Gas Separation Selectivity Prediction Based on Finely Designed Descriptors

Emmanuel Ren[†,‡] and François-Xavier Coudert[*,‡]

†*CEA, DES, ISEC, DMRC, Univ. Montpellier, Marcoule, France*

‡*Chimie ParisTech, PSL University, CNRS, Institut de Recherche de Chimie Paris, 75005 Paris, France*

E-mail: fx.coudert@chimieparistech.psl.eu

Abstract

Adsorption-based techniques for gas separation using nanoporous materials are widely used and hold a promising future, but systematic identification of the best-performing materials for a given application is still an open problem. For that task, we need to estimate selectivity at different operating conditions (temperature, pressure) on a large set of nanoporous structures. To this aim, we have developed a machine learning-assisted screening process based on a fast grid calculation of interaction energies, in addition to newly designed geometrical descriptors to predict ambient-pressure selectivity. As a proof of concept, we tested our methodology for the separation of a 20-80 xenon/krypton mixture at 298 K and 1 atm in the nanoporous materials of the CoRE MOF 2019 database. Based on a standard train/test split of the dataset, our model is promising with an RMSE of 2.5 on the ambient-pressure selectivity values of the test set and 0.06 on their base-10 logarithm. This method can thence be used to pre-select the best performing materials for a more thorough investigation.

1 Introduction

Gas separation and purification are essential processes since they provide key reactants and inert gases for the chemical industry, as well as medical or food grade gases. Among them, we can find easily extractable or synthesizable molecules such as nitrogen, oxygen, carbon dioxide, noble gases, hydrogen, methane, or nitrous oxide. Moreover, gas separation is crucial in mitigating negative environmental impact at the end of industrial processes, such as facilities emitting green house gases (*e.g.* concrete or steel plants) or treating radioactive off-gases like ^{85}Kr. Cryogenic liquefaction or distillation is currently the mainstream technique to achieve industrial gas separation, while adsorbent beds made of nanoporous materials (activated alumina or zeolites) are mostly used as a less energy-intensive pre-purification system.[1]

A wider use of nanoporous materials could reduce the energy consumption of current separation processes since adsorption is way less energy intensive than liquefaction.[2] For instance, some prototypes involving beds of nanoporous materials have been developed for xenon/krypton separation to avoid employing cryogenic distillation.[3] For the process to be viable, materials need to perform even better and many studies focus on synthesizing ever more selective materials by leveraging all chemical intuitions around noble gas adsorption properties.[4–6] In order to speed the discovery process of novel materials with key properties, computational screening can identify factors explaining the performance and pre-select candidates for further experimental studies. As recently conceptualized by Lyu et al., a synergistic workflow combining computational discovery and experimental validation can push material discovery to the next stage.[7,8] But to efficiently guide experimental discoveries, computational chemists are facing two major challenges: generating reliably more structures and evaluating them with fast and accurate models.

The number of nanoporous materials is potentially unlimited; for the metal–organic frameworks (MOFs) alone, over 90,000 structures have been synthesized[9] and 500,000 computationally constructed[10–12]. To deal with this ever-increasing amount of structures, we need to design more efficient screening procedures as well as faster performance evaluation tools. To go beyond the time-consuming calculations over the whole dataset, computational chemists developed funnel-like screening procedures to reduce the need for expensive simulations and introduced machine learning (ML) models to replace them with faster evaluation tools.[13] To further improve the selectivity screening for Xe/Kr separation, we will need to design better performing structural and energy-based descriptors.

Simon et al. published one of the first articles on an ML-assisted screening approach for the separation of a Xe/Kr mixture extracted from the atmosphere.[14] Their model's performance was highly relying on the Voronoi energy, which is basically an average of the interaction energies of a xenon atom at each Voronoi node.[15] To rationalize this increase in performance, we regarded this Voronoi energy as a faster proxy for the adsorption enthalpy. By comparing it to the standard Widom insertion, we found that although it is faster, it is less accurate; and we developed a more effective alternative, the surface sampling (RAESS) using symmetry and non-accessible volumes blocking.[16] Recently, Shi et al. used an energy grid to generate energy histograms as a descriptor for their ML model, which gives an exhaustive description of the infinitely diluted adsorption energies,[17] but can be computationally expensive.

All the approaches described above can have good accuracy in the prediction of low-pressure adsorption (i.e., in the limit of zero loading) but are not suitable for prediction of adsorption in the high-pressure regime, when the material is near saturation uptake. While this later task is routinely performed by Grand Canonical Monte Carlo (GCMC) simulations, there is a lack of methods at lower computational cost for high-throughput screening. To better frame our challenge, in this work we are essentially trying to predict the selectivity in the nanopores of a material at high pressure, where adsorbates are interacting with each other, while only having information on the interaction at infinite dilution. The comparison between the low and high pressure cases gives key information on the origin of the differences of selectivity. For instance, we previously showed that selectivity could drop between the low and ambient pressure cases in the Xe/Kr separation application, and it was mainly attributed to the presence of different pore sizes and potential reorganizations due to adsorbate–adsorbate interactions.[18]

In this article, we combined a grid-based approach with core components of our previously developed RAESS algorithm[16] to design a new adsorption energy sampling technique. Moreover, a statistical characterization of the pore size and energy distributions has been performed to inform the model on a potential selectivity drop. By combining these two approaches, we propose a set of useful ML descriptors for fast and accurate ambient-pressure selectivity prediction, and we highlight its performance on the case of xenon/krypton separation in the CoRE MOF 2019 database[19].

2 Methods

2.1 The machine learning model

We chose to use eXtreme Gradient Boosting (XGBoost) as the machine learning framework for our predictive model because of its accuracy, efficiency and simplicity of use. Its performance has long been proven since 17 out 29 Kaggle challenge winning solutions were based on this algorithm in 2015. The XGBoost system is highly scalable and parallelized, which gives very fast model training.[20] Compared to more standard tree-based algorithms such as random forest (commonly used in the field[14]), the boosting component of the algorithm means that it learns from previous mistakes and puts higher weights on the problematic data points, hence improving the accuracy of the final ML model.

In the next sections, we introduce new descriptors for nanoporous materials, as well as new concepts of feature engineering based on energy and pore size histograms. The ML features presented have been selected by progressively filtering out the less influential ones on the accuracy of the final model, see the complete list in Table S1-3 of Supporting Information (SI). The influence or importance are defined later in a section dedicated to the interpretation of the model. The hyperparameters of the model were fine-tuned using random search to design the best performing final model. Finally, the influence of the pre-selected descriptors on the final model is interpreted using a unified approach.

2.2 Target variable

We want to predict the Xe/Kr ambient-pressure selectivity faster than standard techniques. To obtain reference values (ground truth), we used the Raspa2 software[21] to run grand canonical Monte Carlo (GCMC) calculations of 20-80 Xe/Kr mixtures at 298 K and 1 atm on our cleaned database. The van der Waals interactions are described by a Lennard-Jones (LJ) potential with a cutoff distance of 12 Å. The LJ parameters of the framework atoms are given by the universal force field (UFF),[22] and the guest atoms (xenon and krypton) have their LJ parameters taken from a previous screening study.[23] The study only focuses on a given Xe/Kr composition usually obtained by cryogenic distillation of ambient air[1] as a first step towards predicting other mixtures at different physical conditions (*e.g.* Xe/Kr mixtures out of nuclear off-gases). In the broader scope, this methodology could be adapted to the desired application with some tweaks on the descriptors calculation (*e.g.* CO_2/CH_4 separation).

We decided to use a logarithmic transform of the selectivity instead of the raw value because we are more interested in the order of magnitude of the selectivities than to predict the higher values of selectivity — an ML model that predicts selectivity values can lower down the errors by focusing the prediction more on the higher values than the lower ones. By focusing on the logarithmic transform of the selectivity, we can better separate the different orders of magnitude of the selectivities. This approach distributes more evenly the efforts on all the different values of selectivities. Moreover, this logarithmic transform is related to a thermodynamic quantity that we elaborate later in the section 2.6.3; it can therefore be easily compared with the energy descriptors we introduced in this article.

