PCCP

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Phys. Chem. Chem. Phys.*, 2025, **27**, 23152

Density-aware active learning for materials discovery: a case study on functionalized nanoporous materials

V. Gkatsis,[ab] P. Maratos,[c] C. Rekatsinas,*[b] G. Giannakopoulos[bd] and P. Krokidas*[b]

Received 30th July 2025,
Accepted 26th September 2025

DOI: 10.1039/d5cp02908b

rsc.li/pccp

Machine learning algorithms often rely on large training datasets to achieve high performance. However, in domains like chemistry and materials science, acquiring such data is an expensive and laborious process, involving highly trained human experts and material costs. Therefore, it is crucial to develop strategies that minimize the size of training sets while preserving predictive accuracy. The objective is to select an optimal subset of data points from a larger pool of possible samples, one that is sufficiently informative to train an effective machine learning model. Active learning (AL) methods, which iteratively annotate data points by querying an oracle (*e.g.*, a scientist conducting experiments), have proven highly effective for such tasks. However, challenges remain, particularly for regression tasks, which are generally considered more complex in the AL framework. This complexity stems from the need for uncertainty estimation and the continuous nature of the output space. In this work, we introduce density-aware greedy sampling (DAGS), an active learning method for regression that integrates uncertainty estimation with data density, specifically designed for large design spaces (DS). We evaluate DAGS in both synthetic data and multiple real-world datasets of functionalized nanoporous materials, such as metal–organic frameworks (MOFs) and covalent-organic frameworks (COFs), for separation applications. Our results demonstrate that DAGS consistently outperforms both random sampling and state-of-the-art AL techniques in training regression models effectively with a limited number of data points, even in datasets with a high number of features.

Introduction

Designing materials with specific properties is a challenging and time-consuming task, often relying on trial-and-error (Edisonian) methods.[1] This traditional approach increases both the time and cost of experiments. The main difficulty lies in the complexity of materials design, where the relationship between a material structure and its properties is poorly understood.[2] These challenges are usually addressed using accumulated chemical intuition, which can be limiting: intuition often struggles to navigate the vast and complex landscape of modern materials design spaces, especially when those spaces are high-dimensional, non-linear, or span unfamiliar chemistries. This can lead to biased exploration, favoring familiar structures or compositions and overlooking novel or counterintuitive candidates that could offer superior properties. Furthermore, as materials discovery increasingly integrates high-throughput simulations and data-driven approaches, the scale and diversity of possible candidates often exceed the capacity of intuition-based methods to make meaningful selections.

The emergence of artificial intelligence (AI) and machine learning (ML) offers new opportunities in this area. ML models excel at finding patterns in data, often surpassing human capabilities.[3–6] These models can take information about a material's characteristics (features) and predict how it will perform (properties). By doing so, ML can guide researchers in selecting which materials to study, reducing experimental effort and cost.[7] However, reliable ML models require high-quality data for training, and generating such data is expensive and labor-intensive. This creates a paradox: while we aim to reduce experimental costs, creating the large, diverse datasets needed for training these models remains costly. To address the limitations of conventional data collection methods, researchers are exploring strategies that go beyond random

[a] *Department of Informatics and Telecommunications, National and Kapodistrian University, Athens, Greece*
[b] *Institute of Informatics & Telecommunications, National Center for Scientific Research "Demokritos", Agia Paraskevi 15310, Greece. E-mail: crek@iit.demokritos.gr, p.krokidas@iit.demokritos.gr*
[c] *School of Electrical & Computer Engineering, National Technical University of Athens, Athens, Greece*
[d] *SciFY PNPC, Agia Paraskevi 15310, Greece*

sampling (RS)—the simplest approach for selecting candidate instances to build training datasets in materials science. In RS, samples are chosen randomly and independently from a larger pool of possible materials. This pool, once mapped to a vector space where each dimension represents a structural or compositional property, is commonly referred to as the design space of the materials. For each selected sample from this design space, an experiment—computational or experimental—is conducted to determine the values of the target (dependent) variables. The resulting instance, consisting of input features and corresponding outputs, is then added to the dataset. While RS is straightforward and easy to implement, it does not incorporate any prior knowledge about the distribution or structure of the data. As a result, it may frequently select samples that are redundant or unlikely to improve the model's performance—commonly known as uninformative samples —particularly in low-data regimes where every labeled point carries significant weight.

In such cases, active learning (AL) techniques are more appropriate, as they aim to strategically select the most informative samples, thereby maximizing model improvement while minimizing the number of required labeled instances.[8,9] Active learning is a semi-supervised learning method meaning that target values of the dataset are partially unknown, and the machine learning model is trained by selecting data points one by one and querying their target values to an oracle. This method uses the acquired knowledge about the data space in order to effectively guide the selection of the next data point, usually by evaluating an uncertainty measure. After the query to the oracle that annotates the data samples, the obtained feature-value pair is added to the training set thus updating the model's knowledge of the data space. Using this technique allows researchers to focus on the most informative samples, thus optimizing the process. AL techniques often use concepts like diversity[10] (choosing samples that differ significantly from each other) and representativeness[11] (choosing samples that best represent the dataset) to guide sample selection.

Some active learning methods, known as model-based approaches, rely on the ML model to guide the identification of the samples to annotate, focusing on those most likely to improve predictions. A seminal example is the work by Cohn *et al.*,[12] who proposed selecting samples to reduce model uncertainty. However, it can be computationally intensive—particularly for neural networks—and relies on assumptions (*e.g.*, Gaussian noise, negligible bias) that may not always hold, limiting its scalability and generality. AL has been extensively used for classification tasks, where the selection criterion often relies on entropy-based uncertainty measure,[13,14] vote entropy[15] and expected model change.[16] However, in regression tasks, where the computation of entropy is infeasible, the AL bibliography is limited, and the main techniques require different criteria to substitute the uncertainty measure. Regarding regression tasks, approaches such as greedy sampling (GS), combine diversity and representativeness to improve predictions by greedily selecting the data sample that maximizes on a specified criterion: GSx focuses exclusively on the exploration of the feature space, while GSy prioritizes target property space exploration through the model's predictions.[17] Although both methods manage to adequately learn the design space, their individual predictive performances are hindered due to their lack of insight into each other's data space domain (target property space for GSx and feature space for GSy). To this end, the improved GS method (iGS) was devised to combine both methods, achieving remarkable results.[18,19] Another prominent technique, expected model change maximization (EMCM),[20] evaluates the potential impact of annotating a sample on the current model and selects the sample that leads to the greatest change in the model's parameters, measured as the difference between the current model parameters and the updated parameters after training with the enlarged training set. This method works under the assumption that the greatest parameter change is correlated with significant learning opportunities in the design space. While effective, methods like EMCM can be computationally intensive as the model has to constantly estimate the gradient of the loss and update all model parameters for each new annotated sample of the space. This led to the development of batch strategies such as B-EMCM[21] to address these challenges.

Recently, researchers have explored Mondrian trees,[22,23] which is a type of regression tree that branches randomly rather than based on features. While they can achieve modest improvements over other state-of-the-art methods, the high variance in predicted values within each leaf node and reliance on scaling datasets to a fixed range ([0, 1]) can limit their practicality.[23] Furthermore, emerging AL techniques now incorporate advanced tools like Bayesian models, Gaussian processes (GP), and deep learning. For example, Gaussian processes can model data uncertainty but are mainly used for low-dimensional datasets. Similarly, deep learning-based methods, such as batch model deep active learning (BMDAL),[24] are designed for large datasets and may not suit applications where data annotation is expensive.

In this work, we address a critical limitation of AL which is that there are problem cases where it struggles to significantly outperform baseline sampling methods on finding the most informative data points and efficiently training ML models using them. This happens when the data space is not homogeneous, meaning that the data samples are not uniformly distributed across the feature space hypercube domain and form dense and sparse regions resulting in the decrease of pure exploration AL framework's performance. This is because a pure exploratory AL framework such as iGS mainly selects samples from sparse regions as they are more diverse to the already explored space, while simpler methods such as RS follow the underlying density distribution and select more samples from the denser regions thus optimizing the predictions of the model. To overcome this, we propose an AL framework for regression tasks that incorporates density-awareness to the selection process of improved greedy sampling, called density-aware greedy sampling (DAGS). For classification tasks, modeling the density of the data space is common as the framework has to ensure that the selected sample is both informative and representative of its class.[25–28]

However, density-awareness has been largely overlooked in the active learning literature for regression tasks, and this constitutes the main contribution of our work. Our method explicitly exploits density as a characteristic of the design space, allowing us to balance exploration with representativeness and thereby select more informative samples. In this way, we address the limitations of iGS, which often overemphasizes outliers and expends oracle queries on points that contribute little to model improvement. Our results further show that the proposed density-aware approach can match or even surpass random sampling, which implicitly reflects data density to some extent. Finally, we benchmark DAGS not only against random sampling but also against more sophisticated active learning strategies, including query-by-committee,[29] regression tree-based AL,[30] and plain iGS.[17]

To evaluate the performance of our proposed framework against the aforementioned techniques, we first constructed synthetic datasets based on four distinct formulas. Each formula is examined in two versions: (a) homogeneous and (b) non-homogeneous distributions of data points. Following this controlled evaluation, we apply the framework to a real-world scenario involving complex sample spaces of materials with high correlation complexities and heterogeneity. Specifically, we focus on metal–organic frameworks (MOFs),[31] a class of functionalized materials whose structures can be modulated at the molecular level. MOFs exhibit exceptional potential as adsorbent/storage materials[32] or components in separation membranes.[33] However, understanding how their design influences performance remains a complex challenge, often requiring either labor-intensive experiments or computationally demanding *in silico* simulations. Our proposed framework aims to address these challenges by improving the efficiency and accuracy of predictive modeling in such complex material systems. In both the synthetic data and MOF datasets, our approach consistently outperforms the other methods demonstrating superior performance compared to them.

