
ParisTech

REMERCIEMENTS

En premier lieu, je voudrais adresser ici mes plus vifs remerciements

TABLE OF CONTENTS

GENERAL INTRODUCTION

Nanoporous materials are material

[Just a copy paste from last article]

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 volatile radioactive wastes 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.[kerry2007industrial]

A wider use of nanoporous materials could reduce the energy consumption of current separation processes since adsorption is way less energy intensive than liquefaction.[national2019research] For instance, some prototypes involving beds of nanoporous materials have been developed for xenon/krypton separation to avoid employing cryogenic distillation.[Banerjee2018] 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.[Chen_2014, Li_2019, Pei_2022] 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.[Lyu_2020, Jablonka_2022] 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[Groom_2016] and 500,000 computationally constructed.[Wilmer_2012, Boyd_2016, Colon_2017] 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.[Ren_2022] 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.[Simon_2015] 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.[Rycroft_2009] 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.[Ren_2023] 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,[Shi_2023] 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.[Ren_2021]

[Xe/Kr applications in the industry]

This thesis presents my work on the

1

HIGH-THROUGHPUT COMPUTATIONAL SCREENING OF NANOPOROUS MATERIALS

1.1 Nanoporous Materials 5
1.1.1 Definition of the concept 5
1.1.2 Computational databases 8
1.1.3 Exploring the chemical and structural space 9
1.2 Review of Screening Methodologies 10
1.2.1 Non-adsorption properties 10
1.2.2 Transport adsorption properties 15
1.2.3 Thermodynamic adsorption properties 19
1.2.4 Gas separation 21
1.3 Separation of Xenon from Krypton 25
1.3.1 Industrial applications 25
1.3.2 Promising materials for the separation 26
1.3.3 From the computer to the test tube 27
1.3.4 The future of screening 29

1.1 Nanoporous Materials

1.1.1 Definition of the concept

Adsorption isotherms and geometrical descriptors

Nanoporous materials are defined as materials with a nanoscale structure constituted by pores and cavities, which some are connected by a network of channels. These pores can be empty or filled with a variety of substances called adsorbates. By adhering molecules from a liquid or gas phase into the internal surfaces of the material, we can use it in diverse applications such as gas separation and purification,[Li_2009, Lagorsse_2007] energy storage and conversion,[Morris_2008, Qiu_2020] heterogeneous catalysis[Bell_2003, Singh_2019, Pascanu_2019] drug delivery,[Della_Rocca_2011, Bernini_2014] or sensing.[Breslin_1976] By designing the chemical na-

ture, size, shape and distribution of the pores, we can tailor the physicochemical properties to the targeted application.[Yan_2020]

The process of adhering particles or molecules on surfaces is called adsorption. Adsorption occurs due to attractive forces between adsorbates and the adsorbent surface, such as van der Waals forces, hydrogen bonding, and electrostatic interactions. The adsorption performance depends on the chemical nature of the interface, its exposed surface area and the shape of the pores. We usually characterize adsorption properties of an adsorbate compound by measuring the numbers of adsorbed molecules as a function of its pressure at a given temperature, which is called the adsorption isotherm. These isotherms can possibly be used, among other techniques, to specify the pore size distribution, accessible surface area and pore volume.[Rouquerol_1994] By using fitting models, we can also use adsorption isotherms to characterize the maximum adsorption uptake among other adsorption descriptors.[Wang_2020] Using a set of experimental isotherms at close but different temperatures, we can also retrieve information on the isosteric heat of adsorption q_{st} (the negative differential of the excess enthalpy of adsorption with respect to the excess adsorption).[Nicholson2000] This heat of adsorption (related to the enthalpy of adsorption) can also be directly obtained using calorimetry.[Dunne_1996] Furthermore measurements at infinite dilution can also lead to a linear relation between the adsorbed quantity and the pressure defined by the Henry's law; another key adsorption descriptor, the Henry adsorption constant, is defined as the slope of this linear regime.[Finsy2007] All of these thermodynamic quantities are most valuable in comparing experimental data to computational modeling to compare and characterize the materials suitable for a given gas adsorption process.

Most of the materials studied in this thesis will have pores with a size around the nanometer called "nanopores". The International Union of Pure Applied Chemistry (IUPAC) classifies these pores into three categories according to their size: micropores ($\leq$2 nm), mesopores (2 nm–50 nm) and macropores (>50 nm).[Sing_1985] Here, we will use a single terminology (nanopore) to designate all pores of under a few nanometers. A good characterization of the nanopores of these materials is key to fine-tuning the adsorption properties.[Yan_2020] The pore size distribution (PSD) can be computationally determined if we have resolved the structure of the nanoporous material (using X-ray diffraction on crystallized porous solids). This is the most accurate determination method of the PSD, but it relies on considering that the structure is perfectly rigid and crystalline so that only one structural data can characterize it. Other experimental methods rely on assumptions, model systems (e.g., cylindrical) or adsorption characteristics. For instance, stereological analyses based on plane sections cut through a porous material can evaluate the PSD.[Haynes_1973] The Horvath-Kawazoe (HK) method is a semi-empirical analytic model of adsorption isotherm that can extract PSD. Small angle X-ray and neutron scattering methods are non-destructive methods of pore characterization.[Radlinski_2004] In this thesis, we will rely on computationally analyzing X-ray diffusion data to deduce pore sizes and other geometrical characteristics.

