_4, where it was determined that only four probe atoms were sufficient for obtaining the highest accuracy of the approach.

Various number of probe atoms with different (σ, ε) LJ parameters were combined to the standard structural features (VF, VSA, PLD and LCD) and were used for the development of predictive ML models. Their values are tabulated in Table S2. In brief, we examined values of the LJ parameter ε/k_B in Eq. (1) in the range 10 to 100K and σ in the range 2.0 to 5.0Å. Larger values of ε or σ lead, in some cases, to a high increase of the probe atom–host attractive interactions and in turn in high values of the Boltzmann factor. For that, probes with diameters larger than 5.0 Å were not considered.

In Figures S1 and S2 the values of R^2 for the various sets of descriptors are compared for all systems examined. The results were obtained using the ERT method and the MCCV evaluation protocol described before. In each graph the first red bar corresponds to the results obtained with the same set of probe atoms used in our previous work, [34] while the gray bar to the results obtained by combining all different probes. After visual inspection it is concluded that when the value of LJ parameter ε is low (ε/k_B=10 K) lower accuracy is achieved. However, for all other set of probes examined, the accuracy of the obtained results is very similar. It is observed no systematic trend towards a set of probe descriptors. Even when all descriptors are combined no improved results are obtained.

In the case that instead of the ERT, the RF or the XGB algorithm are employed, the accuracy of the predictions (not presented) is slightly lower. However, no noticeable qualitative differences are observed to the previous trends. Based on the previous results it may be concluded that the accuracy of predictions is not depended to the probe atoms parameters for a wide range of the underlying LJ parameters, (ε, σ) and the number of probes. Since the results so far do not justify any refinement of the probe parameters determined for the simpler case of CH_4, for the rest of the work we will use the initial set of parameters as str+Vprb.

Table 1 summarizes the ERT prediction results obtained using the str+Vprb descriptors alongside the structural baseline (str). Comparison with the str-only model highlights the added value of incorporating Vprb descriptors, with the degree of improvement varying across adsorbates and thermodynamic conditions. Overall, the inclusion of probe-based descriptors significantly enhances performance: while R^2 values for the str-only model range from 0.803 to 0.956, the str+Vprb model achieves notably higher accuracy, with R^2 values between 0.949 and 0.984. These results are consistent with previous studies [33,34], which showed that energy-based descriptors yield greater benefits under elevated temperature and reduced pressure conditions.

Although the accuracy of the results obtained using the RF (Table S3) and XGB (Table S4) algorithms is slightly lower overall compared to the ERT algorithm, the key conclusions remain consistent across all models.

3.2. Evaluation of str+VFn descriptors

In Table 1, the ERT results obtained with str+VFn descriptors and the MCCV protocol are also tabulated. It is evident that the differences with the str+Vprb results are negligible, not favoring either of the two descriptor sets. However, the str+VFn framework provides a more transparent basis for the isolated evaluation of void fraction moments, disentangled from the empirically chosen probe-based descriptors. Unlike Vprb, which rely on abstract probe configurations and tunable interaction parameters, the VFn descriptors possess clear physical meaning rooted in the statistical geometry of the pore space–facilitating both interpretability and reproducibility in descriptor design.

To quantify the value of each $VF^{(n)}$ descriptor, we compared model performance using incremental descriptor sets. As shown in Figures S1

Table 1
ERT predictions for the adsorption of various gases by ToBaCCo MOFs, using models trained with different descriptor sets. MAE and RMSE values are reported in units of v_{STP}/v.

system	str				str+Vprb				str+VFn			
	R^2	MAE	RMSE	WAPE	R^2	MAE	RMSE	WAPE	R^2	MAE	RMSE	WAPE
Kr-1–273	0.827	3.0	7.3	29	0.961	1.0	3.5	10	0.962	1.0	3.4	10
Kr-10–273	0.845	10.7	17.1	18	0.976	3.1	6.7	5	0.977	2.9	6.6	5
Xe-1–273	0.851	8.4	15.6	28	0.968	3.2	7.2	11	0.966	3.1	7.4	11
Xe-10–273	0.921	14.3	19.1	11	0.981	6.2	9.4	5	0.980	6.3	9.6	5
Eth-4–298	0.863	12.6	19.4	20	0.982	4.1	7.1	6	0.982	4.1	7.1	6
Eth-20–298	0.938	10.3	14.0	6	0.984	4.9	7.1	3	0.984	5.0	7.2	3
Eth-40–298	0.970	5.3	7.1	2	0.982	3.7	5.5	2	0.982	3.7	5.5	2
Pro-1–298	0.868	11.6	18.8	24	0.976	4.4	8.1	9	0.975	4.5	8.2	9
Pro-5–298	0.927	9.6	15.7	7	0.980	5.2	8.1	4	0.979	5.3	8.3	4
Pro-10–298	0.956	3.9	7.8	2	0.959	3.5	7.6	2	0.958	3.5	7.6	2
But-0.24–298	0.850	12.0	20.2	23	0.970	5.0	9.0	10	0.965	5.3	9.8	10
But-1.2–298	0.888	10.5	21.0	9	0.964	5.9	11.8	5	0.964	5.9	12.0	5
Hex-0.02–298	0.803	11.7	21.2	25	0.949	5.2	10.8	11	0.951	5.1	10.6	11
Hex-10–495	0.889	5.7	7.7	9	0.980	2.2	3.3	4	0.978	2.4	3.5	4
Hex-25–495	0.951	2.9	3.9	3	0.981	1.6	2.4	2	0.981	1.6	2.5	2
DMB-13–477	0.909	5.5	7.6	9	0.967	2.8	4.6	4	0.965	2.9	4.7	4

and S2 of the Supporting Information, the addition of $VF^{(2)}$ to the str baseline yields a significant improvement in predictive accuracy across all systems studied. Incorporating $VF^{(3)}$ (skewness-related) leads to modest gains in select cases, while the inclusion of $VF^{(4)}$ (kurtosis-related) does not yield statistically meaningful improvements.

To further interpret these effects, SHAP analysis was performed using the reduced descriptor set str + $VF^{(2)}$, presented in Figure S4. $VF^{(2)}$ consistently ranks among the top contributing descriptors, confirming its relevance for capturing meaningful variations in the void space. In contrast, Figure S3 shows strong correlations between $VF^{(2)}$ and higher-order moments — $\rho = 0.79$ between $VF^{(2)}$ and $VF^{(3)}$, and $\rho = 0.98$ between $VF^{(3)}$ and $VF^{(4)}$ — indicating statistical redundancy and explaining the inflated SHAP attribution observed for $VF^{(3)}$ and $VF^{(4)}$ (not shown) in extended models. Additionally, VF (i.e., $VF^{(1)}$), already embedded within the structural descriptors, shows no correlation with $VF^{(2)}$ ($\rho = 0.01$), confirming its complementary role.