2.3 Database and data preparation

To test our methodology on a set of realistic MOFs, we chose to screen the 12,020 all-solvent removed (ASR) structures of the CoRE MOF 2019 database[19]. After removing the disordered and the non-MOF structures as well as the ones with a large unitcell volume of 20 nm^3, we obtained a set of 9,748 structures. Then we analyze the string information given by the Zeo++ software[24] to reduce the number to 9,177 by removing the structures that are not tridimensional, where solvents are still detected (wrongly classified in all-solvent removed), or where the metal is radioactive or fissile (e.g., Pu-MOF TAGCIP[25], Np-MOF KASHUK[26], U-MOF ABETAE[27] or Th-MOF ASAMUE[28]) — this can be a source of risks in a nuclear waste processing plant. Furthermore, we added a condition on the largest cavity diameter (LCD) to keep only the structures that can accept a xenon molecule: 8,529 structures have a LCD higher than 4 Å (approximately the size of a xenon molecule). This is equivalent to removing the structures with very unfavorable adsorption enthalpies, that are not promising for our adsorption-based separation (see previous work[16]).

Then, the descriptors summarized below (and fully detailed in Supporting Information) were calculated on this restrained dataset. At this stage, 370 structures failed to be calculated in GCMC and 82 have no standard deviation for the pore distribution (skewness and kurtosis cannot be retrieved). A final dataset of 8,077 structures was therefore used to perform our ML-assisted method of screening the Xe/Kr adsorption selectivity. Based on this final set, 20% were randomly used for the test set and 80% were used to train our model. The goal is to learn from the training set a relationship between the descriptors and the target ambient-pressure selectivity in order to evaluate the performance on the test set. A CSV file of training and test sets can be found in the data availability section.

2.4 Geometrical and chemical ML descriptors

Looking at a number of different research papers on supervised ML for the prediction of adsorption properties,[14,29–32] we see that some descriptors are recurrent: 1) geometrical descriptors obtained by software like Zeo++[24] such as the surface area (SA), the void fraction (VF), the largest cavity diameter (LCD) and the pore limiting diameter (PLD); and 2) physical and chemical descriptors such as the framework's density, the framework's molar mass, the percentage of carbon (C%), nitrogen (N%), oxygen (O%), hydrogen but also halogen, nonmetals, metalloids and metals, and the degree of unsaturation. Although these descriptors are very versatile and used in many ML models, they however fail to provide specific information for our ML task. As shown by Simon et al., energy descriptors are greatly influential in ML models for selectivity prediction.

The geometric analysis of the crystalline porous materials is typically based on the van der Waals (vdW) radii predefined by the Cambridge Crystallographic Data Centre (CCDC). This force field-independent choice can create a gap between the geometrical descriptors and the thermodynamic values obtained through molecular simulations. Inspired by a recent work on the comparison of PLDs and self-diffusion coefficients,[33] we defined a list of vdW radii to be read by the Zeo++ software (more details in `https://github.com/eren125/zeopp_radtable`). In this study, all Zeo++ calculations use an atomic radius that corresponds to the distance where the LJ potential reaches $3k_\mathrm{B}T/2$, for $T = 298\,\mathrm{K}$.

The SA exposed to different probe sizes (1.2 Å, 1.8 Å and 2.0 Å) were tested. The probe occupiable volume was chosen to measure the void fraction (VF) for different adsorbent by using probe sizes of 1.8 Å (close to the radius of krypton) and 2.0 Å (close to that of xenon). This definition of the pore volume was found to be in better agreement with experimental nitrogen isotherms.[34]

Because our goal is to predict the difference between the low-pressure selectivity and the ambient-pressure (for a given gas mixture composition), some of these descriptors have very little importance, and the key factor is the difference of accessible volume and the affinity of the remaining pore volume with xenon, compared to krypton. The intuition developed in the previous study sketched the role of a diverse distribution of pores with different xenon affinities.[18] For all these reasons, from all the "standard" descriptors taken from the literature, we kept only the following 7 descriptors: C%, N%, O%, LCD ("D_i_vdw_uff298"), PLD ("D_f_vdw_uff298"), SA for a 1.2 Å probe ("ASA_m2/cm3_1.2") and VF for a 2.0 Å probe ("PO_VF_2.0"). We also built a new descriptor ΔVF void fraction values, the difference of volumes occupiable by xenon (2.0 Å) and by krypton (1.8 Å). All these descriptors along with other pore size distribution based geometrical descriptors are presented in detail in the Table S1 of the Supplementary Information (SI).

2.5 Pore size distribution

To generate a histogram of pore sizes (or pore size distribution, PSD), Monte Carlo steps are used to measure the frequency of every accessible pore sizes binned by 0.1 Å.[35] This histogram can then be used to generate descriptors based on statistical parameters that describes the overall location, the dispersion, the shape and the modality of the distribution. In addition to the mean and standard deviation of the distribution, we introduced two additional moments: the skewness (γ) corresponds to the third standardized moment and measures the asymmetry of a distribution; and the kurtosis (k), being the fourth standardized moment, measures the relative weight of the tails of the distribution. Knowing the importance of characterizing the number of different pore sizes suspected to be at the origin of the selectivity drop observed, we tried to find a simple descriptor to measure the number of modes in the distribution. The Sarle's bimodality coefficient, $\mathrm{BC} = (\gamma^2 + 1)/k$, represents a simple quantification of how far we are from the unimodality based only on skewness and kurtosis.[36] Finally to further measure the diversity of pores, we introduced an effective number $n_{\text{eff}} = N^2/\sum n_i^2$ of pore sizes, where N is the total number of points in the histogram and n_i the number of points associated with the i^{th} bin. This number is very similar to a statistical number widely used in other scientific fields: in political science it is used to measure the effective number of political parties,[37], in ecology the inverse Simpson's index evaluates the species diversity in an ecosystem,[38] or in quantum physics the inverse participation number measures the degree of localization of a wave-function.[39] This effective number of pore sizes gives an idea of the diversity of pore sizes (considering a binning of 0.1 Å). A high effective number would mean that multiple pore sizes are highly represented in the structure; this descriptor gives an idea of how scattered the pore sizes are. All these descriptors carries information on the form of the PSD needed to figure out the loading and selectivity situation in the framework near saturation uptake, which is crucial to predict the ambient-pressure selectivity.

2.6 Energy-based descriptors

2.6.1 Grid calculation

Inspired by our recent work on a faster way of calculating the low-pressure adsorption enthalpy and Henry's constant,[16] we designed an approach based on symmetry-respecting grids. These were generated using the Gemmi project's C++ library,[40] using an algorithm implemented with the following steps. First, we loop over the framework atoms and the grid points around a sphere of radius $0.8 \times \sigma_{g-h}$, where σ_{g-h} is the distance at which the LJ potential energy between the guest atom g and the host atom is zero. The LJ potential energy between the guest molecule and the closes host atom is calculated and only the grid points with an energy lower than a predefined threshold (here set to $100\,\mathrm{kJ\,mol^{-1}}$) are considered "unvisited" and will be recalculated in the following loop, the others are considered blocked by the framework and will be considered already "visited". This first loop over the framework atoms aims at filtering out the grid points that are blocked by the framework, and we will refer to this preliminary filtering step as "blocking" in the Table 1. Then, a second loop over the "unvisited" grid points is performed — at each increment, if the point is "unvisited" we calculate the interaction energy between the guest and all the host atoms within the cutoff, then the symmetric images of this point are filled with the same energy value and are considered "visited" by the algorithm. This symmetry-aware grid exploration allows the algorithm to divide the time required by the average number symmetry images — this module will be referred to as "symmetry" in the Table 1. By combining both the "blocking" of the high energy grid points and the "symmetry" based calculation of the interaction energies, we built a "fast" version of the grid calculation algorithm that can compete with our previously developed rapid surface sampling method (RAESS).

To highlight the improvement in performance in this procedure: the average void fraction for a 1.2 Å probe radius is equal to 0.16 and the average number of symmetric images is equal to 5.8 on the CoRE MOF 2019 database (most MOFs present symmetry operations). On average, the "blocking" procedure means that only ∼16% of the grid points really need to be calculated. The "symmetry" procedure means only ∼17% of points need to be considered, and the combination of both reduces the number of useful points to only 2.7% of the grid. This leads to a significant reduction in the CPU time of the calculation, as shown in Table 1.