The pore volume consists in the measure of the volume of "closed" and "open" pores of nanoporous materials. Depending on the way of measuring it, different quantities are probed. Some pores could not be accessed by some adsorbate; depending on the probe size the volume calculated will not be the same. Methods that do not rely on adsorption like scattering or stereology will, however, measure the total pore volume. The porosity or void fraction would be defined as the ratio between the pore volume and the apparent framework volume. Depending

on the method, we can therefore retrieve either the total porosity, the porosity opened or closed to a given probe adsorbate.

The cavities of the nanoporous material lay out an incredibly large adsorbable surface area, which is extremely useful in increasing the number of molecules in a given volume or mass of material, several thousands of square meters can be found in a gram of some nanoporous materials.[Farha_2012] The higher the surface area, the more molecules can be adsorbed for storage, separation or reaction purposes; it is therefore crucial to measure this surface area using experimental and computational methods. The most extensively used method to experimentally measure surface areas from adsorption isotherms is based on the Brunauer–Emmett–Teller (BET) theory.[Detsi_2011] Most BET areas are calculated on N_2 isotherm at its boiling temperature (77 K); although different probe adsorbates can be considered, they are not standard.[Tian_2017] However, the definition of the surface area depends highly on the condition of measurement but also on the fitting methodology; a dozen isotherms were given to 61 labs for BET area calculation, and the statistical experiment yielded to a high level of disparities in the calculation.[Osterrieth_2022]

Beyond the experimental techniques, some software like Zeo++ or PoreBlazer focus on computing pore size distributions, surface areas and void fractions using well-defined structure files.[Zeo++, PoreBlazer] The definition of these values also depends on the probe size chosen to model a given adsorbate, the size of the framework atoms and the quality of the input structure. The computational values do not rely on adsorption models or on isotherm data as in the BET area, they are now relying on more comprehensive structural data. They, however, very much rely on a well-designed definition of the volume, the surface and the pore size we want to evaluate. Moreover, these values also highly depend on the radii of the framework atoms and the adsorbate we consider..[Hung_2021] In this thesis, we will rely on these computational methods to define these geometrical descriptors of nanoporous materials.

Classes of nanoporous materials

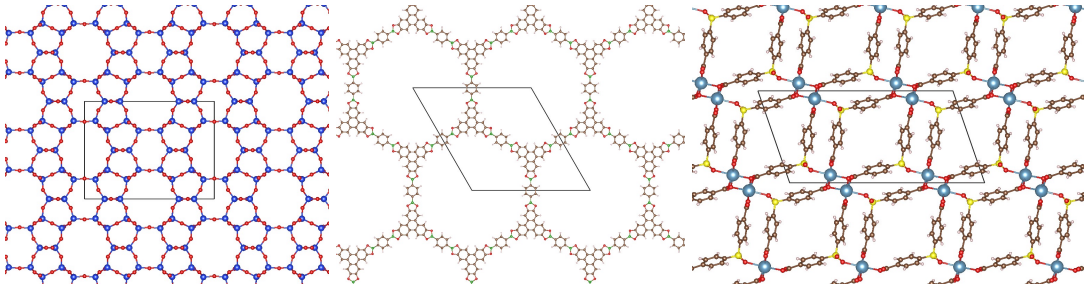


Figure 1.1: Illustration of a zeolite FER,[FER] a COF[Cote_2005] and a MOF.[KAXQIL] Color code: brown for C, white for H, red for O, blue for Si, cyan for Ca, yellow for S and green for B.

Nanoporous materials can have different degrees of crystallinity from perfectly crystalline to completely amorphous. Most of the computational work is focused on crystalline structures, since the atoms are well described within a periodic framework, which enables faster simulations. The presence of defects is also usually neglected, which could explain some discrepancies between simulations and experiments. And amorphous materials are described by thousands of atomic positions in order to grasp their intrinsic non-periodicity.[Thyagarajan_2020] Activated carbons, a famous class of amorphous material, are extensively used in the industry for gas purification, but cannot be rationally studied to characterize their adsorption properties. One

can distinguish roughly three main classes of crystalline nanoporous materials: the inorganic materials like zeolites (aluminosilicates or aluminophosphates), the organic materials like the porous polymer networks (PPNs) or the covalent organic frameworks (COFs) and the metal–organic frameworks (MOFs).

