Density-Aware Active Learning for Materials Discovery: A Case Study on Functionalized Nanoporous Materials

V. Gkatsis,[1,2] P. Maratos,[3] C. Rekatsinas,[2] G. Giannakopoulos,[2,4] P. Krokidas[2]

[1]Department of Informatics and Telecommunications, National and Kapodistrian University, Athens

[2]Institute of Informatics & Telecommunications, National Center for Scientific Research "Demokritos", 15341 Agia Paraskevi Attikis, Greece.

[3]School of Electrical & Computer Engineering, National Technical University of Athens, Athens, Greece

[4]SciFY PNPC, Agia Paraskevi 15310, Greece

Table of Abbreviations

AL	Active Learning
B-EMCM	Batch mode Expected Model Change Maximization
BMDAL	Batch Model Deep Active Learning
COFs	Covalent Organic Frameworks
DAGS	Density Aware Greedy Sampling
DS	Design Space
EMCM	Expected Model Change Maximization
GP	Gaussian Processes
GS	Greedy Sampling
iGS	Improved Greedy Sampling
KDE	Kernel Density Estimation
MAE	Mean Absolute Error
ML	Machine Learning
MOFs	Metal Organic Frameworks

QBC	Query by Committee
RS	Random Sampling
RT-AL	Regression Tree-based Active Learning
SUD	Sampling by Uncertainty and Density

Regression Models

In this section we provide a short description of the different models that we have used in our work. Table S1 shows the parameters that we have used and their corresponding values, as well as the python libraries that we used. The reported hyperparameters for the XGboost model have been taken from previous work of the writers[1]. As for the other models the values have been chosen after heuristic experimentation in small subsets of the dataset. We report the selected values for reproducibility reasons.

XGBoost

XGBoost is a gradient boosting algorithm meaning it works by sequentially building decision trees where each tree attempts to correct the errors made by the previous one. It is widely used in machine learning as it is known for its speed, accuracy, and handling of overfitting through regularization. We selected it to be our main model, that being, the one that gets trained with the selected data samples, in order to test whether a powerful model can give good results with a limited amount of training data. We also used XGBoost as one of the three models used in the query by committee method.

Decision Tree

The Decision Tree model is very simple as it splits the data into subsets based on the most significant feature at each step, forming a tree-like structure. It has been used as the basis for the Regression Tree Active Learning method as described in the manuscript.

Random Forest

A Random Forest algorithm is an ensemble model that uses several decision trees, by training them on various sub-samples of the dataset. It also uses averaging to improve the predictive

accuracy and control over-fitting. Random Forest was used as one of the voting models on the query by committee method.

Gaussian Processes

Gaussian Processes is a probabilistic supervised machine learning algorithm widely used because of its very good predictive capabilities. It is also highly customizable since a variety of kernels can be used, and even custom ones can be made tailored to a specific problem. Finally its probabilistic nature makes it easy to estimate confidence intervals and hence decide if some particular region needs refitting with new observations. We selected it as our last method used in the query by committee method due to its different philosophy compared to the other models.

Table S1. Parameters used for the training of Machine Learning models.

Model	Library	Parameters	Value
XGBoost	xgboost https://xgboost.readthedocs.io/en/release_3.0.0/	n_estimators	500
		max_depth	5
		eta	0.07
		subsample	0.75
		colsample_bytree	0.7
		reg_lambda	0.4
		reg_alpha	0.13
		random_state	6410
Gaussian Process	scikit-learn https://scikit-learn.org/stable/modules/gaussian_process.html	Rational Quadratic Kernel:	
		length_scale	1.0
		alpha	1.5
Decision Trees	scikit-learn https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.ht	random_state	

	ml		
		min_samples_leaf	
Random Forest	scikit-learn https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html	random_state	0
		max_depth	5

Training Features

Tables S2 and S3 show the features that we have extracted from N_2, O_2 and CH_4, H_2 and He datasets and that have been used for the training of our machine learning models. These features have been selected by the material science expert in our team, based on their significance for the structure of the material. More specifically, differences in values of these features are considered enough to distinguish one material from another and hence they are expected to train the machine learning algorithm efficiently.

Table S2. Features considered when training the regression models for the prediction of N_2 and O_2 diffusivities in MOFs.

Group	Description
Geometric	Largest cavity diameter
	Pore-limiting diameter
	Largest sphere along free path
	Volume of unit cell
	Gravimetric accessible surface area (m^2/g)
	Volumetric accessible surface area (m^2/cm^3)
	Gravimetric non-accessible surface area (m^2/g)
	VF
	Accessible volume (cm^3/g)
	Non-accessible volume (cm^3/g)
Atom Type	Metal types present (categorical)
	Number of specified element atoms in unit cell: H, C, N, F, Cl, Br, V, Cu, Zn, Zr
Chemical	Total degree of unsaturation

	Metallic percentage
	Oxygen-to-metal ratio
	Electronegative-to-total ratio
	Weighted electronegativity per atom
	Nitrogen-to-oxygen ratio

Table S3. Features considered when training the regression models for the prediction of CH_4, H_2 and He diffusivities in MOFs.