Table 1: Performance comparison of the new grid method to other standard techniques used to calculate the adsorption enthalpies. The RMSE is calculated by comparing to the values given by a 100k-steps Widom insertion considered as the ground truth. The associated calculations are performed on the CoRE MOF 2019 database with a single Intel Xeon Platinum 8168 core at 2.7 GHz.

Energy sampling method	Average CPU time (s)	RMSE on adsorption enthalpy ($\mathrm{kJ\,mol^{-1}}$)
Grid – naive – 0.1 Å	71.3	0.025
Grid – blocking – 0.1 Å	18.8	0.026
Grid – symmetry – 0.1 Å	16.8	0.024
Grid – fast – 0.1 Å	4.8	0.023
Grid – fast – 0.3 Å	0.16	0.22
RAESS[16]	0.34	0.34
Widom[41] (12k cycles)	150	0.01

From the energy values of this grid, we can now calculate many useful descriptors that are used in our final model. A fully detailed description of these descriptors as well as their labeling names are given in the Table S2 of the SI.

2.6.2 Single component thermodynamic values

From these host–guest interaction energies, we can calculate different thermodynamic quantities corresponding to different statistical averaging. For instance, the Henry's constant K_H

corresponds to the average of the Boltzmann factors $\langle \exp(-\mathcal{E}_{\text{int}}/RT)\rangle$, while the adsorption enthalpies is the Boltzmann average of the interaction energies — all these concepts have been used and summarized in our previous paper on the surface sampling of energies to determine adsorption enthalpy and Henry's constant.[16] The adsorption Gibbs free energy $\Delta_{\text{ads}}G$ can then be deduced from the Henry's constant since $\Delta_{\text{ads}}G = -RT \ln\left(\langle \exp(-\mathcal{E}_{\text{int}}/RT)\rangle\right)$, and finally the adsorption entropy is naturally derived from the Gibbs energy: $\Delta G = \Delta H - T\Delta S$.

2.6.3 Exchange equilibrium and selectivity

The exchange equilibrium corresponds to what occurs in the competitive adsorption process between two adsorbate molecules of a mixture. Adsorption sites can be either occupied by adsorbate A or adsorbate B, leaving the other in the gas phase. This equilibrium can be modeled by the equation $\mathrm{A}_{(\text{ads})} + \mathrm{B}_{(\text{gas})} = \mathrm{A}_{(\text{gas})} + \mathrm{B}_{(\text{ads})}$, and the equilibrium constant corresponds to the selectivity $s^{\mathrm{A/B}} = (q_{\mathrm{A}} y_{\mathrm{B}})/(q_{\mathrm{B}} y_{\mathrm{A}})$. The exchange Gibbs free energy is then simply derived from the selectivity:

$$\Delta_{\text{exc}} G^{\mathrm{A/B}} = -RT \ln s^{\mathrm{A/B}} \tag{1}$$

which is consistent with the relationship between selectivity and Henry's constant at low-pressure. According to Hess's law, the exchange enthalpy is simply the difference between the adsorption enthalpies $\Delta_{\text{exc}} H^{\mathrm{A/B}} = \Delta_{\text{ads}} H^{\mathrm{A}} - \Delta_{\text{ads}} H^{\mathrm{B}}$. Finally the entropic term $-TS$ can also be obtained for our exchange equilibrium $-T\Delta_{\text{exc}} S = \Delta_{\text{exc}} G - \Delta_{\text{exc}} H$. We used these formulas to calculate the Gibbs free energy of the most influential descriptor, the xenon/krypton exchange equilibrium at infinite dilution $\Delta G_0^{\mathrm{Xe/Kr}}$ and most of the energy descriptors presented in Table S2.

2.6.4 Learning from higher temperature thermodynamics

The adsorption enthalpy of xenon at 298 K is very different from the adsorption enthalpy of xenon at ambient pressure given by GCMC calculations. However, when exploring the behavior at higher temperature (such as 900 K), we can find a better correlation with this xenon adsorption enthalpy as we can see in the Figure S1. The R^2 coefficient of determination increases from 0.80 to 0.92, which indicates a better consideration of the ambient-pressure enthalpy using higher temperature averaging. For this reason, we used this temperature to calculate the adsorption Gibbs free energy of xenon and krypton and also the Xe/Kr exchange Gibbs free energy. Differences between the 298 K and 900 K temperatures were then computed for the Xe/Kr exchange Gibbs free energies $\Delta_{\text{exc}}G^{\text{Xe/Kr}}(298\text{K}) - \Delta_{\text{exc}}G^{\text{Xe/Kr}}(900\text{K})$, enthalpies and entropies. We added these differences as descriptors, because they can inform the model on the energy differences between the low and ambient pressure cases which yields to better predictions.

2.6.5 Statistics on the energy distributions

Inspired by the thermodynamic averaging, we introduced other statistical transformations of the Boltzmann weighted energy distribution, like its standard deviation. To describe the multi-modality of the energy distribution we also introduced the Boltzmann weighted skewness and kurtosis; the Sarle's bimodality coefficient of Boltzmann weighted interaction energies can then be deduced from them. We can also retrieve statistical measures from the grid values of interaction energy as descriptors, without weighing by Boltzmann factors, to give a richer description of the distribution. For instance, the mean and standard deviation have been calculated for xenon and krypton.

2.7 Hyperparameter fine-tuning

We used the training data to perform a random search of hyperparameters, with 5-fold cross-validation to evaluate the root mean squared errors (RMSE) of the model. The range of search explored for each hyperparameter is made available in the SI. After this search, a set of optimal hyperparameters were identified, that give an average RMSE of $0.36\,\mathrm{kJ\,mol^{-1}}$; we used it to build our final model. A convergence plot of the model performed using 5-fold cross-validations is given in Figure S6. Given this configuration, the model is tested on the prior defined test-set and interpretation tools are used to better understand the structure-property relationships in play.

2.8 Interpretation of the final model

The final model is trained on the predefined training set using XGBoost with the fine-tuned hyperparameters. By testing it on the test set, we measure the accuracy of our approach, however, it is interesting to extract chemical insight into the hidden relationship between the predicted value and the descriptors, to better understand the thermodynamic origins of the performance. In this work, we used the Shapley values,[42] a game theory concept developed by Shapley in 1953, to measure the contribution of each descriptor in the predicted value. This tool is used locally to understand for a given structure how their characteristics had contributed to the prediction. To draw structure-property relationships, we would need to use a global interpretation methods such as the SHapley Additive exPlanations (SHAP) method thoroughly detailed in the online book *Interpretable Machine Learning* of Christoph Molnar.[43] The SHAP tool developed by Lundberg and Lee[44] is based on a faster algorithm adapted to tree-based ML models like gradient boosting, TreeSHAP, and integrates useful global interpretation modules like SHAP feature importance and dependence plot.

3 Results & Discussions

3.1 From infinite dilution to ambient pressure

The low-pressure selectivity provides a first intuition of the selectivity at higher pressure, as demonstrated in our previous work showing a correlation between the selectivity at both pressures.[18] If we adopt the Gibbs free energy formalism (Equation **??**), which correspond to a logarithmic transform of the selectivity values, this correlation is confirmed and highlighted on Figure 1. We can also note that although a majority of structures have similar selectivity in both pressure conditions, a handful of structures experience a selectivity drop at higher pressure. The zero-loading selectivity is always higher or similar to the ambient-pressure one, it gives therefore a solid ground on which to build an efficient prediction model. The second ingredient for a good prediction model is to build explanatory descriptors related to this selectivity drop phenomenon. One of the main causes to the selectivity drop being the presence of bigger pores that are less attractive xenon, therefore additional information on the pore size distributions or the energy landscape would be helpful for this task.

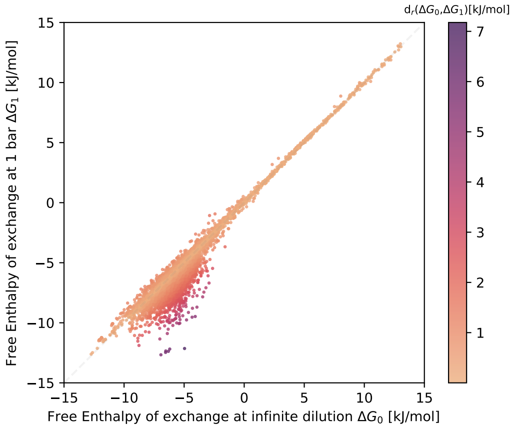


Figure 1: Comparison between the Gibbs free energy of exchange at low pressure ΔG_0 and ambient pressure ΔG_1 labeled by the relative distance between them. This plot is equivalent to a logarithmic plot of the selectivities at these two pressure conditions.