Zeolites are naturally occurring nanoporous aluminosilicate materials that are commonly synthesized to be used in the industry as a commercial adsorbent and heterogeneous catalyst.[Ozin_1989, Ma_2000] It is considered as one of the most mature nanoporous material technology at our disposal. This class of material also leaves a wide room for innovation since different Al/Si ratios of a same zeolite type pan out a wide range of structures. Furthermore, zeolite materials inspired the synthesis of zeolitic frameworks harboring different atoms such as the aluminophosphates or the zeolitic imidazolate frameworks.[Wang_2012, Chen_2014_zeo]

Porous polymer networks (PPNs) are porous materials based on the already very developed polymer material technology.[Lu_2010, Wang_2020, Che_2020] However, one of the major drawback of this type of material is the formation of irreversible covalent bonds, which make the synthesis kinetically controlled leading to difficulties in crystallizing PPNs.[Feng_2012] In order to create crystalline porous materials, Cote et al. figured out a way of using boron-based organic compounds to form reversible bounds, which formed thermodynamically stable materials COF-1 and COF-5.[Cote_2005] This initiative was led by the group of Yaghi who was at the initiative of another very promising and well-known class of materials. A decade earlier, they pioneered a hydrothermal synthesis of a metal–organic framework presenting broad rectangular channels.[Yaghi_1995]

Metal–organic frameworks (MOFs) are a class of nanoporous materials formed by metallic centers connected with organic linkers to form a stable crystalline solid. Even if the first synthesis of such a material was done since the early 90s,[Abrahams_1991] and brought about a sparking interest in the scientific community a couple of decades later.[Kuppler_2009, Furukawa_2013] Because plenty of combinations of linkers and metals are imaginable, an infinite amount of MOFs could theoretically be designed. Their structure can be tuned to our needs to enhance their performance in the targeted application.[Ejsmont_2021] This diversity of nanoporous materials offer a wide range of potential candidates that could be evaluated for any targeted applications.

1.1.2 Computational databases

All the previously described materials have been either synthesized and resolved using X-ray crystallography or computationally constructed. By combining almost all possible nanoporous materials, almost a million structures have been considered for separation or storage applications.[Simon_2015, Si] This extended database can be broken down into the synthesized materials and hypothetical ones for all the above-mentioned classes of material.

The International Zeolite Association (IZA) gave a standardized set of 244 zeolites (in their idealized all-silica form) that can be used for screening purposes. To generate a dataset of structures, existing experimental databases like the Cambridge Structural Database can be exploited. However, the raw structures determined experimentally by X-ray cannot be used directly as is. To obtain a computation-ready dataset, Chung et al. used algorithmic cleaning procedures to build the publicly available Computation-Ready Experimental MOF (CoRE MOF) database.[Chung_2014, Chung_2019] CoRE MOF 2019 contains about 14,000 MOF

structures, which is the biggest experimental database. Similar approach applied to organic frameworks led to the construction of a set of 187 COFs with disorder-free and solvent-free structures.[Tong_2017, Ongari_2019]

These experiment-based databases can already be used in computational screenings to retrieve valuable information, but unknown structures that are yet to be discovered are not represented. To overcome the limits and biases of experimental synthesis, artificial ways of generating nanoporous material datasets can be used, which proved to be extremely efficient. The first *in silico* generated database of about 130,000 MOFs used a recursion-based assembly (or Tinkertoy-like) algorithm to combine 102 building blocks.[Wilmer_2012] Martin and Haranczyk then proposed a topology-specific structure assembly algorithm that leverages the topological information of the structures.[Martin_2014] Inspired by this algorithm, topology-based databases emerged a few years later with the set of 13,000 MOF structures generated using the Topologically Based Crystal Constructor (ToBaCCo) algorithm developed by Colon, Gómez-Gualdrón and Snurr.[Colon_2017] Later, Boyd and Woo proposed another topology-based algorithm using a graph theoretical approach and generated a 300,000-structure database (BW-DB) based on 46 different network topologies.[Boyd_2016] Similar approaches are used for other classes of materials, Deem and co-workers proposed a dataset of nearly 2.6 million hypothetical zeolite structures.[Earl_2006, Deem_2009, Pophale_2011] However, one could wonder if these hypothetical structures are synthesizable and can remain stable under operational conditions (e.g. thermal, mechanical, radioactive constraints). To discuss their synthetic likelihood, Anderson and Gómez-Gualdrón computed the free energies of 8,500 hypothetical structures and compared them to experimentally observed MOF structures.[Anderson_2020] Later, Nandy et al. performed a meta-analysis of thousands of articles associated to the CoRE MOF 2019 database to extract their experimental solvent-removal stability and thermal decomposition temperature.[Nandy_2021] These data are then leveraged in the training of multiple ML models to predict stability properties. These predictions can be very useful to gauge the relative stability of each material and to only consider stable structures. Other types of materials have been explored, Turcani et al. published 60,000 organic cage structures and used machine learning to predict their stability based on the shape persistence metric.[Turcani_2018]

