# Enhancing Gas Separation Selectivity Prediction through Geometrical and Chemical Descriptors

Emmanuel Ren, François-Xavier Coudert

# Enhancing Gas Separation Selectivity Prediction through Geometrical and Chemical Descriptors

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

†*CEA, DES, ISEC, DMRC, Univ. Montpellier, Marcoule, 30207 Bagnols-sur-Cèze, 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 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 the $\log_{10}$ of the selectivity. 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, there are 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 treatment plants for 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]. This ever-increasing amount of structures requires more efficient

screening procedures as well as faster 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.[13] To further improve the selectivity screening for Xe/Kr separation, the research needs to focus on designing 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, i.e., an average of the interaction energies of a xenon atom at each Voronoi node.[15] To rationalize this increase in performance, this Voronoi energy can be regarded as a faster proxy for the adsorption enthalpy. This Voronoi sampling was much faster than a standard Widom insertion, but also much less accurate. Therefore, we recently developed a more effective alternative, a surface sampling algorithm (RAESS) using symmetry and non-accessible volumes blocking to speed up the calculation of relevant interaction energies within a porous framework.[16] Recently, Shi et al. used an energy grid to generate energy histograms as a descriptor for their ML model, providing an exhaustive description of the infinitely diluted adsorption energies.[17]

All the approaches described above can accurately predict the 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 provides clarifications on the origin of the differences in selectivity values. For some materials, selectivity could drop when increasing the pressure in the Xe/Kr separation application. And, this was mainly attributed to the presence of different pore sizes

and potential reorganizations due to adsorbate–adsorbate interactions.[18]

In this article, we develop a new adsorption energy sampling technique using a grid-based approach. Moreover, we perform a statistical characterization of the pore size and energy distributions to inform the model on a potential selectivity drop. By combining these two approaches, we introduce a set of useful ML descriptors for fast and accurate ambient-pressure selectivity prediction, and we highlight its performance in the case of xenon/krypton separation for the CoRE MOF 2019 database[19].

# 2 Methods

## 2.1 The machine learning model

We choose the eXtreme Gradient Boosting (XGBoost) algorithm as the machine learning model architecture due to its accuracy, efficiency, and simplicity of use. Its performance has been extensively demonstrated, as evidenced by 17 out of 29 winning solutions in Kaggle Challenges being based on this algorithm in 2015. The XGBoost system is highly scalable and parallelized, resulting in fast model training.[20] Compared to more conventional tree-based algorithms like random forest (commonly used in the field[14]), the boosting component of the algorithm enables learning from previous mistakes and allocating greater effort to problematic data points, thereby improving the accuracy of the final ML model.

In the following sections, we will introduce new descriptors for nanoporous materials, along with novel concepts of feature engineering based on energy and pore size histograms. We select the ML features through progressive filtering, eliminating less influential features based on the performance on the training set. The complete list can be found in Table S1-3 of the Supporting Information (SI). We will define the influence or importance of these features in a subsequent section dedicated to model interpretation. We also fine-tune the hyperparameters of the model through random searches to design the best-performing final model. Lastly, we will use a unified approach to interpret the influence of the preselected

descriptors on the final model.

## 2.2 Target variable

This study aims at building an ML model to predict the Xe/Kr ambient-pressure selectivity faster than standard techniques. To obtain reference values (ground truth in this study), we use the RASPA2 software[21] to run 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 forcefield (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).

To achieve this, we consider a logarithmic transform of the selectivity instead of the raw value because the goal is rather to predict the order of magnitude of the selectivity values than to directly predict the higher values of selectivity — an ML model that focuses its prediction on raw selectivity values can reach lower errors by simply focusing on the higher values than the lower ones. The use of a logarithmic transform better separates the different selectivity categories through the different orders of magnitude of the selectivity values. This approach distributes more evenly the efforts on the whole spectrum of selectivity values. Moreover, this logarithmic transformation is effectively an exchange Gibbs free energy (defined later in equation 1), so that we can easily compare it with the energy descriptors introduced in this article.

## 2.3 Database and data generation

We test this methodology on a set of realistic MOFs by considering 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 unit cell volume of 20 $nm^3$, the database is reduced to a set of 9 748 structures. Then, with the string information given by the Zeo++ software[24] this number is reduced 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]), which is a source of risks in a nuclear waste processing plant. Furthermore, the structures with pore sizes allowing the adsorption of xenon are selected using a condition on the largest cavity diameter (LCD): this is the case for 8 529 structures with an 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.

Then, we calculate the descriptors summarized below (and fully detailed in SI) on this restrained dataset. At this stage, 140 structures fail in the GCMC calculation due to RAM limitations and 83 have no standard deviation for the pore distribution (skewness and kurtosis cannot be retrieved). In the following training–testing process, we will therefore use a final dataset of 8 300 structures used to perform our ML-assisted method of screening the Xe/Kr adsorption selectivity. Based on this final set, the model is trained on 80% of randomly selected structures (6 640 structures) and tested on the remaining structures (1 660 structures). However, Jablonka et al. have recently criticized the standard train/test split (as used in this study) because of the multiple occurrences of similar structures in the data. Therefore, we also compare the results obtained by the GroupShuffleSplit function of sklearn. In this grouped split, similar structures labeled using the chemical formula of the MOFs are always grouped in either the training or the test set, hence avoiding the aforementioned problem. Doing so, we did not notice any significant alteration of the generalization error, which we attribute to the high number and the diversity of the structures we are deadling with. 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

Examining a number of research papers on supervised ML for the prediction of adsorption properties,[14,29–32] we identify a few recurrent descriptors: (i) geometrical descriptors obtained using software like Zeo++[24] including surface area (SA), void fraction (VF), largest cavity diameter (LCD) and pore limiting diameter (PLD); and (ii) physical and chemical descriptors such as framework density, framework molar mass, percentage of carbon (C%), nitrogen (N%), oxygen (O%), hydrogen, as well as halogen, nonmetals, metalloids and metals, and degree of unsaturation. Although these descriptors are versatile and widely used in ML models, they fail to provide specific information for the ML task of this study. As demonstrated by Simon et al., energy descriptors greatly influence ML models for selectivity prediction.

