Informative Training Data for Efficient Property Prediction in Metal-Organic Frameworks by Active Learning

Ashna Jose, Emilie Devijver, Noel Jakse, Roberta Poloni

Informative Training Data for Efficient Property Prediction in Metal-Organic Frameworks by Active Learning

Ashna Jose,[†] Emilie Devijver,[‡] Noel Jakse,[*,†] and Roberta Poloni[*,†]

†*SIMaP, Grenoble-INP, CNRS, University of Grenoble Alpes, 38042 Grenoble, France*

‡*LiG, Grenoble-INP, CNRS, University of Grenoble Alpes, 38042 Grenoble, France*

Received November 18, 2024; E-mail: noel.jakse@grenoble-inp.fr; roberta.poloni@grenoble-inp.fr

Abstract: In recent data-driven approaches to materials discovery, scenarios where target quantities are expensive to compute or measure are often overlooked. In such cases, it becomes imperative to construct a training set that includes the most diverse, representative, and informative samples. Here, a novel regression tree-based active learning algorithm is employed for such a purpose. It is applied to predict band gap and adsorption properties of metal-organic frameworks (MOFs), a novel class of materials that results from the virtually infinite combinations of their building units. Simpler and low dimensional descriptors, such as those based on stoichiometric and geometric properties, are used to compute the feature space for this model owing to their ability to better represent MOFs in the low data regime. The partitions given by a regression tree constructed on the labeled part of the dataset are used to select new samples to be added to the training set, thereby limiting its size while maximizing the prediction quality. Tests on the QMOF, hMOF, and dMOF data sets, reveal that our method constructs *small* training data sets to learn regression models that predict the target properties more efficiently than existing active learning approaches, and with lower variance. Specifically, our active learning approach is highly beneficial when labels are unevenly distributed in the descriptor space and when the label distribution is imbalanced, which is often the case for real world data. The regions defined by the tree helps revealing patterns in the data, thereby offering a unique tool to efficiently analyze complex structure-property relationships in materials and accelerate materials discovery.

Introduction

Metal-organic frameworks (MOFs),[1,2] formed through coordination bonds between metal ions and organic ligands, are promising materials for efficient gas capture and separation,[3,4] due to their ultrahigh porosity, chemical tunability and large surface area.[5,6] Recently, they have been shown to be potential candidate materials also for energy storage,[7–9] water harvesting,[10,11] catalysis,[12,13] and sensing,[14] thus evoking an interest in the electronic properties of MOFs.[9,15–19] Remarkably, a large variety of properties[20,21] are expected in MOFs as a consequence of these materials resulting from the virtually infinite combinations of their building units. As such, the identification and/or discovery of novel MOFs with specific properties becomes challenging.

To assist in this endeavor, computational techniques such as molecular simulations and density-functional theory[22–28] were used to screen large MOF datasets. Alternatively, machine learning (ML) approaches were exploited to further accelerate MOFs discovery.[29–36] Based on a training sample, a descriptor-based ML model is learned, for *e.g.* kernel ridge regression, random forests, or gradient boosting regression trees,[29,36–41] to predict properties such as electronic and gas adsorption properties of unseen samples. Recently, deep learning methods such as crystal graph convolutional neural networks (CGCNN[42,43]) and transformer-based models[35,44,45] were also investigated. Despite being powerful and well-suited for large and complex data, deep-learning methods require a substantial amount of labeled data and computational resources to train a complex model. They also require accurate hyperparameter optimizations and sometimes pre-training,[35,44,46] which is not feasible when few data are labeled.

In this work, we adopt an opposite strategy to MOFs discovery: we focus on situations where properties are expensive to obtain and therefore large labeled datasets are not available.[47] This calls for a need to optimize the training set. Active learning (AL) algorithms are a class of ML methods that aim at constructing the most informative, diverse and representative training set iteratively. They use the knowledge of the samples labeled in each iteration and select high quality samples, thus avoiding labeling redundant samples as it may occur in random sampling. Many AL algorithms currently used are model-free, *i.e.* they select new samples based solely on diversity[48] and/or good representation[49] of the input space. Their main drawback is that the training sets are not sufficiently diverse in the target space. Intuitively, adding knowledge of the targets assists in understanding its conditional distribution with the features, which ensures selection of better samples. Model-based AL schemes,[48,50,51] such as Query By Committee[52] and active learning using Gaussian processes,[53–57] accomplish this by defining an acquisition criterion based on the knowledge of the samples already labeled. Gaussian processes for instance is a popular choice, due to its interpretabilty and ability to estimate uncertainty in the predictions. While these methods usually construct training sets better than model-free approaches, they exhibit high computational complexity, require precise hyperparameter optimisation and are not transferable to other ML models. Therefore, it becomes imperative to head towards universal and transferable AL methods that reduce labeling costs for any given input description, which is the scope of this work.

We rely on a novel tree-based AL algorithm developed by the present authors.[58] This approach, named Regression Tree-based Active Learning (RT-AL[58]), uses the knowledge of the input and the output smartly to iteratively add the most diverse, representative, and informative samples to the training set. It has proven to be more efficient than other existing AL approaches, and is transferable to different ML models. We show that it efficiently predicts very different

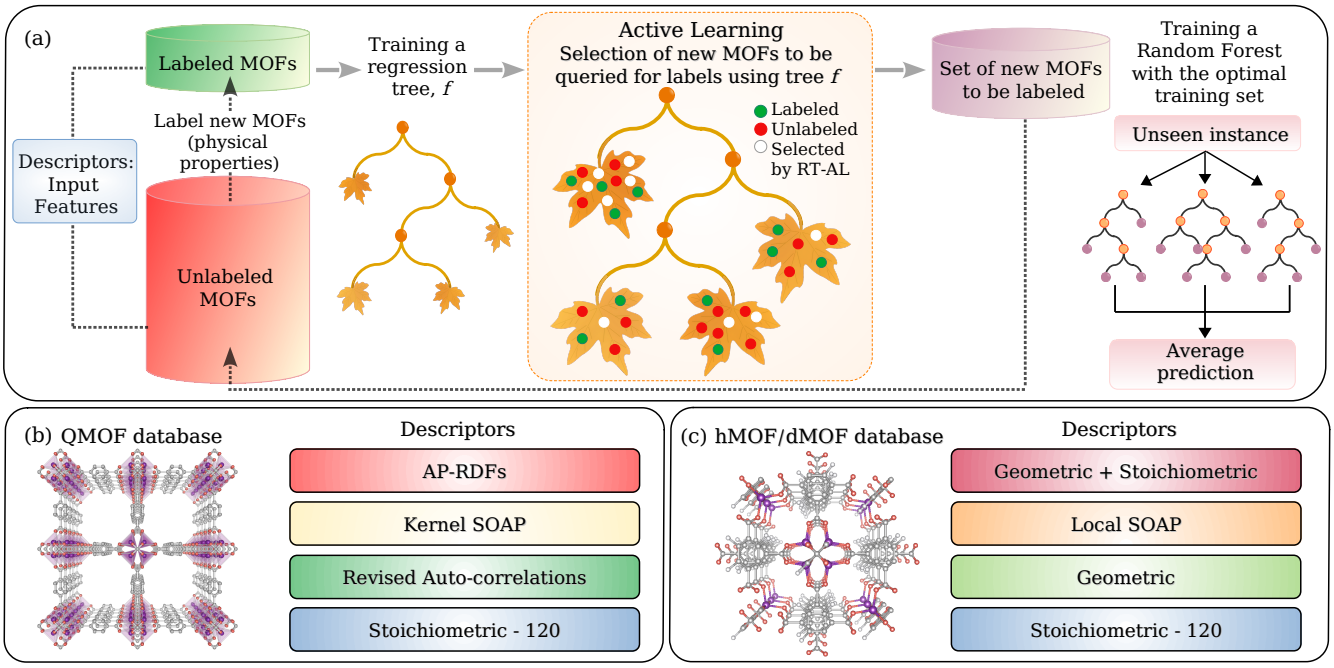


Figure 1. (a) Schematic representation of the workflow of active learning using RT-AL. New samples are added iteratively to the training set using the leaves of the regression tree to form the most informative, diverse and representative set. A random forest is finally trained on the optimal training set. The set of different descriptors studied in this work for the (b) QMOF and (c) hMOF/dMOF database.

properties of MOFs, such as the band gap and gas adsorption, for any given set of descriptors, and with very low variance. We also show that RT-AL succeeds to construct more informative training sets for MOFs compared to other model-free and model-based AL approaches. Through the regions defined by the regression tree, RT-AL is able to identify meaningful patterns in the data. Finally, we show that RT-AL is highly beneficial for label distributions that are peaked or multimodal, which is usually the case for real world data. A schematic representation of the workflow is shown in Figure 1(a).

Besides the integrated approaches provided by deep learning, the description of MOFs is at the core of this AL technology. There is a continued need to explore efficient encoding strategies so as to maximize the predictive power of descriptor-based ML approaches. Although many local and global descriptors[41] were used to represent MOFs previously, their performance in the low data regime has not been tested. We investigate various structural, geometric and stoichiometric descriptors of different dimensions for band gap and gas adsorption properties. A summary of the descriptors used in this work is presented in Figure 1(b) and (c). We find that descriptors apt for large training set sizes are not suitable when less labeled data is available. When data is scarce, simpler and low dimensional stoichiometric descriptors lead to better models for band gap prediction, and a combination of geometric and stoichiometric features train better models for gas adsorption.