Methodology

In this section, we present our proposed density-based active learning method, called density-aware greedy sampling (DAGS), designed as an improvement of the iGS method by incorporating density-based sample selection.

Problem formulation

We consider a process f: $U \rightarrow Y$ generating data $y_i = f(u_i)$, where $u_i \in U$ and $y_i \in Y$ are vectors with dimensions n and m, respectively. This process may represent an experiment or a simulation that returns accurate target values y_i for a given input u_i, but at a high cost, for example, in terms of time, computation, or other resources. The set U denotes the design space, that is, all possible u_i elements that can be represented in the n-dimensional space.

To obtain the target values y_i for a specific input u_i, the user must run one iteration of the expensive process f. However, when the locations of inputs yielding desirable outputs are unknown within the design space, the user may need to evaluate many such inputs, resulting in high overall cost.

To address this, we propose to train a machine learning model M that approximates the mapping f: $U \rightarrow Y$. In this way the user can have a good estimation of target values of each u_i and thus will be able to run targeted iterations of f, for those u_i predicted to have a target value y_i closer to the desired. Training the model M also means acquiring target values for each u_i that will be used as the training dataset. In order to create an efficient model, we want to find the balance between maximizing the model performance and minimizing the dataset creation cost. A low cost means that M should use the least amount of training data, so that the number of iterations of f performed is reduced as much as possible without significant loss of estimation performance. For this reason, we assume a limited budget of N available iterations. Let $L \subseteq U$ be the subset of inputs selected for training, such that:

$$L = \{l_1, l_2, \ldots l_N\}, \ l_i \in U \qquad (1)$$

Initially, L will contain $k < N$ randomly selected samples. For each randomly selected sample, an iteration of f is performed and its target value is acquired, and the machine learning model is trained with L as the training dataset. Then, we define a selection process, s, which given the current state of the model and the training dataset, identifies the next element that should be used for training

$$L_{i+1} = s(L_i, M) \qquad (2)$$

The elements are sampled one by one; for each one, an iteration of f is performed, and after its target value has been acquired, the machine learning model is retrained with L_{i+1} as the training dataset.

Our goal in this research work is the creation of an algorithm serving as a selection process s, which will efficiently achieve this balance between model performance and data creation cost.

Fig. 1 provides a graphical representation of the problem case that we are exploring.

State-of-the-art methods

In this section, we overview current state-of-the-art methods, highlighting strengths and weaknesses that have led to the proposal of our DAGS algorithm.

Random sampling. Random sampling is a very simple selection strategy which, as the name suggests, selects each new sample in a random manner.[34,35] The lack of a complicated mechanism for sample selection makes it a rather fast approach. Although it is not a sophisticated method, it has the unique ability to often outperform other state-of-the-art methods.[36] This observation constitutes random sampling as a nontrivial baseline method for evaluating AL frameworks.

Query-by-committee. Query-by-committee (QBC) is an AL technique that uses the disagreements among a committee of predictors to select informative samples.[37] For a given training

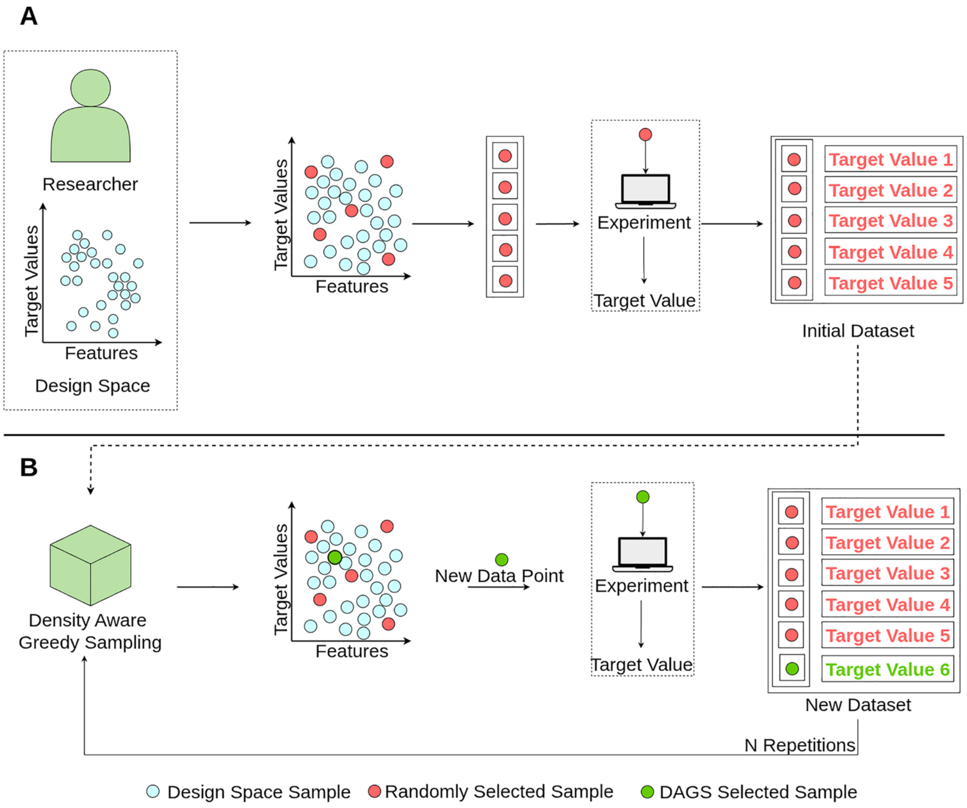

Fig. 1 (A) Given a design space of possible candidates, a researcher selects *k* random samples to conduct experiments and evaluate the value of their target property. (B) Our proposed sampling method iteratively selects the next candidate for experimentation, based on the current state of the dataset, until the budget *N* is completed.

set L and an unlabeled set U, the method trains k diverse predictors $f_1, f_2, \ldots, f_k$ on L and evaluates their predictions $f_i(l)$ for $l \in L$. The sample with the maximum disagreement, quantified as the variance among the predictor's outputs, is selected for annotation. Specifically, the variance for a sample x is computed as follows:

$$\mathrm{Var}(x) = \frac{1}{k}\sum_{i=1}^{k}\left(f_i(x) - \bar{f}(x)\right)^2 \tag{3}$$

where

$$\bar{f}(x) = \frac{1}{k}\sum_{i=1}^{k} f_i(x) \tag{4}$$

is the mean prediction of the committee. The sample x^* selected for querying the oracle is thus:

$$x^* = \arg\max_{x \in U} \mathrm{Var}(x) \tag{5}$$

This strategy assumes that areas of disagreement represent regions with high uncertainty, making them valuable for improving the model's performance. While QBC minimizes overfitting when predictors are diverse, such as using models from different learning paradigms, it suffers from limitations in regression tasks. Specifically, its focus on the target property alone for query selection often leads to suboptimal performance in cases with complex feature–target correlations, such as MOF datasets. In our implementation, the models used were XGBoost,[38] random forest, and Gaussian process regressors – studies have shown that two or three predictors are generally sufficient.[39] Despite its conceptual appeal, QBC's performance is often inferior to more balanced exploration–exploitation approaches like iGS and density-based methods, particularly in high-dimensional regression problems. It is also worth mentioning that the need for multiple regressors (those forming the committee) may make this approach more expensive than others, since each time a new data point is sampled we need to re-train them all. The query-by-committee code used in this work was developed by the authors.