Feature (unit)	Symbol
largest cavity diameter (Å)	LCD
pore limiting diameter (Å)	PLD
pore size ratio	LCD/PLD
density (g/cm³)	P
pore volume (cm³/g)	PV
porosity	Φ
surface area (m²/g)	SA
carbon percentage	C%
hydrogen percentage	H%
nitrogen percentage	N%
oxygen percentage	O%
halogen (Br, Cl, F, I) percentage	halogen%
metalloid (As, B, Ge, Te, Sb, Si) percentage	metalloid%
ametal (Se, S, P) percentage	ametal%
metal percentage	metal%
total degree of unsaturation	TDU
degree of unsaturation	DU
metallic percentage (# of metal/# of C atoms)	MP
oxygen-to-metal ratio	O-to-M
heat of adsorption (kJ/mol)	$Q_{st\text{ZO}}^{0}$

Standard Deviation

We have described our experimental workflow in the subsection *Computational Experimental Evaluation Work-flow* of the manuscript, where we have also presented the average MAE scores of each method for each dataset across the 10 experimental runs. In this section we present the average standard deviation for all sampling methods. For each dataset we report a separate table.

There we show the average standard deviation of each method, across the 10 runs, for 1, 25, 50, 100, 125, and 145 sampling queries.

Real Datasets

CH4							
Method Name	Standard Deviation On #Queries						
	1	25	50	75	100	125	145
DAGS	0.1386	0.0523	0.0355	0.0326	0.0285	0.0269	0.0255
iGS	0.1158	0.0665	0.0609	0.0439	0.0375	0.0339	0.0290
QBC	0.1556	0.0737	0.0640	0.0648	0.0595	0.0584	0.0503
RT	0.1373	0.0531	0.0393	0.0317	0.0279	0.0267	0.0221
Random	0.1336	0.0563	0.0390	0.0336	0.0288	0.0244	0.0244

H2							
Method Name	Standard Deviation On #Queries						
	1	25	50	75	100	125	145
DAGS	0.1264	0.0279	0.0236	0.0213	0.0214	0.0214	0.0210

iGS	0.1047	0.0317	0.0258	0.0241	0.0251	0.0244	0.0249
QBC	0.0958	0.0387	0.0297	0.0310	0.0265	0.0286	0.0273
RT	0.0948	0.0320	0.0282	0.0272	0.0250	0.0219	0.0222
Random	0.1139	0.0311	0.0232	0.0220	0.0216	0.0209	0.0213

He							
Method Name	Standard Deviation On #Queries						
	1	25	50	75	100	125	145
DAGS	0.0753	0.0389	0.0320	0.0303	0.0287	0.0274	0.0269
iGS	0.0716	0.0370	0.0355	0.0345	0.0332	0.0321	0.0321
QBC	0.0710	0.0435	0.0406	0.0367	0.0319	0.0304	0.0305
RT	0.0719	0.0355	0.0315	0.0274	0.0282	0.0291	0.0285
Random	0.0792	0.0366	0.0338	0.0297	0.0281	0.0288	0.0260

N2							
Method Name	Standard Deviation On #Queries						
	1	25	50	75	100	125	145

DAGS	0.1377	0.0446	0.0289	0.0273	0.0246	0.0235	0.0236
iGS	0.1751	0.0542	0.0434	0.0346	0.0326	0.0334	0.0344
QBC	0.1663	0.0709	0.0625	0.0574	0.0510	0.0439	0.0382
RT	0.1902	0.0511	0.0387	0.0320	0.0312	0.0295	0.0282
Random	0.1728	0.0519	0.0355	0.0334	0.0298	0.0292	0.0250

O2							
Method Name	Standard Deviation On #Queries						
	1	25	50	75	100	125	145
DAGS	0.1122	0.0293	0.0241	0.0178	0.0183	0.0167	0.0173
iGS	0.1279	0.0540	0.0431	0.0398	0.0329	0.0279	0.0253
QBC	0.1444	0.0419	0.0324	0.0300	0.0253	0.0244	0.0242
RT	0.1165	0.0453	0.0303	0.0248	0.0214	0.0175	0.0181
Random	0.1336	0.0393	0.0302	0.0273	0.0228	0.0188	0.0190

