Contents lists available at ScienceDirect

Sustainable Chemistry for the Environment

journal homepage: www.editorialmanager.com/scenv

Comparison of machine learning approaches for the identification of top-performing materials for hydrogen storage

Antonios P. Sarikas [a], George S. Fanourgakis [a,*,1], Konstantinos Gkagkas [b], George E. Froudakis [a,*]

[a] Department of Chemistry, University of Crete, Voutes, Campus, GR-700013 Heraklion, Greece
[b] Advanced Technology Division, Toyota Motor Europe NV/SA, Technical Center, Hoge Wei, 33B, Zaventem B-1930, Belgium

ARTICLE INFO

Keywords:
Machine learning
Metal-organic frameworks
Hydrogen storage
Nanomaterials
Nanoporous materials
Gas adsorption

ABSTRACT

Unique properties of Metal-Organic Frameworks (MOFs), such as their extremely high surface areas and porosity, render them as one of, if not the most promising adsorbents for gas storage applications. However, their extremely flexible and tunable nature has resulted in an enormous expansion of the available material pool which in turn begs the question as to whether an efficient identification of the best materials is feasible. In the last years, Machine Learning (ML) techniques have been extensively applied for the exploration of large material databases, since they can significantly accelerate this process. In this work, “traditional” ML models and models based on our recently developed iterative self-consistent approach are compared with respect to their ability to efficiently identify the best materials of a database. As a case study, we have used hydrogen adsorption in MOFs at different thermodynamic conditions. Despite their high accuracy, traditional models struggle to pinpoint the best materials, regarding usable gravimetric uptake, without compromising computational resources. On the other hand, self-consistent models can even reduce by two orders of magnitude the amount of reference data required for the identification of the best gravimetric materials compared to traditional ones. Notably, 300 training samples are enough for the SC models to correctly identify the top-100 gravimetric materials of a database. Nevertheless, both type of models underperform when they are queried for the top-performing materials with regards to usable volumetric uptake.

1. Introduction

Owing to their inherent properties, such as high surface area and void fraction, Metal-Organic Frameworks (MOFs) are prominent candidates for applications involving gas adsorption [1,2]. A prime example is hydrogen storage, where materials exhibiting high H_2 uptakes along with the ability to rapidly adsorb and release the latter are needed. The fast kinetics, reversibility and their exceptional H_2 capacities, mark them as one of the most promising hydrogen sorbents [3].

MOFs are nanoporous crystalline materials consisting of a metal ion or a metal cluster and organic linkers, collectively known as building blocks [4]. The metal corners are connected in space with organic linkers forming a three dimensional network. The amount of potential components is enormous and as a result the number of structures that can be realized, either experimentally [5,6] or in silico, is unlimited [7–9]. Computer-aided design has been adopted in the last years, giving birth to large databases of hypothetical MOFs. One of the first hypothetical databases were constructed by Wilmer et al. [10] where the combination of 102 different building blocks, generated 137,953 structures. In contrast to this bottom-up approach, a top-down generator was also introduced [11], where topological fingerprints are used as a template for the construction of hypothetical MOFs. More recently, over 100 trillions MOFs can be constructed with the help of an advanced porous material generator [12]. This vast expansion of the available material pool unavoidably leads to the following challenge: how the optimal materials for a given application can be identified from this huge ”search space” in an efficient manner?

Over the past few decades, molecular simulations served as the main tool to speed up the discovery and performance characterization of new materials. Particularly, grand canonical Monte Carlo (GCMC)

[*] Corresponding authors.
E-mail addresses: fanourg@uoc.gr (G.S. Fanourgakis), frudakis@uoc.gr (G.E. Froudakis).
[1] Present address: Laboratory of Quantum and Computational Chemistry, Department of Chemistry, Aristotle University of Thessaloniki, GR-54124, Thessaloniki, Greece

https://doi.org/10.1016/j.scenv.2023.100056
Received 22 July 2023; Received in revised form 30 November 2023; Accepted 30 November 2023
Available online 5 December 2023

simulations have been widely performed for the assessment of nanoporous materials regarding uptake capacities of various gas species [13–20]. Despite their efficiency compared to traditional approaches, such as experimental synthesis and characterization, applying brute force screening in databases of the aforementioned size is apparently prohibitive. Notice that, a single GCMC simulation (under certain thermodynamic conditions) for the ZIF-8/H_2O system required 115 days to converge, as reported by [21]. Although simulations for gases such as CH_4 and H_2 require significantly less time (minutes or few hours per material), they are still costly if rapid screening of huge databases is desired.

In the era of big data, supervised ML methods [22–26] can pave the way towards efficient screening as they require a substantially smaller amount of data compared to traditional methods. These data, which can be collected through experiments or simulations, serve to train the ML algorithm. The ML model obtained after the end of training phase, approximates the relationship between some input variables and an output variable. In machine learning jargon, the input variables are called descriptors (or features) and the output variable is called label, which can be either continuous (target) or categorical (class). An important step prior to training, is the selection of the descriptors. The latter should be chosen in a reasonable manner, so the algorithm is able to extract a meaningful structure-property relationship. In the case of nanoporous materials, various structural features such as surface area, void fraction and pore volume have been employed as descriptors, giving rise to ML models of satisfactory accuracy [27]. Performance enhancements can be attained by introducing more complex descriptors, taking into account the chemical environment and the energetic landscape of the pore [28–34]. Employing such a set of descriptors, can reduce the training set size required to exceed a threshold accuracy compared to structural features [35–37]. Once the ML model has been built, screening of large databases can be performed in just few minutes (or even seconds), that is in less than the cost of a single GCMC simulation.

Regarding the identification of materials with the potential of exhibiting high hydrogen capacities, ML predictive models have been developed and employed for the in silico screening of large datasets. Thornton et al. [38] used ML based approaches for the high-throughput screening of nanoporous materials for hydrogen storage at room and at cryogenic temperatures and at pressures between 100 and 1 bar. More than 850,000 hypothetical and synthesized porous crystalline materials were examined. An iterative procedure was employed for the identification of the most promising candidates. During this procedure successive ML models were trained using training sets of incremental sizes. The materials added each time in the training data, were determined by an ML model during the previous iteration. In total, reference data for only a small portion of the dataset (3,000 out of the 850,000 MOFs) were computed by GCMC simulations. One of the most important conclusions of the study was that many of the top-performing materials for hydrogen storage have been already synthesized.

In one of the most recent and extensive studies regarding hydrogen storage in MOFs [39], ML models were employed to screen a vast database of approximately 918,000 MOFs. For 98,695 MOFs GCMC simulations were initially performed and their working capacities at four different thermodynamic conditions were determined. Prior to the exploration of the database, 14 ML algorithms were trained to predict the working capacities at the various conditions. Information for 74,201 MOFs was used for the training of the ML algorithms while the remaining 24,674 MOFs were used for assessing the performance of the predictive models. In all cases 7 structural features of MOFs were used as descriptors. The extremely randomized trees (ERT) algorithm showed the best performance among the algorithms examined. In terms of statistical accuracy the values of the coefficient of determination (R^2) statistical metric were impressively high ranging between 0.967 and 0.997 for the four cases examined. The predictive model was subsequently applied to the remaining unknown MOFs of the database and 8,282 MOFs were identified with the potential of their working capacities to exceed that of the state-of-the-art materials.