Method

Tree-based active learning

Our recently proposed method based on regression trees (RT-AL)[58] is a novel model-based AL algorithm for regression. Regression trees partition the input feature space into a set of K hyper-rectangles, called regions and denoted $\mathcal{R}_k = \prod_{\ell=1}^{p}[a_{k,\ell}, b_{k,\ell}]$ for $1 \leq k \leq K$, and assign a common weight $\gamma_k \in \mathbb{R}$ to each region k:

$$f(\mathbf{x};\Theta) = \sum_{k=1}^{K} \gamma_k \mathbf{1}_{\{\mathbf{x}\in\mathcal{R}_k\}}.$$

The set of parameters $\Theta = ((\mathcal{R}_k, \gamma_k)_{1\leq k\leq K})$ correspond to the set of regions and the associated weights. They are estimated using a labeled set such that the weights minimize the quadratic loss for fixed regions, leading to the empirical mean of observations in each region. The regions are constructed recursively by finding the feature and splitting point to divide a current region into two in such a way that the variance in the prediction is minimized. This can be represented as a tree, in which each node determines the feature to split and its corresponding value, and the final prediction is given by the leaves of the tree.

In our approach, the first few samples, n_{init}, are randomly chosen from a full data set (where all samples are unlabeled) and are labeled, forming the set I_{init}. This is followed by training a decision tree for regression (referred to as regression tree) with K leaves using the labeled set and is used to predict the labels, $(\hat{Y}_i^{I_{\text{init}}})_{i\notin I_{\text{init}}}$, for every unlabeled sample. In the active part of the algorithm, the leaves of the tree are used to add more samples to the training set. Conditionally to the first labeled set, the number of samples to be labeled from each leaf k, n_k^*, are distributed into the different leaves as:

$$n_k^* = n_{\text{act}} \frac{\sqrt{\pi_k \hat{\sigma}_k^2}}{\sum_{\ell=1}^{K} \sqrt{\pi_\ell \hat{\sigma}_\ell^2}}; \quad (1)$$

where n_{act} are the total number of samples to be selected by AL, $\hat{\sigma}_k^2$ denotes the variance computed on the true labels of the labeled samples in leaf k, and π_k the probability that an unlabeled sample $\mathbf{x}_i$ belongs to leaf k, defined formally

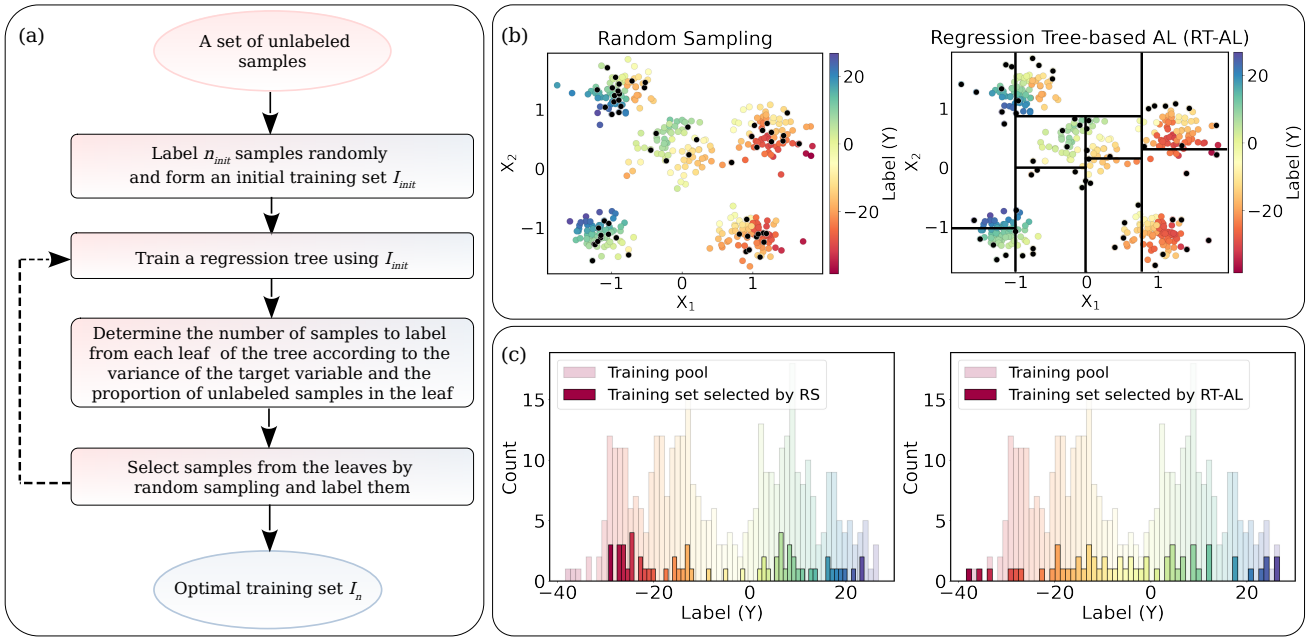


Figure 2. (a) Flowchart representing Regression Tree-based Active Learning (RT-AL), with the acquisition criterion described in detail. (b) Comparison between the training sets (shown as black circles) constructed by Random Sampling (RS) (left) and RT-AL (right) from a generated dataset with 2 features and 500 samples. The colors represent the true values of labels. The black lines correspond to the different regions the regression tree partitions the feature-label space into. (c) Label distributions of the training set constructed by RT-AL and RS, compared to the label distribution of the available training pool.

as follows: for $1 \leq k \leq K$,

$$\hat{\sigma}_k^2 = \frac{\sum_{i \in I_{\text{init}}: \mathbf{x}_i \in \mathcal{R}_k} (\hat{Y}_i^{I_{\text{init}}} - Y_i)^2}{|i \in I_{\text{init}} : \mathbf{x}_i \in \mathcal{R}_k| - 1}, \tag{2}$$

$$\pi_k = \frac{|i \notin I_{\text{init}} : \mathbf{x}_i \in \mathcal{R}_k|}{N}. \tag{3}$$

In other words, the number of samples to be labeled from each leaf depends on (i) the variance computed on the true labels, and (ii) the proportion of unlabeled samples in the leaf. Since the leaves of a decision tree correspond to homogeneous regions in the feature-target space, this acquisition strategy can be seen as a trade-off to select samples diverse in the target but representative of the feature space, thus taking into account maximum possible information. This is an essential point that is missed in most active learning algorithms.

After computing n_k^*, the samples are selected from each leaf using random sampling. Once the new samples are labeled, the tree can be retrained. This routine can then be repeated by adding few samples at each step, until the desired size of the training set or a targeted accuracy is reached. The algorithm is described as a flowchart in Figure 2 (a). Figure 2 (b) illustrates the advantage of RT-AL over random sampling on a generated data set of 2 input features and 500 samples. The colors depict values of the target quantity (labels). The data set is randomly divided into a training pool and a held out test set in the ratio 8:2: 80% being the pool from which the training set is to be chosen, and the remaining 20% being the test set, used to determine the model performance. The training set, shown as black circles, selected from the training pool by random sampling are not diverse or representative of the feature or the target space, which would lead to inefficient predictions when large numbers of labels cannot be afforded (Figure 2 (b) left). Our method, on the other hand, takes into account the response through the regression tree, resulting in a diverse set of samples. This can be seen from the samples that are now spread in the different regions, i.e. leaves, shown as partitions in Figure 2 (b) (right). This is quantified in Figure 2 (c) that shows the label distribution of the full training set and the ones selected by RS (left) and RT-AL (right). While the samples selected by RS disregard regions of the label space where samples are fewer, RT-AL samples evenly from all regions, ensuring a good representation of the target space. This leads to a better performance of RT-AL over RS on the held-out test set (see Figure S1 for more details).

Data sets and technical details

Three publicly available MOF datasets are used in this study, namely the Quantum MOF (QMOF[43]), the hypothetical MOF (hMOF[59]) and the diverse MOF (dMOF[60]) datasets. The version of the QMOF database used for this work[43] consists of 14,482 experimentally synthesized MOFs, with optimised structures and electronic band gaps computed at the PBE-D3(BJ) level using density functional theory (DFT). The hMOF database consists of 137,652 hypothetical MOFs, with data of CO_2 and CH_4 adsorption at 0.05, 0.5, and 2.5 bar pressure obtained using grand canonical Monte Carlo (GCMC) simulations. The dMOF database consists of $\approx$ 20,000 hypothetical MOFs, with H_2 adsorption data at 100 bar and 77 K, also obtained using GCMC simulations. These databases are selected as they contain very different properties of MOFs, such as gas adsorption, which mostly depends on the structure and the pore geometry, and band gap for which structure and chemistry are more relevant.