Overall, these results highlight the importance of $VF^{(2)}$ as a robust and independent descriptor within the ToBaCCo MOF dataset. While $VF^{(3)}$ and $VF^{(4)}$ show statistical redundancy and limited additional impact in this context, their relevance may vary depending on the material class or pore topology under study. Given their negligible computational cost and potential utility in other frameworks, we include all $VF^{(n)}$ descriptors in our modeling and recommend their retention in future applications where broader material diversity or more complex shape distributions may be encountered.

3.3. Comparison of str+Vprb, str+VFn and 2D–EH descriptors

In Fig. 1, the R^2 and RMSE metrics from RF predictions are compared across all systems using three descriptor sets: 2D–EH, str+Vprb, and str+VFn. Corresponding numerical values are provided in Table S5. All models were evaluated using the fixed train/test split previously defined in Ref. [27], ensuring consistency with the original protocol. In particular, comparison of RF results based on str-only descriptors (Table S5) confirms full alignment with the methodology adopted in prior work.

The results indicate a relatively small but consistent improvement in predictive accuracy for the str+Vprb and str+VFn descriptors compared to the 2D–EH approach. Both str+Vprb and str+VFn yield nearly identical performance across all systems examined. Similar trends are observed with the XGB method, as shown in Figure S5 and tabulated in Table S6. In some cases, 2D–EH exhibits slightly better results than str+Vprb and str+VFn, though differences remain marginal.

It is worth emphasizing that this level of predictive accuracy is achieved by the str+Vprb using only 8 descriptors (4 structural and 4 probe-based) and by the str+VFn using only 7 descriptors (4 structural and 3 VF moments), in contrast to the over 200 features required by the 2D–EH framework.

3.4. Comparison of predictions in CoRE MOFs and ToBaCCo MOFs

The fact that the significantly simpler str+VFn and str+Vprb approaches, compared to the 2D–EH, resulted in reliable predictions for all gases and thermodynamic conditions examined motivated us to further investigate the factors that influencing the accuracy of the ML predictive models. In particular, we found surprising that the accuracy for all systems examined is higher than what was found in our previous work [34] for the adsorption by CoRE MOFs of the simpler CH_4 that was treated as a one-particle, spherical molecule. For the CH_4 adsorption at temperature T=298 K and pressure P=1 bar, significantly lower values of R^2 were achieved by the ML models for the str and the str+Vprb descriptors, i.e., 0.686 and 0.930, respectively. Even at the highest pressure examined (P=65 bar), where the structural features have a decisive role in the amount of the adsorbed gas, the statistical accuracy of the predictions is lower (R^2=0.930 and 0.955 without and with the probe atoms) compared to all systems examined in the present work.

For this reason, we decided to re-examine the ML predictions for ethane at T=298 K and P=1 bar but this time using materials from the CoRE 2019 MOFs database as adsorbents. In the first step of this study we proceeded in the calculation of the ethane adsorption by 2093 randomly selected MOFs by performing GCMC simulations using the RASPA2 package [45]. The simulation protocol was exactly the one described by Shi et al. [27] (input files for the RASPA2 are provided by the authors of this work), apart from a smaller number of Monte Carlo cycles. More specifically, due to limited computer resources available, 1×10^4 MC cycles were used for equilibration and 2×10^4 MC cycles for production, compared to the 3×10^4 cycles that were used in both cases in the original work. Structural descriptors were computed with the v3.0 of the Zeo++ [46] program apart from the VF that was computed together with the VFn and Vprb descriptors using the modified version of Poreblazer [38]. The LJ parameters of the Universal Force Field (UFF) [47] were used for the framework atoms, while for ethane the united-atom TraPPE force field [48] was employed.

Before presenting the ML results, we first examine and compare the str+Vprb and str+VFn descriptors of the 2093 CoRE MOFs and 2000 ToBaCCo databases for which the adsorption of ethane has been computed. Since str+Vprb and str+VFn consist of 8 and 7 descriptors, respectively, we employed the t-SNE (t-distributed Stochastic Neighbor Embedding) analysis method [49] to reduce the dimensionality of the

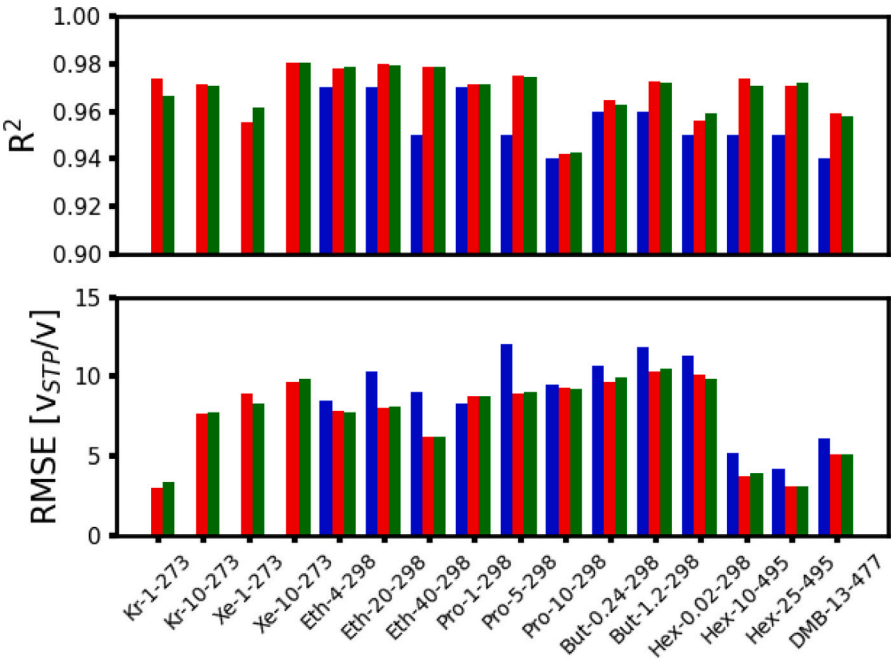


Fig. 1. Graphical comparison of R^2 and RMSE metrics for RF predictions of gas adsorption in the ToBaCCo MOFs, using models trained with different descriptor sets. All models were evaluated using the same fixed train/test split as defined in [27]. Bars represent results from 2D–EH descriptors (blue), str+Vprb descriptors (red), and str+VFn descriptors (green).

original space in 2 dimensions. t-SNE converts similarities between data points into probabilities and then tries to minimize the difference between these probabilities in the high-dimensional and low-dimensional spaces, meaning that similar points in the high-dimensional space remain close together in the low-dimensional representation.

Fig. 2 illustrates t-SNE maps computed using the str+VFn feature space. On the two graphs in left the t-SNE maps of all 4093 MOFs (in gray) together with the 2000 ToBaCCo (top graph, in blue) and CoRE (lower graph, in red) MOFs, are shown separately. It can be easily seen that the two databases have common but also distinct regions. The two graphs on the right are the same as the graphs on the left but the total space of the 4093 MOFs is not shown anymore. Instead, the colors used are now proportional to the amount of ethane adsorbed by ToBaCCo (top graph) and CoRE (lower graph) MOFs, as shown in the vertical bars next to each graph. It is concluded that there is a clear separation of the low- and the high-loading regions with the region on the right of the graphs to correspond to the MOFs with the highest capacity. Also, there are differences between the two databases, however, these are pronounced only on the regions with MOFs of low capacity.