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

We test several values of the surface area exposed to different probe sizes (1.2 Å, 1.8 Å and 2.0 Å). The probe occupiable volume is 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 pore volume demonstrates a better agreement with experimental nitrogen isotherms.[34]

Given the objective of predicting the difference between low-pressure selectivity and ambient-pressure selectivity (for a specific gas mixture composition), some descriptors hold

little importance, and the key factor lies in the difference in accessible volume and affinity of the remaining pore volume with xenon compared to krypton. The intuition developed in the previous study outlines the role of a diverse distribution of pores with different xenon affinities.[18] We test different combinations of geometrical descriptors (along with the following energy and pore size distribution descriptors) using a cross-validation scheme on the training data. Using these accuracy results, from all the "standard" descriptors mentioned in the literature, the following 7 descriptors are retained: 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"). Additionally, we introduce a new descriptor $\Delta$VF to represent the difference in void fraction values, specifically the difference in volumes occupiable by xenon (2.0 Å) and krypton (1.8 Å). A comprehensive presentation of all these descriptors, including other geometrical descriptors based on pore size distribution, can be found in Table S1.

## 2.5 Pore size distribution

We use Monte Carlo steps to measure the frequency of every accessible pore sizes binned by 0.1 Å and to generate a histogram of pore sizes (or pore size distribution, PSD).[35] This histogram then generates descriptors based on statistical parameters describing 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 introduce two additional moments: the skewness ($\gamma$), corresponding to the third standardized moment, measures the asymmetry of a distribution; and the kurtosis ($k$), being the fourth standardized moment, measures the relative weight of the distribution's tails. Recognizing the importance of characterizing the number of different pore sizes that are likely to have contributed to the observed selectivity drop, we try to find a simple descriptor for measuring the number of modes in the distribution. The Sarle's bimodality coefficient, $\mathrm{BC} = (\gamma^2 + 1)/k$, provides a simple quantification of the extent to which the distribution deviates from unimodality by considering only skewness and kurtosis.[36]

Finally, to assess the diversity of pores, we introduce an effective number $n_{\text{eff}} = N^2/\sum n_i^2$ of pore sizes, where $N$ represents the total number of points in the histogram and $n_i$ the number of points associated with the $i^{\text{th}}$ bin. This number bears resemblance to a statistical measure widely used in other scientific domains. In political science, it is used to measure the effective number of political parties,[37] while in ecology, the inverse Simpson's index evaluates the species diversity in an ecosystem[38]. Similarly, in quantum physics, the inverse participation number measures the degree of localization of a wave-function.[39] In our case, this effective number of pore sizes gives an idea of the diversity of pore sizes (considering a binning of 0.1 Å). A large effective number suggests that multiple pore sizes are well represented in the structure. Thus, this descriptor provides insight into the scattering of pore sizes within the distribution.

All these descriptors contain valuable information regarding the form of the PSD required to understand 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 propose another approach based on symmetry-respecting grids. We generate these grids 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 $\sigma_{g-h}$ is the distance at which the LJ potential energy between the guest atom $g$ and the host atom is zero. Then, the LJ potential energy between the guest molecule and the closest host atom is calculated and only the grid points with an energy lower than a predefined threshold (here set to $100\,\text{kJ}\,\text{mol}^{-1}$) 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 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 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 divides the time required by the average number symmetry images — this module will be referred to as "symmetry" in 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). We will refer to this new sampling technique as the GrAED algorithm in the following text.

To highlight the improvement in performance in this procedure: the average void fraction for a 1.2 Å probe radius equals to 0.16 and the average number of symmetric images equals 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" reduces this number of points to 17%, and the combination of both reduces it 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.

From the energy values of this grid, we can now calculate many useful descriptors that are used in our final model. These energy-based descriptors are calculated using the GrAED algorithm except for the ambient-pressure case which is handled using GCMC simulations. A fully detailed description of these descriptors as well as their labeling names are given in Table S2.

Table 1: Performance comparison of the new grid method to other standard techniques used to calculate the xenon adsorption enthalpies. The RMSE is calculated by comparing to the values given by a 100k-step Widom insertion considered as the ground truth. The associated calculations are performed on the structures with the $\mathrm{LCD}_{\mathrm{CCDC}}$ over 3.7 Å of 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 – 0.12 Å | 35.4 | 0.014 |
| Grid – blocking – 0.12 Å | 10.4 | 0.014 |
| Grid – symmetry – 0.12 Å | 8.3 | 0.014 |
| Grid – fast – 0.12 Å | 4.8 | 0.014 |
| Grid – fast – 0.3 Å | 0.13 | 0.21 |
| RAESS[16] | 0.34 | 0.33 |
| Widom[41] (12k cycles) | 150 | 0.01 |

### 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 $K_{\mathrm{H}}$ corresponds to the average of the Boltzmann factors $\langle \exp(-\mathcal{E}_{\mathrm{int}}/RT) \rangle$, 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 $\Delta_{\mathrm{ads}}G$ can then be deduced from the Henry's constant since $\Delta_{\mathrm{ads}}G = -RT \ln \left( \langle \exp(-\mathcal{E}_{\mathrm{int}}/RT) \rangle \right)$, and finally the adsorption entropy naturally derives 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 are either occupied by adsorbate A or adsorbate B, leaving the other in the gas phase. We model this equilibrium by the equation $\mathrm{A_{(ads)}} + \mathrm{B_{(gas)}} = \mathrm{A_{(gas)}} + \mathrm{B_{(ads)}}$, 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 then simply derives from

the selectivity:

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

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

### 2.6.4 Learning from higher temperature thermodynamics

The adsorption enthalpy of xenon at infinite dilution 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 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 use this temperature to calculate the adsorption Gibbs free energy of xenon and krypton and also the Xe/Kr exchange Gibbs free energy. Then, we also compute differences between the 298 K and 900 K temperatures 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 add 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 introduce other statistical transformations of the Boltzmann weighted energy distribution, like its standard deviation. To describe

the multi-modality of the energy distribution, we also introduce the Boltzmann weighted skewness and kurtosis; we can then deduce the Sarle's bimodality coefficient of Boltzmann weighted interaction energies. 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 model uses the mean and standard deviation of this distribution calculated for xenon and krypton.