Before selecting the MOFs to be added to the training set for each of these 3 databases, the full data set is split in the ratio 8:2 randomly, as described before. To construct the training set using RT-AL, the first 20 MOFs are chosen

randomly in the training set, and the initial regression tree is trained using Scikit-learn.[61] The depth of the tree is controlled through the hyperparameter that sets the minimum samples in a leaf. It is set to 5, as suggested in our previous work,[58] keeping it high enough to avoid over-fitting and to get meaningful variance among the labeled samples, but sufficiently low for the tree to give accurate predictions (hyperparameter optimisation is shown in Figure S2). This is followed by iterative additions of MOFs to the training set using RT-AL until approximately 10% of the total available training pool is labeled. At each iteration, a random forest (RF) of 50 regression trees (results with different number of trees are shown in Figure S3) is trained using the training set at the given iteration and its performance is measured by making predictions on the held out test set and computing the Mean Absolute Error (MAE) given by:

$$\mathrm{MAE} = \frac{1}{T}\sum_{i=1}^{T} |(f(\mathbf{x}_i) - y_i)|, \tag{4}$$

where T and y_i are the size and true labels of the test set, and $f(\mathbf{x}_i)$ is the prediction for a test sample $\mathbf{x}_i$ using the RF. Note that RF is used as the final predictor, but other ensemble tree-based methods like Gradient Boosting Regression Tree (GBRT) and XGBoost,[62] as well as other ML approaches such as deep learning models can also be trained.[58] Tree-based models are known to be highly accurate when training data is scarce, which is why they are used here. RFs were also trained on samples selected using random sampling and other AL approaches described hereafter, to determine the degree of improvement of our algorithm over them.

Descriptors

In order to find the best representation for band gaps and gas adsorption in the low data regime, we set up various descriptors, that are listed below. All features (except for the geometric descriptors) are computed in this work, and the technical details and parameters used to compute these are reported in S3.

Stoichiometric-120 (ST-120): The stoichiometric descriptors (ST-120),[63] which consists of 103 features specifying elemental fractions, and 17 statistical attributes of elemental properties were computed using Pymatgen and Matminer.[64] The attributes consist of averages and ranges of atomic properties: mass, number, radii, electronegativity, group and period numbers, along with fractional and average s, p, d, f electron information.

Smooth Overlap of Atomic Positions (SOAP): SOAP[65] encodes regions of atomic geometries by using a local expansion of a Gaussian-smeared atomic density with orthonormal functions based on spherical harmonics and radial basis functions. Although commonly used as a local representation, it can be used as a global descriptor by computing a similarity kernel[66] that estimates similarity between pairs of local atomic environments among different structures. For the QMOF database, SOAP kernel is used because local SOAP features scale quickly with the number of types of chemical species, which is large for QMOF dataset. Since our focus is on the low data regime, high dimensional features are not ideal. For the hMOF and dMOF databases, local SOAP features are computed as the number of features are only moderately high (2772 and 4284, respectively).

Atomic Property Weighted Radial Distribution Functions (AP-RDFs): AP-RDFs[67] describe MOFs by the weighted probability distribution of finding an atom pair in a spherical volume of radius R inside the unit cell. The atomic properties used to weigh the RDFs are atomic mass, number, group number and period number, which form a set of 164 features (see Figure S4 for more details).

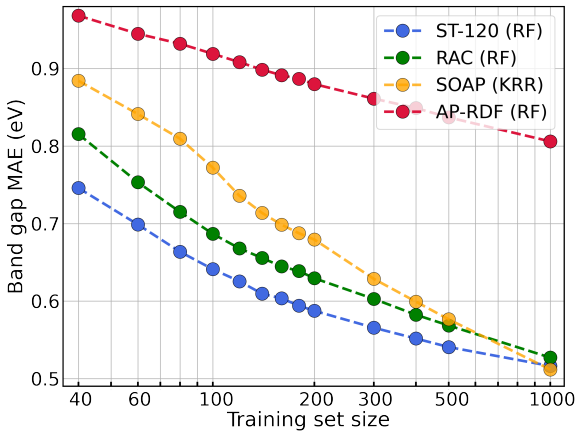


Figure 3. QMOF database: Mean absolute error (MAE) for band gap prediction on the test set as a function of training set size using a random forest. RFs are trained using descriptors ST-120, RACs and AP-RDFs, while a KRR model is trained using kernel SOAP. The sub-samples are selected using random sampling and each point is an average over 40 runs with different seeds for the train-test split.

Revised Auto-Correlations (RACs): RACs[37] are products and differences of heuristic atomic properties on graphs. They have been shown to be valuable descriptors for transition metal complexes[37] as well as for MOFs.[68] Metal-centered, linker and functional-group descriptors are generated using molSimplify,[69] weighted by atomic properties that include atom identity, connectivity, Pauling electronegativity, covalent radii, nuclear charge and polarisability. Averaging over all atoms in each MOF produces 156 features.

Geometric descriptors: A set geometric properties of the MOFs is curated for the hMOF and dMOF database, with features like Pore Limiting Diameter (PLD), Largest Cavity Diameter (LCD), void fraction, gravimetric and volumetric surface area etc (refer to Table S1 for the full list). These features (referred to as Geom-5 for the hMOF, and Geom-15 for the dMOF dataset) are cheap to compute and are known to directly impact the adsorption properties of transition metal complexes.[29,39,70]

Geometric + Stoichiometric (GS): GS is a combination of the non-zero features from ST-120 and the geometric features described above. For the hMOF database, three sets of GS descriptors are computed: Geom-5 combined with (i) element fractions of carbon, nitrogen and oxygen (these elements are present in most MOFs), named GS1; (ii) all the non-zero element fractions, named GS2, and with (iii) all non-zero features of ST-120 (both the element fractions and the statistical attributes), named GS3.

Alternative active learning approaches

As introduced in the first section, active learning methods can be grouped into model-free and model-based approaches. Model-free methods select new samples based on feature-

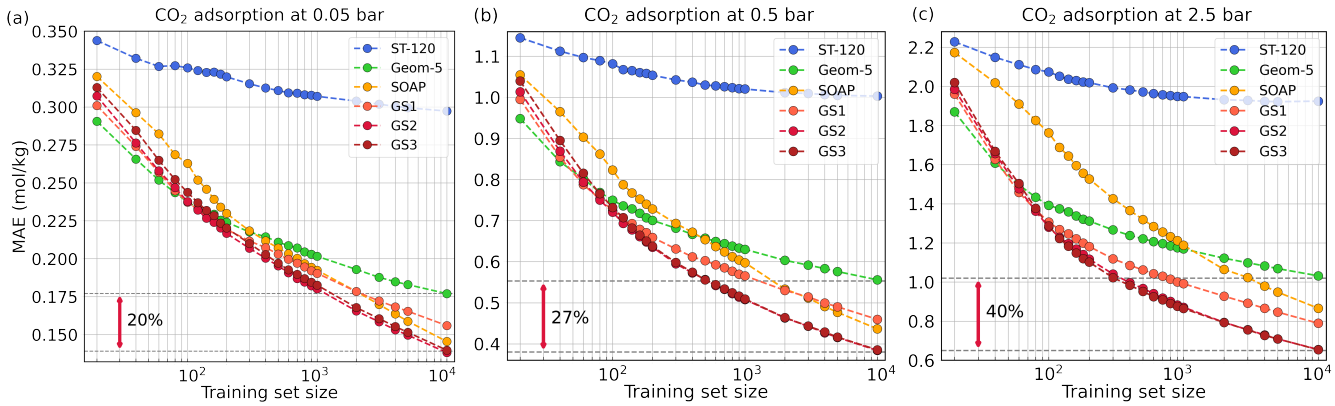


Figure 4. hMOF database: MAE for CO_2 adsorption prediction at (a) 0.05, (b) 0.5 and (c) 2.5 bar on the test set as a function of training set size using a random forest. RFs are trained using descriptors ST-120, SOAP, Geom-5, GS1, GS2 and GS3. The sub-samples are selected using random sampling and each point is an average over 40 runs with different seeds for the train-test split. The horizontal dotted line compares the labeling cost for different descriptors for a fixed accuracy. Percentage improvement of GS3 with respect to Geom-5 are given and indicated by a red double-ended arrow for each gas pressure.

space diversity[48] and representativity,[49] while model-based methods[52,53] use an initial model trained on a small set of labeled samples to increase the size of the training set. We recall that RT-AL was shown to be the best performer for all datasets studied in our previous work.[58] To validate the transferability of RT-AL to MOF databases, the performance of RT-AL is compared here with AL methods that were found to be the most competitive,[58] and those that are commonly used in materials informatics:

Greedy Sampling (GSx[48]): it is a model-free AL method that selects the sample closest to the centroid of the feature space as the first one in the training set, followed by the one farthest from it, based on L_2 distance of the descriptor vector. The next samples to be labeled are those farthest from all samples that have been previously selected, to ensure diversity in the feature space.

Iterative Representativeness Diversity Maximization (iRDM[49]): it is a model-free AL method that uses k-means clustering to partition the feature space into a number of clusters equal to the number of samples to be labeled. It subsequently selects the samples closest to the centroids of these clusters as the starting points, and over the course of the algorithm, combines it with the basic idea of feature space diversity from GSx[48] to update the centroids to samples which are representative and diverse.