Although highly accurate ML models are desirable, minimizing the amount of reference data required for the training of the ML algorithms is of equal importance if we are interested in fast screening of large databases. In the present work, we will argue that although ML models with excellent performance—in terms of statistical accuracy—may be constructed, they should be cautiously used for screening purposes during the identification of the most promising materials. For this reason we will evaluate two different approaches: in the first most commonly used approach, ML algorithms will be trained on a portion of the available data and will be evaluated based on their predictions for the top-performing materials. In the second approach we will employ our previously developed iterative, self-consistent (SC) approach [40] which attempts to directly identify promising candidates, using the lesser possible information in a fashion similar to that used by [38].

Our main conclusion is that even though ML models based on the first approach can be extremely accurate, they still struggle to pinpoint the best materials of a database, at least if they are challenged to do it efficiently. The second approach is an efficient alternative to the "traditional" one, but there are still limitations that bound the identification of the best materials.

2. Methodology

2.1. Computational algorithm

The construction of accurate generic ML models requires large training sets since many regions of the feature space must be covered. However, if we are interested in the identification of the top-performing materials (for a given application), computational cost should be spent wisely in order their regions to be efficiently explored. Based on this idea, our algorithm dictates the way the simulations are performed, in such manner that the best structures are identified with the least possible computational effort. Our proposed approach is schematically presented in Fig. 1.

As a first step, a small random subset of materials from a large database is selected and their adsorption uptakes are calculated (e.g. through GCMC simulations), forming an initial training set. The latter serves to build (train) an ML model which in turn makes predictions for all the materials in the database. Next, the training set is enriched by a number of top-performing structures as predicted by the ML model. The capacity determination, training set augmentation and capacity

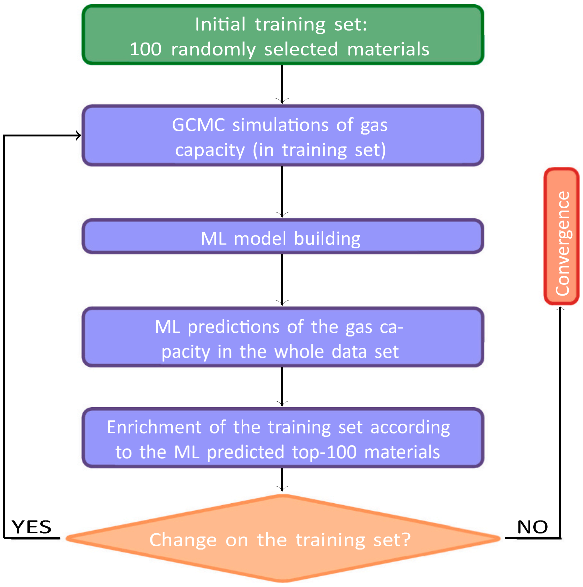
Initial training set:
100 randomly selected materials
GCMC simulations of gas capacity (in training set)
ML model building
ML predictions of the gas capacity in the whole data set
Enrichment of the training set according to the ML predicted top-100 materials
Change on the training set?
YES
NO
Convergence

Fig. 1. Workflow of the proposed SC approach.

prediction steps are coupled in an iterative manner. This process continues until all the top-performing structures as predicted by the ML model are already in the training set, where the procedure is said to have converged.

The number of materials in the initial training set and the number of materials that are added at each iteration must be specified prior to the start of the iterative procedure. Information regarding performance variation with respect to the aforementioned parameters can be found on our previous work [40]. In this study, the same value was chosen for both parameters, namely 100. Ideally, the top-100 materials should be included in the converged training set. It is worth noting that the number of materials that are added in each iteration may be less than 100, since some structures may be already added from previous iterations.

2.2. Dataset and ML descriptors

In this work, a labeled dataset sourced from 19 databases (including experimental ones such as the CoRE 2019 [5]), was employed. The dataset was developed and used by Siegel and his coworkers [39], and it is deposited at the HyMARC data hub [41]. This dataset was used throughout this work for the training and test phase of the various ML algorithms. It contains information for seven crystallographic features and four H_2 usable capacities of 98,695 MOFs.

The crystallographic properties include: density, gravimetric surface area, volumetric surface area, void fraction, pore volume, largest cavity diameter and pore limiting diameter. On the other hand, usable capacities include gravimetric and volumetric capacities under two operating conditions:

1. pressure swing (PS), between 100 and 5 bar at 77 K
2. temperature-pressure swing (TPS), between 77 K, 100 bar and 160 K, 5 bar

Gravimetric capacities are measured in wt% while volumetric capacities in gL^{-1}.

3. Results and discussion

3.1. Evaluation of ML models

An extensive study for the predictive performance of 15 different algorithms has been already performed by Ahmed et al. [39]. Here, for the sake of discussion we developed as well a number of predictive models using 4 different algorithms, namely, the decision trees (DT), random forest (RF), extremely randomized trees (ERT) and gradient boosted trees (GBT), as implemented in the scikit-learn package [42] (version 0.22.2). Notice that in the previous work was concluded that the ERT algorithm provides overall the most accurate predictions. The protocol we used for the model development is the following: we first create training datasets for the ML algorithms by randomly selecting a number of MOFs from the pool of the 98,695 materials. The size of the training set varies between 100 and 20,000 while the remaining materials are used as test sets for the evaluation of the model performance. Predictions are made to the test data that were kept unknown to the ML algorithm during the training phase. In order to avoid any bias from the choice of the training data (in particular for small training sizes) we repeated the previous procedure 100 times using different random splits. The reported results correspond to the average of the 100 results obtained from this procedure.

The results of all ML algorithms for the training set size 10,000 are tabulated in Table 1. It can be seen that the performances are very similar to those reported by Siegel, while the small differences should be assigned to the slightly different evaluation protocols and to the different training set sizes employed in the two studies. In any case, it is seen that the ML algorithms are more accurate for the UG instead of UV.

Table 1
Performance metrics for various ML algorithms, trained on 10,000 structures. Metrics are calculated from predictions for the remaining 98,695–10,000 = 88,695 structures. c.u., capacity units.

UG at PS	R^2	MAE (c.u.)	RMSE (c.u.)	WAPE (%)
DT	0.994	0.203	0.285	5.08
GBT	0.996	0.157	0.219	3.93
ERT	0.997	0.148	0.210	3.69
RF	0.997	0.148	0.208	3.69
UG at TPS				
DT	0.993	0.249	0.341	3.72
GBT	0.995	0.201	0.271	3.00
ERT	0.996	0.183	0.255	2.74
RF	0.996	0.182	0.251	2.724
UV at PS				
DT	0.965	1.429	2.07	6.42
GBT	0.979	1.111	1.583	4.99
ERT	0.981	1.056	1.541	4.74
RF	0.981	1.049	1.511	4.71
UV at TPS				
DT	0.922	2.014	2.965	4.99
GBT	0.953	1.602	2.296	3.97
ERT	0.958	1.469	2.165	3.64
RF	0.959	1.467	2.146	3.63

Since the ERT algorithm exhibited excellent performance both in our study and that of [39], we choose this algorithm for comparison with our proposed method. From now on, the ERT models (for each target capacity) will be referred to as the classical models.