To incorporate information on the pore size diversity of the materials, we carried out statistical measurements on the PSD. By analyzing them, we detected explanatory factors at the origin of the observed selectivity drop. A high degree of multi-modality in the distribution would mean a diverse set of pores, which can lead to a selectivity drop if the pores are significantly different one from another. The more distant is the average pore size from the largest cavity diameter the higher the chance of observing a selectivity drop, because a big difference between the pore sizes bring about a lower selectivity. All these statistics are designed to give as much knowledge as possible on a hypothetical selectivity drop and on the quantitative estimation of its magnitude.

To better quantify the change of selectivity, it could be interesting to give statistics on the distribution of interaction energies for xenon and krypton calculated by our grid algorithm. These statistics include moments of different orders (up to 4) of the energy distribution, which informs on the adsorbate–adsorbent interaction energies in the nanopores at higher loading. The shape of the energy distribution can help assess quantitatively the change in selectivity. We can consider this as a way of compressing the whole energy distribution into a few statistical values, which is a standard method used in the field of data science to tackle distribution data. The same approach has also been applied to the Boltzmann weighted distributions to generate temperature specific descriptors for the energy distributions.

By using different temperatures, we noted that the infinite dilution adsorption enthalpies at higher temperatures can be better correlated to the adsorption enthalpy at ambient pressure. The minimum error was found for the adsorption enthalpy at 900 K, which gives an RMSE of $1.76\,\mathrm{kJ\,mol^{-1}}$ instead of $2.87\,\mathrm{kJ\,mol^{-1}}$ for the 298 K case. This new type of descriptor is very interesting since it better performs around the high selectivity region, where the standard Boltzmann average at 298 K loses its accuracy (see Figure S1). As we can see in the Figure S7, the exchange free energy at 900 K and the excess of free energy compared to the 298 K case are the second and third most influential descriptors of our ML model. They are complementary to the exchange free energy at 298 K to predict selectivities at higher pressures.

By combining the above-mentioned features with more standard geometrical descriptors, we trained an ML model for the ambient pressure selectivity that identifies the origins of the selectivity drop and gives promising prediction results.

By combining the above-mentioned features with more standard geometrical descriptors, we trained an ML model for the ambient pressure selectivity that identifies the origins of the selectivity drop and gives promising prediction results.

3.2 ML model performance

In this section, we present the performance of the ML model that learned the joint effects of all the newly introduced descriptors to detect and evaluate the observed drop between the easily accessible low-pressure selectivity and the more computationally demanding ambient-pressure selectivity. A GCMC simulation of a 20-80 xenon/krypton gas mixture takes in average 2.400 s per structure on the CoRE MOF 2019 database, while our grid-based adsorption calculation only takes about 5 s per structure (on a single Intel Xeon Platinum 8168 core at 2.7 GHz). To compute all features needed for our prediction, we would need less than a minute per structure, which is way faster than the 40 minutes required for a GCMC calculation. The ML-based approach has a very clear speed advantage over standard molecular simulations. But to be a good substitute, it needs to keep a good level of accuracy on an unseen set of structures.

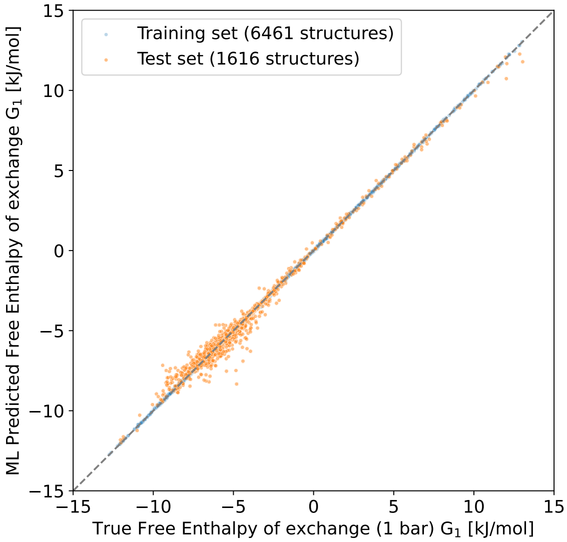


Figure 2: Scatter plot of the exchange free energy predicted by the model. There is a good agreement between the predicted and true values. On the test set, there is an RMSE of $0.37\,\mathrm{kJ\,mol^{-1}}$ and an MAE of $0.21\,\mathrm{kJ\,mol^{-1}}$. This plot is equivalent to the scatter plot between the logarithm of the ambient-pressure selectivities (see Figure S5 of the SI). The corresponding errors for the ambient-selectivity are 2.5 and 1.1 for respectively the RMSE and MAE of the selectivity, and 0.065 and 0.038 for the RMSE and MAE of its base-10 logarithm.

We defined a set of 80% randomly chosen structures out of the final dataset to train and fine-tune the parameters of our model. A randomized search over a range of maximum depths, learning rates, sizes of feature samples used by tree and by level, sizes of data sample and alpha regularization parameters has been performed and a set of hyperparameters have been chosen to minimize the average RMSE computed using a 5-fold cross-validation. The ranges used in the randomized search as well as the final hyperparameters set are given in SI. By using this parameterization, our XGBoost model has an RMSE of $0.37\,\mathrm{kJ\,mol^{-1}}$ and an MAE of $0.21\,\mathrm{kJ\,mol^{-1}}$ on the exchange Gibbs free energies of the test set of 1,616 structures. If we convert back these results to the selectivity values, the RMSE on the selectivity values would be 2.5 and 0.07 on the logarithm base 10 of the selectivity, which means that the order of magnitude of the selectivity is known with a very good accuracy. To prove that this good performance is not fortuitous, we used a 5-fold cross-validation procedure on the whole dataset and found an average RMSE of $0.36\,\mathrm{kJ\,mol^{-1}}$ with a standard deviation of $0.01\,\mathrm{kJ\,mol^{-1}}$, which is consistent with the performance given by a standard train/test split.

This method can later be used in a screening procedure that relies on cheap descriptors to skim off obviously undesirable structures to only keep the promising structures for the final ML model evaluation. For this is the reason, as previously explained in the methods, only the 3D MOF structures with an LCD above 4 Å are kept because they have a positive xenon affinity, which is a necessary condition for a good Xe/Kr selectivity. Our model being very good at predicting the ambient pressure selectivity of structures with good xenon affinity, the proposed screening procedure, illustrated Figure 3, would include (i) a check of the nature of the structure to ensure it is a 3D MOF structure, (ii) then a filter on the LCD value (above 4 Å), (iii) a pre-evaluation of the Xe/Kr selectivity at infinite dilution using the grid-based method, and (iv) finally the ML evaluation to keep only structures above a certain threshold of ambient-pressure selectivity (*e.g.* 30). We could eventually evaluate more thoroughly the top structures using GCMC simulations, *ab initio* calculations or adsorption experiments.