Variance-based Query By Committee (QBC[52]): it is a model-based AL method that selects the samples with the highest variance among the predictions from a committee of models (models used here are decision trees). The committee is constructed by bootstrapping on an initial set of randomly labeled samples.

Gaussian Process Regression (GP[53]): it is a model-based AL method that uses uncertainty (given by standard deviation) or relative uncertainty (ratio of standard deviation and the prediction) in the predictions given by GPs trained on the labeled part of the dataset, and subsequently adds the most uncertain sample to the training set.

The hyperparameters and more details on these methods can be found in SI.

Results

Descriptors for low data regime

QMOF database

For the QMOF database, we compute the ST-120, RACs, AP-RDFs and kernel SOAP descriptors, and RF models were trained using the first three. For SOAP kernel, a Kernel Ridge Regression (KRR) model performs better than RF (see Figure S5), as the descriptor itself is based on a similarity kernel. Samples were selected using random sampling. Note that only 11,799 MOFs are featurizable using RACs features. Yet, as the distribution of the target quantity remains the same for both the full QMOF (14,482) and this subset (see Figure S6), it is reasonable to compare the performance of the model trained using RACs to the others used in this work.

Figure 3 shows the MAE for band gap predictions on the test set as a function of training set size for each descriptor, averaged over 40 runs. For training set sizes up to 1,000, ST-120 is the best descriptor. RACs, that also encode local atomic properties of MOFs, perform only slightly worse than ST-120. As already pointed out,[43] the performance of the model based on kernel SOAP is very poor in the low data regime, but it learns much faster than those based on ST-120 and RACs, making this descriptor informative when large training data is available. This is due to the fact that kernel SOAP is a complex and global descriptor, thus mostly advantageous for large sets of labeled MOFs. ST-120, on the other hand, is a simpler descriptor (fewer number of features) and performs significantly better than kernel SOAP for small training set sizes. Interestingly, AP-RDFs perform the worst among this set of representations, for both low data and the full training set, while the other descriptors reach similar prediction accuracy at 1,000 samples.

hMOF database

For the hMOF database, we compute ST-120, Geom-5, local SOAP, and the three sets of GS descriptors. Note that RACs were not computed for this data set as they were found to give models with low accuracy[35] for the prediction of gas adsorption. RF models were trained using RS to predict CO_2

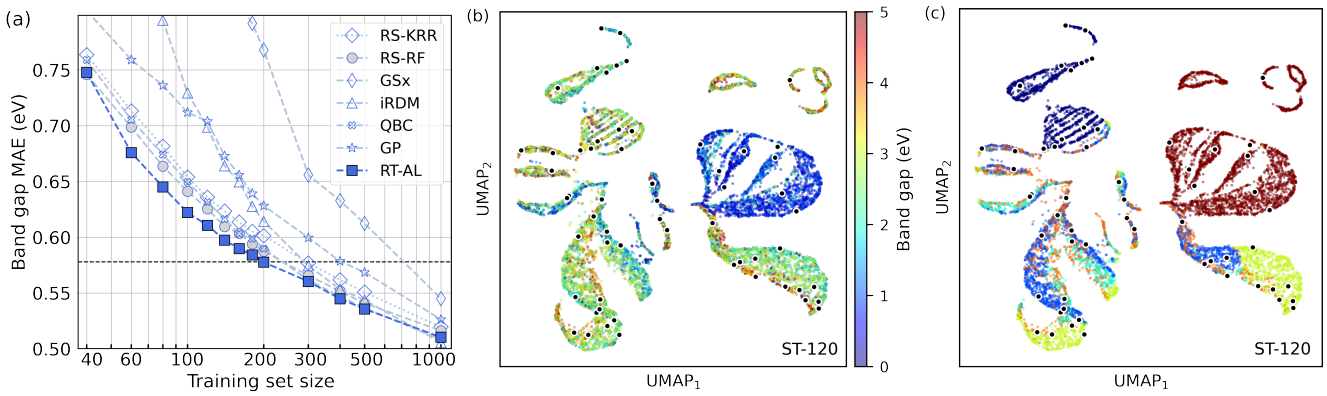


Figure 5. (a) MAE for predicting band gaps on the test set as a function of training set size for ST-120 descriptor, using random sampling with KRR (RS-KRR) and RF (RS-RF), and the active learning methods GSx, iRDM, QBC, GP and RT-AL with RF. Each point is an average over 40 runs with different seeds for the train-test split. The horizontal dotted line is a guide to the eye to compare the reduction in labeling cost for RT-AL over other sampling methods. (b) Dimensionality reduction of the training pool of the QMOF data set performed using UMAP, with a distance matrix obtained from the ST-120 feature-set of MOFs in the data set with color code for band gap. (c) Same as (b) with different colors representing the regions the regression tree partitions the descriptor space into (leaves). 60 samples selected by RT-AL are shown as black circles in (b) and (c).

and CH_4 adsorption at 0.05, 0.5, and 2.5 bar for training set sizes starting from as low as 20, until approximately 10% of the available training pool (10,000 here). The training curves for CO_2 adsorption are shown in Figure 4 (for CH_4 these are provided in S6.1). Contrary to the QMOF case, RFs trained using ST-120 lead to high values of MAE, consistent with previous studies.[35] Purely geometric descriptors (Geom-5) perform much better, which can be understood from the fact that they represent a key factor driving the adsorption process. For very low data, Geom-5 performs the best, owing to its low dimensions. Adding the non-zero features from ST-120 (GS descriptors) to it enhances the prediction quality. As more training data becomes available, the GS feature sets lead to better models. Local SOAP features give high values of MAE for very small training set sizes, as expected, due to its substantially larger dimension. For larger training data, it outperforms Geom-5, but fails to outdo GS2/GS3. This is likely due to the fact that the fine local structural details here computed from local SOAP features are not sufficient for predicting adsorption properties in the low data regime. Finally, the best descriptor for predicting CO_2 adsorption at both low and high pressures is GS3.

It is interesting to analyze in detail the effect of the descriptors for adsorption at different gas pressures. Specifically, we compare the percentage of improvement in the performance when using GS3 instead of Geom-5, at 10,000 training set size (see Figure 4). This improvement progressively increases for increasing gas pressure, and ranges from ≈20% at 0.05 bar (Figure 4a) to ≈40% at 2.5 bar (Figure 4c). Because GS2 and GS3 give essentially the same performance, we can say that the element fractions (which are also present in GS1, but only partially) are the dominant factor in the improvement over Geom-5. These elemental attributes therefore provide a larger enhancement in the prediction performance at high gas pressure where geometry is assumed to be more relevant for adsorption. This is consistent with (i) Geom-5 features yielding a progressively better description than SOAP features for increasing pressure, and (ii) the trend observed on comparing SOAP and GS3: the detailed structural description provided by the former is more relevant for adsorption at low gas pressure. These results also indicate that the importance of stoichiometric aspects can not be ruled out even at high pressure. A similar analysis of the descriptors for predicting H_2 adsorption at 100 bar and 77 K in the dMOF database is shown in Figure S16.

Active Learning

QMOF: Band gap prediction

The performance of our active learning method, RT-AL, is assessed using the ST-120 descriptor. The MAE for band gap predictions on the held out test set as a function of training set size is reported in Figure 5 (a) for RT-AL and the active learning approaches GSx, iRDM, QBC, and GP, introduced previously. RT-AL is the best performer for all training set sizes. Importantly, the model-based methods, QBC and GPs, give a higher MAE than RT-AL. The discontinuity in the learning curve of GP at around 120 occurs due to the limitation of Gaussian processes to train when the feature size (120 in this case) is greater than the training set size, requiring different parameter settings and optimisation. GSx and iRDM perform rather poorly for low training set sizes, confirming that feature space information alone is not sufficient for informative sampling from the QMOF dataset. For comparison, the performance of random sampling using Kernel Ridge Regression (RS-KRR) and Random Forest (RS-RF) is also reported in Figure 5 (a) and shows a superior performance of the latter.