## 2.7 Hyperparameter fine-tuning

The search for hyperparameter values aims at finding the best model to optimize the generalization error. The most common strategy is to perform cross-validations to evaluate different model configurations, known as hyperparameter search or optimization. In this case, the randomized search algorithm with 5-fold cross-validation is used to find the best parameters within a predefined reasonable range. The Supporting Information provides the range of hyperparameters explored by the algorithm. After this search, we identify a set of optimal hyperparameters, which gives an average RMSE of $0.37\,\mathrm{kJ\,mol^{-1}}$, which defines our final model. A convergence plot of the model performed using 5-fold cross-validations is given in Figure S6. The hyperparameter search is carried out on the training set to avoid any data leakage in the final model and ensure an accurate evaluation of the generalization error of the model.

Given this configuration, we test the model on the test set and use interpretation tools to understand better the structure–property relationships in play.

## 2.8 Interpretation of the final model

We then train the final model 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 apprehend the thermodynamic origins

of the performance. In this work, we use the Shapley values,[42] a game theory concept developed by Shapley in 1953, to measure the contribution of each descriptor in the predicted value. We can evaluate, locally on a nanoporous material, the contribution of each descriptor to the prediction using this tool. 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 a faster algorithm adapted to tree-based ML models like gradient boosting, TreeSHAP, which allows the calculation on large databases and integrates useful global interpretation modules like feature importance evaluation and dependence plot.

# 3 Results & Discussions

This study presents a prediction of the exchange Gibbs free energy at 1 bar and 298 K, which represents an energetic interpretation of the ambient-pressure selectivity (Equation 1), using geometrical, chemical and energy descriptors presented in Tables S1 and S2. The most correlated descriptor is the exchange Gibbs free energy at infinite dilution and 298 K. We will begin by studying the correlation between these two quantities since the exchange energy calculated by the GrAED algorithm already gives a very fast first evaluation of the selectivity. As shown in a previous study,[18] the difference of values between the low-pressure and ambient-pressure case is mainly a selectivity drop effect due to the near-saturation loading of adsorbates in the nanoporous material. To improve the accuracy of the evaluation, we train a model that integrates features that could help detect and quantify the selectivity drop that affects some highly selective materials. The ML model uses computationally cheaper descriptors to predict the computationally expensive ambient-pressure selectivity. Finally, we interpret the model to see how each feature contributed to the improved prediction compared to the simple infinite-dilution baseline.

## 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 1), which corresponds to a logarithmic transform of the selectivity values, Figure 1 confirms and highlights this correlation. 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. On top of this, in order to build a good prediction model we need to add 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.

To incorporate information on the pore size diversity of the materials, we carry out statistical measurements on the PSD. By analyzing them, we detect 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 brings 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.

The statistics on the distribution of interaction energies for xenon and krypton calculated by our grid algorithm can help quantify the change of selectivity. 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

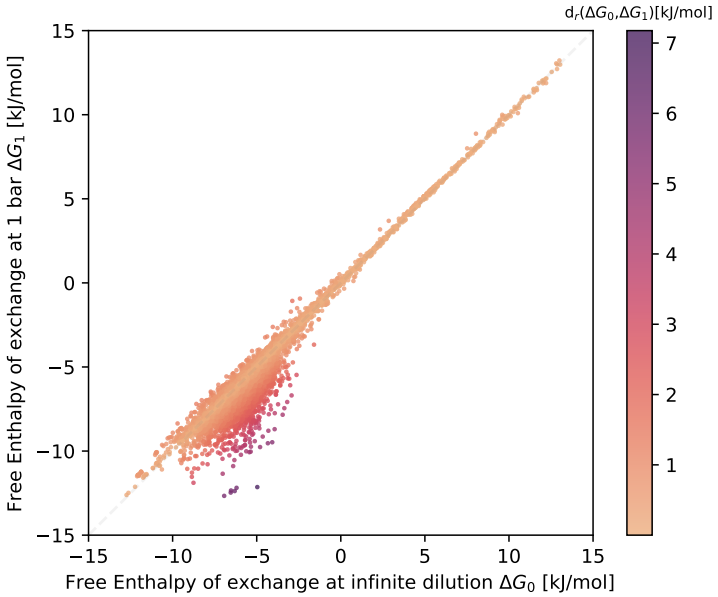



Figure 1: Comparison between the Gibbs free energy of exchange at low pressure $\Delta G_0$ (calculated by the GrAED algorithm) and ambient pressure $\Delta G_1$ (calculated by GCMC) labeled by the relative distance between them. This plot is equivalent to a logarithmic plot of the selectivity values at these two pressure conditions. The RMSE between these quantities is equal to 0.81 kJ mol$^{-1}$ and the MAE 0.49 kJ mol$^{-1}$.

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. We also apply the same approach to the Boltzmann weighted distributions to generate temperature specific descriptors for the energy distributions.