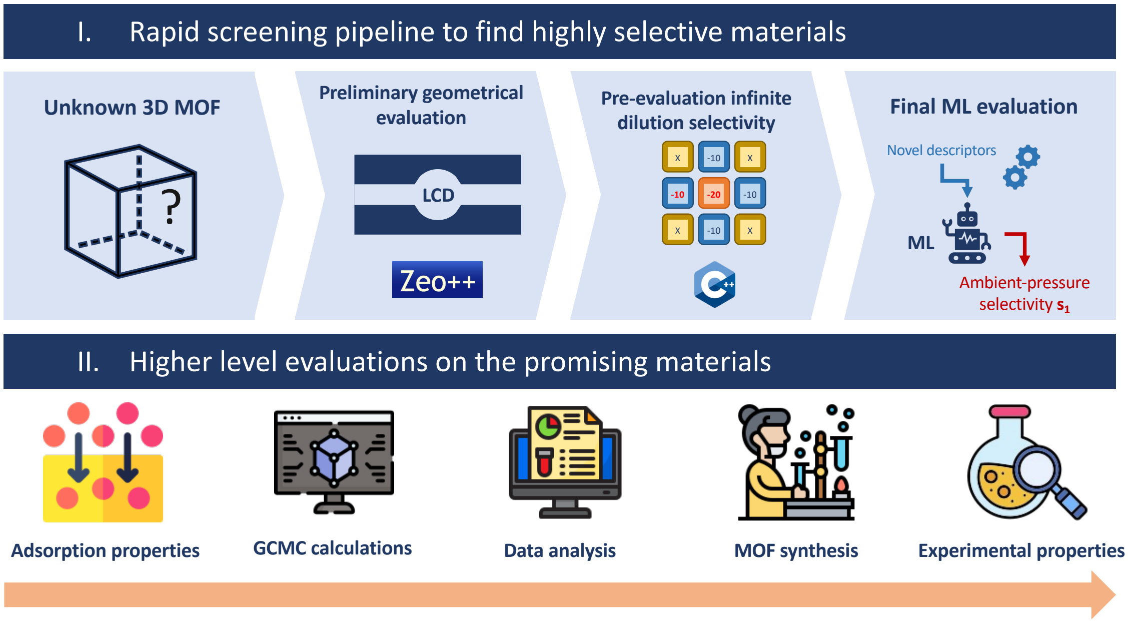

I. Rapid screening pipeline to find highly selective materials
Unknown 3D MOF
?
Preliminary geometrical evaluation
LCD
Zeo++
Pre-evaluation infinite dilution selectivity
x
-10
x
-10
-20
-10
x
-10
x
Final ML evaluation
Novel descriptors
ML
Ambient-pressure selectivity s_1
II. Higher level evaluations on the promising materials
Adsorption properties
GCMC calculations
Data analysis
MOF synthesis
Experimental properties

Figure 3: An illustration of the screening procedure that could be used to find highly selective materials.

3.3 Opening the black box

To better understand the intuition behind this selectivity drop, we used the SHAP[43,44] library of interpretation models to draw relationships between the descriptors and the predicted ambient-pressure selectivity. This code library is based on the calculation of Shapley values[42] that measure the contribution of each descriptor to the prediction to locally interpret our ML model. This interpretation model untangles the interdependence between the descriptors to extract an individual contribution. To go beyond the local interpretation, we can rapidly compute the Shapley values for the whole dataset using faster algorithms;[44] scatter plots of the contribution as a function of the descriptor values called SHAP dependence plots can then be drawn to make a more global interpretation of our ML model. Knowing a descriptor value, we could then infer, with a certain level of uncertainty, how it changes the final predicted value, which highlights unknown structure–property relationships. Finally, we can use the mean absolute Shapley values of each feature on the training set to measure the feature importance (see Figure S7 and S8).

3.3.1 Global interpretability

To rank the descriptors according to their average impact on the magnitude of the model output, we can use the mean absolute Shapley values of each descriptor. The importance plot associated with these values are presented in Figure S8. Even if the low-selectivity exchange Gibbs free energy has a SHAP importance value way above the others, it only serves as a baseline where a correlation close to the one presented on Figure 1 can be reached; the other descriptors play a major role in moving the outliers of the figure closer to the diagonal line. Energy descriptors play a major role in the model's prediction, and the geometry-based new descriptors, while playing a more secondary role, are key in evaluating the gaps between the low-pressure case with the ambient-pressure one that we are interested in. To dig deeper into the mechanisms that allow the model to predict the selectivity with a very good accuracy — the RMSE and MAE on the test set's selectivity being respectively 2.5 and 1.1 — we are now going to look into the SHAP dependence plots of each interesting descriptor that plots the contribution to the predicted value as a function of the actual descriptor value.

To make a global interpretation, we applied the partial dependence module provided by the SHAP library on our model. Although other methods to compute dependence plots exist (*e.g.* partial dependence plots),[43] we can keep a good level of consistency between our global and local interpretations by using the same underlying theory. The SHAP dependence plots of all the descriptors of the Figures S9 and S10, these plots have a rather distinct form, directions and shape, which is encouraging for the interpretability of our model. By looking at the profile of the dependence plots, we can extract valuable information on how the ML model predicts the ambient-pressure selectivity.

The most important descriptor is obviously the exchange free energy "G_0" associated to the low-pressure selectivity, its contribution has a very strong positive linear correlation (see Figure 4), which gives a base value on top of which the other contributions will either reduce the free energy (more selective) or increase it (less selective). The model can be interpreted as the combination of a baseline combined with smaller tweaks that estimate the magnitude of the deviation from the ideal low dilution case. For instance, the next two descriptors "G_900K" (900 K low-pressure exchange free energy) and "G_Xe_900K" (900 K low-pressure xenon adsorption free energy) continue to build up the baseline by providing information on the low-pressure selectivity, but they start giving a glimpse of deviations needed to differentiate between the structures experiencing a drop with the ones that keep their selectivity. As we can see in the SI (Figure S1 and S2), the thermodynamic quantities at high pressure is closer to the 900 K case than to the ambient temperature one, these two descriptors inform naturally on the selectivity at higher pressure. For "G_900K" (see Figure 4), blue points (corresponding to a "G_0" of around $-8\,\mathrm{kJ\,mol^{-1}}$) can have either negative or negligible contributions depending on the value; values below $-4\,\mathrm{kJ\,mol^{-1}}$ give a negative contribution with a linear relation, whereas values between -4 and $5\,\mathrm{kJ\,mol^{-1}}$ give constantly almost zero contributions. This type of domain differentiation illustrates how the model can identify structures with a selectivity drop based on the values of a descriptor. We will see more telling examples of how the contribution to the selectivity values are determined using the values of the remaining descriptors.

The U-shape of some SHAP dependence plots can highlight optimal values for the associated descriptors. For instance, the optimal value of "D_i_vdw_uff298" is around 5.1 (see Figure 4) and the optimal average of pore sizes is around 5.6. These optimal values match with the physical need of having pores of the size of a xenon to be more attractive to it, which was identified in several papers in the literature. We can note that these values are a bit higher than the ones mentioned in the literature due to the different definition of the atom radii.[33] Moreover, values of "delta_G0_298_900" between 4 and $6\,\mathrm{kJ\,mol^{-1}}$ (see Figure 4) have a higher chance of giving a negative contribution, which means a lower ambient-pressure selectivity. These sweet spots constitute valuable hints to tell the truly selective materials from the others. Some SHAP dependence plots have a rather linear domain for the most selective structures (in blue) — the difference of pore volumes between Xe and Kr sized probes "delta_VF_18_20" have a good linear contribution (see Figure 4), which means that the lower the more selective the structure will be. The same can be said for the standard deviations of the PSD "pore_dist_std" and of the Boltzmann weighted krypton interaction energies distribution "enthalpy_std_krypton". The optimal values for these descriptors are zero, the closest to zero it is the more negative the contribution will be and the more selective the structure at ambient pressure.

Sometimes the optimal values are not around well-identified values but are contained within larger domains with threshold values separating them. For instance, the difference between the LCD and the average pore size "delta_pore" has a threshold value around 0.3 Å below which the contribution for the most selective structures (blue) is negative (see Figure 4); even though no clear correlations can be found, we can at least find a threshold value (about 0.23) below which there is higher probability of having a high ambient-pressure selectivity. The same type of domain splits can be found for the average of krypton interaction energies distribution "mean_grid_krypton" (at around 15), the Boltzmann weighted xenon interaction energies distribution "enthalpy_std_xenon" (at around 2.5), the difference of exchange entropic term between the ambient temperature "delta_TS0_298_900" (at around 3) and high temperature and the effective number associated to the PSD "pore_dist_neff" (at around 2.3). These domains separate structures that are selective at low pressure, which is key to telling apart the structures with a selectivity drop at ambient pressure from the ones without.

3.3.2 Local interpretability

To put into practice our previous analysis, let's look at some archetypal structures and how the model predicted the selectivity based on the descriptor values. We chose two MOF structures from the test set, their CSD code being respectively VIWMIZ and BIMDIL. Both structures are selective at low pressure but the first one decreases in selectivity while the other maintains it at ambient pressure. It will be interesting to see what the model does to tell apart these two completely different behaviors.