This result shows that through the use of informative and training sets, RT-AL samples better than other AL methods and random sampling. For instance, to achieve the performance level shown by the horizontal dotted line, we observe that the most commonly used methods, GSx and Gaussian processes require a much higher number of samples (700 and 400, respectively) to be labeled and included in the training set, while our model requires only 200. This significant decrease is extremely important in situations where labeling data is very expensive and saving on 200-500 labels (half to two third of the budget) could mean reducing substantially resources and time. Importantly, our method exhibits low variance which is crucial in active learning settings. The standard deviation of RT-AL for 100 training samples is 0.031 eV, while that for Gaussian process and QBC is 0.052

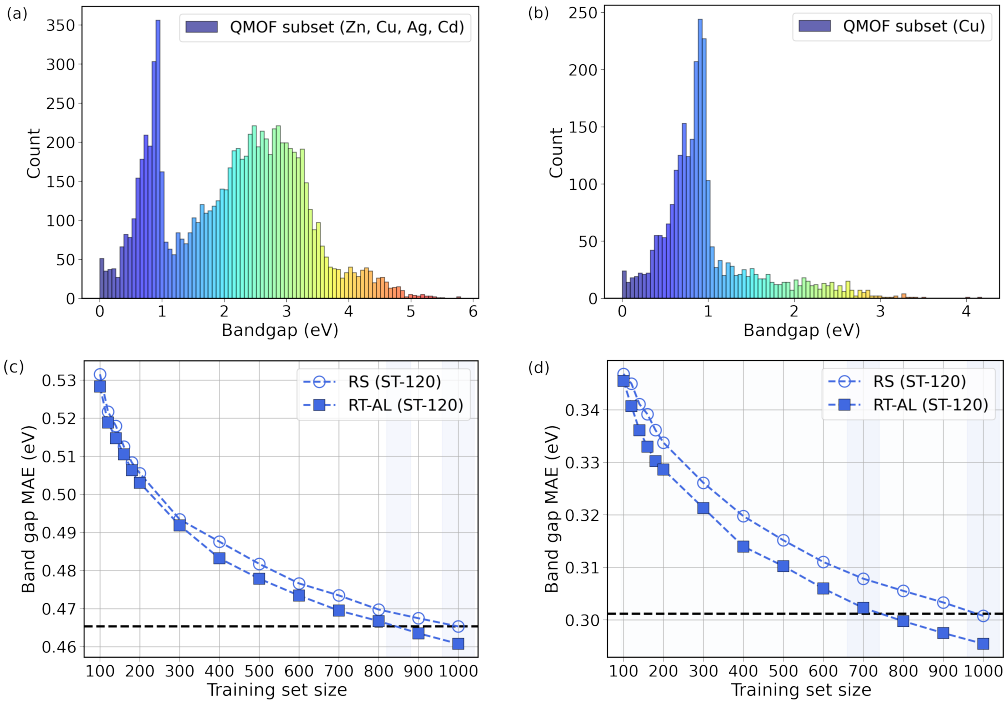


Figure 6. Histograms depicting distributions of band gap of two subsets of the QMOF data set, (a) $S_{Zn,Cu,Ag,Cd}$ and (b) S_{Cu}, showing a more balanced nature for $S_{Zn,Cu,Ag,Cd}$ and imbalanced for S_{Cu}. For these subsets, (c) and (d) show MAE for predicting band gaps on the respective test set as a function of training set size for ST-120 descriptor using random sampling and RT-AL, with Random Forest as the ML model. Each point is an average over 100 runs with different seeds for the train-test split. The horizontal dotted line is to compare the reduction in labeling cost by using RT-AL instead of RS for a fixed accuracy.

and 0.041 eV, respectively.

To further understand the reason behind the good performance of RT-AL, we compute an unsupervised structural dimensionality reduction performed using the Uniform Manifold Approximation and Projection (UMAP),[71] with a distance matrix obtained from the ST-120 feature-set of the QMOF training pool (80% of the data set). The result is reported in Figure 5(b) and (c) and the colors on the UMAP represent the values of band gaps and the leaves, respectively. Although some clusters in the UMAP space are carried forward to the target space, some others have a hint of all colors, as shown in Figure 5 (b). This implies that the data is not well clustered in the target space, and neither evenly distributed. In these figures, we also show 60 samples selected using RT-AL as black circles. RT-AL uses both the input and the target information through the structure of the regression tree and thus it selects MOFs from every region of the target space, and is eventually able to give better predictions for all band gap values. Importantly, RT-AL ensures to sample from all regions of the target space also for very small training sets. In addition, Figure 5 (c) shows that the samples selected by RT-AL are well distributed in the feature space, as well as among the leaves. The tree succeeds to find meaningful patterns in the data as shown by how the leaves are distributed in the UMAP. We further note that the leaves identify regions by grouping MOFs that are similar according to both specific chemical features and the target space, as illustrated in the tree structure reported in Figure S9.

RT-AL for imbalanced datasets

As seen in the previous section, taking the information of the labels into account is necessary, especially in the low data regime. To further stress on its significance, subsets of the QMOF database, which have different distributions of band gap, are constructed. Three subsets are created, based on the metal present in each structure: a subset of MOFs with at least one of Zn, Cu, Ag or Cd present, another of MOFs that contain Zn, and a third one that contains Cu. These subsets have 8,491, 2,410 and 2,587 MOFs respectively, and are hereafter referred to as $S_{Zn,Cu,Ag,Cd}$, S_{Zn} and S_{Cu}.

RF models are trained using ST-120 for each of these subsets. Figure 6(a) and (b) show histograms that correspond to distributions of band gaps for $S_{Zn,Cu,Ag,Cd}$ and S_{Cu} (S_{Zn} shown in Figure S10). Figure 6(c) shows the evolution of the MAE for band gap predictions using RS and RT-AL. RT-AL improves random sampling in both cases, but the degree of improvement differs. For the set $S_{Zn,Cu,Ag,Cd}$, the improvement of RT-AL is similar to that achieved on the full data set. This is because the label distribution of this subset and the full data set are very similar (see Figures S6 and S10). Because of the multimodal nature of the distribution, the probability that RS selects samples from regions with few data is low. RT-AL, on the other hand, does sample such regions through the structure of the regression tree. This improvement is further enhanced for the S_{Cu} subset. Here, the performance achieved by RS for 1,000 samples in the training set is achieved with only 700 samples by RT-AL. This is attributed to the sharp and therefore imbalanced distribution of band gaps in this subset. RS picks up large numbers of samples from this peak and fewer elsewhere. RT-AL, on the contrary, samples from different regions of the target space: as the band gap is more evenly distributed for $S_{Zn,Cu,Ag,Cd}$, the improvement of RT-AL over RS is lower here. This inference is of utmost importance as most real world data sets are imbalanced.

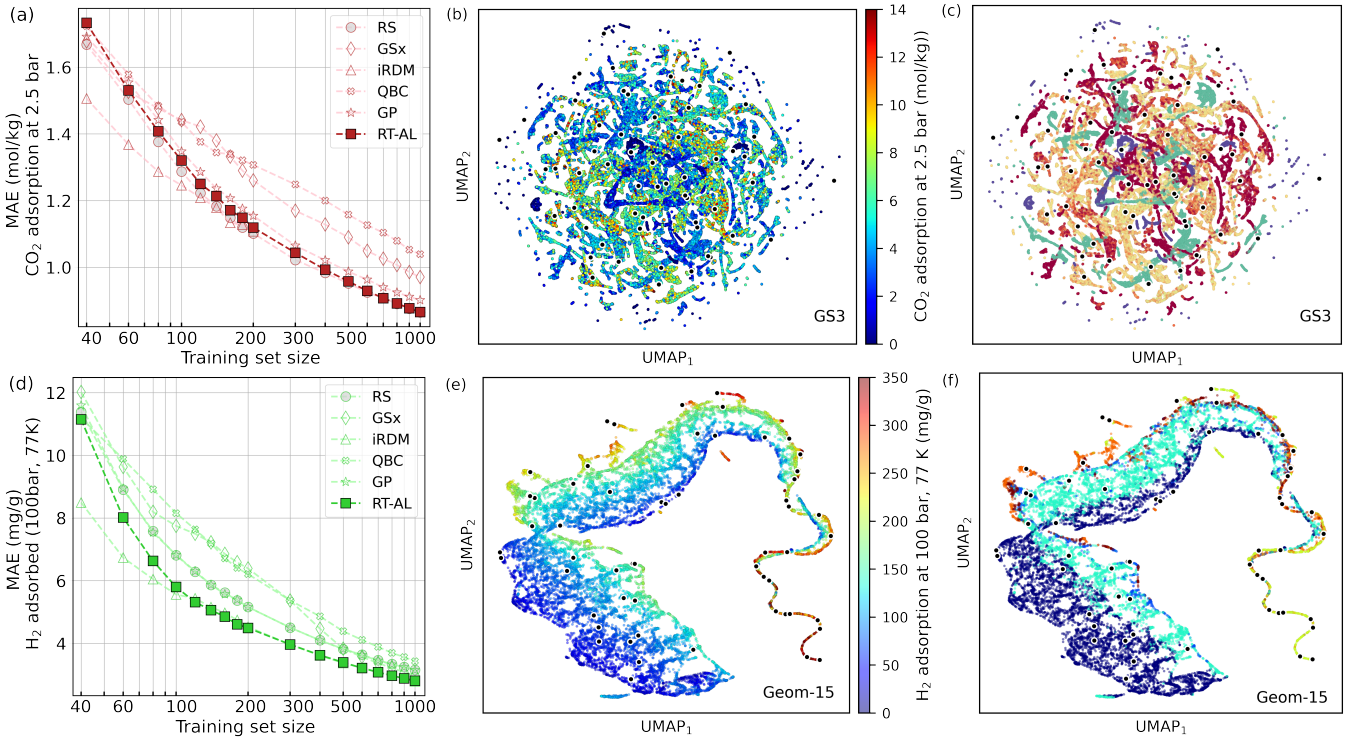


Figure 7. (a) hMOF: MAE for predicting CO_2 adsorption on the test set as a function of training set size for GS3 descriptor, using sampling methods RS, GSx, iRDM, QBC, GP and RT-AL, with random forest model. Each point is an average over 40 runs with different seeds for the train-test split. (b) Dimensionality reduction performed using UMAP, with a distance matrix obtained from the GS3 feature-set. Colors represent CO_2 adsorption at 2.5 bar, and (c) the different regions defined by the leaves of the regression tree. (d) dMOF: MAE for predicting H_2 adsorption on the test set as a function of training set size for Geom-15 descriptor, using RS, GSx, iRDM, QBC, GP and RT-AL with random forest model. Each point is an average over 40 runs with different seeds for the train-test split. (e) Dimensionality reduction performed using UMAP, with a distance matrix obtained from the Geom-15 feature-set. Colors represent H_2 adsorption at 100 bar, 77 K, and (f) the different regions defined by the leaves of the regression tree. 60 samples selected by RT-AL are shown as black circles in (b), (c), (e) and (f).