By using different temperatures, we note that the infinite dilution adsorption enthalpies at higher temperatures can be better correlated to the adsorption enthalpy at ambient pressure. The minimum error is found for the adsorption enthalpy at 900 K, which gives an RMSE of 1.76 kJ mol$^{-1}$ instead of 2.87 kJ mol$^{-1}$ for the 298 K case. This new type of descriptor is very interesting since it performs better around the high selectivity region, where the standard Boltzmann average at 298 K loses its accuracy (see Figure S1). As shown on 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 selectivity values at higher pressures.

By combining the above-mentioned features with more standard geometrical descriptors, we train 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 learns the joint effects of all the newly introduced descriptors to detect and evaluate the 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). Computing all the necessary features for the prediction would take less than a minute per structure, significantly faster than the 40 minutes required for a GCMC calculation. The ML-based approach clearly demonstrates a speed advantage over standard molecular simulations. However, it needs to maintain a high level of accuracy on an unseen set of structures to be a good substitute to GCMC.

We perform 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. And a set of hyperparameters 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 the SI. By using this parameterization, our XGBoost model has an RMSE of $0.37\,\mathrm{kJ\,mol^{-1}}$ and an MAE of $0.22\,\mathrm{kJ\,mol^{-1}}$ on the exchange Gibbs free energies of the test set containing 1 660 structures, which corresponds to a good correlation as shown in Figure 2.[1] 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 $\log_{10}$ of the selectivity, which means that the order of magnitude of the selectivity is known with a very good accuracy.

[1] With a split that groups similar structures in the same set, the performance obtained are very similar with an RMSE of $0.38\,\mathrm{kJ\,mol^{-1}}$ and an MAE of $0.23\,\mathrm{kJ\,mol^{-1}}$.

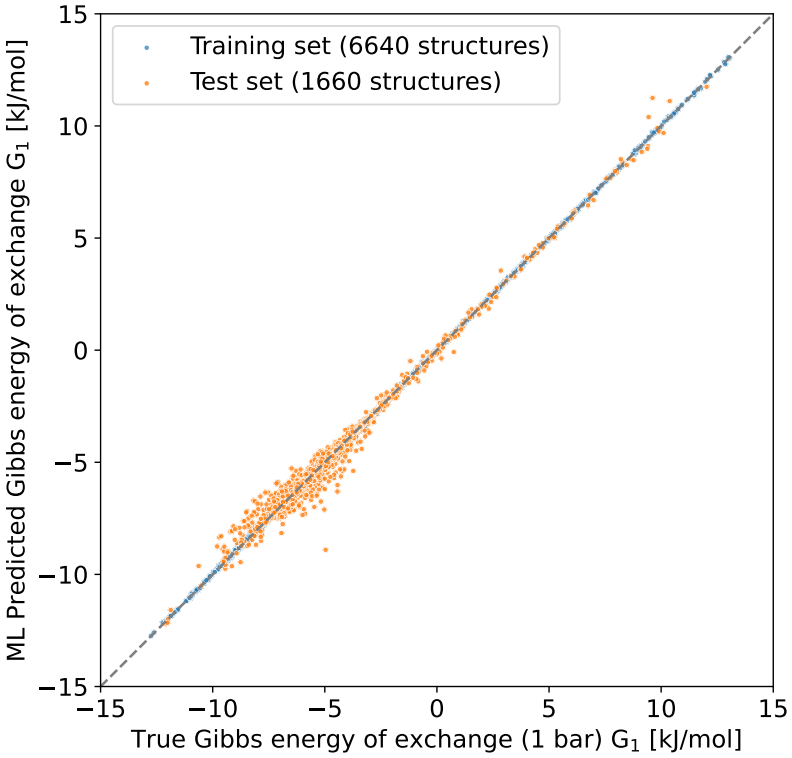



Figure 2: Scatter plot of the exchange free energy predicted by the model. There is a good agreement between the predicted and true values. On the test set, there is an RMSE of $0.37\,\mathrm{kJ\,mol^{-1}}$ and an MAE of $0.21\,\mathrm{kJ\,mol^{-1}}$. This plot is equivalent to the scatter plot between the logarithm of the ambient-pressure selectivity values (see Figure S5). 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 $\log_{10}$.

To prove that this good performance is not fortuitous, we use a 5-fold cross-validation procedure on the whole dataset and found an average RMSE of $0.37\,\mathrm{kJ\,mol^{-1}}$ with a standard deviation of $0.01\,\mathrm{kJ\,mol^{-1}}$, which is consistent with the performance given by the train/test split performed.

To see if it would be possible to train a better model with more training data, we train different models with different fractions of the training set as shown on Figure 3. The RMSE unsurprisingly decreases as we increase the amount of data, but it seems to start stabilizing for a fraction of 95% of the training set. This means that the model has a sufficient amount of training data to achieve what seems to be the minimum error on this test set, although it

may still be improved if we had a larger data set.

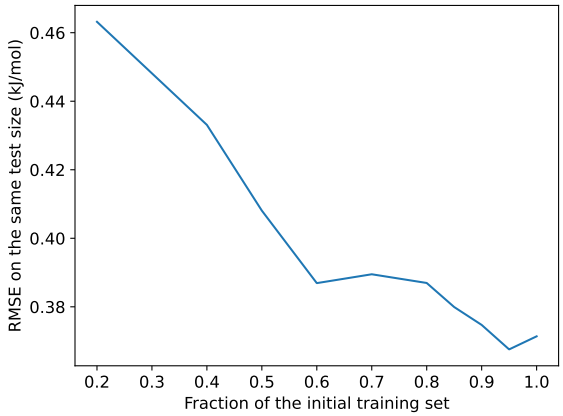



Figure 3: Root mean squared errors on the same test set (20% of all data) as a function of the fraction of the training set used to train smaller models. The error decreases as the amount of data increases.