Improved greedy selection across the feature space (iGS). iGS[17] is an improved version of GS which is a combination of methods known in the literature as GSx and GSy. GSx is greedy selection performed across the feature space x. In this method, the learner iteratively selects to query the unlabeled sample x_t that maximizes the minimum distance from the samples that exist in the training set x_n. If the training set L contains k labeled samples, and the design space contains in total $/U/$ samples, then the learner selects the $k + 1$ sample from the remaining $/U/ - k$ as:

$$x_{k+1} = \arg\max_{x \in U} d_x(x) \tag{6}$$

where

$$d_x(x) = \min_{n=1,\ldots,k;t=k+1,\ldots,|U|} \|x_t - x_n\| \tag{7}$$

GS*y* is greedy selection across the target value space, *Y*. Predictions of the target property are produced for all unlabeled data, and the sample $\hat{y}_t$ whose predicted target value is most "foreign" compared to the evaluated target values in the training set y_n is selected:

$$x_{k+1} = \arg\max_{x\in U} d_y(x) \tag{8}$$

where

$$d_y(x) = \min_{n=1,\ldots,k;t=k+1,\ldots,|U|} \|\hat{y}_t - y_n\| \tag{9}$$

While these methods are effective in their respective domains, they are limited in scope: GS*x* focuses solely on the feature space, and GS*y* focuses only on the target space. Both approaches ignore the correlation between the two spaces, which is critical for representing the underlying process. The iGS method combines GS*x* and GS*y* by incorporating information from both the feature and target spaces. It selects the next sample using the following criterion:[17]

$$x_{k+1} = \arg\max_{x\in U} d_x(x)d_y(x) \tag{10}$$

where $d_x(x)$ and $d_y(x)$ are computed by (7) and (9), respectively. By using the product of these metrics, iGS is immune to scaling differences between the feature and target spaces. The main drawback of this method is that it does not inherently guarantee that the selected samples follow the design space distribution, which increases the sensitivity of the method in querying outliers, thus creating sets of samples that are not representative of the design space. The iGS code used in this work was developed by the authors, based on the work of Wu *et al.*[17]

Regression-tree based active learning. The regression tree-based active learning (RT-AL) method, introduced by Jose *et al.*,[30] focuses on constructing optimal training sets for regression tasks where labeling data is expensive. This approach employs regression trees to partition the feature–response space into homogeneous regions, using a splitting criterion that minimizes response variance within each partition. Specifically, the variance reduction is calculated using the following equation:

$$\Delta(t) = \mathrm{Var}(t) - \left(\frac{|t_\mathrm{L}|}{|t|}\mathrm{Var}(t_\mathrm{L}) + \frac{|t_\mathrm{R}|}{|t|}\mathrm{Var}(t_\mathrm{R})\right) \tag{11}$$

where t represents a node, and t_L and t_R are its left and right child nodes. Following tree construction, RT-AL selects samples based on their representativeness and diversity using a diversity-based query, which balances exploration of the feature response spaces.[30]

Extensive benchmarking demonstrates the ability of RT-AL to achieve lower error rates with reduced sample sizes compared to other state-of-the-art methods, particularly in datasets with complex distributions. The method's robustness and low variance make it a reliable choice for regression tasks across diverse application domains. For our implementation of RT-AL, we have adapted and used the code provided by Jose *et al.*[30]

The aforementioned methods work well in scenarios where data are uniformly distributed across the design space. However, many real-world datasets exhibit imbalances, with dense and sparse regions in the design space. In such cases, purely explorational AL techniques may underperform, and even random sampling can outperform these methods.

Our proposed method: density-aware greedy sampling

To address the aforementioned limitation of the state-of-the-art AL approaches, we propose a density-based AL method, called density-aware greedy sampling (DAGS), that combines iGS with a weighting factor representing space density. The density factor for an unlabeled sample x is calculated using the following equation:

$$D(x) = \frac{1}{\dfrac{\sum_{i=1}^{n} \mathrm{dist}(x_i, x)}{n}} \tag{12}$$

where $x_1, \ldots, x_n$ are the n-nearest unlabeled neighbors of x, and $\mathrm{dist}(x_i, x)$ is the Euclidean distance between x and x_i. This density factor measures the average proximity of a sample to its neighbors.

Using this factor, the next sample is selected by:

$$x_{k+1} = \arg\max_{x\in U} \mathrm{iGS}(x)D(x) \tag{13}$$

where $\mathrm{iGS}(x)$ is the uncertainty produced by the iGS method for the unlabeled sample x.

This approach has been inspired by the work of Zhu *et al.*,[8] where a similar strategy was proposed for classification tasks. Specifically, the authors devise an uncertainty-based active learning framework named sampling by uncertainty and density (SUD), where the selection criterion consists of the multiplication of an uncertainty and a density factor. In this method, the uncertainty for each unlabeled sample is modeled as the entropy of the estimated probabilities for the sample to belong in each class. The density factor is computed as the average cosine similarity of the sample x with its K-nearest neighbors. The two factors are then multiplied to produce the final selection metric for the unlabeled set. The main drawback of this approach when implemented for regression tasks is that calculating the entropy for each sample is challenging as there are no well defined classes (we can either model each sample as a separate class or rely on clustering methods which make the entropy computation inefficient and inaccurate).

To tackle this problem, in our method, we substitute the entropy with the iGS factor which adequately represents the uncertainty of a design space and is suitable for regression applications. Another difference between the two methods is that we define density as the average inverse of the sample's Euclidean distances with its neighbors and not the average of their cosine similarities. This decision has been made based on the assumption that, after querying a sample, the model gains knowledge of the target property's behaviour on a small area around it, as samples with almost identical feature space values will probably exhibit approximate target property's values. Conforming with this assumption, a dense area should consist of samples that have absolute proximity and not necessarily the same direction of feature vectors. In general, the Euclidean distance provides a more intuitive and reliable measure of the density "neighborhood", particularly in continuous spaces where absolute distances are

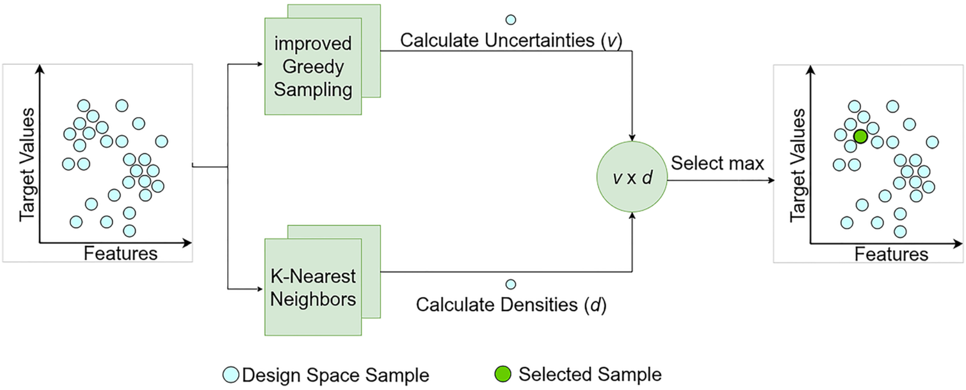

Fig. 2 Given a design space, iGS computes uncertainties (*v*) and K-NN computes densities (*d*) for all samples. DAGS selects the sample that maximizes the product of *v* and *d*.

critical as it directly measures spatial proximity. In conclusion, our density-based method selects samples that maximize exploration while prioritizing dense areas, ensuring the selected samples provide the most significant knowledge about the design space. This helps reduce the average prediction error, as sparse areas often represent outliers with little relevance.

Computational experimental evaluation work-flow

Our experimental evaluation methodology can be described as follows. First, we train the XGBoost model using the current training dataset L_i and we evaluate its performance. Second, we employ the density aware greedy sampling (DAGS) method to select the next sample. DAGS calls the improved greedy sampling (iGS) method to compute uncertainty values (v) and the density factor (d) for each unlabeled element in U. Third, the selection method selects the element which maximizes the product of v and d and adds it to the training dataset thus creating L_{i+1}. The workflow of DAGS is shown in Fig. 2. These three steps are repeated until we reach the end of our N query budget (the maximum number of samples we can query to the oracle). We refer to this entire process as one experiment.

In order to mitigate concerns that the final results are due to dataset peculiarities, we use a k-fold routine where we shuffle and divide the design space in k consecutive folds. Then, we select $k-1$ folds as the training set while the remaining fold becomes the test set. In our work-flow, we set $k = 10$.

As mentioned before, the training set L_{i+1} is built by evaluating the model on the previous training set L_i and then adding the next sample proposed by the selection method. To bootstrap this process, we initialize L_0 by randomly selecting 5 samples from the design space. We ensure that, throughout our experiments, the five initially selected random samples remain the same for each dataset. Maintaining this consistency prevents random chance from significantly influencing our results.

The five selection methods are evaluated using the mean absolute error (MAE) metric, which is expressed as follows:

$$\mathrm{MAE} = \frac{1}{n}\left(\sum_{i=1}^{n} |y_i - \hat{y}_i|\right) \tag{14}$$

where y_i is the ground truth and $\hat{y}_i$ is the prediction of the model, on a test set of size n.