hMOF and dMOF: Adsorption properties prediction

To further validate the method, its performance on the hMOF and dMOF databases is compared with other AL methods and random sampling. For the former, the target quantities are CH_4 and CO_2 adsorption while the latter reports H_2 adsorption. Figure 7(a) shows the evolution of the MAE for CO_2 adsorption at 2.5 bar on the held out test set of the hMOF database as a function of training set size, using RT-AL, GSx, iRDM, QBC, GP and RS, with the GS3 descriptor (see S6 for training curves all the descriptors, and for CO_2 and CH_4 adsorption at different pressures). Figure 7(d) shows training curves of the different sampling methods using the Geom-15 descriptor for predicting H_2 adsorption in the dMOF database.

First, we compare RT-AL with RS. While for hMOF, RT-AL shows no improvement over RS, for dMOF, it substantially outperforms it: the accuracy achieved by RS at 1,000 samples is achieved by RT-AL in only 600 samples. This difference can be understood from the UMAP of the two databases reported in Figure 7(b) and Figure 7(e) for hMOF and dMOF, computed using the distance matrix of the GS3 and Geom-15 descriptors, respectively. For hMOF, the labels are evenly spread out in the feature space, making it easier for RS to pick samples from all regions, and finally resulting in a similar performance using AL. In contrast, the dMOF database is well structured in the descriptor space (see Figure 7(d)), and the labels are unevenly distributed throughout. RS is therefore unable to sample efficiently in this case as small regions with uniform values of the target may be missed, for example the tail of the UMAP which corresponds to high values of H_2 adsorption.

Regarding the performance of the other AL methods, GSx (and QBC) performs poorly for both datasets, similar to the QMOF case. iRDM, on the other hand, performs better than RT-AL for very low data. This is possibly due to the target distribution being more uniformly distributed in hMOF and dMOF as compared to QMOF, as illustrated in Figures S12 and S15, respectively, resulting in a higher benefit of using representativity in feature space for low data. Although RT-AL is designed to obtain a uniform target distribution in the training set, it also ensures diversity and representativity through the structure of the tree, which is why even though it starts from a random set of samples, it learns faster (see Figure S18). Because the target distribution is not known *a priori* in active learning settings, there is no guarantee that iRDM will perform well, while RT-AL will always provide a meaningful sampling. Moreover, the different regions defined by the leaves of the regression tree in Figures 7 (c) and (f) show that RT-AL identifies interesting patterns in the feature space that are related not only to the adsorption properties, but also to specific geometric features (see Figures S14 and S17 for more details).

Finally, iRDM is highly computationally expensive; for instance it takes 4.63 seconds to label 100 MOFs from the QMOF dataset using RT-AL, while for iRDM it takes 235.10 seconds (average over 5 runs in both cases). Table S3 reports the average computational time for 5 runs for the different AL methods for the QMOF dataset. To conclude, RT-AL is

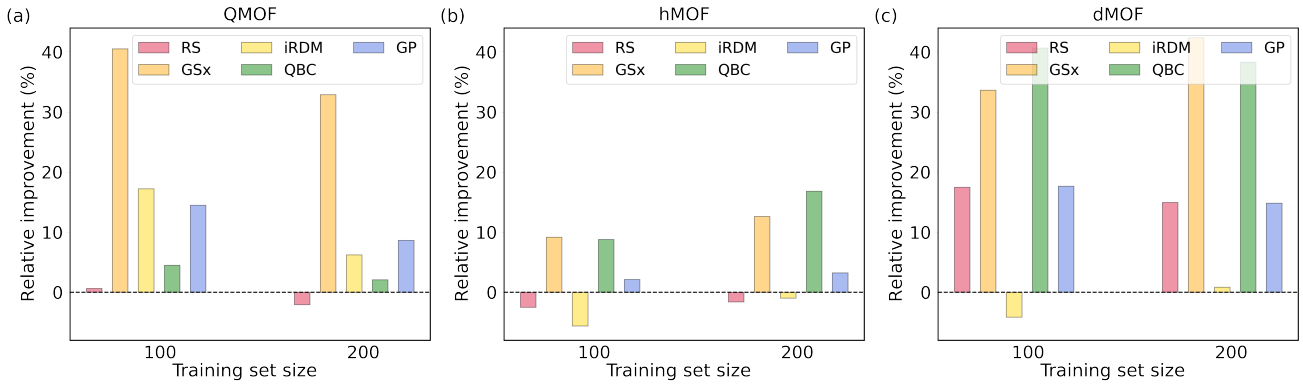


Figure 8. Relative improvement in the performance of RT-AL over other sampling methods for predicting (a) band gap (QMOF), (b) CO_2 adsorption at 2.5 bar (hMOF) and (c) H_2 adsorption (dMOF) on the respective test sets for training set sizes 100 and 200. The descriptors used are (a) ST-120, (b) GS3 and (c) Geom-15.

more general, computationally less expensive and constructs more informative training sets than other active learning methods.

Discussion and Conclusion

With the goal of improving the predicting performance when less labeled data is available, the crucial choice of descriptors well suited when training data is scarce is first addressed. It is found that in the low data regime, simpler and low dimensional descriptors are more efficient, as opposed to refined and higher dimensional ones, that are suitable for large amounts of training data. When only up to 10% labeled data is available, the stoichiometric descriptors perform the best for predicting band gaps (QMOF database). For gas adsorption (hMOF database), this is the case when a combination of purely geometric and stoichiometric descriptors are used.

Our novel regression tree-based active learning algorithm, RT-AL, is then applied to QMOF, hMOF, and dMOF databases, using these descriptors. It is shown that informative sampling is achieved using RT-AL especially in situations in which (i) the label distribution is imbalanced (multimodal and/or peaked), and (ii) the labels are unevenly distributed in the descriptor space. By selecting new samples from each leaf of the regression tree, built on the labeled part of the dataset, the new samples contain information of both the features and the labels. RT-AL significantly outperforms other sampling methods for very different properties such as band gaps and gas adsorption, on benchmark data sets of different sizes. It selects the most informative samples, forming optimal training sets of smaller sizes, without compromising on the prediction quality and helps identifying meaningful patterns in the data.

In Figure 8, we show the relative improvement of RT-AL over other sampling methods at 100 and 200 training set sizes. Our method gives a large relative improvement in most cases. There are a few cases in which the relative improvement of RT-AL is negative, but the absolute value is very low. To further quantify this, we report the MAE values (averaged over 40 runs) for each dataset in Table S2, and compare the statistical significance using t-test at level 0.05. Interestingly, even though iRDM gives a lower MAE for dMOF at 100 labeled samples, the MAE from RT-AL is statistically equivalent to it.

We also highlight the low variance of RT-AL, which is vital when doing active learning. Taking the QMOF database as a test case, the reduction of labeling cost by using RT-AL is shown to be higher when the label distribution is peaked and less balanced. This is crucial as most *real world data* is imbalanced. As one aims to be able to delve into the unexplored regions as much as possible to discover new and exceptional chemical and electronic properties of MOFs, utilizing superior sampling methods becomes essential. The effect of the distribution of the labels in the descriptor space is also investigated for the hMOF and dMOF datasets, with the conclusion that active learning performs better for the dMOF database, where labels are unevenly distributed in the descriptor space.

Importantly, as opposed to deep learning approaches that require substantial hyperparameter optimizations and fine tuning, simpler and more interpretable models trained using RT-AL can achieve similar prediction accuracy, thus highlighting the importance of smart sampling. Specifically, the performance of our simple model (RF) trained on a smart dataset constructed by RT-AL is very close to that of a CGCNN reported by Rosen et al.[43] in the low data regime. For instance, at 200 samples, our approach yields the same MAE of 0.57 eV as the CGCNN trained in ref.[43] using the same number of samples. Further, tree-based methods have been shown to train models with accuracy close to that of many deep learning models in the past, like graph neural networks,[72] with far less computational complexity and more explainability. More generally, this promising approach opens the door to a deeper understanding of complex structure-property relationships in materials.