3.2. Application of SC approach

We apply the SC approach using the dataset of 98,695 MOFs aiming to identify the 100 top-performing MOFs at the four different conditions previously specified. Fig. 2 illustrates how frequently the 100 top-performing MOFs were identified during the 100 runs performed. This frequency can be obtained by examining how many times these materials were among the list of the top-100 predicted materials. In the ideal case, all top-100 MOFs should be present on the converged training set. It can be readily seen that the materials with the highest gravimetric working capacities under PS and TPS have been successfully identified by the SC algorithm in almost all runs. This behavior is not seen when the materials with the highest volumetric-based working capacities are examined. In this case it is observed that the frequency of appearance of these materials is significantly lower, especially at TPS conditions, while some of them were never observed.

Table 2 summarizes the performance metrics of the SC approach, averaged across 100 runs. Similar to the behavior of the classical ML models, SC models' accuracy is lower for the volumetric capacities compared to their gravimetric counterparts. This discrepancy in performance which is present on both approaches (classical and SC), will be discussed later on. It should be noted that although the SC based models are built such as to maximize efficiency (see Table 3 for their average training set size N_{train}^{SC}), they are still able to achieve satisfactory generalization performance. In order to make comparisons between the classical and SC models on equal footing, the latter were constructed using the ERT algorithm as base model.

3.3. Comparison of SC and traditional approach

Regarding the gravimetric capacities, when both approaches are challenged to identify the top-100 structures, the SC model outperforms the classical one in terms of efficiency, since it *identifies correctly the top-100 materials with less training samples*. As shown in the top panel of Fig. 3, for the classical approach a training set of two orders of magnitude larger ($N_{train}^{ERT} \approx$ 3,000) is required compared to the SC approach (≈ 300), for identifying roughly the same number of top-100 (N_{100}^{SC}) materials. This can be seen by projecting the crossing point of the blue and

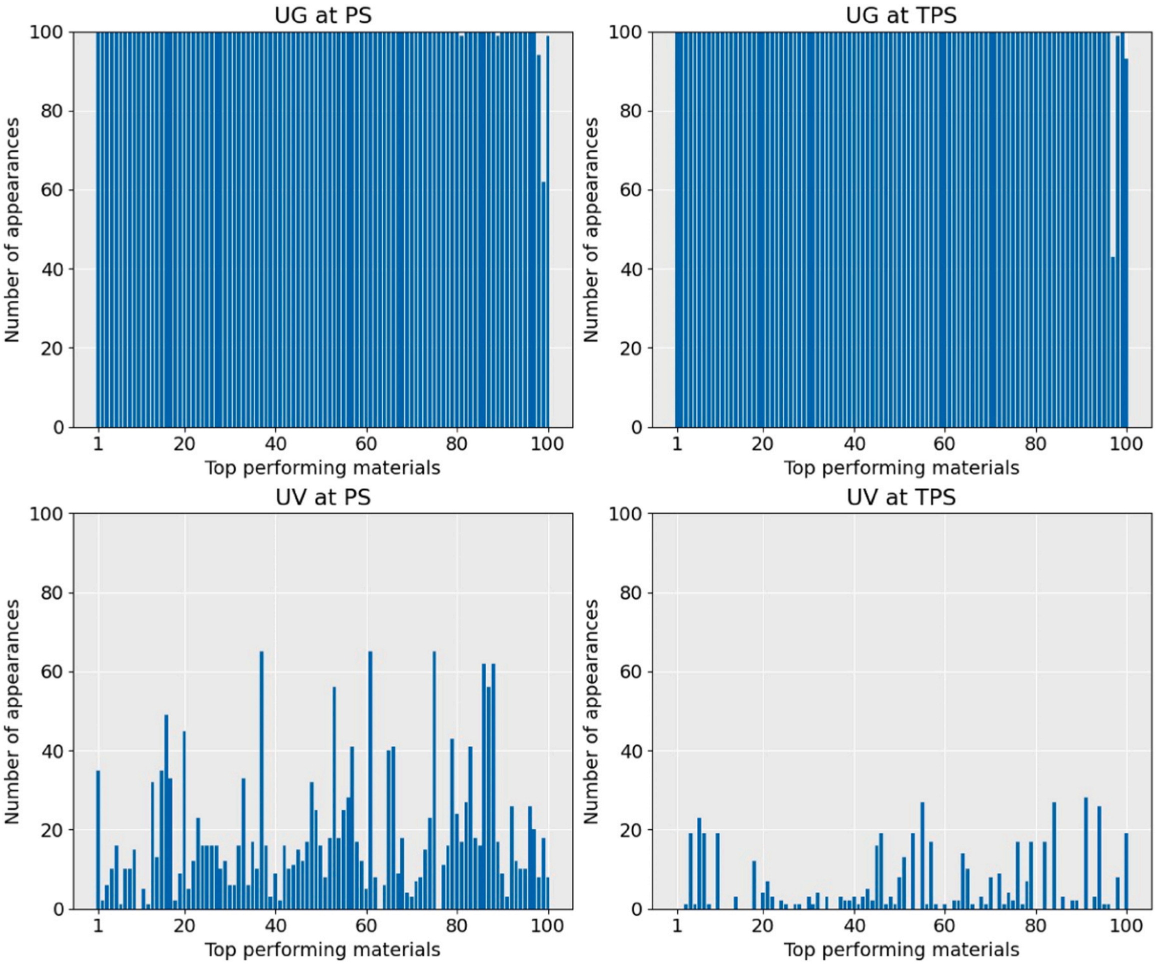

Fig. 2. Number of appearances for the top-100 materials in the 100 runs performed (*y*-axis). The structures have been sorted from left to right (*x*-axis) in descending order (MOFs indexed as 1 and 100 represent the materials with the highest and lowest capacities, respectively).

Table 2
Performance metrics of the SC approach using the ERT algorithm as base model. Metrics are calculated from predictions for the whole database. c.u., capacity units.

Capacity	R^2	MAE (c.u.)	RMSE (c.u.)	WAPE (%)
UG at PS	0.986	0.269	0.417	6.712
UG at TPS	0.987	0.303	0.459	4.529
UV at PS	0.963	1.547	2.124	6.956
UV at TPS	0.909	2.156	3.186	5.342

Table 3
N_{100}^{SC} denotes the number out of the top-100 MOFs that the SC model identifies while $N_{\mathrm{train}}^{\mathrm{SC}}$ denotes the training set size of the latter. $N_{\mathrm{train}}^{\mathrm{ERT}}$ stands for the training set size of the classical model required to reach the limit specified byN_{100}^{SC}. When this is not possible, the “optimum” size is reported (see text). Results reported for each target and method are averaged across 100 runs.