This method can later be used in a screening procedure that relies on inexpensive descriptors to filter out obviously undesirable structures, retaining only the promising structures for the final ML model evaluation. To achieve this, as previously explained in the methods, only 3D MOF structures with an LCD above 4 Å are retained, as they possess an affinity for xenon, which is a necessary condition for a good Xe/Kr selectivity. Given the excellent predictive performance of the model regarding the ambient-pressure selectivity in structures with good xenon affinity, the proposed screening procedure, illustrated Figure 4, 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.

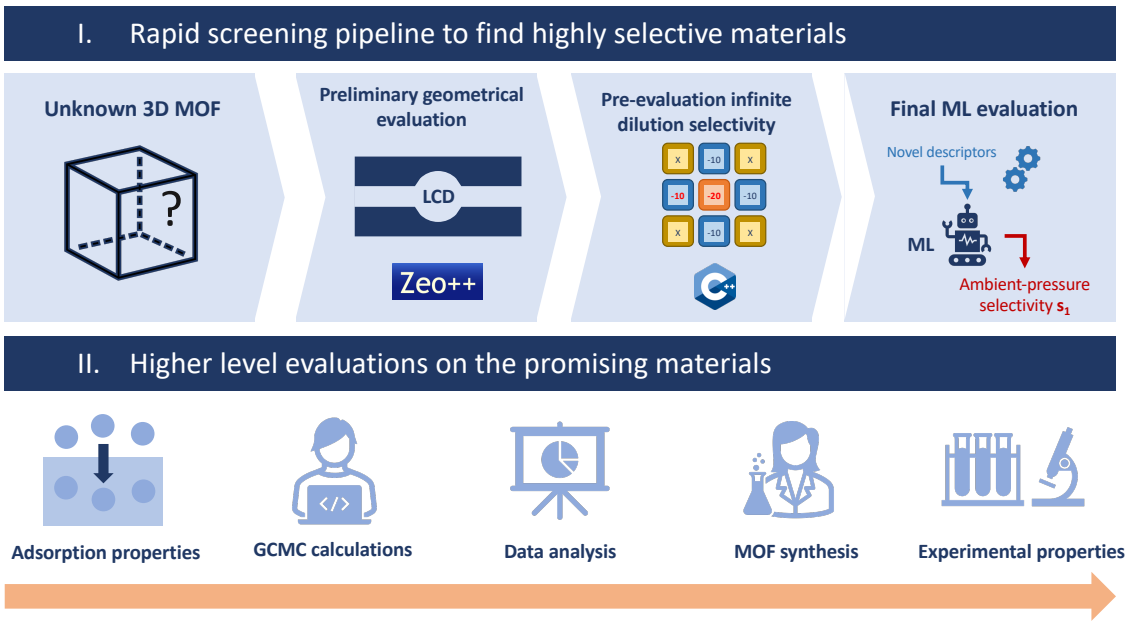



Figure 4: An illustration of the screening procedure that could be used to find highly selective materials.

## 3.3 Opening the black box

To understand the intuition behind this selectivity drop, we use 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. In game theory, the Shapley value is used to equally distribute a bounty according to the contribution of each player in a collaborative game. In machine learning, these values are used to break down the predicted values into a set of contributions for every feature (the sum of the contributions is equal to the predicted value). This interpretation model untangles the interdependence between the descriptors to extract an individual contribution.[2]

To go beyond the local interpretation, we can rapidly compute the approximate Shapley values for the whole dataset using faster algorithms, [44] and then use all them to make a global interpretation of the model. The global interpretation is based on multiple Shapley values

[2]The Shapley values does not depend on the units of the input data.

that can be aggregated using an averaging or by simply plotting them and look at their dependence to the feature value. If we plot the Shapley values as a function of the feature values for each structure of the dataset, we can see the contribution value depending on the feature value. This plot is called a SHAP dependence plot, which has a similar role as the partial dependence plot usually used for this purpose. Using the dependence plot, we can then infer, with a certain level of uncertainty, the level of contribution to the final predicted value of a feature, which highlights model-related 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). This mean value corresponds to the average magnitude of the contribution to the predicted value, which is a measure of the influence of the feature on the model output.

### 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 describing the materials without selectivity drops as shown in Figure 1; 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.

The partial dependence module offered by the SHAP library provide a comprehensive

interpretation of the model. Although, we can use other methods, such as partial dependence plots, to compute dependence plots (*e.g.* partial dependence plots),[43], it is preferable to maintain a good level of consistency between global and local interpretations by utilizing the same underlying theory. The SHAP dependence plots for all descriptors in Figures S9 and S10 exhibit distinct forms, directions, and shapes, which bodes well for the interpretability of the model. Valuable information regarding how the ML model predicts ambient-pressure selectivity is gleaned from the profile of these dependence plots.

The most important descriptor is the exchange free energy "G_0" associated with low-pressure selectivity. Its contribution displays a very strong positive linear correlation (Figure 5). This descriptor establishes a baseline, on top of which other contributions either decrease the free energy (more selective) or increase it (less selective). The model can be interpreted as a combination of a baseline and smaller adjustments estimating the deviation magnitude 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), further contribute to the baseline by providing information on low-pressure selectivity. Moreover, they also offer insights into the deviations necessary to differentiate structures experiencing a drop in selectivity from those maintaining their selectivity. As we can see in Supporting Information (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 5), 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 $-4\,\mathrm{kJ\,mol^{-1}}$ contribute negatively to the prediction 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. In the following, we will present further examples highlighting the determination of selectivity contributions using the remaining descriptors.