The Materials Genome Initiative, 100 million dollar effort from the White House that aims to "discover, develop, and deploy new materials twice as fast", led to the creation of the "Materials Project", a centralized database containing all the above-mentioned structures.[kalil2011national, Matgenome, Jain] The fast development of this nanoporous materials genome motivated Boyd et al. to write a comprehensive review on all the initiatives on generating new data for computational analysis.[Boyd_2017]

Yet, the sole increase in size of the databases is not enough. One needs to add diversity to have more general knowledge on the maximum performance and the explanatory features of such performance. Moreover, the diversity of structures ensure the quality of the predicted best materials for a given application. To qualitatively or quantitatively assess the diversity of a database, inventive methodologies have been developed. For instance, Martin, Smit and Haranczyk proposed a Voronoi hologram representation as a way of measuring similarities between structures to generate geometrically diverse subsets of a database.[Martin_2011] Moosavi et al. made a comparative study of the diversity of three well-known databases CoRE MOF 2019,[Chung_2019] BW-DB[Boyd_2016] and ToBaCCo[Gomez_Gualdron_2016, Colon_2017] using ge-

ometrical and chemical descriptors to design a theoretical strategy for generating the most diverse set of materials.[Moosavi_2020] Another approach consists in searching for similarities instead of differences in the materials by studying topological patterns in the data.[Lee_2017] These investigations on the data structures give a solid ground to develop novel materials by objectively defining similarity, diversity and novelty. From the analysis gathered so far, one would need to radically change the approach by proposing materials with new chemistry, topology or mechanism (e.g. flexibility) in order to significantly improve the diversity of the current databases.

1.1.3 Exploring the chemical and structural space

With the development of ever-increasing nanoporous material databases, computational chemists proposed more and more inventive methods to evaluate or screen thousands of structures. Other challenges arose, such as the design of more efficient methods than the brute force screening or the analysis of big data. Two research groups in Northwestern University led by R. Snurr and J. Hupp began to address those questions, they used a "funnel-like" approach to efficiently screen about 130,000 hypothetical MOF structures.[Wilmer_2012] To do so, they performed a first screening involving fewer steps of simulation on the whole dataset, then they extracted a subset of top-performing structures to perform a second round with more simulation steps. This procedure is repeated until a few materials are selected by a final round of simulations with reasonable accuracy. Similar "funnel-like" procedures have then been used in other fields of applications as described in the Figure **??**. This type of screening saves precious computation time by balancing the complexity of the calculation with the amount of data to be screened. The most demanding simulations or experiments are only applied to the few most promising structures. This method can rather efficiently identify top candidates, but it can't draw quantitative structure-property relationships (QSPR), beside facing scalability issues above a critical dataset size.

To overcome these new challenges, people are looking increasingly towards transferable models trained by a machine learning (ML) algorithm on a diverse and size-limited subsample. Ideally, such a model is transferable to potentially millions of structures and can provide valuable QSPR. For instance, Fernandez et al.[Fernandez_2013] used multiple linear regression analysis, decision tree regression, and nonlinear support-vector machine models to extract QSPR and establish rules of designing well-performing MOFs for methane storage, while identifying promising structures. In this first work, they only used geometrical descriptors to describe methane storage,[Fernandez_2013] but realizing the importance of chemical descriptors, they proposed the atomic property weighted radial distribution function as a powerful descriptor to predict CO_2 uptakes.[Fernandez_2013_rdf] More importantly, they proved that ML can be used as a pre-screening tool to avoid running time-costly simulations by correctly identifying around 95 % of the top 1000 best performing materials. Recently, the same group used similar techniques to predict CO_2 working capacity as well as CO_2/H_2 selectivity in MOFs for pre-combustion carbon capture.[Dureckova_2019]