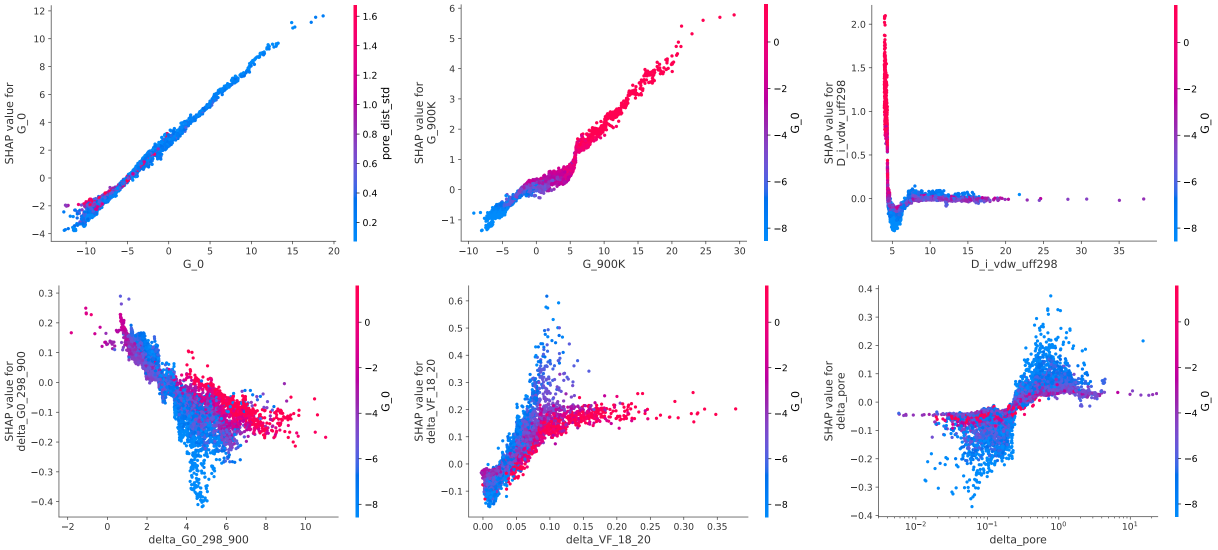
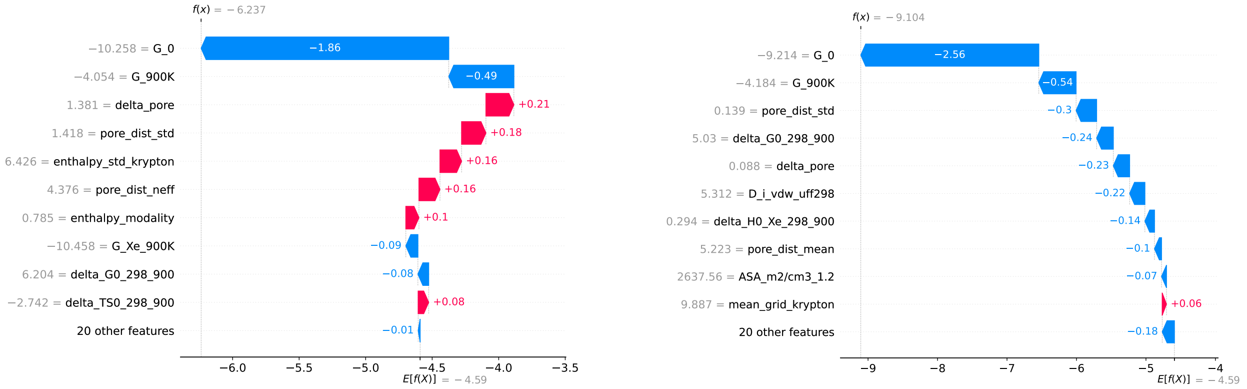


Figure 4: Some SHAP dependence plots that are analyzed in the main article. The 18 top descriptors' SDPs can be found in the SI.

VIWMIZ is part of the highly selective structures that experience a selectivity drop at ambient pressure. If we convert back the free energy values to selectivity values, its selectivity is 62.8 at infinite dilution and 14.5 at ambient pressure. The ML model manages to give a close prediction of 12.0 for the ambient-pressure selectivity based on the given values of the descriptors. If we only look at "G_0", it has one of the most negative values, which explains the rather high negative contribution of -1.81. However, the -0.57 contribution of "G_900K" is rather low compared to other materials (see Figure 4), since a value of -4.05 is not the most negative considering all structures. On the other hand, the remaining descriptors have values in the domain of positive contributions, which lead to the drop of the selectivity. For example, the difference of pore sizes "delta_pore" has a value of 1.38 Å (above the threshold of 0.23 Å), which contributes +0.25 to the predicted selectivity and is consistent with the value ranges of the associated dependence plot. By reporting the values to the dependence plots, the same analyses can be made on the other positive contributions of the Figure 5: "pore_dist_std" is above the threshold of 0.4, "enthalpy_std_krypton" is above 2.5 $\mathrm{kJ\,mol^{-1}}$, "pore_dist_neff" is above 2.3, "delta_TS0_298_900" is below 3 $\mathrm{kJ\,mol^{-1}}$ and "enthalpy_modality" is around 0.75 where positive contributions are more commonly observed. However, the "delta_G0_298_900" value is a bit too close to its optimal value, which explains its negative contribution in this particular prediction. The rest of the features have almost negligible contributions and are detailed in the Figure S11. By analyzing the contributions of each descriptor to the prediction given by our model, we can understand the underlying features of the VIWMIZ structure that explains the selectivity drop at higher pressure. The shape of the xenon and krypton energy distributions ("enthalpy_std_krypton" and "enthalpy_modality") and of the PSD ("pore_dist_std" and "pore_dist_neff") as well as the void fraction difference "delta_pore" are key descriptors at the origin of the lower selectivity at ambient pressure compared to the ideal infinite dilution case. Intuitively, one can easily understand that effective number of pores exceeding 2 can mean the presence of different pore sizes, which is consistent with the presence of pores that are less attractive to the xenon and leads necessarily to less selectivity. The previous statement is also very much consistent with a high standard deviation of the PSD or the Boltzmann weighted krypton interaction energy distribution. One can also conceive that a much larger difference between the average pore size and the LCD could mean a high disparity in pore sizes that leads to the presence of larger pores more and more loaded as the pressure rises. The entropic term is however way more complex to interpret and opens unexplored ways of tackling the problem of selectivity drop at higher pressure unraveled by our previous study[18].



(a) VIWMIZ: true $\Delta_{\mathrm{exc}}G_1 = -6.63$

(b) BIMDIL: true $\Delta_{\mathrm{exc}}G_1 = -9.20$

Figure 5: Main contributions of the descriptors on the selectivity prediction of two archetypal examples. The descriptor labels used are detailed in the Table S1 and S2 of the SI.

The second structure BIMDIL is also among the most selective with a selectivity at low pressure of 41.0, while maintaining it to 41.2 at ambient pressure. The model manages to predict this stability of the selectivity by giving a value of 40.0. Consequently, the first contribution of "G_0" is among the most negative ones and set a baseline of -2.4 for the upcoming contributions. The contributions of "G_900K" and "G_900K" are not the highest possible but they continue to lower down the value of the predicted selectivity. It is the joint contributions of the other descriptors that will really discriminate between the two structures and decide why this one will keep its selectivity. Unlike the previously analyzed structure, this one has a "delta_pore" value below 0.3 Å, which explains the negative Shapley value it has for our prediction. The contribution of "delta_G0_298_900" that was only a little negative for the other one, is now playing a major role since it is right within the range of between 4 and 6 $\mathrm{kJ\,mol^{-1}}$ (see Figure 5). We can also verify that "pore_dist_std" is now below the threshold instead of being above for the other structure. We can confirm that the other contributions are also following the rules implied by the SHAP dependence plots, no apparent anomalies are detected, and the joint efforts of all the descriptors tend to give a lower free energy value, which leads to the conservation of the selectivity value at higher pressure. The set of descriptor values is clearly very different from the previous structure, many values are in opposite contribution domains, which explains how the model manages to disentangle the highly selective structures to find out the ones that would keep their selectivity at higher pressure.