The whole evaluation workflow as described above is performed 10 times. Finally, we plot the average MAE across the 10 experiments for increasing training dataset sizes. The code used for the experiments is openly accessible in our GitHub repository.†

The predictive model being used in our experiments is the XGBoost regressor.[38] Details regarding the Python libraries and the hyperparameters of these ML regression models are provided in the SI. For density calculations in the feature space, we used Euclidean distances without applying prior normalization. We acknowledge that omitting normalization can be problematic in very high-dimensional spaces or when feature values differ by several orders of magnitude. In our datasets, however, the number of features is modest (up to 20), and their ranges vary only within a few orders of magnitude. Under these conditions, we chose to focus on demonstrating the impact of incorporating density itself into sampling strategies. Nonetheless, feature normalization remains an important consideration for future work, particularly within a more generalized framework.

Results

The Results section is divided into two subsections: in the first subsection, we compare the selected sampling methods on synthetic data spaces, while in the second subsection we move to actual design spaces, in the setting of MOFs.

Synthetic data spaces

Four synthetic data spaces were prepared, each one with a homogeneous and a heterogeneous version of the distribution of its points across the same range of available data points. We examine the DAGS method and compare it with the baseline models in order to highlight the effect of heterogeneity on the performance of the pure exploration AL methods. For each synthetic dataset (homogeneous and heterogeneous), we

† https://github.com/insane-group/Density_Aware_Greedy_Sampling.

provide a plot of the data points which depicts the density across the design space, meaning how close the samples are located to each other. In the case of 1-d design spaces, we used the kernel density estimation (KDE) method to create the color bar which intuitively shows dense and sparse areas of the dataset. For 2-d design spaces, where it is easier to show distances between data points, we plotted the two dimensions along with a color bar showing the variability of values of the target property.

The first space is modelled after the 1d Forrester benchmark[40] (Fig. 3), which is commonly used for evaluating Bayesian optimization methods, as we want to examine the learning capabilities of the AL frameworks on a continuous yet complex data space, where we select 1000 x samples within the range [0, 1]. The target property can be calculated using the following formula:

$$y(x) = (6x - 2)^2 \sin(12x - 4) \tag{15}$$

The next space is a variation of the first, called 1d Jump Forrester (Fig. 4) which inserts a discontinuity at the target function as we want to capture the effect of non-continuous target properties on the performance of the AL frameworks.

$$y(x) = \begin{cases} (6x-2)^2 \sin(12x-4), & 0 \le x \le 0.5 \\ (6x-2)^2 \sin(12x-4) + 10, & 0.5 < x \le 1 \end{cases} \tag{16}$$

The third one is modelled after a 2d Gaussian (Fig. 5) in order to simulate an area of interest at the center of the space and evaluate the degree that each method effectively learns the space when data samples are uniformly scattered or create an extremely dense area at the center, with $x_1, x_2 \in [-3, 3]$, and the target value is produced by:

$$y(x) = \exp\left(-\frac{{x_1}^2 + {x_2}^2}{2\sigma^2}\right) \tag{17}$$

Finally, the last space has $x_1, x_2 \in [0, 1]$ and y is an exponential form of x (Fig. 6) as we want to model a design space that has (complementary to the Gaussian space) the area of interest at the border of the space, examining if a pure exploration AL method performs well in this context. The y is expressed through:

$$y(x) = 1 - \exp(-0.6((x_1 - 0.5)^2 + (x_2 - 0.5)^2)) \tag{18}$$

In the following figures, we showcase the performance of various sampling methods, measuring the mean absolute error (MAE) as a function of the number of samples annotated (we designate this number as "# of queries" in the figures, since these annotations are essentially queries towards an "oracle"). The results show that the iGS and DAGS methods outperform random sampling in all homogeneous spaces (Fig. 7(a), 8(a), 9(a) and 10(a)), because they operate in a strategic and exploratory manner, efficiently identifying the most informative data points based on their position in the design space. An important observation is that in homogeneous spaces, where the density of data points is nearly uniform across the entire space, our method effectively reduces to iGS, as the density factor is almost identical for every unknown point.

In heterogeneous spaces, however, the performance of AL frameworks compared to random sampling is less straightforward. Notably, AL methods that disregard the density distribution of the design space, such as iGS, often fail to outperform RS. Specifically, iGS exhibits reduced performance in the Forrester space (Fig. 7(b)) when compared to the density-based method, shows nearly identical performance to RS in

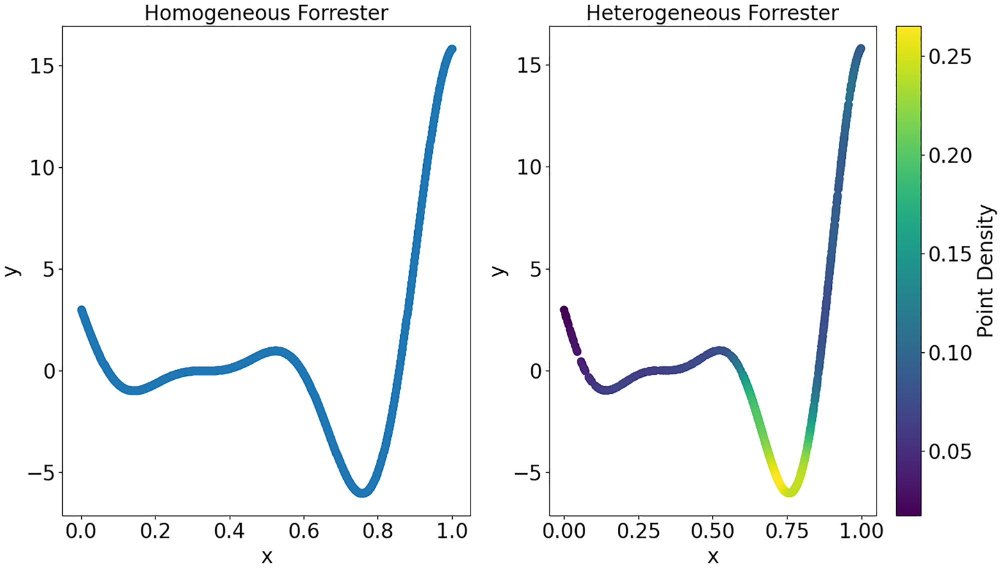

Fig. 3 Plots of homogeneous and heterogeneous Forrester datasets.

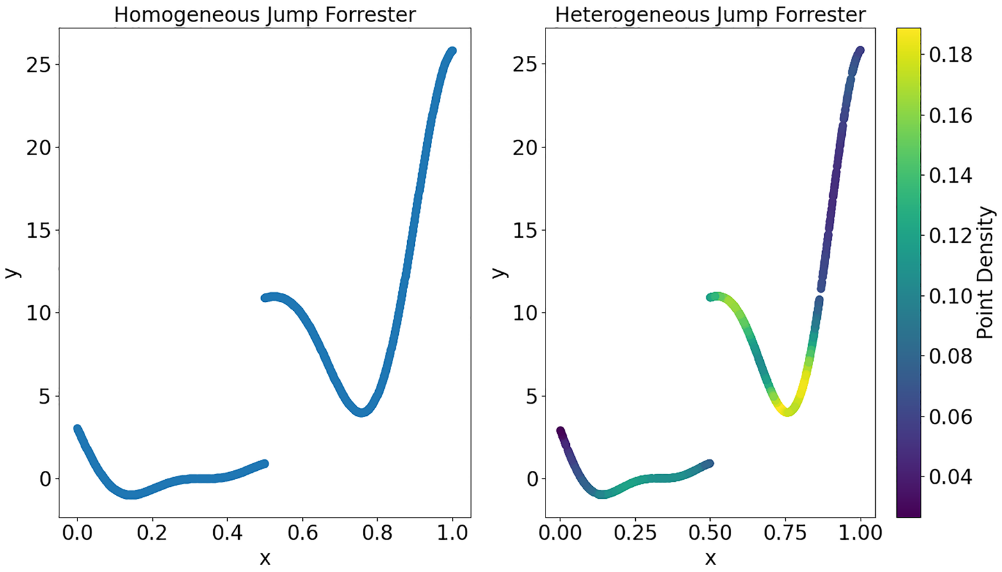

Fig. 4 Plots of Jump Forrester homogeneous and heterogeneous datasets.