In Figure S6 in SI, a similar diagram appears using the str+Vprb descriptors. The conclusions drawn are the same as before. A similar to the t-SNE analysis is presented in Figures S7 and S8 in SI for the str+Vprb and str+VFn descriptors respectively, using this time the Uniform Manifold Approximation and Projection (UMAP) method [50]. Compared to the t-SNE, the UMAP tends to preserve more of the global structure of the data, meaning it can maintain the overall shape and relationships between clusters better. However, no additional insights to the t-SNE are provided in this case by UMAP.

While analyses like the previous one can be useful, especially when the number of descriptors is large, for the task at hand, in which the number of features is relatively small, it may be easier to simply visualize and compare for the two databases the distributions of the feature spaces and of the ethane adsorption capacities. The latter distribution appears in the density plot of Figure S9 in SI, while the structural features, the Vprobes and the VF moments in Figures S10, S11 and S12, respectively.

In Figure S9 it appears that although the range of adsorption values is similar, a larger number of MOFs with high capacities exist in the CoRE 2019 database compared to the ToBaCCo database. Examining the structural features (Figure S10), it is seen that the ToBaCCo MOFs have in general, larger-cavity and pore limiting diameters. On the other hand, the volumetric surface area is quite similar for the two databases. Finally, for the majority of the CoRE MOFs, the VF ranges from 0.2 to 0.8, while for the ToBaCCo MOFs the majority of MOFs have values > 0.8. Based on the latter point and the fact that the VF distribution of MOFs is wider for the CoRE MOFs, it is not surprising that the distributions of the energy based descriptors Vprobes (Fig. S11) and VFn (Fig. S12) are also wider for the CoRE MOFs.

Overall, the feature spaces of the two databases are quite different. It is interesting therefore, to examine if this fact affects the accuracy of the ML predictive models. The statistical metrics previously described (Eqs. (5)–(8)) obtained with the 3 ML predictive models (RF, ERT, XGB) are tabulated in Table 2 for various combinations of the structural and energy-based descriptors and are compared for the ToBaCCo and the CoRE MOFs separately. The R^2 for the previous cases are also illustrated in Fig. 3. Based on them we can draw several conclusions: (i) Despite the quite different structural features of the two databases, when the structural descriptors are employed, the 3 ML algorithms predict similar values of R^2. For example for the CoRE 2019 MOFs the R^2=0.810–0.839 while for the ToBaCCo MOFs, R^2=0.844–0.867. We can consider therefore that the predictive models for the two databases are of the same accuracy. (ii) For the same database and the same ML algorithm, it can be seen that the predictions of the str+Vprb and the str+VFn descriptors are similar. In some cases the str+Vprb descriptors appear to be slightly more accurate than the str+VFn, while in fewer cases the opposite is true. (iii) In contrast to the two previous observations, when using the same ML algorithm and the str+Vprb or the str+VFn descriptors, the results for the CoRE and ToBaCCo MOFs exhibit significant differences. Clearly, the predictions for the ToBaCCo MOFs are by far more accurate than the predictions for CoRE MOFs. For example, the R^2 is 0.07 to 0.08 higher in ToBaCCo MOFs, and at the same time, the WAPE is almost half compared to the CoRE MOFs.

One possible reason investigated was the fact of the quite different distributions of descriptors for the two databases: since the distributions of the energy based descriptors (Vprb and VFn) are narrower in the ToBaCCo MOFs, we expect that fewer data are required for the proper

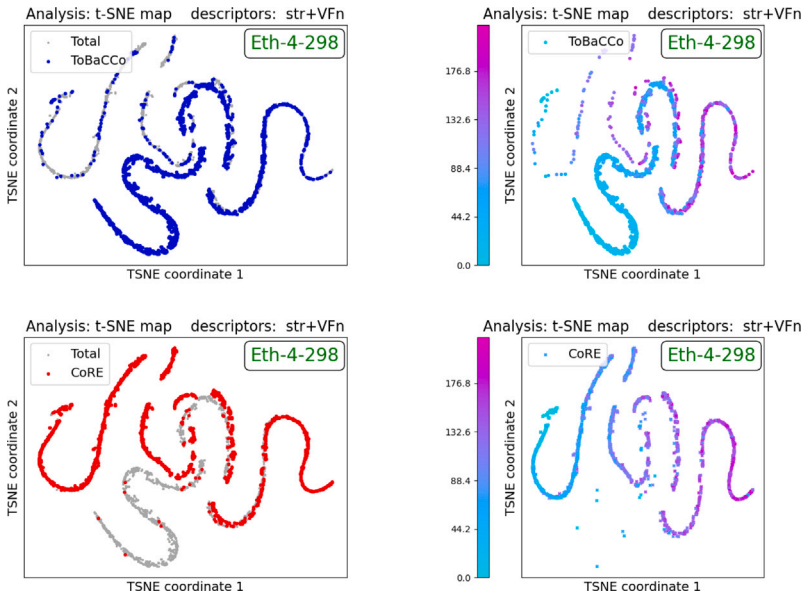


Fig. 2. t-SNE maps illustrating the diversity of the str+VFn descriptors of 2000 ToBaCCo and 2093 CoRE MOFs for which the adsorption of ethane at T=298 K and P=4 bar was calculated by GCMC simulations. On the two graphs in the left the t-SNE maps of ToBaCCo MOFs (upper graph) and CoRE MOFs (lower graph) are shown in blue and red respectively, over the space of all MOFs (shown in gray). On the corresponding graphs to the right the colors are proportional to the ethane loading.

Table 2
Accuracy of the RF, XGB and ERT algorithms in the prediction of ethane adsorption by MOFs at T=298 K and P=4 bar for various set of descriptors. MAE and RMSE values are reported in units of v_{STP}/v. The number in parentheses is the uncertainty in the last digit(s) of the average.