Capacity	N_{100}^{sc}	$N_{\mathrm{train}}^{\mathrm{sc}}$	$N_{\mathrm{train}}^{\mathrm{ERT}}$
UG at PS	99	285	3000
UG at TPS	99	298	3000
UV at PS	19	369	5000
UV at TPS	5	379	2000

black line onto the *x*-axis. Although these lines meet at a training set size $\approx$ 20,000, a substantial number of the top-100 materials are identified by the classical model using smaller training set sizes. For this reason and the fact that at training set size of 3,000 the standard deviation has significantly decreased compared to smaller training sets, the aforementioned size is assigned as the “optimum” for the classical method, at both PS and TPS conditions. Furthermore, if the classical model is trained with the same amount of reference data as the SC model, it identifies only half of the top-100 materials on average and suffers from large standard deviation. That is, the identification of the top materials depends on the sampling while this is not the case for the SC model. The latter shows nearly zero standard deviation since all top-100 MOFs are included in the converged training set almost for all runs (see Fig. 2). In other words, the random sampling step required for the initialization of the iterative procedure, has no effect on the identification of the best 100 structures. That is, the performance-directed selection of the training samples enables the efficient tracing of the best materials.

When volumetric capacities are considered, both approaches fail to identify efficiently a substantial number of the best structures as depicted in the bottom panel of Fig. 3. In the case of classical models, the number of top-100 materials identified, particularly at TPS conditions, increases much slower as the training set size is increased compared to their gravimetric counterparts. On the other hand, the number of top-100 materials found by the SC models has significantly decreased, while their converged training set size is slightly raised in comparison with their gravimetric counterparts. Nevertheless, similar to the case of gravimetric capacities, classical models require a greater training set size—approximately 5,000 at PS (compared to 390 for SC) and 2,000 at TPS (compared to 419 for SC) at TPS—to reach the threshold defined byN_{100}^{SC}. Moreover, for the same training size a smaller number of the

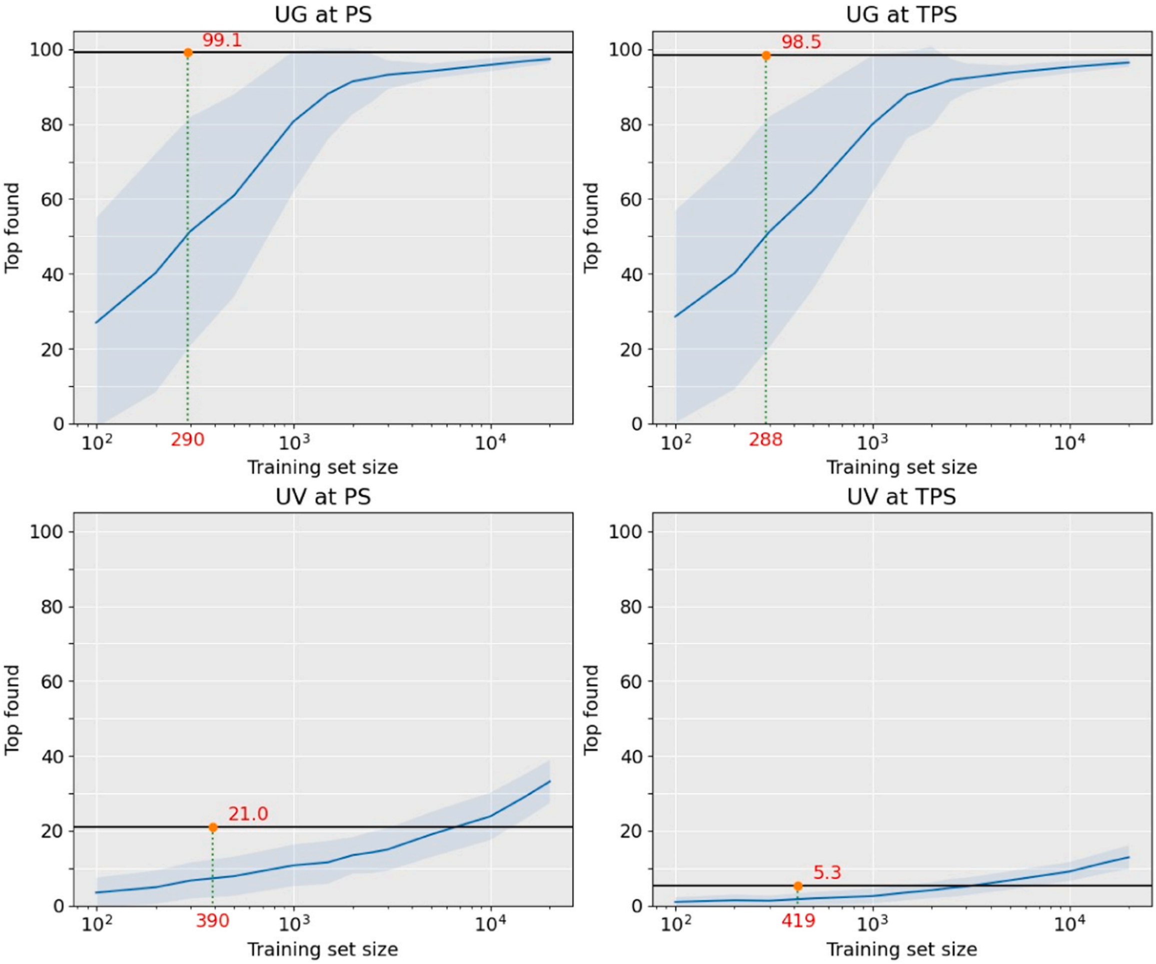

Fig. 3. Blue line represents the number of the top-100 materials (calculation is analogous to that of the SC approach, see text) identified by the classical model (*y*-axis) as function of the training set size (*x*-axis, logarithmic scale), while blue shaded area shows the respective standard deviation. The number of the top-100 materials that the SC model identifies and its converged training set size are represented by the black and green line, respectively. Results for each approach are averaged across 100 runs.

top-100 structures is found on average compared to the SC models. However, this time the profits of SC models in terms of computational efficiency, especially at TPS conditions, are less pronounced. The results of the previous comparisons for each usable capacity are summarized in Table 3.

The observation that for all target capacities the classical models need a greater training set size to achieve the threshold defined byN_{100}^{SC} should be attributed to the different character that each method adopts. The SC approach prioritizes the identification of the top-performing materials and it achieves it by carefully selecting the training samples. On the other hand, the classical approach focuses on building a model that is able to generalize well for all the regions, including those for which the performance of the materials is mediocre or low. By trying to capture the structure-property relationship as accurately as possible, a considerable computational effort is wasted in modeling regions that are not of interest.

3.4. Why methods fail for UV?

To understand the discrepancy in performance between the gravimetric and volumetric capacities we take a closer look on the feature space of MOFs. In order to visualize the seven-dimensional feature space (seven structural descriptors were used), we computed a reduced two-dimensional space using the t-distributed stochastic neighbour embedding (t-SNE) algorithm, as implemented in the scikit-learn package (version 0.22.2). The t-SNE is a statistical method for visualizing high-dimensional data by assigning each data point a location in a low-dimensional space (e.g. R^2 or R^3). The mapping is performed in such a way that similar points in the high-dimensional (original) space are represented by nearby points in the reduced space whereas dissimilar points are represented by distant points.