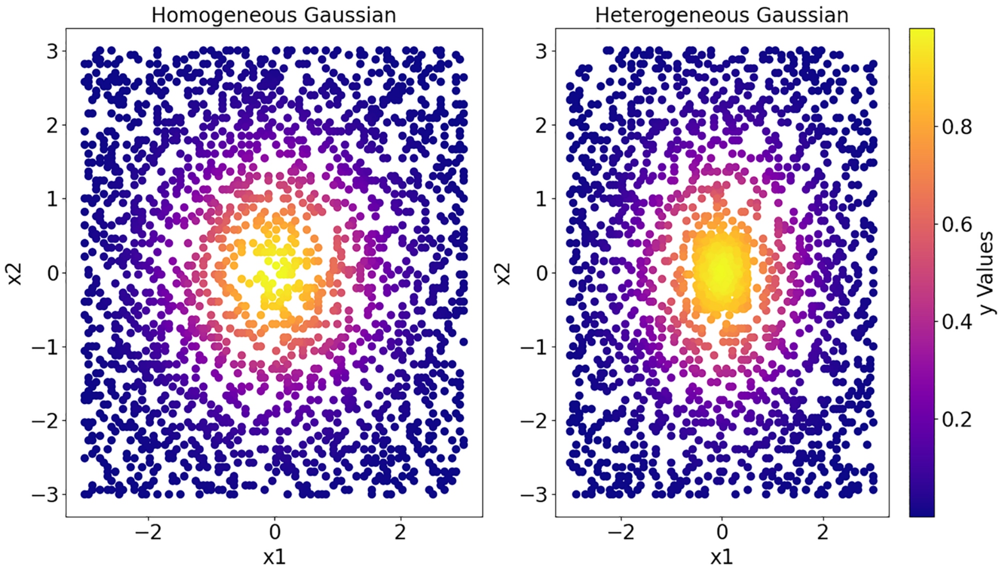

Fig. 5 Plots of homogeneous and heterogeneous Gaussian datasets.

the Jump Forrester benchmark (Fig. 8(b)), and suffers complete performance degradation in the Gaussian space (Fig. 9(b)). The poor performance of iGS in the Gaussian benchmark can be explained by its tendency to select points far from the center, as it prioritizes coverage of the entire design space. This approach neglects the fact that, in the heterogeneous case, more than half of the data samples are concentrated in the central region, where selecting points is critical for achieving a significant reduction in mean absolute error (MAE). The only heterogeneous design space where iGS performs well is the exponential space (Fig. 10(b)), where it predominantly selects samples from the edges of the space, focusing on modeling the area of interest rather than the central region.

In contrast, the DAGS method performs robustly across both homogeneous and heterogeneous spaces. In homogeneous spaces, it operates in a purely exploratory manner, similar to iGS. In heterogeneous spaces, however, it effectively captures the underlying density distribution of the design space. This adaptability enables our method to consistently outperform RS across all synthetic benchmarks. In rare cases, such as the exponential space (Fig. 10(b)), its performance is comparable to iGS which indicates that at extreme data space heterogeneity scenarios where the area of interest requires a purely exploration criterion, and the density factor in our method leads to the selection of some suboptimal points. Overall, the density-based AL framework demonstrates superior versatility

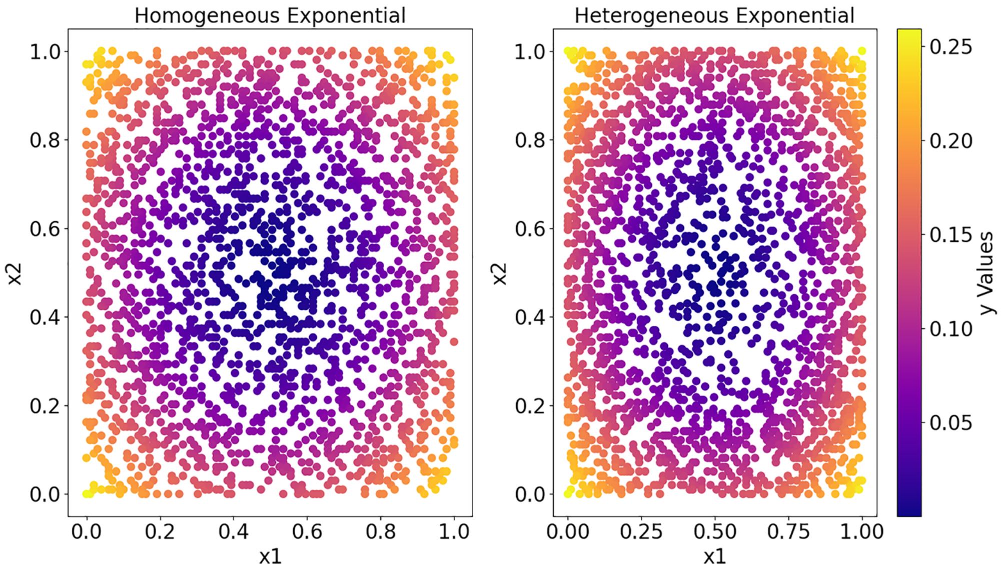

Fig. 6 Homogeneous and heterogeneous exponential datasets.

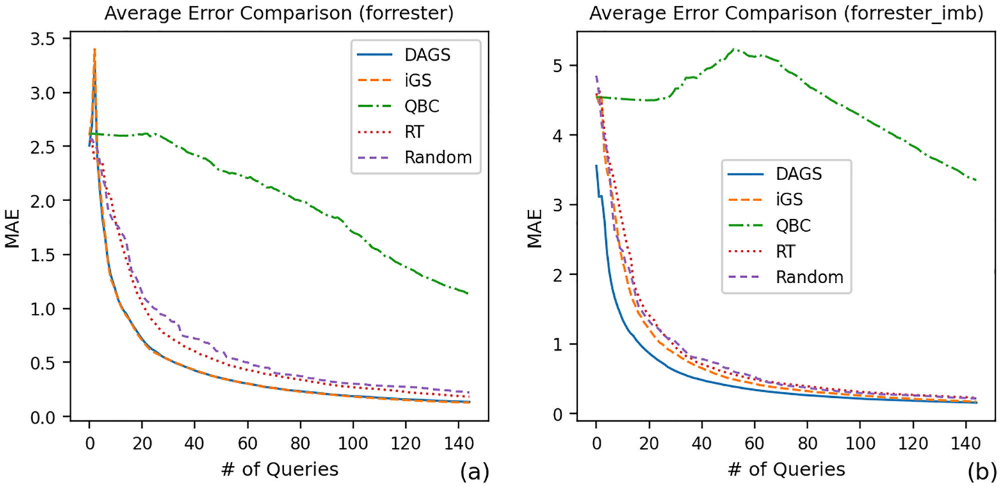

Fig. 7 MAE as a function of the number of queries for Forrester (a) homogeneous and (b) heterogeneous datasets. A lower MAE means better predictive capabilities of the model.

and effectiveness, making it a more reliable choice for diverse design spaces.[41] In all cases we set our query budget N at 150 data points, at which point we stopped the sampling. Out of those, 5 were initially randomly selected and 145 selected by each method. For design spaces of 1000 or 2000 samples, 150 queries represent 15% and 7.5% of the whole space, respectively, as we opt for simulating realistic training size – design space size ratios in order to test the efficiency of the proposed method.

MOF design spaces

In this section, we evaluate the performance of all the sampling methods in real-world scenarios drawn from materials chemistry, specifically focusing on functionalized nanoporous materials known as metal–organic frameworks (MOFs). This domain is particularly relevant due to the well-recognized challenge of establishing accurate structure–performance correlations,[42] which hinders the development of MOFs for applications such as separation membranes and storage materials.

We utilize five datasets from the literature, each comprising thousands of MOFs characterized by structural and chemical descriptors (or attributes) as input features for model training. These datasets were chosen not only for their size and availability but also because they address the relatively underexplored property of gas diffusivity, as opposed to the more commonly studied sorption capacity or uptake. The target property, diffusivity (D_i), typically measured in either $m^2\ s^{-1}$ or $cm^2\ s^{-1}$, represents the rate at which penetrants (guest molecules) of species i (commonly gases such as CO_2, CH_4, N_2, and O_2) propagate through the porous structure of a material. Target values for diffusivity were obtained through *in silico* experiments, specifically molecular simulations.

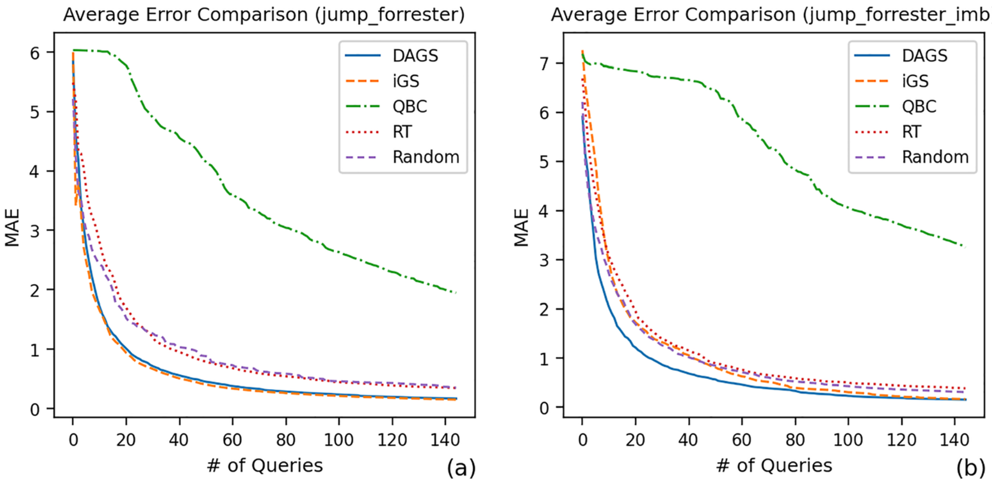

Fig. 8 MAE as a function of the number of queries for Jump Forrester (a) homogeneous and (b) heterogeneous datasets. A lower MAE means better predictive capabilities of the model.