Method	Descriptors	CoRE-2019 MOFs				ToBaCCo MOFs			
		R^2	MAE	RMSE	WAPE	R^2	MAE	RMSE	WAPE
RF	str	0.839(11)	14.0(3)	20.7(5)	14.3(4)	0.867(7)	12.5(3)	19.2(5)	19.4(5)
	str+Vprb	0.923(4)	10.0(3)	14.3(4)	10.2(3)	0.982(2)	4.1(2)	7.1(3)	6.4(2)
	str+VFn	0.904(5)	11.3(3)	16.0(3)	11.5(3)	0.982(1)	4.2(2)	7.1(3)	6.5(2)
XGB	str	0.810(16)	15.3(4)	22.5(8)	15.6(4)	0.844(7)	13.4(4)	20.8(5)	20.8(6)
	str+Vprb	0.912(6)	10.9(3)	15.3(4)	11.1(3)	0.976(3)	4.6(2)	8.1(5)	7.1(3)
	str+VFn	0.888(8)	12.4(3)	17.3(5)	12.6(4)	0.976(2)	4.7(2)	8.1(3)	7.3(2)
ERT	str	0.834(13)	14.3(3)	21.0(7)	14.6(4)	0.863(6)	12.7(3)	19.5(5)	19.8(4)
	str+Vprb	0.919(5)	10.3(3)	14.7(4)	10.5(3)	0.976(3)	4.5(2)	8.1(5)	7.0(3)
	str+VFn	0.899(7)	11.7(3)	16.4(5)	11.9(4)	0.977(2)	4.6(2)	7.9(4)	7.1(2)

training of an ML algorithm. For that reason, in Fig. 4 the R^2 is shown as a function of the training set size for the two databases and the three different sets of descriptors. Visual inspection reveals that indeed for the ToBaCCo MOFs converged results are obtained faster than for the CoRE MOFs. However, it is not likely that even if additional training data will be provided for the CoRE MOFs R^2 values similar to the ToBaCCo MOFs will be eventually reached. Further investigation will be conducted in the future, incorporating additional reference data (GCMC results) to clarify this point. It is also very surprising that the R^2 value obtained for the ToBaCCo MOFs using just 50 training data is higher to the R^2 obtained for the CoRE MOFs using the maximum number (1750) of training data.

One of the main findings of this work is that ML predictions on the gas adsorption by MOFs is more challenging in CoRE MOFs, compared to the ToBaCCo MOFs. Notably, the SHAP analysis presented in Figure S4 of the Supporting Information indicates that, for the Eth-4-298 system, VF and $VF^{(2)}$ consistently emerge as the most influential descriptors across both databases. In the parity plots of Fig. 5 the reference (GCMC computed) training and test data of the ToBaCCo and CoRE MOFs are compared to the predictions of the RF models. The latter have been computed using the str, str+Vprb, and str+VFn descriptors. In all cases 1000 randomly selected MOFs have been used for the training of the algorithms and the remaining for testing. The statistical metrics denoted in each graph and a simple comparison after visual inspection verifies the previous statement. XGB and ERT algorithms (shown in Figures S14 and S16 of SI, respectively), are also in agreement.

Another important point for the two databases is the range of their applicability, namely, under what circumstances ML models trained using data from the one or the other database are expected to provide reliable predictions. The importance of this point was extensively discussed in our previous work [51]. To investigate it, results of additional calculations are presented in Fig. 6. The parity plots on the three upper graphs demonstrate the accuracy of a ML model that was trained using 2000 ToBaCCo MOFs (blue symbols) for the prediction of the Eth-4-298 adsorption in the 2093 CoRE MOFs. Similar to the Fig. 5, the same 3 sets of descriptors were examined, while the RF algorithm was employed as well. The accuracy achieved, in this first scenario, is overall very low: The R^2 values are 0.783, 0.792, and 0.831 for the str, str+Vprb, and str+VFn sets of descriptors, respectively. Surprisingly, the addition of the energy-based descriptors (probe atoms or VF moments) to the structural features of MOFs leads only to minor improvements of the predictions.

A second scenario is examined at the three bottom graphs of Fig. 6: the 2093 CoRE MOFs are used for the training of the RF algorithm and the 2000 ToBaCCo MOFs for the evaluation of the models' accuracy. In this case the accuracy achieved for the R^2 is much higher when the energy-based descriptors are considered (0.951, and 0.961 for the str+Vprb and str+VFn descriptors, respectively, compared to the 0.792 and 0.831 reported before for the first scenario). This is not surprising,

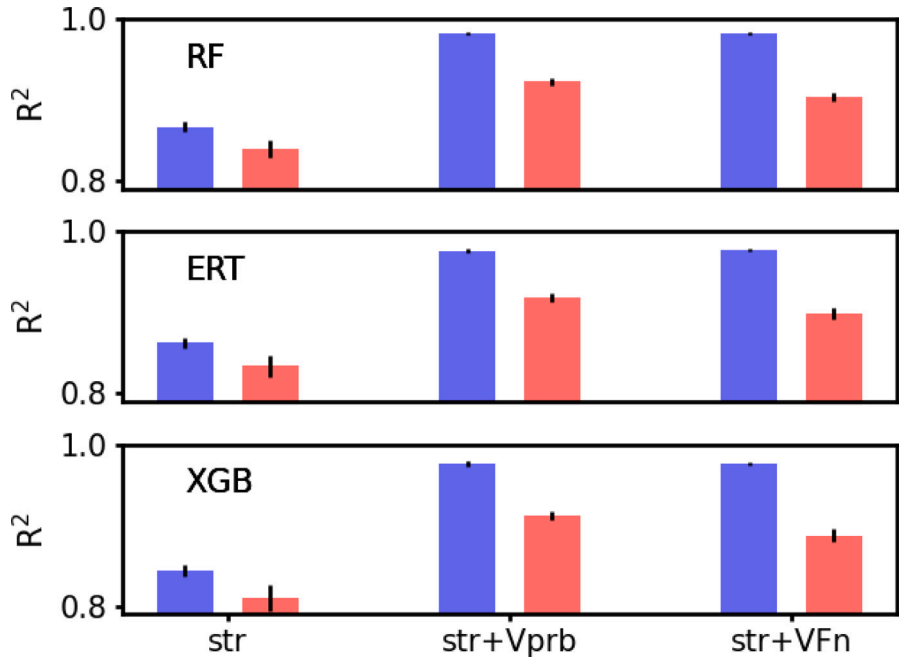


Fig. 3. R^2 values of the predictions of the RF (top), ERT (middle) and XGB (bottom) algorithms for the ethane adsorption by ToBaCCo (blue bars) and CoRE (red bars) MOFs. Various sets of descriptors are examined in each graph. Black vertical lines on the top of the bars indicate the standard deviation of the R^2.

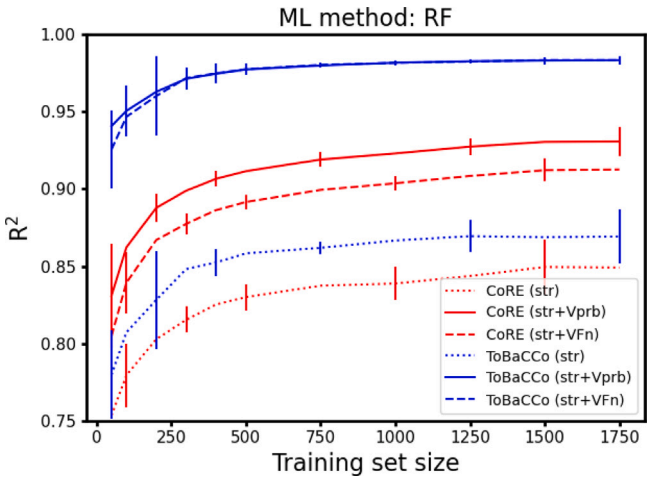


Fig. 4. Convergence of the RF algorithm with the training set size. ToBaCCo MOFs are shown in blue and CoRE MOFs in red. Dotted, dashed and solid lines correspond to the str, str+VFn and str+Vprb descriptors, respectively. Vertical lines on data points indicate the standard deviation of the corresponding values. The corresponding results with the XGB and ERT algorithms are shown in Figure S13 in SI together with some technical details.

according to our previous discussion, since the distribution of the Vprobes and $VF^{(n)}$ descriptors are much narrower in ToBaCCo MOFs and within the value range of the much wider distributions of CoRE MOFs (see Figures S11 and S12 in the SI).