Associated Content

Data availability statement

The python code to use regression tree-based active learning algorithm for MOF datasets can be found on GitHub: https://github.com/AshnaJose/Regression-Tree-based-Active-Learning-for-MOFs. The repository includes a comprehensive example on using RT-AL with the QMOF dataset, although it can be used for any dataset, with any descriptor. The different descriptors, mean absolute errors (MAE) for random sampling, RT-AL and other active learning methods used for comparison in this work, along with the MOFs selected in the training set by each method are available on Zenodo (DOI:10.5281/zenodo.10511345).

Supporting Information

The Supporting Information is available free of charge at: Details and hyperparameters of descriptors, and other active learning approaches. Distributions of the labels for the three databases. Training curves of RT-AL compared to that of RS for different descriptors for each dataset. Descriptor and active learning analysis for CH_4 adsorption in the hMOF dataset.

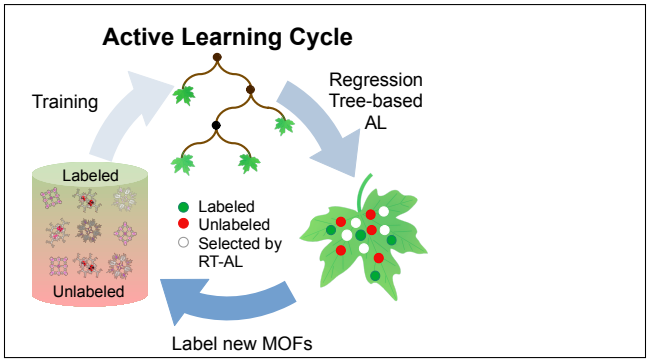
Acknowledgement We acknowledge the CINES, IDRIS and TGCC under project No. INP2227/72914/gen7211, as well as CIMENT/GRICAD for computational resources. This work has been partially supported by MIAI@Grenoble Alpes (ANR-19-P3IA-0003).

References

(1) Introduction to Metal–Organic Frameworks. *Chem. Rev.* **2012**, *112*, 673–674.
(2) Wang, C.; Liu, D.; Lin, W. Metal–Organic Frameworks as A Tunable Platform for Designing Functional Molecular Materials. *J. Am. Chem. Soc.* **2013**, *135*, 13222–13234.
(3) Ding, M.; Flaig, R. W.; Jiang, H.-L.; Yaghi, O. M. Carbon capture and conversion using metal–organic frameworks and MOF-based materials. *Chem. Soc. Rev.* **2019**, *48*, 2783–2828.
(4) Li, H.; Eddaoudi, M.; O'Keeffe, M.; Yaghi, O. M. Design and synthesis of an exceptionally stable and highly porous metal-organic framework. *Nature* **1999**, *402*, 276–279.
(5) Li, H.; Wang, K.; Sun, Y.; Lollar, C. T.; Li, J.; Zhou, H.-C. Recent advances in gas storage and separation using metal–organic frameworks. *Mater. Today* **2018**, *21*, 108–121.
(6) Li, J.-R.; Kuppler, R. J.; Zhou, H.-C. Selective gas adsorption and separation in metal–organic frameworks. *Chem. Soc. Rev.* **2009**, *38*, 1477–1504.
(7) Baumann, A. E.; Burns, D. A.; Liu, B.; Thoi, V. S. Metal-organic framework functionalization and design strategies for advanced electrochemical energy storage devices. *Commun. Chem.* **2019**, *2*.
(8) Zhao, Y.; Song, Z.; Li, X.; Sun, Q.; Cheng, N.; Lawes, S.; Sun, X. Metal organic frameworks for energy storage and conversion. *Energy Storage Mater.* **2016**, *2*, 35–62.
(9) Mariano, A. L.; Fernández-Blanco, A.; Poloni, R. Perspective from a Hubbard U-density corrected scheme towards a spin crossover-mediated change in gas affinity. *J. Chem. Phys.* **2023**, *159*, 154108.
(10) Almassad, H. A.; Abaza, R. I.; Siwwan, L.; Al-Maythalony, B.; Cordova, K. E. Environmentally adaptive MOF-based device enables continuous self-optimizing atmospheric water harvesting. *Nat. Commun.* **2022**, *13*, 4873.
(11) Hanikel, N.; Prévot, M. S.; Yaghi, O. M. MOF water harvesters. *Nat. Nanotechnol.* **2020**, *15*, 348–355.
(12) Lee, J.; Farha, O. K.; Roberts, J.; Scheidt, K. A.; Nguyen, S. T.; Hupp, J. T. Metal–organic framework materials as catalysts. *Chem. Soc. Rev.* **2009**, *38*, 1450–1459.
(13) Huang, Y.-B.; Liang, J.; Wang, X.-S.; Cao, R. Multifunctional metal–organic framework catalysts: synergistic catalysis and tandem reactions. *Chem. Soc. Rev.* **2017**, *46*, 126–157.
(14) Gamonal, A.; Sun, C.; Mariano, A. L.; Fernandez-Bartolome, E.; Guerrero-SanVicente, E.; Vlaisavljevich, B.; Castells-Gil, J.; Marti-Gastaldo, C.; Poloni, R.; Wannemacher, R.; Cabanillas-Gonzalez, J.; Sanchez Costa, J. Divergent adsorption-dependent luminescence of amino-functionalized lanthanide metal-organic frameworks for highly sensitive NO2 sensors. *J. Phys. Chem. Lett.* **2020**, *11*, 3362–3368.
(15) Xie, L. S.; Skorupskii, G.; Dincă, M. Electrically Conductive Metal–Organic Frameworks. *Chem. Rev.* **2020**, *120*, 8536–8580.
(16) Poloni, R.; Lee, K.; Berger, R. F.; Smit, B.; Neaton, J. B. Understanding Trends in CO_2 Adsorption in Metal–Organic Frameworks with Open-Metal Sites. *J. Phys. Chem. Lett.* **2014**, *5*, 861–865.
(17) Johnson, E. M.; Ilic, S.; Morris, A. J. Design strategies for enhanced conductivity in metal-organic frameworks. *ACS Cent. Sci.* **2021**, *7*, 445–453.
(18) Zhang, H.; Nai, J.; Yu, L.; Lou, X. W. D. Metal-Organic-Framework-Based Materials as Platforms for Renewable Energy and Environmental Applications. *Joule* **2017**, *1*, 77–107.
(19) Kshirsagar, A. R.; Blase, X.; Attaccalite, C.; Poloni, R. Strongly Bound Excitons in Metal–Organic Framework MOF-5: A Many-Body Perturbation Theory Study. *J. Phys. Chem. Lett.* **2021**, *12*, 4045–4051.
(20) Moosavi, S. M.; Nandy, A.; Jablonka, K. M.; Ongari, D.; Janet, J. P.; Boyd, P. G.; Lee, Y.; Smit, B.; Kulik, H. J. Understanding the diversity of the metal-organic framework ecosystem. *Nat. Commun.* **2020**, *11*, 4068.
(21) Falcaro, P. et al. A new method to position and functionalize metal-organic framework crystals. *Nat. Commun.* **2011**, *2*, 237.
(22) Rosen, A. S.; Notestein, J. M.; Snurr, R. Q. Identifying promising metal-organic frameworks for heterogeneous catalysis via high-throughput periodic density functional theory. *J. Comput. Chem.* **2019**, *40*, 1305–1318.
(23) Ren, E.; Guilbaud, P.; Coudert, F.-X. High-throughput computational screening of nanoporous materials in targeted applications. *Digital Discovery* **2022**, *1*, 355–374.
(24) Nazarian, D.; Ganesh, P.; Sholl, D. S. Benchmarking density functional theory predictions of framework structures and properties in a chemically diverse test set of metal–organic frameworks. *J. Mater. Chem. A Mater. Energy Sustain.* **2015**, *3*, 22432–22440.
(25) Jose, R.; Bangar, G.; Pal, S.; Rajaraman, G. Role of molecular modelling in the development of metal-organic framework for gas adsorption applications. *J. Chem. Sci. (Bangalore)* **2023**, *135*, 19.
(26) Duan, C.; Nandy, A.; Meyer, R.; Arunachalam, N.; Kulik, H. J. A transferable recommender approach for selecting the best density functional approximations in chemical discovery. *Nat. Comput. Sci.* **2022**, *3*, 38–47.
(27) Cho, Y.; Nandy, A.; Duan, C.; Kulik, H. J. DFT-based multireference diagnostics in the solid state: Application to metal-organic frameworks. *J. Chem. Theory Comput.* **2023**, *19*, 190–197.
(28) Kulik, H. J. et al. Roadmap on Machine learning in electronic structure. *Electron. Struct.* **2022**, *4*, 023004.
(29) Burner, J.; Schwiedrzik, L.; Krykunov, M.; Luo, J.; Boyd, P. G.; Woo, T. K. High-Performing Deep Learning Regression Models for Predicting Low-Pressure CO2 Adsorption Properties of Metal–Organic Frameworks. *J. Phys. Chem. C* **2020**, *124*, 27996–28005.
(30) Altintas, C.; Altundal, O. F.; Keskin, S.; Yildirim, R. Machine Learning Meets with Metal Organic Frameworks for Gas Storage and Separation. *J. Chem. Inf. Model.* **2021**, *61*, 2131–2146.
(31) Lee, S.; Kim, B.; Cho, H.; Lee, H.; Lee, S. Y.; Cho, E. S.; Kim, J. Computational Screening of Trillions of Metal–Organic Frameworks for High-Performance Methane Storage. *ACS Appl. Mater. Interfaces* **2021**, *13*, 23647–23654.
(32) Choudhary, K.; Yildirim, T.; Siderius, D. W.; Kusne, A. G.; McDannald, A.; Ortiz-Montalvo, D. L. Graph neural network predictions of metal-organic framework CO2 adsorption properties. *Comput. Mater. Sci.* **2022**, *210*, 111388.
(33) Nandy, A.; Duan, C.; Kulik, H. J. Using Machine Learning and Data Mining to Leverage Community Knowledge for the Engineering of Stable Metal–Organic Frameworks. *J. Am. Chem. Soc.* **2021**, *143*, 17535–17547.
(34) Moosavi, S. M.; Novotny, B. Á.; Ongari, D.; Moubarak, E.; Asgari, M.; Kadioglu, Ö.; Charalambous, C.; Ortega-Guerrero, A.; Farmahini, A. H.; Sarkisov, L.; Garcia, S.; Noé, F.; Smit, B. A data-science approach to predict the heat capacity of nanoporous materials. *Nat. Mater.* **2022**, *21*, 1419–1425.
(35) Cao, Z.; Magar, R.; Wang, Y.; Barati Farimani, A. MOFormer: Self-Supervised Transformer Model for Metal–Organic Framework Property Prediction. *J. Am. Chem. Soc.* **2023**, *145*, 2958–2967.
(36) Demir, H.; Daglar, H.; Gulbalkan, H. C.; Aksu, G. O.; Keskin, S. Recent advances in computational modeling of MOFs: From molecular simulations to machine learning. *Coordin. Chem. Rev.* **2023**, *484*, 215112.
(37) Janet, J. P.; Kulik, H. J. Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure–Property Relationships. *J. Phys. Chem. A* **2017**, *121*, 8939–8954.
(38) Fumanal, M.; Capano, G.; Barthel, S.; Smit, B.; Tavernelli, I. Energy-based descriptors for photo-catalytically active metal–organic framework discovery. *J. Mater. Chem. A* **2020**, *8*, 4473–4482.
(39) Ren, E.; Coudert, F.-X. Enhancing Gas Separation Selectivity Prediction through Geometrical and Chemical Descriptors. *Chem. Mater.* **2023**, *35*, 6771–6781.
(40) Orhan, I. B.; Le, T. C.; Babarao, R.; Thornton, A. W. Accelerating the prediction of CO2 capture at low partial pressures in metal-organic frameworks using new machine learning descriptors. *Commun. Chem.* **2023**, *6*, 214.
(41) Jablonka, K. M.; Rosen, A. S.; Krishnapriyan, A. S.; Smit, B. An Ecosystem for Digital Reticular Chemistry. *ACS Central Science* **2023**, *9*, 563–581.
(42) Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **2018**, *120*, 145301.
(43) Rosen, A. S.; Iyer, S. M.; Ray, D.; Yao, Z.; Aspuru-Guzik, A.; Gagliardi, L.; Notestein, J. M.; Snurr, R. Q. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter* **2021**, *4*, 1578–1597.
(44) Kang, Y.; Park, H.; Smit, B.; Kim, J. A multi-modal pre-training transformer for universal transfer learning in metal–organic frameworks. *Nat. Mach. Intell.* **2023**, *5*, 309–318.
(45) Park, H.; Kang, Y.; Kim, J. Enhancing Structure–Property Relationships in Porous Materials through Transfer Learning and Cross-Material Few-Shot Learning. *ACS Appl. Mater. & Inter.* **2023**, *15*, 56375–56385, PMID: 37983088.
(46) Shoghi, N.; Kolluru, A.; Kitchin, J. R.; Ulissi, Z. W.; Zitnick, C. L.; Wood, B. M. From Molecules to Materials: Pre-