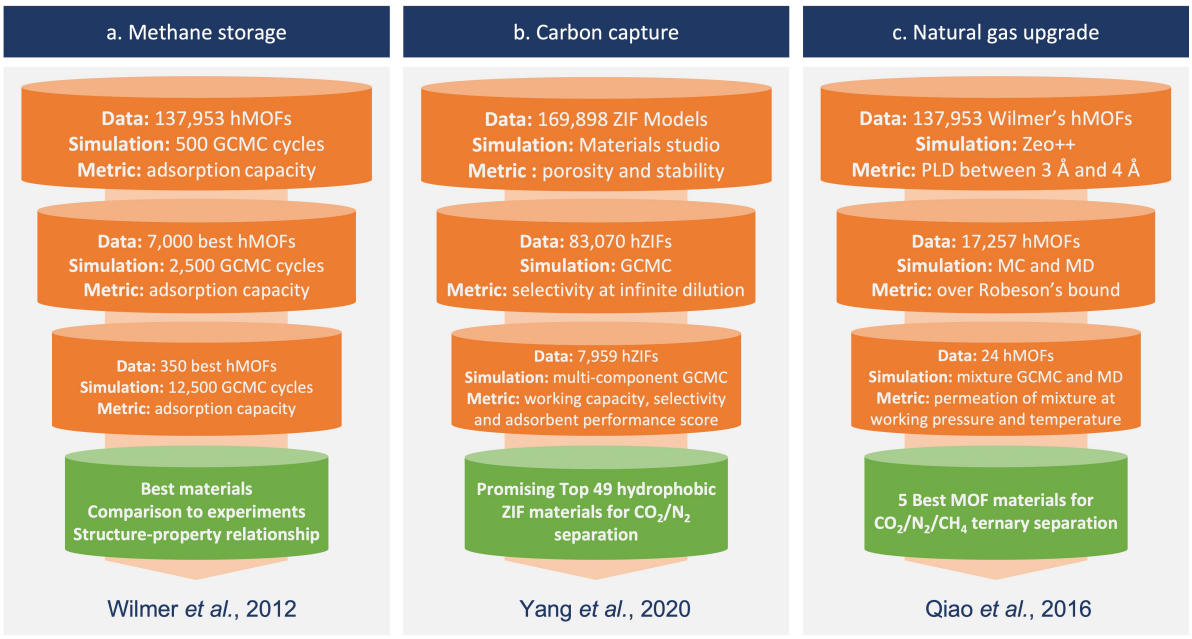


Figure 1.2: Simplified representation of typical funnel-type screening procedures, exemplified on three different applications from the published literature. (a) Wilmer et al.[Wilmer_2012] *used a series of bi-component Grand Canonical Monte Carlo (GCMC) calculations at different levels of complexity to screen a large dataset of hypothetical MOFs for methane storage application. (b) Yang et al.*[Yang_2020] *used simulations at infinite dilution to prescreen the dataset before using computationally demanding simulations and multiple metrics to find the most promising ZIFs for carbon capture. (c) In Qiao et al.,*[Qiao_2016] *transport properties were screened along standard adsorption properties to find the best materials for the targeted $CO_2/N_2/CH_4$ ternary separation; similarly, cheaper calculations at infinite dilution were carried out in a first step, before using more expensive calculations at working pressure and temperature.*

1.2 Review of Screening Methodologies

1.2.1 Non-adsorption properties

Due to their high internal surface area, adsorption applications were a natural outlet for nanoporous materials. However, these materials can be used in many other applications. This section is dedicated to the physical and chemical properties not directly related to the adsorption process inside nanoporous materials such as catalytic activity,[Singh_2015, Greeley_2006, Back_2020] mechanical properties,[Chibani_2019, Gaillac_2020] or thermal properties.[Toher_2014, Sarikurt_2020, Ducamp_2021] These properties require a more refined description of the atomic interactions within the material. DFT simulations are usually performed to accurately retrieve these properties. However, the computational cost required is multiplied by several orders of magnitude compared to classical simulations. The size of the datasets screened is therefore much smaller (a few hundreds maximum), and the use of ML can potentially speed up the whole process. ML is based on lower-cost descriptors,[Evans_2017, Ducamp_2022] or it can be used in ML potentials for molecular simulations.[Eckhoff_2019, Friederich_2021]

Catalytic activity

Beyond adsorption properties, screening procedures have been applied to chemical properties such as catalytic activities. Heterogeneous catalysis is generally performed using metallic nonporous structures, the use of nanoporous materials can increase dramatically the active surface area and the catalytic activity. Consequently, MOFs have been demonstrated to show catalytic properties for several chemical reactions. Just to cite a few, one can think of hydrogenation, hydrolysis, oxidation, among others explicitly covered by McCarver et al. in their review.**McCarver_2021** Considering the sheer number of possible materials, computational studies are potentially more effective than experimental ones. Therefore, computational screenings evolved in the last decade aiming at studying more sizable datasets.

Although the vast majority of computational screenings have been done on small series, there are a few systematic screenings of bigger datasets. The scarcity of the latter can be explained by the high level of computational cost required. Here, we show some examples of such attempts by focusing on the example of C–H bond activation for the conversion of alkanes into alcohols in the presence of nitrous oxide.