On the other hand, for the same reasons, one also expects comparable or even higher accuracy in the first scenario (training data from the ToBaCCo and test data from the CoRE database) when structural features alone are used as descriptors by the ML algorithms. In this case, although the VF distribution is wider in CoRE MOFs, the distributions of the pore sizes, PLD and LCD is much wider in the ToBaCCo database (see Figure S10 in SI). Therefore, the R^2=0.718 found in this case is lower to what was found in the first scenario (R^2=0.783).

Similar to the RF results in Figures 5 and 6 are shown in SI, for the XGB (Figures S14 and S15) and the ERT algorithms (Figures S16 and S17). When the str+Vprb or the str+VFn descriptors are employed, the predictions for the CoRE MOFs of the ERT algorithm that was trained using the ToBaCCo MOFs are more accurate compared to the XGB and the RF algorithms. Instead, the most accurate predictions for the ToBaCCo MOFs from models trained using the CoRE MOFs are obtained by the RF algorithm. Despite some small quantitative differences on the results of the 3 ML algorithms that were examined, in all cases the accuracy of the ML models that were trained using the CoRE MOFs is by far higher compared to the models trained using the ToBaCCo MOFs. We may conclude therefore that predictive models developed using CoRE MOFs lead to more reliable predictions than the ToBaCCo MOFs for a wider range of materials.

4. Conclusions

Using a small set of descriptors, namely by combining 4 structural features of MOFs with 4 probe atoms or with 3 moments of VF, we constructed accurate ML predictive models for the adsorption by MOFs of various gases at different thermodynamic conditions. The accuracy of the obtained results is similar or slightly higher to the reported results obtained with the 2D energy histograms [27] in which a significantly larger number of descriptors was used.

Regarding the str+Vprb set of descriptors, an extensive search on the number and the parameters of the probe atoms did not reveal any substantial improvements with respect to the original development [34] which could justify any refinements. It is clear that, for all gases and conditions examined, the results are not very sensitive, at least for a wide range of the probe parameters. The accuracy of the results obtained with the str+Vprb descriptors is similar to that obtained with the newly developed str+VFn descriptors. However, the latter set has a clear physical meaning, i.e., it indicates that not only the available space, as it is expressed by the VF, is important for the amount of the adsorbed gas, but also the distribution of the voids in the material is a decisive factor. In that sense, we demonstrated that a simple statistical description of the voids' distribution, namely the first few moments of the distribution, leads to accurate ML predictive models. The fact that the $VF^{(n)}$ are computed using the same equation (Eq. (4)) to what used for the probe atoms (Eq. (1)) leads to the conclusion that $VF^{(n)}$ are a special case of Vprobes.

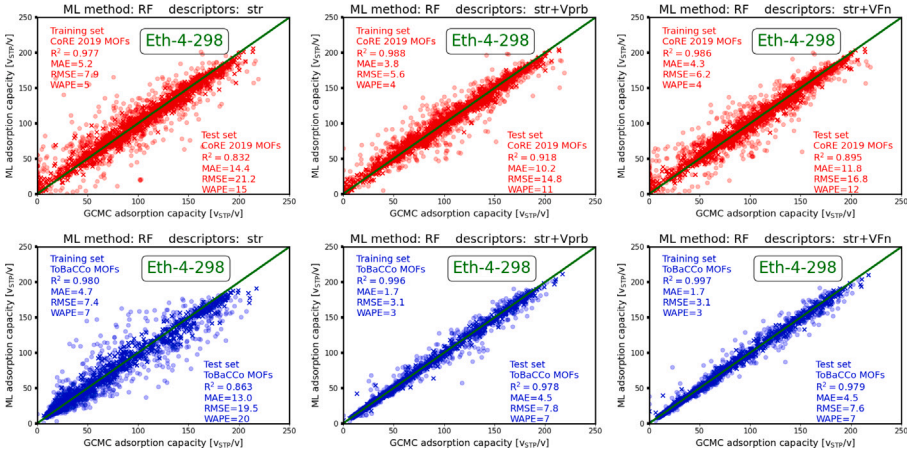


Fig. 5. RF predictions for the adsorption of ethane at P =4 bar and T =298 K, using various set of descriptors (structural alone (str) or combined with Vprobes (str+Vprb) or VF moments (str+VFn). CoRE 2019 MOFs are shown with red color, while ToBaCCo MOFs with blue. The training data (1000 MOFs in all cases) are shown with "×" while the test data (1000 ToBaCCo or 1063 CoRE MOFs) with "∘". Various statistical metrics for the training and test data are denoted inside the graphs.

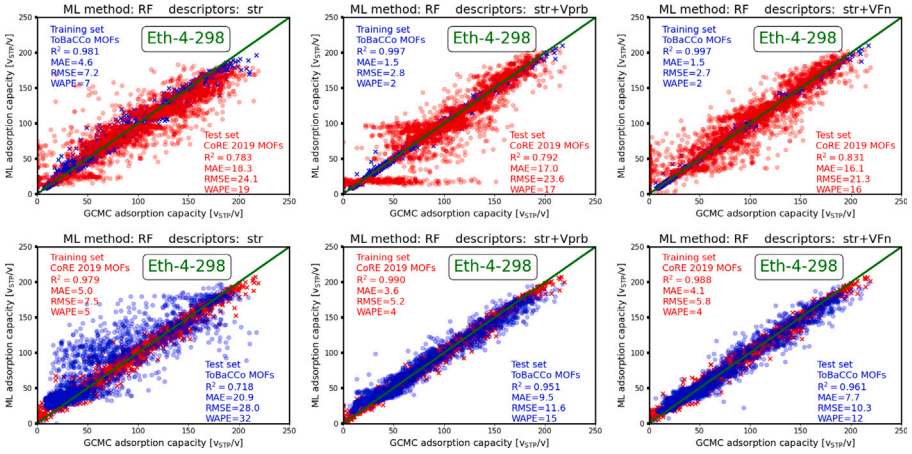


Fig. 6. Similar with the Fig. 5. In this case the in each graph the training and test data belong to a different database of MOFs (either ToBaCCo or CoRE 2019 MOFs).