These two examples allow us to understand a bit more how the model tells apart the structures that will lose selectivity at higher pressure from the ones that will not. Most of the dependence plots can give very strong association between the descriptors and their effects; the outliers are rare enough that the inner logic of our model can be understood. As developed previously, the first three descriptors set a baseline on few information on the eventual drop of selectivity; then the other descriptors contribution is either positive, negligible or negative depending on the domain of values the descriptor is in. For instance, the average pore size and the largest cavity diameter need to be around very specific values to maximize the chance of keeping the selectivity at higher pressure, which was highlighted by previous works that emphasize on the importance of pore sizes close to the size of xenon for Xe/Kr separation.[14,18] The difference of entropy between the ambient temperature and 900 K is surprising descriptor that separates selective structures depending on whether its value is within a given range. The difference of void fraction occupied by xenon and krypton is also very interesting since it affects the selectivity differently depending on whether it is highly selective or not, and the contribution is more or less proportional to its value. Different ways of measuring the disparity of the PSD and interaction energy distribution are key in sorting highly selective structures (in blue on the dependence plot Figure 4) between the ones maintaining their performance and the ones decreasing in selectivity. Among others, we can find the difference between the average pore size and the LCD, as well as the standard deviation of the PSD or of the Boltzmann weighted energy distribution that would behave very differently according to the domain in which the value lies. The SHAP dependence plots, partially plotted in the main text and entirely available in the SI, are very valuable reading grid to understand the mechanisms behind our ML model and more broadly to what it understood from the origins of Xe/Kr separation.

4 Conclusions and perspectives

In order to better understand separation processes inside nanoporous materials, we performed a machine learning prediction of Xe/Kr ambient-pressure selectivity that is faster than standard GCMC calculations. For MOF structures of the CoRE MOF 2019 database, a xenon/krypton selectivity evaluation would take less than a minute, while an equivalent GCMC calculation takes around 40 min. Unlike most of the selectivity predictions of the literature, we chose to predict a selectivity in the logarithmic scale, because it focuses more on the order magnitude than the exact value of the selectivity of highly selective materials.

Moreover, the conversion to an exchange Gibbs free energy allows a more thermodynamic approach based on enthalpy, entropy and free energy values. The challenge was then to predict a free energy equivalent of the ambient-pressure selectivity by using the low-pressure selectivity along with key energy, geometrical and chemical descriptors. The final, fully optimized ML model performs very well with an RMSE of $0.36\,\mathrm{kJ\,mol^{-1}}$, which corresponds to a 0.06 RMSE on the base-10 log of the selectivity.

One of our more specific goals was to uncover underlying reasons of a selectivity drop at high pressure observed on some highly selective materials at low pressure. Previous studies found that a high diversity of pore sizes and channel sizes that favor adsorbate reorganizations could be at the origin of this phenomenon.[18] By applying interpretability tools, we found quantitative factors that explain the conservation or the drop of the selectivity for highly selective materials. Depending on energy averaging at 900 K, on statistical characterizations of the energy or pore size distributions, and on the difference of volumes occupiable we have a structure either with a selectivity similar to the low-pressure case or that is less selective at higher pressure. All the quantitative rules are contained in a complex ensemble of decision trees constructed by our XGBoost model, and they can be extracted to build rule of thumbs in order to back our intuition on the Xe/Kr selectivity in MOF structures.

The final ML model can be used in a well-designed workflow to find the best performing materials. For instance, we could filter out the structures with pores that cannot fit a xenon in, then we could use a first calculation of the low-pressure selectivity to filter out the selectivity below a given threshold. Finally, we can use the model to remove the structures that would experience a selectivity drop. We tested our methodology on the Xe/Kr separation as proof of concept since it is one of the simplest adsorption systems (monoatomic species with no electrostatic interactions). A similar approach can be generalized to other separation applications by calculating the infinite dilution energies with a more standard method (*e.g.* Widom's insertion) and by adjusting the descriptor definitions to fit the adsorbates of interest.

This study ambitions to add new descriptor ideas to help the development of ever more efficient screening methodologies to find the best materials for target applications. However, like many other studies on the topic, this one also relies on a few strong assumptions — the simulations are performed in rigid frameworks with non-polarized classical force fields. As suggested in the literature, the most selective materials ever synthesized for Xe/Kr separation are all based on the effect of open-metal sites that uses the difference of polarizability between the two molecules to efficiently separate them.[5,6] Moreover, the structures can be made flexible using flexible force fields with adapted simulation methodologies[45] or by using multiple rigid simulations of snapshots from NPT simulations[46]. It would be possible to improve the simulations at the cost of CPU times, if we coupled it with a reduction of simulation time like the one presented in this article. The quest of ever-faster evaluation tools will allow us to investigate more complex properties and uncover structures with ever more interesting characteristics.

Conflicts of interest

There are no conflicts to declare.

Acknowledgement

The authors thank Philippe Guilbaud and Isabelle Hablot for discussions on the topic of adsorption-based separation.

Funding

This work was financially supported by Orano.

Supporting Information Available

Additional information in the supporting information file, raw data available online at `https://github.com/fxcoudert/citable-data`, the Grid Adsorption Energy Sampling code available on `https://github.com/coudertlab/GrAED` and `https://github.com/eren125/ml-selectivity`

References

(1) Kerry, F. G. *Industrial gas handbook: gas separation and purification*; CRC press, 2007.

(2) National Academies of Sciences, Engineering, and Medicine, *A Research Agenda for Transforming Separation Science*; The National Academies Press: Washington, D.C., 2019.

(3) Banerjee, D.; Simon, C. M.; Elsaidi, S. K.; Haranczyk, M.; Thallapally, P. K. Xenon Gas Separation and Storage Using Metal-Organic Frameworks. *Chem* **2018**, *4*, 466–494.

(4) Chen, L. et al. Separation of rare gases and chiral molecules by selective binding in porous organic cages. *Nature Mater.* **2014**, *13*, 954–960.

(5) Li, L.; Guo, L.; Zhang, Z.; Yang, Q.; Yang, Y.; Bao, Z.; Ren, Q.; Li, J. A Robust Squarate-Based Metal–Organic Framework Demonstrates Record-High Affinity and Selectivity for Xenon over Krypton. *J. Am. Chem. Soc.* **2019**, *141*, 9358–9364.

(6) Pei, J.; Gu, X.-W.; Liang, C.-C.; Chen, B.; Li, B.; Qian, G. Robust and Radiation-Resistant Hofmann-Type Metal–Organic Frameworks for Record Xenon/Krypton Separation. *J. Am. Chem. Soc.* **2022**, *144*, 3200–3209.

(7) Lyu, H.; Ji, Z.; Wuttke, S.; Yaghi, O. M. Digital Reticular Chemistry. *Chem* **2020**, *6*, 2219–2241.

(8) Jablonka, K. M.; Rosen, A. S.; Krishnapriyan, A. S.; Smit, B. An Ecosystem for Digital Reticular Chemistry. *ACS Central Science* **2023**,

(9) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Cryst. B* **2016**, *72*, 171–179.

(10) Wilmer, C. E.; Leaf, M.; Lee, C. Y.; Farha, O. K.; Hauser, B. G.; Hupp, J. T.; Snurr, R. Q. Large-scale screening of hypothetical metal–organic frameworks. *Nature Chem.* **2011**, *4*, 83–89.

(11) Boyd, P. G.; Woo, T. K. A generalized method for constructing hypothetical nanoporous materials of any net topology from graph theory. *CrystEngComm* **2016**, *18*, 3777–3792.

(12) Colón, Y. J.; Gómez-Gualdrón, D. A.; Snurr, R. Q. Topologically Guided, Automated Construction of Metal–Organic Frameworks and Their Evaluation for Energy-Related Applications. *Cryst. Growth Des.* **2017**, *17*, 5801–5810.

(13) Ren, E.; Guilbaud, P.; Coudert, F.-X. High-throughput computational screening of nanoporous materials in targeted applications. *Digital Discovery* **2022**, *1*, 355–374.

(14) Simon, C. M.; Mercado, R.; Schnell, S. K.; Smit, B.; Haranczyk, M. What Are the Best Materials To Separate a Xenon/Krypton Mixture? *Chem. Mater.* **2015**, *27*, 4459–4475.