Based on the t-SNE plots for the four different targets (Fig. 4), it becomes apparent that materials with high gravimetric capacities are localized in well-defined regions of the feature space. More specifically for both PS and TPS a narrow area of features contains 98 out of the top-100 MOFs, while a second area contains the remaining 2 (Fig. 5). The SC approach very efficiently locates these two regions enabling the identification of the best materials in them.

In contrast, materials with high volumetric capacities span different regions of the feature space, and especially at TPS conditions they appear almost everywhere. This in turn makes the identification of the top-performing materials a lot harder for the SC model since at each iteration the algorithm updates its predictions based on the top performing materials already included in the training set. If the neighborhoods of these materials in the feature space are not rich in the top-performing materials, as it is the case when the latter appear all over the feature space, then it is unlikely that they will be included in the training set during the next iteration. It should be noted that the SC models show no bias towards the top-100 materials they identify. As depicted in Fig. 5, there is no preferential exploration of the top-

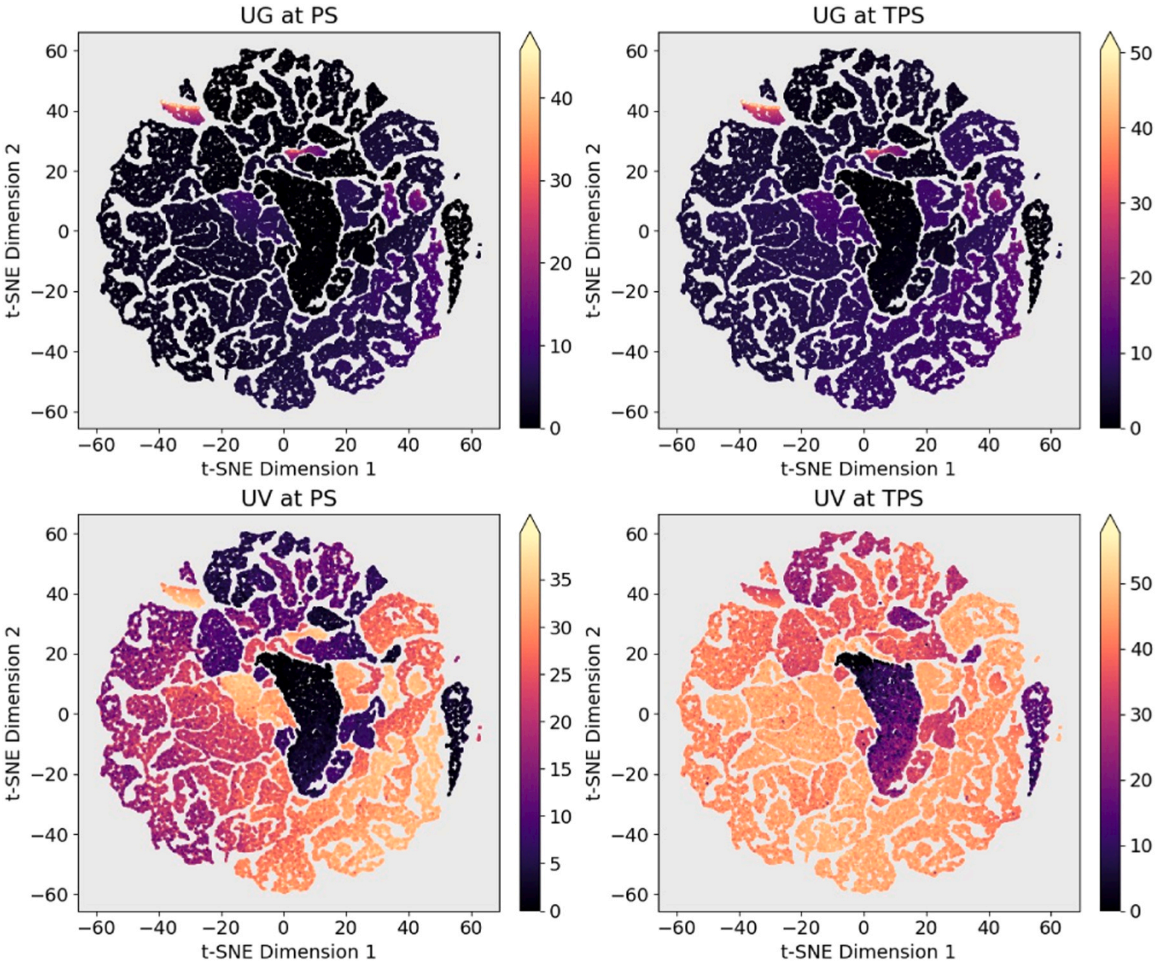

Fig. 4. A two-dimensional representation of the original seven-dimensional feature space. Each point corresponds to a structure on the dataset with the color bar denoting their usable capacities.

performing regions and many of the top-100 materials were identified at least once. Although not shown here, visualization of specific runs reveals that, in contrast to the case of UG, the results are very sensitive to the selection of the initial training set.

It should be reminded that both the SC and the classical approach employ ML algorithms for their predictions, and as a result they are both characterized by limited extrapolation capabilities. In the case of SC approach, this means that the algorithm does not have a good sense of what the structure-property landscape looks like (outside the region spanned by the training samples), limiting its ability to better control the way the requested simulations are performed. In the case of classical approach, the aforementioned weakness is reflected on the lower accuracy of the models for UV capacities compared to UG. As shown in Fig. 4, the structure-capacity relationship is more complex in the case of UV compared to UG, as intense alterations in terms of capacity along the feature space are present. As such, larger training sets are required in order the structure-property relationship to be sufficiently captured by the classical models and the top-100 materials to be identified.

Performance improvements for both approaches, especially for the case of volumetric capacities, can be attained by augmenting the set of descriptors with energy-based descriptors, i.e. descriptors that take into account host-guest interactions. This kind of descriptors is important when modelling gases with non-negligible electrostatic interaction such as CO_2 and H_2, since it allows ML algorithms to extract more complex structure-property relationships, which in turn can lead to ML models of higher generalization ability.

4. Conclusions

Two approaches were examined, namely a traditional one that builds a general predictive model using random samples and a self-consistent approach where the samples are chosen aiming to improve predictions in the region of interest.

Surprisingly, while accurate traditional ML models were constructed, when the latter were asked to identify the most promising materials, their high accuracy—according to various statistical metrics—did not translated to fruitful predictions in all cases. While their results were satisfactory for materials with high gravimetric capacities, this was not the case for the volumetric ones. Although this behavior became apparent thanks to the large number of available data that allowed us for an extensive evaluation of both approaches, this will not be possible when only a limited amount of data is available. *Judging an ML model based solely on its statistical accuracy may be misleading if we are interested in tracing the best materials of a database*. In our previous work for methane [40] while ML predictive models of lower accuracy were constructed ($R^2 = 0.940 - 0.983$ depending on the materials and the thermodynamic conditions examined compared to the $R^2 = 0.959 - 0.997$ in this work) more than 70 out of the top-100 materials were correctly identified by the SC procedure.