training Large Generalizable Models for Atomic Property Prediction. 2023.
(47) Nandy, A.; Duan, C.; Kulik, H. J. Audacity of huge: overcoming challenges of data scarcity and data quality for machine learning in computational materials discovery. *Curr. Opin. Chem. Eng.* **2022**, *36*, 100778.
(48) Wu, D.; Lin, C.-T.; Huang, J. Active learning for regression using greedy sampling. *Inf. Sci.* **2019**, *474*, 90–105.
(49) Liu, Z.; Jiang, X.; Luo, H.; Fang, W.; Liu, J.; Wu, D. Pool-based unsupervised active learning for regression using iterative representativeness-diversity maximization. *Pattern Recognit. Lett.* **2021**, *142*, 11–19.
(50) Cai, W.; Zhang, Y.; Zhou, J. Maximizing Expected Model Change for Active Learning in Regression. 2013 IEEE 13th International Conference on Data Mining. 2013; pp 51–60.
(51) Goetz, J.; Tewari, A.; Zimmerman, P. Active Learning for Non-Parametric Regression Using Purely Random Trees. Adv. Neur. In. 2018.
(52) Burbidge, R.; Rowland, J. J.; King, R. D. Active Learning for Regression Based on Query by Committee. Intelligent Data Engineering and Automated Learning - IDEAL 2007. Berlin, Heidelberg, 2007; pp 209–218.
(53) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; MIT Press, 2006.
(54) Mukherjee, K.; Osaro, E.; Colón, Y. J. Active learning for efficient navigation of multi-component gas adsorption landscapes in a MOF. *Digital Discovery* **2023**, *2*, 1506–1521.
(55) Osaro, E.; Mukherjee, K.; Colón, Y. J. Active Learning for Adsorption Simulations: Evaluation, Criteria Analysis, and Recommendations for Metal–Organic Frameworks. *Industrial & Engineering Chemistry Research* **2023**, *62*, 13009–13024.
(56) Mukherjee, K.; Dowling, A. W.; Colón, Y. J. Sequential design of adsorption simulations in metal–organic frameworks. *Mol. Syst. Des. Eng.* **2022**, *7*, 248–259.
(57) Jablonka, K. M.; Jothiappan, G. M.; Wang, S.; Smit, B.; Yoo, B. Bias free multiobjective active learning for materials design and discovery. *Nat. Commun.* **2021**, *12*.
(58) Jose, A.; de Mendonça, J. P. A.; Devijver, E.; Jakse, N.; Monbet, V.; Poloni, R. Regression tree-based active learning. *Data Min. Knowl. Discov.* **2023**,
(59) Wilmer, C. E.; Leaf, M.; Lee, C. Y.; Farha, O. K.; Hauser, B. G.; Hupp, J. T.; Snurr, R. Q. Large-scale screening of hypothetical metal-organic frameworks. *Nat. Chem.* **2011**, *4*, 83–89.
(60) Majumdar, S.; Moosavi, S. M.; Jablonka, K. M.; Ongari, D.; Smit, B. Diversifying Databases of Metal Organic Frameworks for High-Throughput Computational Screening. *ACS Appl. Mater. Interfaces* **2021**, *13*, 61004–61014.
(61) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
(62) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA, 2016; p 785–794.
(63) Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J. E.; Doak, J. W.; Thompson, A.; Zhang, K.; Choudhary, A.; Wolverton, C. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B* **2014**, *89*, 094104.
(64) Ward, L. et al. Matminer: An open source toolkit for materials data mining. *Comput. Mater. Sci.* **2018**, *152*, 60–69.
(65) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, 184115.
(66) De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754–13769.
(67) Fernandez, M.; Trefiak, N. R.; Woo, T. K. Atomic Property Weighted Radial Distribution Functions Descriptors of Metal–Organic Frameworks for the Prediction of Gas Uptake Capacity. *J. Phys. Chem. C* **2013**, *117*, 14095–14105.
(68) Moosavi, S. M.; Nandy, A.; Jablonka, K. M.; Ongari, D.; Janet, J. P.; Boyd, P. G.; Lee, Y.; Smit, B.; Kulik, H. J. Understanding the diversity of the metal-organic framework ecosystem. *Nat. Commun.* **2020**, *11*, 4068.
(69) Ioannidis, E. I.; Gani, T. Z. H.; Kulik, H. J. molSimplify: A toolkit for automating discovery in inorganic chemistry. *J. Comput. Chem.* **2016**, *37*, 2106–2117.
(70) Nandy, A.; Terrones, G.; Arunachalam, N.; Duan, C.; Kastner, D. W.; Kulik, H. J. MOFSimplify, machine learning models with extracted stability data of three thousand metal-organic frameworks. *Sci. Data* **2022**, *9*, 74.
(71) McInnes, L.; Healy, J.; Saul, N.; Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **2018**, *3*, 861.
(72) Jiang, D.; Wu, Z.; Hsieh, C.-Y.; Chen, G.; Liao, B.; Wang, Z.; Shen, C.; Cao, D.; Wu, J.; Hou, T. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *J. Cheminform.* **2021**, *13*, 12.

TOC Graphic


Active Learning Cycle
Training
Regression
Tree-based
AL
Labeled
Unlabeled
Labeled
Unlabeled
Selected by
RT-AL
Label new MOFs