(15) Rycroft, C. H. VORO++: A three-dimensional Voronoi cell library in C++. *Chaos* **2009**, *19*, 041111.

(16) Ren, E.; Coudert, F.-X. Rapid adsorption enthalpy surface sampling (RAESS) to characterize nanoporous materials. *Chem. Sci.* **2023**,

(17) Shi, K.; Li, Z.; Anstine, D. M.; Tang, D.; Colina, C. M.; Sholl, D. S.; Siepmann, J. I.; Snurr, R. Q. Two-Dimensional Energy Histograms as Features for Machine Learning to Predict Adsorption in Diverse Nanoporous Materials. *J. Chem. Theory Comput.* **2023**,

(18) Ren, E.; Coudert, F.-X. Thermodynamic exploration of xenon/krypton separation based on a high-throughput screening. *Faraday Discuss.* **2021**, *231*, 201–223.

(19) Chung, Y. G.; Haldoupis, E.; Bucior, B. J.; Haranczyk, M.; Lee, S.; Zhang, H.; Vogiatzis, K. D.; Milisavljevic, M.; Ling, S.; Camp, J. S.; Slater, B.; Siepmann, J. I.; Sholl, D. S.; Snurr, R. Q. Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: CoRE MOF 2019. *J. Chem. Eng. Data* **2019**, *64*, 5985–5998.

(20) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA, 2016; pp 785–794.

(21) Dubbeldam, D.; Calero, S.; Ellis, D. E.; Snurr, R. Q. RASPA: molecular simulation software for adsorption and diffusion in flexible nanoporous materials. *Mol. Simulat.* **2016**, *42*, 81–101.

(22) Rappé, A. K.; Casewit, C. J.; Colwell, K.; Goddard III, W. A.; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024–10035.

(23) Ryan, P.; Farha, O. K.; Broadbelt, L. J.; Snurr, R. Q. Computational screening of metal-organic frameworks for xenon/krypton separation. *AIChE Journal* **2010**, *57*, 1759–1766.

(24) Willems, T. F.; Rycroft, C. H.; Kazi, M.; Meza, J. C.; Haranczyk, M. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Microporous Mesoporous Mater.* **2012**, *149*, 134–141.

(25) Diwu, J.; Nelson, A.-G. D.; Wang, S.; Campana, C. F.; Albrecht-Schmitt, T. E. Comparisons of Pu(IV) and Ce(IV) Diphosphonates. *Inorg. Chem.* **2010**, *49*, 3337–3342.

(26) Martin, N. P.; März, J.; Volkringer, C.; Henry, N.; Hennig, C.; Ikeda-Ohno, A.; Loiseau, T. Synthesis of Coordination Polymers of Tetravalent Actinides (Uranium and Neptunium) with a Phthalate or Mellitate Ligand in an Aqueous Medium. *Inorg. Chem.* **2017**, *56*, 2902–2913.

(27) Jouffret, L.; Rivenet, M.; Abraham, F. Linear Alkyl Diamine-Uranium-Phosphate Systems: U(VI) to U(IV) Reduction with Ethylenediamine. *Inorg. Chem.* **2011**, *50*, 4619–4626.

(28) Liang, L.; Zhang, R.; Zhao, J.; Liu, C.; Weng, N. S. Two actinide-organic frameworks constructed by a tripodal flexible ligand: Occurrence of infinite $\{(UO_2O_2(OH)_3\}_{4n}$ and hexanuclear $\{Th_6O_4(OH)_4\}$ motifs. *J. Solid State Chem.* **2016**, *243*, 50–56.

(29) Fernandez, M.; Woo, T. K.; Wilmer, C. E.; Snurr, R. Q. Large-Scale Quantitative Structure–Property Relationship (QSPR) Analysis of Methane Storage in Metal–Organic Frameworks. *J. Phys. Chem. C* **2013**, *117*, 7681–7689.

(30) Fanourgakis, G. S.; Gkagkas, K.; Tylianakis, E.; Froudakis, G. E. A Universal Machine Learning Algorithm for Large-Scale Screening of Materials. *J. Am. Chem. Soc.* **2020**, *142*, 3814–3822.

(31) Anderson, R.; Gómez-Gualdrón, D. A. Large-Scale Free Energy Calculations on a Computational Metal–Organic Frameworks Database: Toward Synthetic Likelihood Predictions. *Chem. Mater.* **2020**, *32*, 8106–8119.

(32) Pardakhti, M.; Nanda, P.; Srivastava, R. Impact of Chemical Features on Methane Adsorption by Porous Materials at Varying Pressures. *J. Phys. Chem. C* **2020**, *124*, 4534–4544.

(33) Hung, T.-H.; Lyu, Q.; Lin, L.-C.; Kang, D.-Y. Transport-Relevant Pore Limiting Diameter for Molecular Separations in Metal–Organic Framework Membranes. *J. Phys. Chem. C* **2021**, *125*, 20416–20425.

(34) Ongari, D.; Boyd, P. G.; Barthel, S.; Witman, M.; Haranczyk, M.; Smit, B. Accurate Characterization of the Pore Volume in Microporous Crystalline Materials. *Langmuir* **2017**, *33*, 14529–14538.

(35) Pinheiro, M.; Martin, R. L.; Rycroft, C. H.; Jones, A.; Iglesia, E.; Haranczyk, M. Characterization and comparison of pore landscapes in crystalline porous materials. *J. Mol. Graph. Model.* **2013**, *44*, 208–219.

(36) Tarbă, N.; Voncilă, M.-L.; Boiangiu, C.-A. On Generalizing Sarle's Bimodality Coefficient as a Path towards a Newly Composite Bimodality Coefficient. *Mathematics* **2022**, *10*, 1042.

(37) Laakso, M.; Taagepera, R. "Effective" Number of Parties. *Comparative Political Studies* **1979**, *12*, 3–27.

(38) Simpson, E. H. Measurement of Diversity. *Nature* **1949**, *163*, 688–688.

(39) Kramer, B.; MacKinnon, A. Localization: theory and experiment. *Rep. Prog. Phys.* **1993**, *56*, 1469–1564.

(40) Wojdyr, M. GEMMI: A library for structural biology. *JOSS* **2022**, *7*, 4200.

(41) Widom, B. Some Topics in the Theory of Fluids. *J. Chem. Phys.* **1963**, *39*, 2808–2812.

(42) Shapley, L. S., et al. A value for n-person games. **1953**,

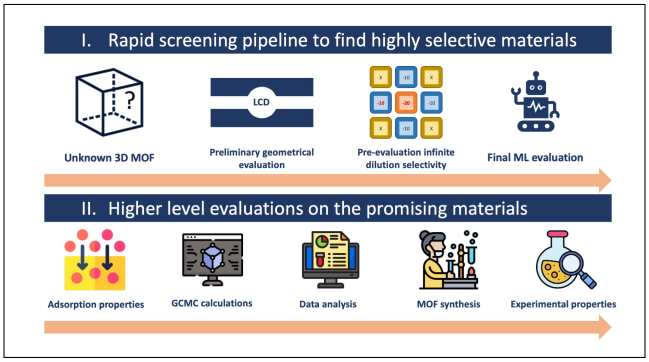
(43) Molnar, C. Interpretable machine learning. Available online at `https://christophm.github.io/interpretable-ml-book/`, 2023.

(44) Lundberg, S. M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems. 2017.

(45) Bousquet, D.; Coudert, F.-X.; Boutin, A. Free energy landscapes for the thermodynamic understanding of adsorption-induced deformations and structural transitions in porous materials. *J. Chem. Phys.* **2012**, *137*, 044118.

(46) Witman, M.; Ling, S.; Jawahery, S.; Boyd, P. G.; Haranczyk, M.; Slater, B.; Smit, B. The Influence of Intrinsic Framework Flexibility on Adsorption in Nanoporous Materials. *J. Am. Chem. Soc.* **2017**, *139*, 5547–5557.

TOC Graphic


I. Rapid screening pipeline to find highly selective materials
?
Unknown 3D MOF
LCD
Preliminary geometrical evaluation
X
-10
X
-10
-20
-10
X
-10
X
Pre-evaluation infinite dilution selectivity
Final ML evaluation
II. Higher level evaluations on the promising materials
Adsorption properties
GCMC calculations
Data analysis
MOF synthesis
Experimental properties