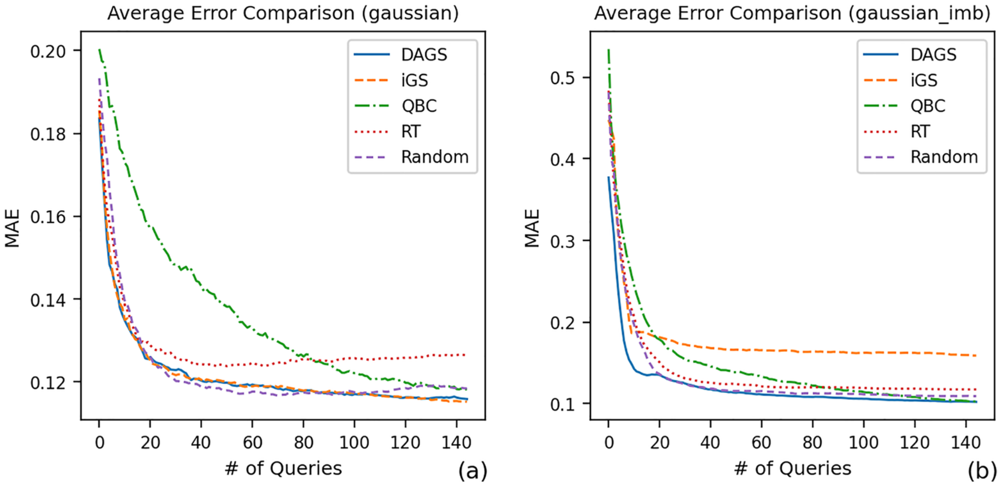

Fig. 9 MAE as a function of the number of queries for Gaussian (a) homogeneous and (b) heterogeneous datasets. A lower MAE means better predictive capabilities of the model.

Gas diffusivity is underrepresented in high-throughput simulation schemes due to its higher computational cost relative to sorption properties. This is primarily because calculating diffusivity requires smaller time steps for numerical integration of the equations of motion, resulting in significantly longer simulation times.[43] By addressing this property, we aim to highlight the applicability and efficiency of active learning frameworks in domains with high computational complexity.

N_2 and O_2 diffusion in MOFs (two datasets)

The first two real-world datasets used in our study are drawn from the work of Orhan *et al.*,[44] where the authors employed high-throughput computational screening and machine learning to predict O_2/N_2 selectivity in 5632 metal–organic frameworks (MOFs). These datasets, derived from the CoRE MOF 2019 database, includes detailed geometric and chemical descriptors alongside the simulated target properties of diffusivity of oxygen (O_2) and nitrogen (N_2), respectively (information about the features of the datasets in our work can be found in Tables S2 and S3 of the SI). It represents a robust and well-characterized collection for exploring structure–performance relationships in MOFs. By querying a total of 145 data samples, we acquire the 2.6% of the data space which simulates a realistic scenario of experimental setup with expensive annotated samples as we want to highlight the importance of performance improvement with minimum data space knowledge.

Similarly to the synthetic dataset, in Fig. 11, we showcase the mean absolute error score as a function of the number of samples annotated for O_2 and N_2 datasets. A general observation across both datasets (N_2 and O_2) is that random sampling, despite its simplicity, performs surprisingly well and serves as a challenging competitor for many state-of-the-art methods. Among the tested methods, query-by-committee struggles to perform well, showing higher MAE values throughout. The iGS

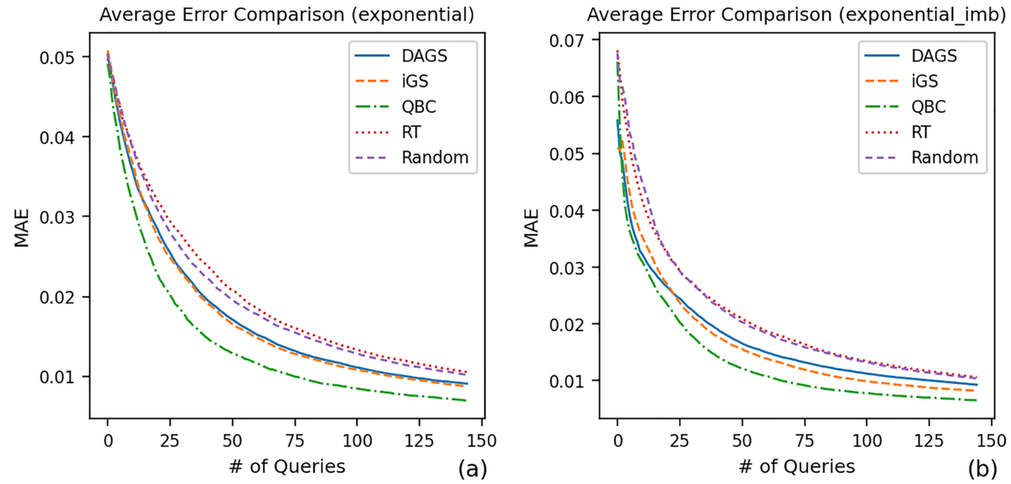

Fig. 10 MAE as a function of the number of queries for exponential (a) homogeneous and (b) heterogeneous datasets. A lower MAE means better predictive capabilities of the model.

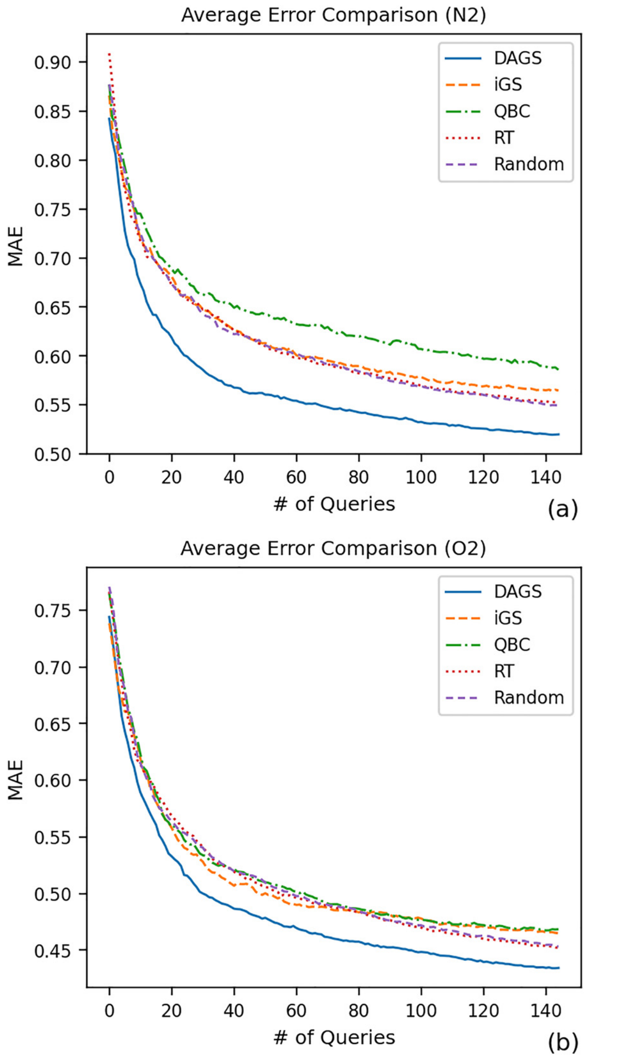

Fig. 11 MAE as a function of the number of queries for (a) N_2 and (b) O_2 diffusion in MOFs. A lower MAE means better predictive capabilities of the model.

method performs better than QBC but still lags behind both random and RT, with the latter closely matching the performance of random. The moderate performance of RT can be attributed to its reliance on randomness for sample selection, which assumes that points with similar characteristics (MOFs with similar chemical and structural properties) will exhibit similar diffusion performance. However, this assumption does not appear to hold true for diffusivity.

In contrast, the DAGS method demonstrates a clear advantage over all other methods. It not only achieves significantly lower MAEs in the early stages of sampling but also reaches a much lower plateau. For example, with 145 samples, RT and random both achieve a MAE of approximately 0.55 (N_2) and 0.45 (O_2). In comparison, our approach reaches the same MAE values with only approximately 60 samples (N_2) and 90 samples (O_2), drastically reducing the number of queries required to achieve similar accuracy by a factor of 2.4 and 1.6, respectively.

This efficiency translates directly to significant time and cost savings in the lab. By requiring fewer annotations to achieve the same prediction accuracy, DAGS minimizes experimental effort and resource use, making it a powerful and practical choice for active learning in materials science applications.