While this study primarily focuses on gas adsorption under moderate to high pressure conditions, as defined by the scope of the GCMC dataset employed, we acknowledge that low-pressure regimes — particularly those governed by Henry's law — are also of significant scientific interest. Although such data were not included in the original simulations, the descriptor framework proposed here is general and may be readily extended to alternative adsorption conditions. Moreover, in a recent study [52] investigating the adsorption of CO_2, H_2S, and H_2 in the low-pressure regime, it was found that ML predictions based on probe atom descriptors achieved accuracy comparable to that of Henry's coefficient-based models, while offering superior generalization and requiring significantly lower computational cost (by approximately two orders of magnitude).

Based on a previous work [33] in which the CO_2, H_2S and H_2 adsorption by CoRE MOFs [35] was studied, it was seen that the use of the probe atoms did not substantially improve the results obtained using the standard structural features alone as descriptors. The important difference between the three previous systems and the systems examined here are the nature of guest-host interactions. In the present systems, only weak van der Waals interactions are considered between the gas and the framework, which are described by simple LJ potentials. In contrast, for the case of the polar CO_2, H_2S and H_2 molecules, we have considered electrostatic interactions as well. Therefore, since the VF does not consider at all these interactions it does not provide sufficient information about the potential energy surface of the material. Instead, the proposed use of neutral atoms that carry a small permanent dipole moment (Dprobes) [33] which account for electrostatic interactions leads to significantly improved predictive models. To conclude, while VF in conjunction with VFn moments provides meaningful insight into gas adsorption capacity, the use of probe atoms — with or without permanent dipole moments — appear to offer greater flexibility.

For the case of the ethane adsorption at low pressure, we examined the accuracy of the present ML models in two different, well established databases of MOFs: the database of hypothetical ToBaCCo MOFs and the database of experimentally synthesized CoRE MOFs. Analysis of the features of these databases revealed that the two datasets differ significantly. CoRE MOFs are having larger chemical diversity, while the structures of ToBaCCo MOFs are characterized by larger pores. As a result, the value range of energy based descriptors Vprobes and $VF^{(n)}$ as well as of the VF is much wider in CoRE MOFs compared to the ToBaCCo MOFs. The opposite holds for the value range of pores. We found that under the same thermodynamic conditions and the same set of descriptors, the accuracy of the ML models is higher for the ToBaCCo database. One reason may be that more training data are required in CoRE MOFs due to the wider range of Vprobes, VF and $VF^{(n)}$. This point will be further investigated in the future. It is safe, however, to conclude that development of accurate ML models is more challenging in CoRE MOFs. At the same time, the main advantage is that the ML models are more general and can provide accurate predictions for a wider range of materials.

Overall, we found that the use of simple and physically motivated energy-based descriptors can lead to very accurate ML predictive models, while for a comparison of the predictive accuracy of the various ML

approaches the materials used for the development of the ML models should be carefully considered.

In recent years, deep learning models such as MOFTransformer, [53] ALIGNN, [54] and AIdsorb [55,56] have demonstrated strong performance in predicting MOF properties by directly leveraging raw structural representations — including atomistic graphs, energy grids, and molecular point clouds. These approaches rely on complex architectures and typically require very large training datasets (ranging from 130,000 to over 1 million structures), along with substantial computational resources. While their predictive accuracy is impressive, they often lack interpretability and may be less accessible to researchers without advanced GPU infrastructure.

In contrast, our descriptor-based framework emphasizes clarity, efficiency, and transferability. It achieves competitive accuracy using only 2000 MOFs, relying on physically meaningful descriptors that require minimal preprocessing and offer straightforward interpretation. This makes the approach well-suited for data-scarce scenarios, targeted screening, and high-throughput applications on standard computing resources.

We acknowledge, however, that further comparative evaluation is warranted. To this end, we are actively benchmarking our approach against the aforementioned methods, aiming to systematically assess trade-offs in predictive performance, data efficiency, and model transparency.

CRediT authorship contribution statement

Loukas Manitsas: Writing – review & editing, Methodology, Formal analysis. **George S. Fanourgakis:** Writing – review & editing, Writing – original draft, Supervision, Software, Resources, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was financially supported by the Research Committee of Aristotle University of Thessaloniki (Project No. 10356). Use of computational resources of the AUTh High Performance Computing (HPC) infrastructure of the Aristotle University of Thessaloniki is also acknowledged.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.micromeso.2025.113796. Figures for algorithms and targets not presented in the original manuscript.

Data availability

Data will be made available on request.

References

[1] Z. Ali Sandhu, M. Asam Raza, N.S. Awwad, H.A. Ibrahium, U. Farwa, S. Ashraf, A. Dildar, E. Fatima, S. Ashraf, F. Ali, Metal–organic frameworks for next-generation energy storage devices; a systematic review, Mater. Adv. 5 (2024) 30–50, http://dx.doi.org/10.1039/D3MA00822C.

[2] L. Zhang, M.D. Allendorf, R. Balderas-Xicohténcatl, D.P. Broom, G.S. Fanourgakis, G.E. Froudakis, T. Gennett, K.E. Hurst, S. Ling, C. Milanese, P.A. Parilla, D. Pontiroli, M. Riccò, S. Shulda, V. Stavila, T.A. Steriotis, C.J. Webb, M. Witman, M. Hirscher, Fundamentals of hydrogen storage in nanoporous materials, Prog. Energy 4 (2022) 042013, http://dx.doi.org/10.1088/2516-1083/ac8d44.

[3] M. Shanmugam, N. Agamendran, K. Sekar, T.S. Natarajan, Metal–organic frameworks (MOFs) for energy production and gaseous fuel and electrochemical energy storage applications, Phys. Chem. Chem. Phys. 25 (2023) 30116–30144, http://dx.doi.org/10.1039/D3CP04297A.

[4] C. Pettinari, A. Tombesi, MOFs for Electrochemical Energy Conversion and Storage, Inorganics 11 (2023) 65, http://dx.doi.org/10.3390/inorganics11020065.

[5] C. Pettinari, A. Tombesi, Metal–organic frameworks for carbon dioxide capture, MRS Energy Sustain. 7 (2020) 35, http://dx.doi.org/10.1557/mre.2020.30.

[6] E. Mahmoud, Mitigating Global Methane Emissions Using Metal-Organic Framework Adsorbents, Appl. Sci. 10 (2020) 7733, http://dx.doi.org/10.3390/app10217733.

[7] L. Ali, E. Mahmoud, Recent advances in the design of metal–organic frameworks for methane storage and delivery, J. Porous Mater. 28 (2021) 213–230, http://dx.doi.org/10.1007/s10934-020-00984-z.

[8] A.U. Czaja, N. Trukhan, U. Müller, Industrial applications of metal–organic frameworks, Chem. Soc. Rev. 38 (2009) 1284–1293, http://dx.doi.org/10.1039/B804680H.