Comparison of the classical and SC approach, revealed that in all cases the latter was more efficient. Notably, the SC approach required a two orders of magnitude smaller training set compared to the classical one in order to identify the best materials in terms of gravimetric performance, at both PS and TPS conditions. Although the profits of the SC

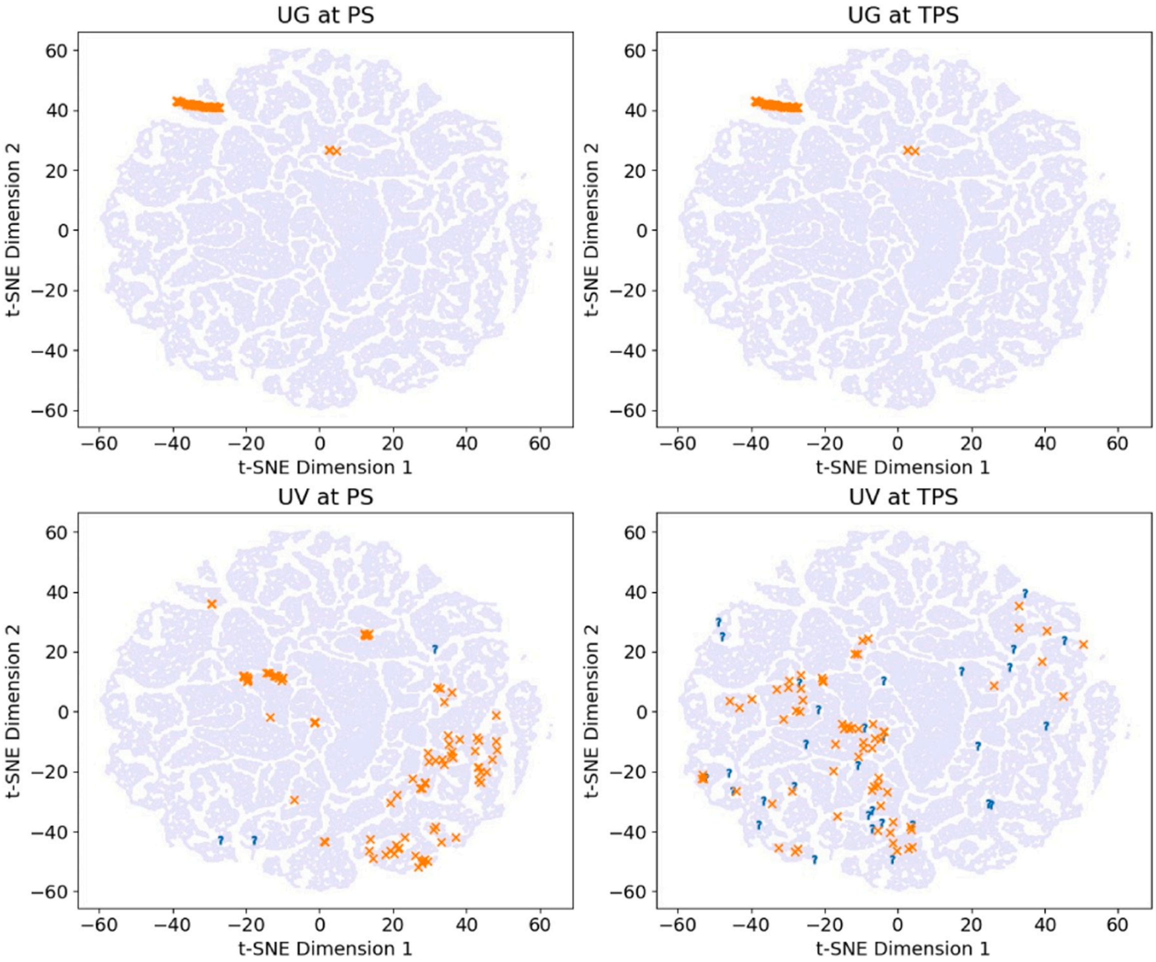

Fig. 5. A two-dimensional representation of the original seven-dimensional feature space. Each point corresponds to a structure on the dataset. Superimposed on that are the top-100 structures with the x-marker denoting the materials that were identified at least once by the SC models while the ?-marker stands for the top-100 materials that were never seen during the 100 runs.

method in terms of computational efficiency were not that prominent when volumetric capacities were examined, still our method was able to set a lower bound in terms of computational cost. The performance-directed selection of the training samples renders the SC method more efficient if we prioritize the discovery of novel materials.

Declaration of Competing Interest

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

Acknowledgements

Authors would like to acknowledge financial support from European Union: Horizon Europe (project MOST-H_2; Grant agreement no. 101058547) and Toyota Motor Europe NV/SA.

References

[1] Jarad A. Mason, Mike Veenstra, Jeffrey R. Long, Evaluating metal–organic frameworks for natural gas storage, Chem. Sci. 5 (2014) 32–51, doi: 10.1039/C3SC52633J. URL https://doi.org/10.1039/ C3SC52633J.

[2] Myunghyun Paik Suh, Hye Jeong Park, Thazhe Kootteri Prasad, DaeWoon Lim, Hydrogen storage in metal–organic frameworks, Chem. Rev. 112 (2) (2012) 782–835, 10.1021/cr200274s. URL https:// doi.org/10.1021/cr200274s.

[3] Jun Yang, Andrea Sudik, Christopher Wolverton, Donald J. Siegel, High capacity hydrogen storage materials: attributes for automotive applications and techniques for materials discovery, Chem. Soc. Rev. 39 (2010) 656–675, doi: 10.1039/B802882F. URL https://doi.org/10.1039/B802882F.

[4] Hiroyasu Furukawa, Kyle E. Cordova, Michael O'Keeffe, Omar M. Yaghi, The chemistry and applications of metal-organic frameworks, Science 341 (6149) (2013) 1230444, 10.1126/science.1230444. URL https://www.science.org/doi/abs/10.1126/science.1230444.

[5] Yongchul G. Chung, Emmanuel Haldoupis, Benjamin J. Bucior, Maciej Haranczyk, Seulchan Lee, Hongda Zhang, Konstantinos D. Vogiatzis, Marija Milisavljevic, Sanliang Ling, Jeffrey S. Camp, Ben Slater, J. Ilja Siepmann, David S. Sholl, Randall Q. Snurr, Advances, updates, and analytics for the computation-ready, experimental metal extendashorganic framework database: CoRE MOF, J. Chem. Eng. Data 64 (12) (2019) 5985–5998, 11 2019. doi: 10.1021/acs.jced.9b00835. URL https://doi.org/10.1021/acs.jced.9b00835.

[6] Peyman Z. Moghadam, Aurelia Li, Seth B. Wiggin, Andi Tao, Andrew G.P. Maloney, Peter A. Wood, Suzanna C. Ward, David Fairen-Jimenez, . Development of a cambridge structural database subset: a collection of metal extendashorganic frameworks for past, present, and future, Chem. Mater. 29 (7) (2017) 2618–2625.