Synthetic Dataset

Forrester							
Method Name	Standard Deviation On #Queries						
	1	25	50	75	100	125	145
DAGS	0.3518	0.0913	0.0593	0.0393	0.0270	0.0208	0.0209
iGS	0.3563	0.0926	0.0595	0.0456	0.0324	0.0236	0.0202
QBC	0.3582	0.3985	0.4080	0.4968	0.4552	0.3091	0.2643
RT	0.3643	0.2184	0.1190	0.091	0.0647	0.0505	0.0349
Random	0.3348	0.2092	0.1288	0.0772	0.0602	0.0547	0.0431

Forrester Imbalanced							
Method Name	Standard Deviation On #Queries						
	1	25	50	75	100	125	145
DAGS	0.4528	0.1354	0.0534	0.0391	0.0310	0.0256	0.0223
iGS	0.4406	0.2153	0.0849	0.0731	0.0561	0.0438	0.0386
QBC	0.4420	0.4757	0.6838	0.7833	0.7260	0.6621	0.5858
RT	0.4489	0.3472	0.1265	0.0905	0.0611	0.0546	0.0493
Random	0.4919	0.1707	0.100	0.0551	0.0417	0.0353	0.0271

Jump Forrester	
Method Name	Standard Deviation On #Queries

	1	25	50	75	100	125	145
DAGS	0.6411	0.1931	0.1230	0.0941	0.0859	0.0649	0.0625
iGS	0.6552	0.1750	0.1016	0.0759	0.0597	0.0558	0.0528
QBC	0.6631	0.9701	1.2064	1.1135	0.9654	0.9706	0.9019
RT	0.6408	0.4754	0.3334	0.2541	0.2160	0.1931	0.1795
Random	0.5708	0.2578	0.2001	0.1488	0.1317	0.1290	0.1166

Jump Forrester Imbalanced							
Method Name	Standard Deviation On #Queries						
	1	25	50	75	100	125	145
DAGS	0.9606	0.5048	0.2527	0.1702	0.0778	0.0448	0.0420
iGS	0.7503	0.8401	0.5290	0.3317	0.1918	0.0894	0.0536
QBC	0.9833	1.2259	1.4751	1.5652	0.9841	0.8869	0.8351
RT	1.084	0.7468	0.3978	0.2785	0.2507	0.2125	0.1965
Random	0.8165	0.4300	0.3019	0.1823	0.1502	0.1203	0.1162

Gaussian							
Method Name	Standard Deviation On #Queries						
	1	25	50	75	100	125	145
DAGS	0.0439	0.0093	0.0083	0.0089	0.0080	0.0082	0.0090
iGS	0.0476	0.0108	0.0092	0.0098	0.0093	0.0087	0.0087
QBC	0.0503	0.187	0.138	0.0083	0.0064	0.0057	0.0056
RT	0.0361	0.0132	0.0101	0.0095	0.0082	0.0084	0.0073
Random	0.0527	0.0101	0.0086	0.0075	0.0084	0.0087	0.0086

Gaussian Imbalanced							
Method Name	Standard Deviation On #Queries						
	1	25	50	75	100	125	145
DAGS	0.0366	0.0142	0.0096	0.0102	0.0100	0.0091	0.0087
iGS	0.0612	0.0190	0.0162	0.0153	0.0153	0.0158	0.0150
QBC	0.0617	0.0222	0.0126	0.0090	0.0079	0.0067	0.0067
RT	0.0612	0.0284	0.0169	0.0104	0.0095	0.0086	0.0077
Random	0.0536	0.0145	0.0122	0.0088	0.0090	0.0086	0.0086

Exponential

Method Name	Standard Deviation On #Queries						
	1	25	50	75	100	125	145
DAGS	0.0053	0.0028	0.0017	0.0010	0.0009	0.0007	0.0007
iGS	0.0063	0.0026	0.0015	0.0012	0.0010	0.0009	0.0008
QBC	0.0057	0.0022	0.0011	0.0006	0.0006	0.0005	0.0005
RT	0.0049	0.0033	0.0027	0.0018	0.0014	0.0012	0.0010
Random	0.0053	0.0036	0.0027	0.0019	0.0014	0.0011	0.0009

Exponential Imbalanced							
Method Name	Standard Deviation On #Queries						
	1	25	50	75	100	125	145
DAGS	0.0057	0.0022	0.0016	0.0011	0.0009	0.0007	0.0007
iGS	0.0052	0.0023	0.0014	0.0009	0.0008	0.0007	0.0006
QBC	0.0106	0.0022	0.0011	0.0008	0.0006	0.0005	0.0005
RT	0.0079	0.0041	0.0030	0.0023	0.0017	0.0014	0.0012
Random	0.0072	0.0037	0.0027	0.0017	0.0014	0.0012	0.0010

Bibliography

1. Krokidas, P., Kainourgiakis, M., Steriotis, T. & Giannakopoulos, G. Inverse design of ZIFs through artificial intelligence methods. *Phys. Chem. Chem. Phys.* **26**, 25314–25318 (2024).