CH_4, H_2 and He diffusion in MOFs (three datasets)

The next three datasets used in our study are drawn from the work of Daglar *et al.*,[45] where the authors calculated the diffusivity of CH_4, H_2 and He in 5215, 2715 and 677 MOFs, respectively. By querying 150 training samples, the training size – design space size ratios are 2.9%, 5.5% and 22% which adds another experimental parameter for our results. Our XGBoost is trained on descriptors carrying chemical and structural information about each MOF, as shown in Table S2.

The results for CH_4, H_2 and He datasets (Fig. 12) largely align with the trends observed in the N_2 and O_2 datasets, as described previously. Random sampling continues to exhibit strong performance, proving itself as a robust baseline method. Query-by-committee (QBC), however, consistently

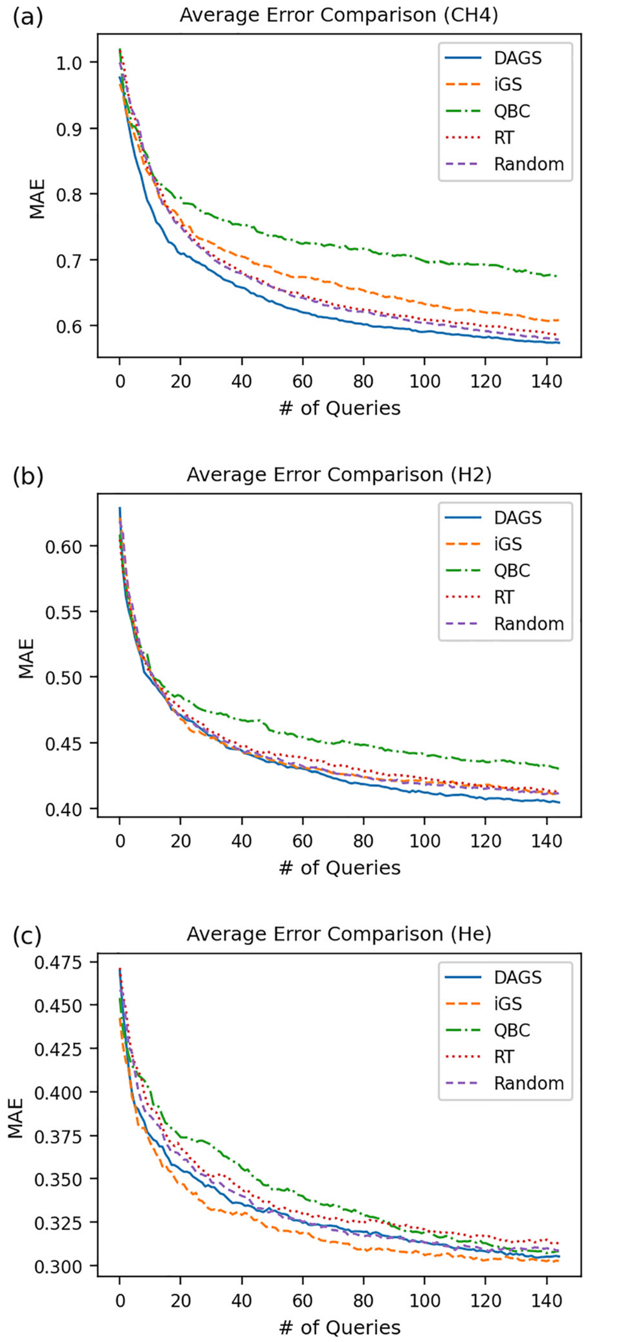

Fig. 12 MAE as a function of the number of queries for (a) CH_4, (b) H_2 and (c) He diffusion in MOFs. A lower MAE means better predictive capabilities of the model.

underperforms, showing the highest MAE values across all datasets. The iGS and RT methods again demonstrate improved performance, with RT closely matching random in most cases. Notably, the He dataset is the only scenario where iGS outperforms all other methods, achieving the lowest MAE, while our density-based method comes second, alongside random.

For the CH_4 and H_2 datasets, our method consistently outperforms all other approaches. It achieves a significantly lower MAE in the early stages of sampling and maintains its advantage throughout. For example, for CH_4, with 145 samples, RT and random reach a MAE of approximately 0.59 and 0.58, respectively, whereas the DAGS method achieves the same MAE with just 100 and 120 samples. This is a reduction of queries by 1.5 and 1.2 times, respectively. Similarly, for H_2, RT and random both achieve final MAEs of 0.41 at 145 samples, while our method reaches these values with only 90 samples, reducing the number of queries performed by a factor of 1.6. This significant reduction in required annotations translates directly to time and cost savings in the lab.

In the He dataset, however, iGS achieves the best performance, highlighting that certain methods may excel in specific scenarios. Nevertheless, the DAGS method still performs competitively, achieving a MAE score of 0.31 at approximately 90 samples, while RT and random achieve the same score with 145 samples and the iGS method gets the same score at 65 samples. These results rank DAGS second among the state-of-the-art methods as it requires 1.6 times less queries than RT and random and 1.5 times more queries than iGS for this specific task. We note that the He dataset has the highest ratio of training size to design space, with the training set covering nearly one quarter of the space. In such cases, a purely explorational method like iGS can afford to query all outliers and still have sufficient budget to cover the denser regions. In smaller design spaces, therefore, incorporating outlier information can improve performance relative to methods that prioritize dense regions. Overall, these findings highlight the robustness and versatility of the density-based approach, especially in datasets with high complexity and imbalanced distributions.

Conclusions

In this paper, we introduced DAGS, a density-aware active learning (AL) method designed to account for the heterogeneity of the design space during data selection. We compared DAGS against state-of-the-art AL frameworks that do not consider the underlying density distribution, as well as the non-trivial baseline of random sampling (RS). We first evaluated our method in four synthetic data spaces, demonstrating that while AL techniques significantly outperform RS in homogeneous spaces, their effectiveness diminishes when feasible data points are unevenly distributed. In heterogeneous synthetic spaces, DAGS exhibited consistent performance improvements over other frameworks due to its ability to prioritize sampling from denser regions. After establishing the importance of incorporating density distribution awareness in AL frameworks through synthetic datasets, we tested DAGS on real-world datasets. Specifically, we applied it to multiple MOF design spaces, which are characterized by pronounced non-uniformity in the distribution of feasible material feature points. Our method achieved remarkable results in four out of five design spaces, yielding lower mean absolute error (MAE) values compared to both random sampling and exploratory techniques like iGS. However, in smaller design spaces, such as that of He, DAGS exhibited slightly reduced performance. In such cases, informative data points may reside in sparser regions, which are less frequently explored by our method by design. This explains why other methods produced marginally better results in this specific scenario.

The objective of this work was to highlight the importance of incorporating spatial characteristics, such as heterogeneity, into the selection criterion of AL frameworks and to provide an initial step in this direction. As future work, we propose the development of density metrics for design spaces to better capture and exploit inherent heterogeneity. Such metrics could serve as valuable tools for selecting the most suitable AL framework for a given problem, ultimately improving predictive performance and data efficiency. We also propose further experimentation with our proposed method in new areas of application as well as in real experimental campaigns.

Conflicts of interest

There are no conflicts to declare.

Data availability

Supplementary information (SI): information and parameters of the regression models used in this work; features for training the regression models on the datasets of this work. Standard deviation of all methods in each dataset. See DOI: **https://doi.org/10.1039/d5cp02908b**.

The code for all AL methodologies demonstrated in this work can be found at **https://github.com/insane-group/Density_Aware_Greedy_Sampling**. The version of the code employed for this study is version 1.0.0. This study was carried out using publicly available data from ibarisorhan/MOF-O2N2 at **https://github.com/ibarisorhan/MOF-O2N2/blob/main/mofScripts/MOFdata.csv** with [accession number], for the O_2 and N_2 diffusion in MOFs cases, and from hdaglar/MOF-basedMMMs_ML at **https://github.com/hdaglar/MOF-basedMMMs_ML/blob/main/rawdata.zip** with [accession number], for the CH_4, H_2 and He diffusion in MOFs cases.

Acknowledgements

This work was supported by the European Union's Horizon Europe research and innovation programme under grant agreement no. 101135927 (NOUS).

References

1 C. L. Essmann, M. Elmi, C. Rekatsinas, N. Chrysochoidis, M. Shaw, V. Pawar, M. A. Srinivasan and V. Vavourakis, The influence of internal pressure and neuromuscular agents on C. elegans biomechanics: an empirical and multi-compartmental in silico modelling study, *Front. Bioeng. Biotechnol.*, 2024, **12**, 1–14.

2 J. Schrier, A. J. Norquist, T. Buonassisi and J. Brgoch, In Pursuit of the Exceptional: Research Directions for Machine Learning in Chemical and Materials Science, *J. Am. Chem. Soc.*, 2023, **145**, 21699–21716.