Inspired by enzymatic catalysis of the reaction of small alkanes with N_2O into alcohols, Vogiatzis et al. identified seven iron-containing MOF structures out of 5,000 structures from the CoRE MOF database.**Vogiatzis_2016** They found two descriptors that govern the catalytic activity: (i) the N–O dissociation energy of N_2O on the adsorption site and (ii) the energy difference between two spin states of the intermediate. Using a screening on these descriptors, three structures were identified as promising for further experimental studies. The best one has been computationally demonstrated to catalytically and selectively oxidize ethane to ethanol in presence of N_2O. Moreover, the authors found that defects played a major role in the observed catalytic activity.

Later, Rosen et al. enlarged the scope of materials screened to other metals.**Rosen_2019** From an 838 DFT-optimized MOFs subset of CoRE MOF 2014, the authors selected 168 MOFs that were likely to have open metal sites and pore-limiting diameters that allows the diffusion of the reactants. They then used a fully automated workflow to place the reactants in the adsorption site and relaxed the system using periodic DFT calculations. As shown in Figure **??**, using the bond activation energy $E_{a,C-H}$ and the metal-oxo formation energy ΔE_O as key parameters, they classified the materials according to their relative stability and reactivity to find the best materials for the application. These energies were then analyzed using physicochemical descriptors such as the spin density on the oxygen and the metal–oxygen distance.

This type of brute force screening can be quickly cumbersome, as a result many researchers in the field are trying to find essential structure-activity relationships to accelerate future computational screenings. Several descriptors have been developed for high-throughput screenings: Butler et al. used electron removal energies to explain photocatalytic behaviors of MOFs;**Butler_2014** Rosen et al. showed that the energy required to form the metal oxide intermediate was a major descriptor of the thermal catalysis of alkane oxidation by N_2O;**Rosen_HTPDFT_2019** and Fumanal et al. show a screening protocol based on two energy-based descriptors to predict photocatalytic properties of MOFs.**Fumanal_descriptor_2020** Lately, Rosen et al. screened thousands of MOF structures to compare different DFT functionals and leveraged the data calculated to train machine learning models that can rapidly predict MOF band gaps.**Rosen_2022_high**

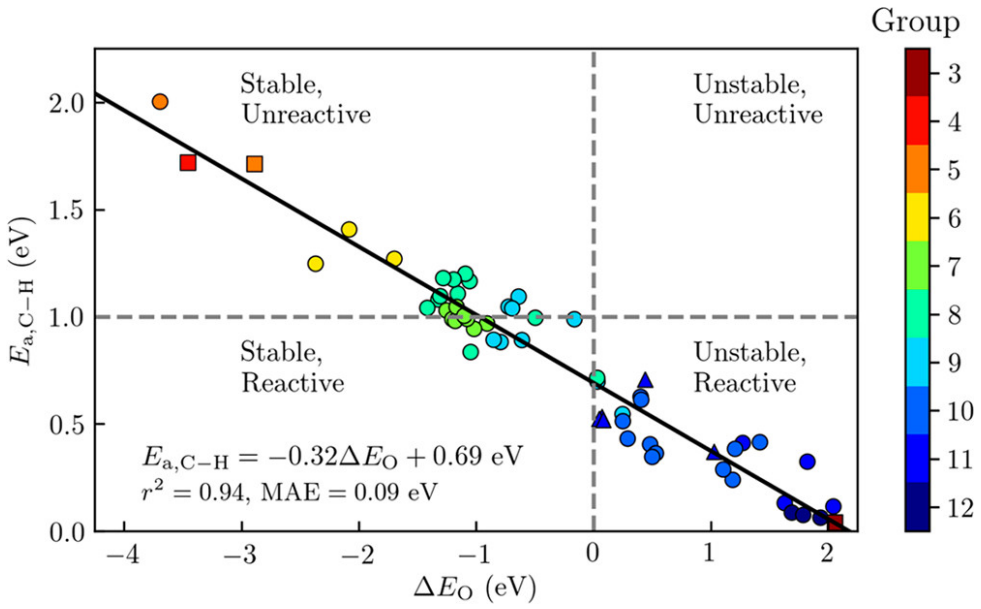


Figure 1.3: Analysis of a diverse set of experimentally derived metal–organic frameworks (MOFs) with accessible metal sites for the oxidative activation of methane. The graph shows the predicted barrier for the C–H bond activation of methane, E_a, as a function of the metal-oxo formation energy, ΔE_O. For each material, the symbol color refers to the group number of the metal in the periodic table. The best-fit line has been plotted in black, and has a mean absolute error (MAE) of 0.09 eV. *MOFs with E_a < 1 eV are classified as being reactive towards C–H bond activation and MOFs with ΔE_O < 0 as having thermodynamically favored active sites when using O_2 as the reference state. Reprinted with permission from Ref. [***Rosen_2019***]. Copyright © 2019 American Chemical Society.*