The optimal values for the associated descriptors is characterized by the U-shape of some SHAP dependence plots. For instance, we observe an optimal value of around 5.1 for "D_i_vdw_uff298" (Figure 5), while the optimal average pore size is approximately 5.6. These optimal values align with the physical requirement of having xenon-sized pores to enhance xenon's attraction, as identified in various literature papers. However, it should be noted that these values are slightly higher than those mentioned in the literature due to differences in atom radius definitions.[33] Moreover, values of "delta_G0_298_900" between 4 and 6 $\mathrm{kJ\,mol^{-1}}$ (Figure 5) have a higher likelihood of contributing negatively, indicating lower ambient-pressure selectivity. These sweet spots provide valuable insights for distinguishing truly selective materials from others. Some SHAP dependence plots have a rather linear domain for the most selective structures (in blue) — a good linear contribution is observed for the difference of pore volumes between Xe and Kr sized probes "delta_VF_18_20" (Figure 5). This implies that a lower void fraction difference corresponds to a more selective structure. The same trend is observed for the standard deviations of the PSD, denoted as "pore_dist_std", and the Boltzmann weighted krypton interaction energies distribution, referred to as "enthalpy_std_krypton". Optimal values for these descriptors tend to be zero. As the value approaches zero, the contribution becomes more negative, indicating a more selective 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 5); even though there is no clear correlations, we can at least find a threshold value (about 0.23) below which there is higher probability of having a high ambient-pressure selectivity. There is the same type of domain splits 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.

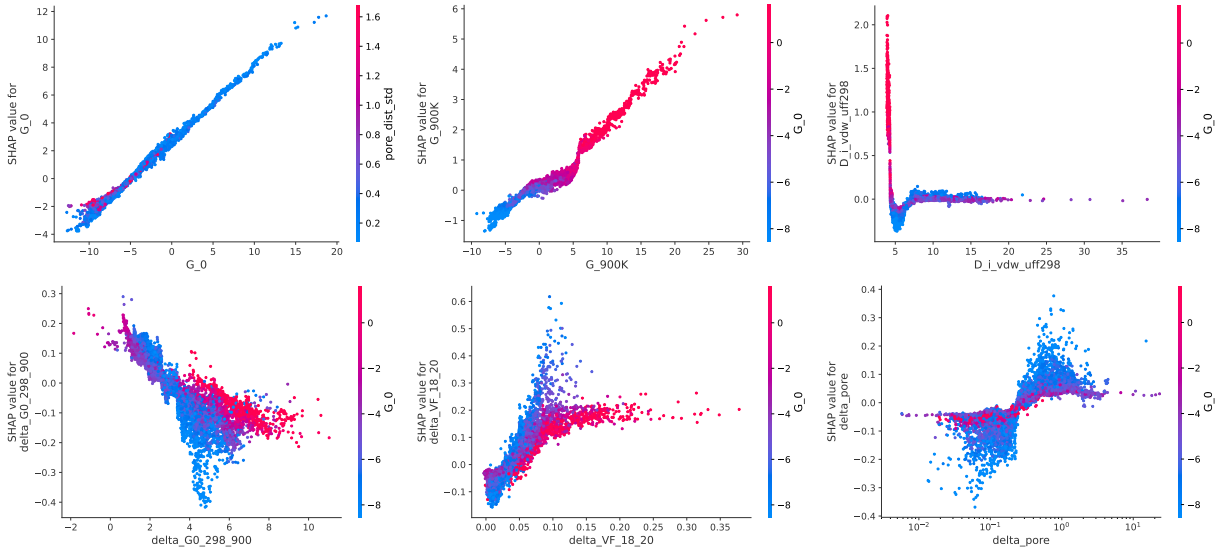


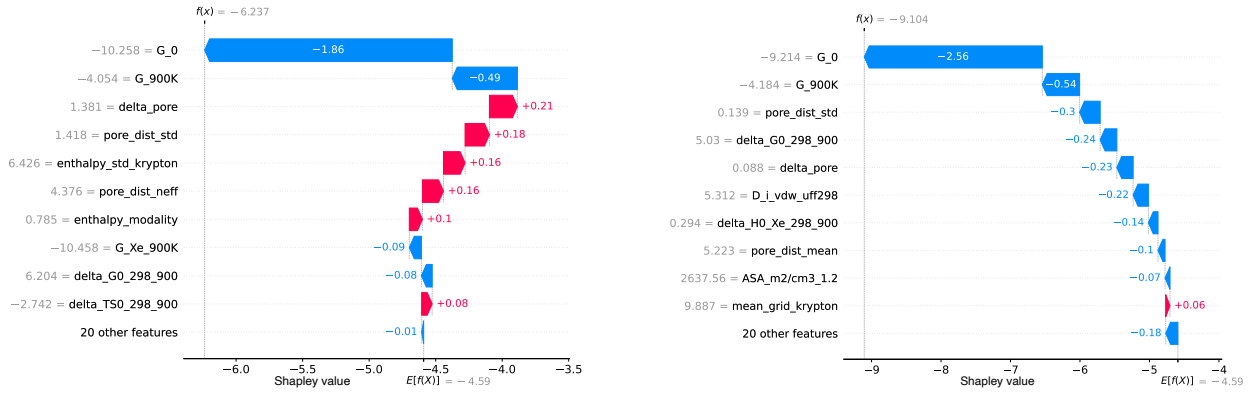

Figure 5: Some SHAP dependence plots that are analyzed in the main text. The 18 top descriptors' SDPs can be found in the SI. A SHAP dependence plot corresponds to the Shapley values as a function of the feature values for every structures. The feature values are value names given in Tables S1 and S2. These SHAP plots show the contribution of the features to the prediction given by the ML model. Each Shapley value depends not only on the value of the feature itself but also on the other features, for this reason, the plots are labeled by a relevant second feature.

### 3.3.2 Local interpretability

To apply the previous analysis in practice, archetypal structures and their selectivity predictions based on descriptor values will be examined. Two MOF structures from the test set, with CSD codes `VIWMIZ` and `BIMDIL`, are chosen. Both structures are selective at low pressure, but the first one decreases in selectivity while the second one maintains it at ambient pressure. The focus will be on understanding how the model distinguishes between these two completely distinct behaviors.