[9] C.E. Wilmer, M. Leaf, C.Y. Lee, O.K. Farha, B.G. Hauser, J.T. Hupp, R.Q. Snurr, Large-scale screening of hypothetical metal–organic frameworks, Nat. Chem. 4 (2012) 83–89, http://dx.doi.org/10.1038/nchem.1192.

[10] M.Z. Aghaji, M. Fernandez, P.G. Boyd, T.D. Daff, T.K. Woo, Quantitative Structure-Property Relationship Models for Recognizing Metal Organic Frameworks (MOFs) with High CO_2 Working Capacity and CO_2/CH_4 Selectivity for Methane Purification, Eur. J. Inorg. Chem. 2016 (2016) 4505–4511, http://dx.doi.org/10.1002/ejic.201600365.

[11] P.G. Boyd, A. Chidambaram, E. García-Díez, C.P. Ireland, T.D. Daff, R. Bounds, A. Gładysiak, P. Schouwink, S.M. Moosavi, M.M. Maroto-Valer, J.A. Reimer, J.A. Navarro, T.K. Woo, S. Garcia, K.C. Stylianou, B. Smit, Data-driven design of metal–organic frameworks for wet flue gas CO2 capture, Nature 576 (2019) 253–256, http://dx.doi.org/10.1038/s41586-019-1798-7.

[12] S. Lee, B. Kim, H. Cho, H. Lee, S.Y. Lee, E.S. Cho, J. Kim, Computational Screening of Trillions of Metal–Organic Frameworks for High-Performance Methane Storage, ACS Appl. Mater. Interfaces 13 (2021) 23647–23654, http://dx.doi.org/10.1021/acsami.1c02471.

[13] P.Z. Moghadam, A. Li, S.B. Wiggin, A. Tao, A.G.P. Maloney, P.A. Wood, S.C. Ward, D. Fairen-Jimenez, Development of a Cambridge Structural Database Subset: A Collection of Metal-Organic Frameworks for Past, Present, and Future, Chem. Mater. 29 (2017) 2618–2625, http://dx.doi.org/10.1021/acs.chemmater.7b00441.

[14] P.Z. Moghadam, A. Li, X.-W. Liu, R. Bueno-Perez, S.-D. Wang, S.B. Wiggin, P.A. Wood, D. Fairen-Jimenez, Targeted classification of metal–organic frameworks in the Cambridge structural database (CSD), Chem. Sci. 11 (2020) 8373–8387, http://dx.doi.org/10.1039/D0SC01297A.

[15] A.L. Mullen, T. Pham, K.A. Forrest, C.R. Cioce, K. McLaughlin, B. Space, A Polarizable and Transferable PHAST CO2 Potential for Materials Simulation, J. Chem. Theory Comput. 9 (2013) 5421–5429, http://dx.doi.org/10.1021/ct400549q.

[16] T.M. Becker, L.C. Lin, D. Dubbeldam, T.J. Vlugt, Polarizable Force Field for CO2 in M-MOF-74 Derived from Quantum Mechanics, J. Phys. Chem. C 122 (2018) 24488–24498, http://dx.doi.org/10.1021/acs.jpcc.8b08639.

[17] T.M. Becker, J. Heinen, D. Dubbeldam, L.C. Lin, T.J. Vlugt, Polarizable Force Fields for CO2 and CH4 Adsorption in M-MOF-74, J. Phys. Chem. C 121 (2017) 4659–4673, http://dx.doi.org/10.1021/acs.jpcc.6b12052.

[18] N.S. Bobbitt, K. Shi, B.J. Bucior, H. Chen, N. Tracy-Amoroso, Z. Li, Y. Sun, J.H. Merlin, J.I. Siepmann, D.W. Siderius, R.Q. Snurr, MOFX-DB: An Online Database of Computational Adsorption Data for Nanoporous Materials, J. Chem. Eng. Data 68 (2023) 483–498, http://dx.doi.org/10.1021/acs.jced.2c00583.

[19] A. Ahmed, S. Seth, J. Purewal, A.G. Wong-Foy, M. Veenstra, A.J. Matzger, D.J. Siegel, Exceptional hydrogen storage achieved by screening nearly half a million metal–organic frameworks, Nat. Commun. 10 (2019) 1568, http://dx.doi.org/10.1038/s41467-019-09365-w.

[20] A. Ahmed, D.J. Siegel, Predicting hydrogen storage in MOFs via machine learning, Patterns 2 (2021) 100291, http://dx.doi.org/10.1016/j.patter.2021.100291.

[21] C.E. Wilmer, O.K. Farha, Y.-S. Bae, J.T. Hupp, R.Q. Snurr, Structure-property relationships of porous materials for carbon dioxide separation and capture, Energy Environ. Sci. 5 (2012) 9849, http://dx.doi.org/10.1039/c2ee23201.

[22] Z. Li, B.J. Bucior, H. Chen, M. Haranczyk, J.I. Siepmann, R.Q. Snurr, Machine learning using host/guest energy histograms to predict adsorption in metal–organic frameworks: Application to short alkanes and Xe/Kr mixtures, J. Chem. Phys. 155 (2021) 014701, http://dx.doi.org/10.1063/5.0050823.

[23] R. Mercado, R.-s. S. Fu, A.V. Yakutovich, L. Talirz, M. Haranczyk, B. Smit, In Silico Design of 2D and 3D Covalent Organic Frameworks for Methane Storage Applications, Chem. Mater. 30 (2018) 5069–5086, http://dx.doi.org/10.1021/acs.chemmater.8b01425.

[24] K. Mukherjee, Y. Colón, Machine learning and descriptor selection for the computational discovery of metal–organic frameworks, Mol. Simul. 47 (2021) 1–21, http://dx.doi.org/10.1080/08927022.2021.1916014.

[25] C. Yang, J. Qi, A. Wang, J. Zha, C. Liu, S. Yao, Application of machine learning in MOFs for gas adsorption and separation, Mater. Res. Express 10 (2023) 122001, http://dx.doi.org/10.1088/2053-1591/ad0c07.

[26] I.-T. Sung, Y.-H. Cheng, C.-M. Hsieh, L.-C. Lin, Machine Learning for Gas Adsorption in Metal–Organic Frameworks: A Review on Predictive Descriptors, Ind. Eng. Chem. Res. (2025) http://dx.doi.org/10.1021/acs.iecr.4c03500.

[27] K. Shi, Z. Li, D.M. Anstine, D. Tang, C.M. Colina, D.S. Sholl, J.I. Siepmann, R.Q. Snurr, Two-Dimensional Energy Histograms as Features for Machine Learning to Predict Adsorption in Diverse Nanoporous Materials, J. Chem. Theory Comput. 19 (2023) 4568–4583, http://dx.doi.org/10.1021/acs.jctc.2c00798.

[28] Y.J. Colón, D.A. Gómez-Gualdrón, R.Q. Snurr, Topologically Guided, Automated Construction of Metal-Organic Frameworks and Their Evaluation for Energy-Related Applications, Cryst. Growth Des. 17 (2017) 5801–5810, http://dx.doi.org/10.1021/acs.cgd.7b00848.