The development of ML methods are also critical in the field,[Rosen_2021] but the lack of centralized database with high precision descriptors is a challenge for the future of these methods. The influence of defects, the different ways of modeling MOFs as periodic structures or clusters, the diversity of structures and the stability of such structures remain open problems. Yet, it does not threaten the major role of high-throughput screenings in the early design process of any nanoporous materials for catalysis. To conclude this brief overview, we point the readers to a more exhaustive presentation of the matter.[Rosen_2022]

Mechanical properties

In the past decade, there has been a growing interest in the systematic study of physical properties of various classes of materials, including inorganic materials and framework materials. Among these physical properties, mechanical properties have been a topic of particular interest, as they are crucial for many applications, and at the same time can be computed by relatively standard methodologies. In particular, is it possible to calculate linear elastic constants (the second-order elastic tensor) in the zero-Kelvin limit by strain/stress or strain/energy approaches, performing a series of DFT calculations of strained structures and calculating the elastic constants. From these constants, all other mechanical properties can be evaluated by tensorial analysis,[Marmier_2010] including the bulk modulus, Young's modulus, shear modulus, Poisson's ratio, etc. This type of calculation can be coupled with any available quantum chemistry code,[Golesorkhtabar_2013] and is even integrated in some packages, like CRYSTAL17.[Dovesi_2018]

One of the first studies that investigated systematically the elastic properties of a family of materials was a 2013 study of all-silica zeolites,[Coudert_2013] i.e., crystalline and porous SiO_2 polymorphs. While this dealt with only 121 zeolitic frameworks out of 244 known structures, it showed that systematic studies at the DFT level were computationally tractable, and that they provided physical insight into the link between microscopic structure and macroscopic physical properties. This study demonstrated, among other things, that a few zeolites presented large negative linear compressibility (NLC), which could be linked to the wine-rack motif of their frameworks.

Outside the specific case of zeolites, other groups have applied DFT calculations of elastic constants in a high-throughput manner. de Jong et al. leveraged the structures of the Materials Project,[Matgenome, Jain_2013] trying to chart the diversity of elastic properties across the whole space of inorganic crystalline compounds.[deJong_2015] As shown in the Figure **??**, they provided a database containing the full elastic information of 1,181 inorganic compounds initially, and has grown steadily since then, containing more almost 14,000 records to date.[MaterialsProject] This dataset has been used in two different ways by researchers in the field.

Firstly, the exploration of the database of elastic properties by tensorial analysis has allowed studying quantitatively the occurrence of certain "anomalous" or rare mechanical behavior, including negative linear compressibility, very high anisotropy, or negative Poisson's ratio (also called *auxeticity*). Indeed, such properties are considered rare and usually sought after — the materials exhibiting these anomalous behaviors are mechanical metamaterials.[Coudert_2019] In addition to their fundamental interest, such materials have applications in materials engineering: for example in energy dissipation (as shock absorbers and for bulletproofing), energy storage, as well as acoustics.[Surjadi_2018] However, it was not possible until now to quantify exactly "how rare" they are. Chibani et al. showed through a systematic exploration of available mechanical properties of crystalline materials that general mechanical trends, which hold for isotropic (noncrystalline) materials at the macroscopic scale, also apply on average for crystals. Moreover, they could quantify the presence of materials with rare anomalous mechanical properties: 3% of the crystals were found to feature negative linear compressibility, and only 0.3% to exhibit complete auxeticity (negative Poisson's ratio in all directions of space).

Secondly, the datasets of mechanical properties were used as a basis to accelerate the discovery of novel materials with targeted behavior. Dagdelen et al. used search algorithms to identify 38 candidate materials exhibiting features correlating with auxetic behavior, from more than 67,000 materials in the Materials Project database.[Dagdelen_2017] Performing DFT calculations on these 38 structures, they could identify 7 new auxetic compounds. In a more complex setup, Gaillac et al.[Gaillac_2020] have used a multiscale modeling strategy for the fast exploration and identification of novel auxetic materials. They combined classical force fields MD simulations with DFT calculations on candidate materials, and then used this reference DFT data to train an ML algorithm. They found that the accuracy of this multiscale method exceeds the current low-computational-cost approaches for screening. In a similar work, Moghadam et al. used molecular simulation to train an artificial neural network (ANN) for the prediction of the bulk modulus of metal–organic frameworks.[Moghadam_2019] This shows the potential of such methodologies to treat very different (chemically as well as structurally) classes of materials.