VIWMIZ belongs to the category of highly selective structures that undergo a selectivity drop at ambient pressure. When converting the free energy values to selectivity values, VIWMIZ has a selectivity of 62.8 at infinite dilution and 14.5 at ambient pressure. The ML model successfully predicts a close value of 12.0 for the ambient-pressure selectivity based on the given descriptor values. Specifically, the descriptor "G_0" has a highly negative value, which explains its relatively high negative contribution of −1.81. However, the contribution of "G_900K" is relatively low at −0.57 compared to other materials (Figure 5), as a value of −4.05 is not the most negative among all structures. Conversely, the remaining descriptors have positive contributions, which lead to the selectivity drop. For instance, the difference in 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. This value aligns with the value ranges observed in the associated dependence plot. We performed similar analyses on the positive contributions of other descriptors shown in Figure 6 by referring to the dependence plots: "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" falls below 3 $\mathrm{kJ\,mol^{-1}}$ and "enthalpy_modality" is around 0.75 where positive contributions are more recurrent. However, the "delta_G0_298_900" value is somewhat close to its optimal value, resulting in a negative contribution in this specific prediction. The remaining features have negligible contributions. Analyzing the contributions of each descriptor to the prediction given by the model of this work helps understanding the underlying features of the VIWMIZ structure that explain the selectivity drop at higher pressure. Descriptors such as the shape of the xenon and krypton energy distributions ("enthalpy_std_krypton" and "enthalpy_modality") and the PSD ("pore_dist_std" and "pore_dist_neff" ) as well as the void fraction difference "delta_pore" play key roles in the lower selectivity at ambient pressure compared to the ideal infinite dilution case. Intuitively, an effective number of pores exceeding 2 suggests the presence of different pore sizes, which aligns with the presence of less attractive pores for xenon, ultimately leading to decreased selectivity. This observation is consistent with a high

standard deviation of the PSD or the Boltzmann-weighted krypton interaction energy distribution. Furthermore, a significant difference between the average pore size and the LCD indicates a disparity in pore sizes, resulting in larger pores that become increasingly loaded as pressure rises. However, interpreting the entropic term is more complex and presents unexplored ways of addressing the selectivity drop at higher pressure, as revealed in the previous study.[18]




(a) VIWMIZ: true $\Delta_{\text{exc}}G_1 = -6.63$ (b) BIMDIL: true $\Delta_{\text{exc}}G_1 = -9.20$

Figure 6: Main Shapley contributions of the ML features on the selectivity prediction of two archetypal examples. The feature labels used are detailed in Tables S1 and S2. The ML predicted values is shown using $f(x) =$ and $E[f(x)]$ is the average predicted values used by SHAP to define the initial value so that $f(x) = E[f(x)] + \sum_{\text{feature}} \text{contribution(feature)}$.

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 predict accurately the stability of the selectivity by assigning a value of 40.0. Consequently, the first contribution of "G_0" is one of the most negative contributions, establishing a baseline of $-2.4$ for subsequent contributions. Although the contributions of "G_900K" and "G_900K", they continue to decrease the predicted selectivity value. The joint contributions of other descriptors will discriminate between the two structures and determine why this particular structure will maintain its selectivity. In contrast to the previously analyzed structure, this structure has a "delta_pore" value below 0.3 Å, explaining its negative Shapley value in the prediction of this study. The contribution of "delta_G0_298_900", which had only a

slightly negative impact on the other structure, now plays a significant role as it falls within the range of 4 to 6 $\mathrm{kJ\,mol^{-1}}$ (Figure 6). Additionally, it is observed that "pore_dist_std" is below the threshold, in contrast to the previous structure where it was above the threshold. Furthermore, the other contributions align with the rules suggested by the SHAP dependence plots, and no apparent anomalies are detected. The combined effects of all the descriptors result in a lower free energy value, leading to the conservation of selectivity at higher pressure. The set of descriptor values for this structure significantly differs from the previous one, with many values contributing to opposite domains. This disparity allows the model to differentiate between highly selective structures and identify those that will maintain their selectivity at higher pressure.

These two examples provide a deeper understanding of how the model distinguishes structures that lose selectivity at higher pressure from those that do not. Most dependence plots exhibit a strong association between descriptors and their effects, with outliers being rare enough to comprehend the internal logic of the model. As previously discussed, the first three descriptors establish a baseline for the observed selectivity drop with limited information. Subsequently, the contributions of other descriptors can be positive, negligible, or negative depending on the domain where the values of the descriptor lie. For instance, the average pore size and largest cavity diameter need to be within specific ranges to maximize the likelihood of maintaining selectivity at higher pressure, aligning with previous studies emphasizing the importance of pore sizes similar to xenon for Xe/Kr separation. The difference in entropy between ambient temperature and 900 K is a surprising descriptor that separates selective structures based on whether its value falls within a specified range. Similarly, the difference in void fraction occupied by xenon and krypton is intriguing as it impacts selectivity differently depending on whether the structure is highly selective or not, with the contribution being more or less proportional to its value. Various methods of measuring the disparity of the PSD and interaction energy distribution play a key role in identifying highly selective structures (indicated in blue on the dependence plot in Figure 5) that either maintain or decrease in

selectivity. These methods include calculating the difference between the average pore size and the LCD, as well as the standard deviation of the PSD or Boltzmann-weighted energy distribution, which exhibits distinct behaviors based on the domain in which the value lies. The SHAP dependence plots provide valuable insights into the mechanisms underlying the ML model presented in this article and, more broadly, shed light on the understanding of Xe/Kr separation origins.