[7] David J. Tranchemontagne, Jose ́ L. Mendoza-Corte ́s, Michael O'Keeffe, Omar M. Yaghi, Secondary building units, nets and bonding in the chemistry of metal–organic frameworks, Chem. Soc. Rev. 38 (2009) 1257–1283, doi: 10.1039/B817735J. URL https://doi.org/10.1039/B817735J.

[8] Michael O'Keeffe, Nets, tiles, and metal-organic frameworks, APL Mater. 2 (12) (2014), 124106 doi: 10.1063/1.4901292. URL https://doi.org/ 10.1063/1.4901292.

[9] P.N. Trikalitis, D.P. Broom, C.J. Webb, G.S. Fanourgakis, G.E. Froudakis, M. Hirscher, Concepts for improving hydrogen storage in nanoporous materials, Int. J. Hydrogen Energy 44 (15) (2019) 7768–7779. ISSN 0360-3199. doi: https://doi.org/10.1016/j.ijhydene.2019.01. 224. A special issue on hydrogen-based Energy storage.

[10] Christopher E. Wilmer, Michael A. Leaf, Chang Yeon Lee, Omar K. Farha, Brad G. Hauser, Joseph T. Hupp, Randall Q. Snurr, Large-scale screening of hypothetical metal-organic frameworks, Nat. Chem. 4 (2) (2011) 83–89.

[11] Yamil J. Colo ́n, Diego A. Go ́mez-Gualdro ́n, Randall Q. Snurr, Topologically guided, automated construction of metal–organic frameworks and their evaluation for energy-related applications, Cryst. Growth Des. 17 (11) (2017) 5801–5810,

https://doi.org/10.1021/acs.cgd.7b00848 (URL), ⟨https://doi.org/10.1021/acs.cgd.7b00848⟩.

[12] Sangwon Lee, Baekjun Kim, Hyun Cho, Hooseung Lee, Sarah Yunmi Lee, Eun Seon Cho, Jihan Kim, Computational screening of trillions of metal–organic frameworks for high-performance methane storage, ACS Appl. Mater. Interfaces 13 (20) (2021) 23647–23654, doi: 10.1021/ acsami.1c02471. URL https://doi.org/10.1021/acsami.1c02471.

[13] Tina Dur̈en, Youn-Sang Bae, Randall Q. Snurr, Using molecular simulation to characterise metal–organic frameworks for adsorption applications, Chem. Soc. Rev. 38 (5) (2009) 1237, doi: 10.1039/b803498m. URL https://doi.org/10.1039%2Fb803498m.

[14] Christopher E. Wilmer, Michael A. Leaf, Chang Yeon Lee, Omar K. Farha, Brad G. Hauser, Joseph T. Hupp, Randall Q. Snurr, Large-scale screening of hypothetical metal-organic frameworks, Nat. Chem. 4 (2) (2011) 83–89.

[15] Cory M. Simon, Jihan Kim, Diego A. Gomez-Gualdron, Jeffrey S. Camp, Yongchul G. Chung, Richard L. Martin, Rocio Mercado, Michael W. Deem, Dan Gunter, Maciej Haranczyk, David S. Sholl, Randall Q. Snurr, and Berend Smit. The materials genome in action: identifying the performance limits for methane storage, Energy Environ. Sci. 8 (2015) 1190–1199, doi: 10.1039/C4EE03515A. URL https://doi.org/10.1039/C4EE03515A.

[16] Diego A. Gomez-Gualdŕon, Yamil J. Coĺon, Xu Zhang, Timothy C. Wang, Yu-Sheng Chen, Joseph T. Hupp, Taner Yildirim, Omar K. Farha, Jian Zhang, Randall Q. Snurr, Evaluating topologically diverse metal–organic frameworks for cryo-adsorbed hydrogen storage, Energy Environ. Sci. 9 (2016) 3279–3289, doi: 10.1039/C6EE02104B. URL https://doi.org/10.1039/C6EE02104B.

[17] Peyman Z. Moghadam, Timur Islamoglu, Subhadip Goswami, Jason Exley, Marcus Fantham, Clemens F. Kaminski, Randall Q. Snurr, Omar K. Farha, David Fairen-Jimenez, Computer-aided discovery of a metal–organic framework with superior oxygen uptake, Nat. Commun. 9 (1) (2018) 1378.

[18] WooSeok Jeong, Dae-Woon Lim, Sungjune Kim, Aadesh Harale, Minyoung Yoon, Myunghyun Paik Suh, Jihan Kim, Modeling adsorption properties of structurally deformed metal–organic frameworks using structure–property map, Proc. Natl. Acad. Sci. USA 114 (30) (2017) 7923–7928.

[19] Cory M. Debasis Banerjee, Anna M. Simon, Radha K. Plonka, Jian Motkuri, Xianyin Liu, Berend Chen, John B. Smit, Maciej Parise, Haranczyk, Praveen K. Thallapally, Metal–organic framework with optimally selective xenon adsorption and separation, Nat. Commun. 7 (1) (2016).

[20] Aaron W. Thornton, Cory M. Simon, Jihan Kim, Ohmin Kwon, Kathryn S. Deeg, Kristina Konstas, Steven J. Pas, Matthew R. Hill, David A. Winkler, Maciej Haranczyk, Berend Smit, Materials genome in action: Identifying the performance limits of physical hydrogen storage, Chem. Mater. 29 (7) (2017) 2844–2854, doi: 10.1021/acs.chemmater.6b04933. URL https://doi.org/10.1021/acs.chemmater.6b04933.

[21] Hongda Zhang, Randall Q. Snurr, Computational study of water adsorption in the hydrophobic metal–organic framework zif-8: adsorption mechanism and acceleration of the simulations, J. Phys. Chem. C 121 (43) (2017) 24000–24010, doi: 10.1021/acs.jpcc.7b06405. URL https://doi.org/10.1021/acs.jpcc.7b06405.

[22] Keith T. Butler, Daniel W. Davies, Hugh Cartwright, Olexandr Isayev, Aron Walsh, Machine learning for molecular and materials science, Nature 559 (7715) (2018) 547–555, https://doi.org/10.1038/s41586-018-0337-2. URL https://doi.org/10.1038%2Fs41586-018-0337-2.

[23] N. Scott Bobbitt, Randall Q. Snurr, Molecular modelling and machine learning for high-throughput screening of metal-organic frameworks for hydrogen storage, Mol. Simul. 45 (14–15) (2019) 1069–1081, https://doi.org/10.1080/08927022.2019.1597271. URL https://doi.org/10.1080/08927022.2019.1597271.