3 N. Kühl, M. Goutier, L. Baier, C. Wolff and D. Martin, Human vs. supervised machine learning: Who learns patterns faster?, *Cognit. Syst. Res.*, 2022, **76**, 78–92.

4 P. Schoenegger, S. Greenberg, A. Grishin, J. Lewis and L. Caviola, AI can outperform humans in predicting correlations between personality items, *Commun. Psychol.*, 2025, **3**, 23.

5 C. López, Artificial Intelligence and Advanced Materials, *Adv. Mater.*, 2023, **35**, 2208683.

6 L. Gao, J. Lin, L. Wang and L. Du, Machine Learning-Assisted Design of Advanced Polymeric Materials, *Acc. Mater. Res.*, 2024, **5**, 571–584.

7 G. Huang, Y. Guo, Y. Chen and Z. Nie, Application of Machine Learning in Material Synthesis and Property Prediction, *Materials*, 2023, **16**, 5977.

8 J. Zhu, H. Wang, B. K. Tsou and M. Ma, Active learning with sampling by uncertainty and density for data annotations, *IEEE Trans. Audio Speech Lang. Process.*, 2010, **18**, 1323–1331.

9 K. Mukherjee, E. Osaro and Y. J. Colón, Active learning for efficient navigation of multi-component gas adsorption landscapes in a MOF, *Digital Discovery*, 2023, **2**, 1506–1521.

10 D. Wu, Pool-Based Sequential Active Learning for Regression, *IEEE Trans. Neural Networks Learn. Syst.*, 2019, **30**, 1348–1359.

11 Z. Liu, X. Jiang, H. Luo, W. Fang, J. Liu and D. Wu, Pool-based unsupervised active learning for regression using iterative representativeness-diversity maximization (iRDM), *Pattern Recognit. Lett.*, 2021, **142**, 11–19.

12 D. A. Cohn, Z. Ghahramani and M. I. Jordan, Active learning with statistical models, *J. Artif. Int. Res.*, 1996, **4**, 129–145.

13 C. E. Shannon, A Mathematical Theory of Communication, *Bell Syst. Tech. J.*, 1948, **27**, 379–423.

14 B. Settles and M. Craven, in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, ed. M. Lapata and H. T. Ng, Association for Computational Linguistics, Honolulu, Hawaii, 2008, pp. 1070–1079.

15 I. Dagan and S. P. Engelson, in *Machine Learning Proceedings 1995*, ed. A. Prieditis and S. Russell, Morgan Kaufmann, San Francisco (CA), 1995, pp. 150–157.

16 B. Settles, M. Craven and S. Ray, *Proceedings of the 21st International Conference on Neural Information Processing Systems*, Curran Associates Inc., Red Hook, NY, USA, 2007, pp. 1289–1296.

17 D. Wu, C. T. Lin and J. Huang, Active learning for regression using greedy sampling, *Inf. Sci.*, 2019, **474**, 90–105.

18 D. Wu, C.-T. Lin and J. Huang, Active learning for regression using greedy sampling, *Inf. Sci.*, 2019, **474**, 90–105.

19 H. Liu, B. Yucel, B. Ganapathysubramanian, S. R. Kalidindi, D. Wheeler and O. Wodo, Active learning for regression of structure–property mapping: the importance of sampling and representation, *Digital Discovery*, 2024, **3**, 1997–2009.

20 W. Cai, Y. Zhang and J. Zhou, in *2013 IEEE 13th International Conference on Data Mining*, IEEE, Dallas, TX, USA, 2013, pp. 51–60.

21 W. Cai, M. Zhang and Y. Zhang, Batch Mode Active Learning for Regression With Expected Model Change, *IEEE Trans. Neural Networks Learn. Syst.*, 2017, **28**, 1668–1681.

22 B. Lakshminarayanan, D. M. Roy and Y. W. Teh, *Proceedings of the 27th International Conference on Neural Information Processing Systems*, MIT Press, Cambridge, MA, USA, 2014, vol. 2, pp. 3140–3148.

23 J. Goetz, A. Tewari and P. Zimmerman, Montréal, Canada, 2018.

24 D. Holzmüller, V. Zaverkin, J. Kästner and I. Steinwart, A framework and benchmark for deep batch active learning for regression, *J. Mach. Learn. Res.*, 2023, **24**, 1–81.

25 M. Wang, F. Min, Z.-H. Zhang and Y.-X. Wu, Active learning through density clustering, *Expert Syst. Appl.*, 2017, **85**, 305–317.

26 S. Kee, E. Del Castillo and G. Runger, Query-by-committee improvement with diversity and density in batch active learning, *Inf. Sci.*, 2018, **454–455**, 401–418.

27 T. Wang, X. Zhao, Q. Lv, B. Hu and D. Sun, in *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, IEEE, Dalian, China, 2021, pp. 156–161.

28 Y. Kim and B. Shin, in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ACM, Washington DC USA, 2022, pp. 804–812.

29 R. Burbidge, J. J. Rowland and R. D. King, in *Intelligent Data Engineering and Automated Learning – IDEAL 2007*, ed. H. Yin, P. Tino, E. Corchado, W. Byrne and X. Yao, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007, vol. 4881, pp. 209–218.

30 A. Jose, J. P. A. de Mendonça, E. Devijver, N. Jakse, V. Monbet and R. Poloni, Regression tree-based active learning, *Data Min. Knowl. Discovery*, 2024, **38**, 420–460.

31 Z. Chen, M. C. Wasson, R. J. Drout, L. Robison, K. B. Idrees, J. G. Knapp, F. A. Son, X. Zhang, W. Hierse, C. Kühn, S. Marx, B. Hernandez and O. K. Farha, The state of the field: from inception to commercialization of metal–organic frameworks, *Faraday Discuss.*, 2021, **225**, 9–69.

32 P. Peng, H. Z. H. Jiang, S. Collins, H. Furukawa, J. R. Long and H. Breunig, *Long Duration Energy Storage Using Hydrogen in Metal–Organic Frameworks: Opportunities and Challenges*, DOI: **10.1021/acsenergylett.4c00894**.

33 Y. Zhang, H. Ben Yin, L. Huang, L. Ding, S. Lei, S. G. Telfer, J. Caro and H. Wang, MOF membranes for gas separations, *Prog. Mater. Sci.*, 2025, 101432.

34 I. Guyon, G. C. Cawley, G. Dror and V. Lemaire, in *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, ed. I. Guyon, G. Cawley, G. Dror, V. Lemaire and A. Statnikov, PMLR, Sardinia, Italy, 2011, vol. 16, pp. 19–45.

35 G. C. Cawley, in *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, ed. I. Guyon, G. Cawley, G. Dror, V. Lemaire and A. Statnikov, PMLR, Sardinia, Italy, 2011, vol. 16, pp. 47–57.

36 A. Tifrea, J. Clarysse and F. Yang, in *Proceedings of the 40th International Conference on Machine Learning*, ed. A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato and J. Scarlett, PMLR, 2023, **vol. 202**, pp. 34222–34262.

37 R. Burbidge, J. J. Rowland and R. D. King, *Active learning for regression based on query by committee*, *Lect. Notes Comput. Sci. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinforma.*, 2007, 4881 LNCS, pp. 209–218.

38 T. Chen and C. Guestrin, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, **vol. 13**, pp. 785–794.

39 B. Settles, *Active Learning Literature Survey*, University of Wisconsin–Madison, Wisconsin, 2009.

40 F. Di Fiore, M. Nardelli and L. Mainini, Active Learning and Bayesian Optimization: A Unified Perspective to Learn with a Goal, *Arch. Comput. Methods Eng.*, 2024, **31**, 2985–3013.

41 P. Krokidas, S. Moncho, E. N. Brothers and I. G. Economou, Defining New Limits in Gas Separations Using Modified ZIF Systems, *ACS Appl. Mater. Interfaces*, 2020, **12**, 20536–20547.

42 G. Ignacz, L. Bader, A. K. Beke, Y. Ghunaim, T. Shastry, H. Vovusha, M. R. Carbone, B. Ghanem and G. Szekely, Machine learning for the advancement of membrane science and technology: A critical review, *J. Membr. Sci.*, 2024, 123256.

43 P. Krokidas, M. Castier and I. G. Economou, Computational Study of ZIF-8 and ZIF-67 Performance for Separation of Gas Mixtures, *J. Phys. Chem. C*, 2017, **121**, 17999–18011.

44 I. B. Orhan, H. Daglar, S. Keskin, T. C. Le and R. Babarao, Prediction of O2/N2Selectivity in Metal-Organic Frameworks via High-Throughput Computational Screening and Machine Learning, *ACS Appl. Mater. Interfaces*, 2022, **14**, 736–749.

45 H. Daglar and S. Keskin, Combining Machine Learning and Molecular Simulations to Unlock Gas Separation Potentials of MOF Membranes and MOF/Polymer MMMs, *ACS Appl. Mater. Interfaces*, 2022, **14**, 32134–32148.