[29] Y.G. Chung, E. Haldoupis, B.J. Bucior, M. Haranczyk, S. Lee, H. Zhang, K.D. Vogiatzis, M. Milisavljevic, S. Ling, J.S. Camp, B. Slater, J.I. Siepmann, D.S. Sholl, R.Q. Snurr, Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: CoRE MOF 2019, J. Chem. Eng. Data 64 (2019) 5985– 5998, http://dx.doi.org/10.1021/Acs.Jced.9b00835.

[30] B.J. Bucior, N.S. Bobbitt, T. Islamoglu, S. Goswami, A. Gopalan, T. Yildirim, O.K. Farha, N. Bagheri, R.Q. Snurr, Energy-based descriptors to rapidly predict hydrogen storage in metal–organic frameworks, Mol. Syst. Des. Eng. 4 (2019) 162–174, http://dx.doi.org/10.1039/C8ME00050F.

[31] H.A. Lorentz, Ueber die Anwendung des Satzes vom Virial in der kinetischen Theorie der Gase, Ann. Phys., Lpz. 248 (1881) 127–136, http://dx.doi.org/10.1002/andp.18812480110.

[32] D. Berthelot, Sur le mélange des gaz, Compt. Rendus 126 (1898) 1703–1706.

[33] G.S. Fanourgakis, K. Gkagkas, E. Tylianakis, G. Froudakis, A Generic Machine Learning Algorithm for the Prediction of Gas Adsorption in Nanoporous Materials, J. Phys. Chem. C 124 (2020) 7117–7126, http://dx.doi.org/10.1021/acs.jpcc.9b10766.

[34] G.S. Fanourgakis, K. Gkagkas, E. Tylianakis, E. Klontzas, G. Froudakis, A Robust Machine Learning Algorithm for the Prediction of Methane Adsorption in Nanoporous Materials, J. Phys. Chem. A 123 (2019) 6080–6087, http://dx.doi.org/10.1021/acs.jpca.9b03290.

[35] Y.G. Chung, J. Camp, M. Haranczyk, B.J. Sikora, W. Bury, V. Krungleviciute, T. Yildirim, O.K. Farha, D.S. Sholl, R.Q. Snurr, Computation-ready, experimental metal–organic frameworks: A tool to enable high-throughput screening of nanoporous crystals, Chem. Mater. 26 (2014) 1–10, http://dx.doi.org/10.1021/cm502594j.

[36] D. Ongari, P.G. Boyd, S. Barthel, M. Witman, M. Haranczyk, B. Smit, Accurate Characterization of the Pore Volume in Microporous Crystalline Materials, Langmuir 33 (2017) 14529–14538, http://dx.doi.org/10.1021/acs.langmuir.7b01682.

[37] S.C. Gupta, V.K. Kapoor, Fundamentals of Mathematical Statistics, Sultan Chand & Sons, 2020.

[38] L. Sarkisov, A. Harrison, Computational structure characterisation tools in application to ordered and disordered porous materials, Mol. Simul. 37 (2011) 1248–1257, http://dx.doi.org/10.1080/08927022.2011.592832.

[39] L. Breiman, Random Forests, Mach. Learn. 45 (2001) 5–32, http://dx.doi.org/10.1023/A:1010933404324.

[40] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach Learn. 63 (2006) 3–42, http://dx.doi.org/10.1007/s10994-006-6226-1.

[41] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[42] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco California USA,2016, pp. 785–794, http://dx.doi.org/10.1145/2939672.2939785.

[43] G. Shan, Monte Carlo cross-validation for a study with binary outcome and limited sample size, BMC Med. Inform. Decis. Mak. 22 (2022) 270, http://dx.doi.org/10.1186/s12911-022-02016-z.

[44] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc. Red Hook, NY, USA,2017, pp. 4768–4777.

[45] D. Dubbeldam, S. Calero, D.E. Ellis, R.Q. Snurr, RASPA: Molecular simulation software for adsorption and diffusion in flexible nanoporous materials, Mol. Simul. 42 (2016) 81–101, http://dx.doi.org/10.1080/08927022.2015.1010082.

[46] T.F. Willems, C.H. Rycroft, M. Kazi, J.C. Meza, M. Haranczyk, Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials, Microporous Mesoporous Mater. 149 (2012) 134–141, http://dx.doi.org/10.1016/j.micromeso.2011.08.020.

[47] A.K. Rappe, C.J. Casewit, K.S. Colwell, W.A. Goddard, W.M. Skiff, UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations, J. Am. Chem. Soc. 114 (1992) 10024–10035, http://dx.doi.org/10.1021/ja00051a040.

[48] M.G. Martin, J.I. Siepmann, Transferable Potentials for Phase Equilibria. 1. United-Atom Description of N-Alkanes, J. Phys. Chem. B 102 (1998) 2569–2577.

[49] L. van der Maaten, G. Hinton, Visualizing Data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.

[50] L. McInnes, J. Healy, J. Melville, UMAP: uniform manifold approximation and Projection for dimension reduction, 2020, http://dx.doi.org/10.48550/arXiv.1802.03426.

[51] G.S. Fanourgakis, K. Gkagkas, G. Froudakis, Introducing artificial MOFs for improved machine learning predictions: Identification of top-performing materials for methane storage, J. Chem. Phys. 156 (2022) 054103, http://dx.doi.org/10.1063/5.0075994.

[52] A.P. Sarikas, G.S. Fanourgakis, E. Tylianakis, K. Gkagkas, G.E. Froudakis, Comparison of Energy-Based Machine Learning Descriptors for Gas Adsorption, J. Phys. Chem. C 127 (2023) 20995–21005, http://dx.doi.org/10.1021/acs.jpcc.3c04223.

[53] Y. Kang, H. Park, B. Smit, J. Kim, A multi-modal pre-training transformer for universal transfer learning in metal–organic frameworks, Nat. Mach. Intell. 5 (2023) 309–318, http://dx.doi.org/10.1038/s42256-023-00628-2.

[54] K. Choudhary, T. Yildirim, D.W. Siderius, A.G. Kusne, A. McDannald, D.L. Ortiz-Montalvo, Graph neural network predictions of metal organic framework CO2 adsorption properties, Comput. Mater. Sci. 210 (2022) 111388, http://dx.doi.org/10.1016/j.commatsci.2022.111388.

[55] A.P. Sarikas, K. Gkagkas, G.E. Froudakis, Gas adsorption meets deep learning: Voxelizing the potential energy surface of metal–organic frameworks, Sci. Rep. 14 (2024a) 2242, http://dx.doi.org/10.1038/s41598-023-50309-8.

[56] A.P. Sarikas, K. Gkagkas, G.E. Froudakis, Gas adsorption meets geometric deep learning: Points, set and match, Sci. Rep. 14 (2024b) 27360, http://dx.doi.org/10.1038/s41598-024-76319-8.