[24] D.P. Broom, C.J. Webb, K.E. Hurst, P.A. Parilla, T. Gennett, C.M. Brown, R. Zacharia, E. Tylianakis, E. Klontzas, G.E. Froudakis, Th.A. Steriotis, P. N. Trikalitis, D.L. Anton, B. Hardy, D. Tamburello, C. Corgnale, B.A. van Hassel, D. Cossement, R. Chahine, M. Hirscher, Outlook and challenges for hydrogen storage in nanoporous materials, Appl. Phys. A 122 (3) (2016), https://doi.org/10.1007/s00339-016-9651-4. URL https://doi.org/10.1007%2Fs00339-016-9651-4.

[25] Yongchul G. Song Li, Cory M. Chung, Simon, Randall Q. Snurr, Highthroughput computational screening of multivariate metal–organic frameworks (mtv-mofs) for co2 capture, J. Phys. Chem. Lett. 8 (24) (2017) 6135–6141, https://doi.org/10.1021/acs.jpclett.7b02700. URL https://doi.org/10.1021/acs.jpclett.7b02700. PMID: 29206043.

[26] Xuanjun Wu, Sichen Xiang, Jiaqi Su, Weiquan Cai, Understanding quantitative relationship between methane storage capacities and characteristic properties of metal–organic frameworks based on machine learning, J. Phys. Chem. C 123 (14) (2019) 8550–8559, https://doi.org/10.1021/acs.jpcc.8b11793. URL https://doi.org/10.1021%2Facs.jpcc. 8b11793.

[27] Michael Fernandez, Tom K. Woo, Christopher E. Wilmer, Randall Q. Snurr, Large-scale quantitative structure–property relationship (qspr) analysis of methane storage in metal–organic frameworks, J. Phys. Chem. C 117 (15) (2013) 7681–7689, https://doi.org/10.1021/jp4006422. URL https://doi.org/10.1021/jp4006422.

[28] Michael Fernandez, Nicholas R. Trefiak, Tom K. Woo, Atomic property weighted radial distribution functions descriptors of metal–organic frameworks for the prediction of gas uptake capacity, J. Phys. Chem. C 117 (27) (2013) 14095–14105, doi: 10.1021/jp404287t. URL https://doi.org/10.1021/jp404287t.

[29] Maryam Pardakhti, Ehsan Moharreri, David Wanik, Steven L. Suib, Ranjan Srivastava, Machine learning using combined structural and chemical descriptors for prediction of methane adsorption performance of metal organic frameworks (mofs), ACS Comb. Sci. 19 (10) (2017) 640–645, doi: 10.1021/acscombsci.7b00056. URL https://doi.org/10.1021/ acscombsci.7b00056.

[30] Maryam Pardakhti, Pariksheet Nanda, Ranjan Srivastava, Impact of chemical features on methane adsorption by porous materials at varying pressures, J. Phys. Chem. C 124 (8) (2020) 4534–4544, 10.1021/acs.jpcc.9b09319. URL https://doi.org/10.1021/ acs.jpcc.9b09319.

[31] Benjamin J. Bucior, N. Scott Bobbitt, Timur Islamoglu, Subhadip Goswami, Arun Gopalan, Taner Yildirim, Omar K. Farha, Neda Bagheri, Randall Q. Snurr, Energy-based descriptors to rapidly predict hydrogen storage in metal–organic frameworks, Mol. Syst. Des. Eng. 4 (2019) 162–174, doi: 10.1039/C8ME00050F. URL https://doi.org/10.1039/ C8ME00050F.

[32] Ryther Anderson, Achay Biong, Diego A. Gómez-Gualdŕon, Adsorption isotherm predictions for multiple molecules in MOFs using the same deep learning model, J. Chem. Theory Comput. 16 (2) (2020) 1271–1283, doi: 10.1021/acs.jctc.9b00940. URL https://doi.org/10.1021%2Facs.jctc.9b00940.

[33] Ioannis Tsamardinos, George S. Fanourgakis, Elissavet Greasidou, Emmanuel Klontzas, Konstantinos Gkagkas, George E. Froudakis, An automated machine learning architecture for the accelerated prediction of metal-organic frameworks performance in energy and environmental applications, Microporous Mesoporous Mater. 300 (2020), 110160. ISSN 1387-1811. doi: https://doi.org/10.1016/j.micromeso.2020.110160. URL https://www.sciencedirect.com/science/article/pii/ S1387181120301633.

[34] Giorgos Borboudakis, Taxiarchis Stergiannakos, Maria G. Frysali, Emmanuel Klontzas, I. Tsamardinos, George E. Froudakis, Chemically intuited, large-scale screening of mofs by machine learning techniques, npj Comput. Mater. 3 (2017) 1–7.

[35] George S. Fanourgakis, Konstantinos Gkagkas, Emmanuel Tylianakis, Emmanuel Klontzas, George Froudakis, A robust machine learning algorithm for the prediction of methane adsorption in nanoporous materials, J. Phys. Chem. A 123 (28) (2019) 6080–6087, doi: 10.1021/acs.jpca.9b03290. URL https://doi.org/10.1021/acs.jpca. 9b03290.

[36] George S. Fanourgakis, Konstantinos Gkagkas, Emmanuel Tylianakis, George Froudakis, A generic machine learning algorithm for the prediction of gas adsorption in nanoporous materials, J. Phys. Chem. C 124 (13) (2020) 7117–7126, doi: 10.1021/acs.jpcc.9b10766. URL https://doi.org/10.1021/acs.jpcc.9b10766.

[37] George S. Fanourgakis, Konstantinos Gkagkas, Emmanuel Tylianakis, George E. Froudakis, A universal machine learning algorithm for large-scale screening of materials, J. Am. Chem. Soc. 142 (8) (2020) 3814–3822, doi: 10.1021/jacs.9b11084. URL https://doi.org/10.1021/jacs.9b11084.

[38] Aaron W. Thornton, Cory M. Simon, Jihan Kim, Ohmin Kwon, Kathryn S. Deeg, Kristina Konstas, Steven J. Pas, Matthew R. Hill, David A. Winkler, Maciej Haranczyk, Berend Smit, Materials genome in action: identifying the performance limits of physical hydrogen storage, Chem. Mater. 29 (7) (2017) 2844–2854. ISSN 15205002. doi: 10.1021/acs.chemmater.6b04933. URL http://pubs.acs.org/doi/10.1021/acs.chemmater.6b04933.

[39] Alauddin Ahmed, Donald J. Siegel, Predicting hydrogen storage in MOFs via machine learning, Patterns 2 (7) (2021), 100291. ISSN 26663899. doi: 10.1016/j.patter.2021.100291. URL https://linkinghub. elsevier.com/retrieve/pii/S2666389921001240.

[40] George S. Fanourgakis, Konstantinos Gkagkas, Emmanuel Tylianakis, George Froudakis, Fast screening of large databases for top performing nanomaterials using a self-consistent, machine learning based approach, J. Phys. Chem. C 124 (36) (2020) 19639–19648. ISSN 19327455. doi: 10. 1021/acs.jpcc.0c05491. URL https://pubs.acs.org/doi/10.1021/acs. jpcc.0c05491.

[41] Alauddin Ahmed and Donald J. Siegel. Hymarc datahub. https://datahub.hymarc.org/dataset/computational-prediction-of-hydrogen-storage-capacities-in-mofs, 2019.

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