# 4 Conclusions and perspectives

To gain a deeper understanding of separation processes within nanoporous materials, a machine learning prediction of Xe/Kr ambient-pressure selectivity was performed, aiming for faster results compared to standard GCMC calculations. The CoRE MOF 2019 database was utilized for MOF structures, enabling the evaluation of xenon/krypton selectivity in less than a minute, whereas an equivalent GCMC calculation typically requires approximately 40 min. Unlike the majority of selectivity predictions in the literature, the decision was made to predict selectivity on a logarithmic scale that focuses on the order of magnitude rather than the exact value of selectivity for highly selective materials. Moreover, converting to an exchange Gibbs free energy allowed for a more thermodynamic approach based on enthalpy, entropy, and free energy values. The challenge consisted of predicting the free energy equivalent of ambient-pressure selectivity using low-pressure selectivity alongside key energy, geometrical and chemical descriptors. The resulting fully optimized ML model exhibited high performance, yielding an RMSE of $0.37\,\mathrm{kJ\,mol^{-1}}$, which corresponds to an RMSE of 0.06 on the base-10 log of selectivity. This represented an improvement compared to the $0.81\,\mathrm{kJ\,mol^{-1}}$ RMSE of a baseline model that is solely based on the low-pressure selectivity calculated by the GrAED algorithm. The energy descriptors along with statistical quantities greatly contributed to the performance of the final model.

One specific objective was to validate the previously highlighted underlying reasons for

the observed selectivity drop at high pressure in certain highly selective materials at low pressure. Previous studies found that high diversity of pore sizes and channel sizes that favor adsorbate reorganizations could be at the origin of this phenomenon. Through the application of interpretability tools, quantitative factors explaining the conservation or decrease in selectivity for highly selective materials are identified. Depending on energy averaging at 900 K, statistical characterizations of energy or pore size distributions, and differences in occupiable volumes, a structure could exhibit either a selectivity similar to the infinite dilution case or a substantially lower selectivity at higher pressure. The XGBoost model employed in this study utilizes a complex ensemble of decision trees to capture the quantitative rules that can be extracted from the model and used to establish heuristics supporting the intuition about Xe/Kr selectivity in MOF structures.

The final ML model could be used in a well-designed workflow to find the best performing materials. For instance, structures with pores unable to accommodate xenon could be filtered out, followed by the application of a low-pressure selectivity calculation to eliminate selectivity values below a specified threshold. Finally, the structures that would encounter a drop in selectivity could be removed using the model. As a proof of concept, the methodology was tested on Xe/Kr separation, which represented one of the simplest adsorption systems (monoatomic species and the absence of electrostatic interactions). A similar approach could be generalized to other separation applications by calculating the infinite dilution energies with a more conventional method (*e.g.* Widom's insertion), while adjusting the definitions of descriptors to suit the adsorbates of interest.

The ambition of this study was to introduce new descriptor ideas that contribute to the development of increasingly efficient screening methodologies for identifying optimal materials for specific applications. However, similar to other studies in this field, the simulations in this study relied on a set of strong assumptions, wherein rigid frameworks and non-polarized classical forcefields were employed. Previous literature suggested that the most selective materials for Xe/Kr separation were designed and synthesized based on the effect of open-metal

sites, leveraging the difference in polarizability between the two molecules to achieve efficient separation.[5,6] Moreover, the flexibility of structures could be achieved by employing flexible forcefields with appropriate simulation methodologies[45] or by conducting multiple rigid simulations using snapshots from NPT simulations[46]. The simulations could be enhanced at the cost of CPU time by coupling them with a reduction in simulation time, such as the one presented in this article. The pursuit of ever-faster evaluation tools enabled the exploration of more complex properties and the discovery of structures with increasingly relevant characteristics.

## Conflicts of interest

There are no conflicts to declare.

## Acknowledgement

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

## Funding

This work was financially supported by Orano.

## Supporting Information Available

Additional discussion and results on the exploration of relevant descriptors, detailed list of features and their selection, details on ML model training. Raw data are available online at `https://github.com/fxcoudert/citable-data`, and the Grid Adsorption Energy Sampling code is available at `https://github.com/coudertlab/GrAED` and `https://github.`

com/eren125/xe_kr_selectivity_xgb

# References


(1) Kerry, F. G. *Industrial gas handbook: gas separation and purification*; CRC Press, 2007; pp 129–168.

(2) National Academies of Sciences, Engineering, and Medicine *A Research Agenda for Transforming Separation Science*; The National Academies Press: Washington, D.C., 2019; pp 7–14.

(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 Cent. Sci.* **2023**, *9*, 563–581.

(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**, *14*, 1797–1807.

(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**, *19*, 4568–4583.

(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. In *Contributions to the Theory of Games (AM-28), Volume II*; Kuhn, H. W., Tucker, A. W., Eds.; Princeton University Press: Princeton, 1953; pp 307–318.

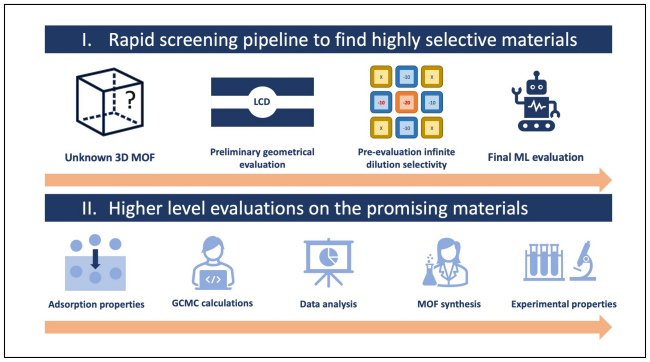
(43) Molnar, C. *Interpretable machine learning*; Self-published online at `https://christophm.github.io/interpretable-ml-book/`, 2023; Chapter 9.5–9.6.

(44) Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. 2017; `https://arxiv.org/abs/1705.07874`, Version v2 submitted on 2017-11-25. Accessed on 2023-07-05.

(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
LCD
Unknown 3D MOF
Preliminary geometrical evaluation
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