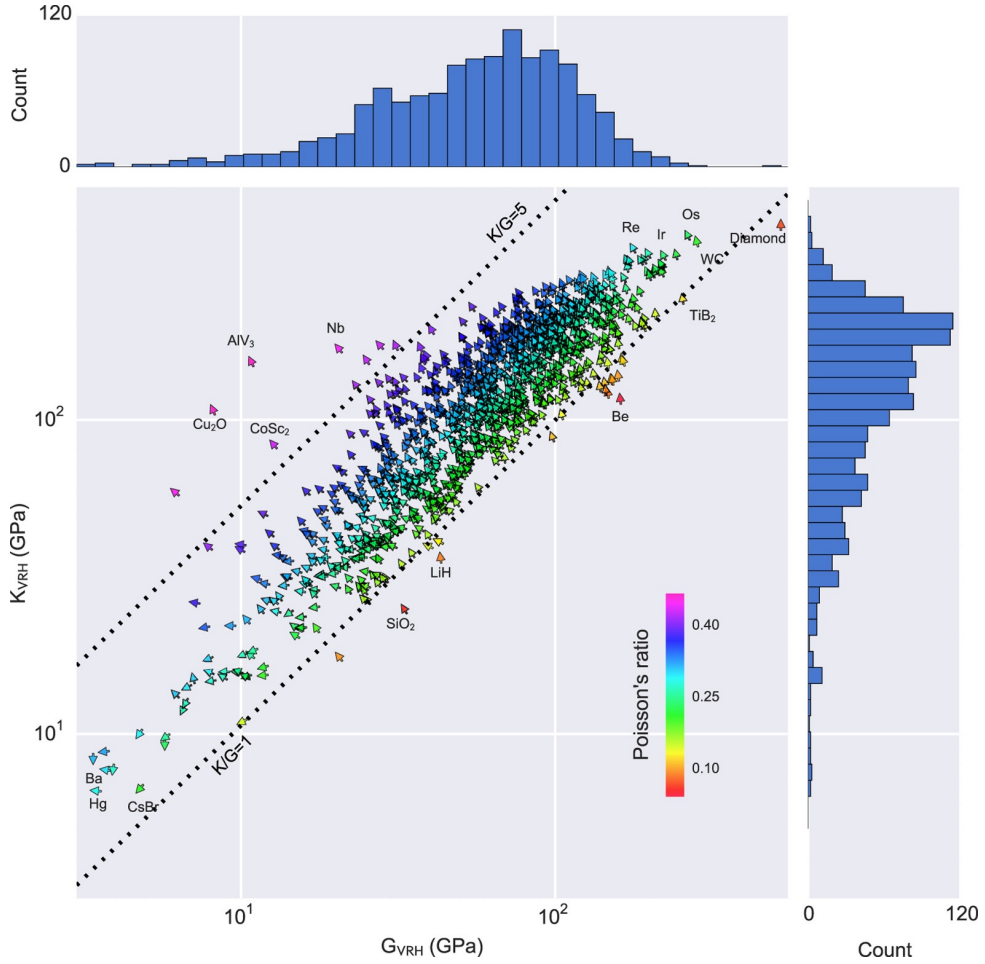


Figure 1.4: Statistical analysis of the calculated volume per atom, Poisson's ratio, bulk modulus K_{VRH} *and shear modulus* G_{VRH} *of 1,181 compounds in the Materials Project database. In the vector field-plot, arrows pointing at 12 o'clock correspond to minimum volume-per-atom and move anti-clockwise in the direction of maximum volume-per-atom, which is located at 6 o'clock. Reprinted from Ref. [***deJong_2015***] under CC-BY license. Copyright © 2015 de Jong* et al.

Thermal properties

While mechanical properties (in the elastic regime) have been by far the most studied physical property in nanoporous materials, others have also been occasionally screened. We can cite, in particular, the systematic study of piezoelectric tensors by de Jong et al., on almost a thousand crystalline compounds, by first-principle calculations based on density functional perturbation theory.[deJong2015_piezo] We can also cite efforts to calculate thermal properties in a high-throughput setup, using the quasi-harmonic approximation (QHA).[Togo_2010] This method requires the calculation of each structure's phonon modes at various volumes, and can be coupled to any electronic structure program.[Togo_2015] It is, however, quite computationally intensive, and sensitive to the parameters of the QHA methodology (range of volume, range of temperature, precision of the frequency calculation, etc.). Therefore, it has been limited so far to modest numbers of structures: a dataset of 75 inorganic structures by Toher et al.,[Toher_2014] and more recently a dataset of 134 pure SiO_2 zeolites by Ducamp et al.[Ducamp_2021] Very recent work in our group on the prediction of thermal properties through machine learning based on structural features alone indicates that thermal behavior is more difficult than mechanical

behavior to predict, and might require the use of a wider set of structural descriptors or more advanced ML models.[Ducamp_2022]

1.2.2 Transport adsorption properties

The thermodynamic properties, we will be presenting in the next section, only describes the state of equilibrium of the adsorption process. But sometimes the transient state can last long before reaching the equilibrium, which makes the process more time-consuming. Thus, the transport properties complete the thermodynamic description of the adsorption process inside a nanoporous material. For example, a low diffusion rate would mean for storage applications more time and energy needed to fill-up the tanks, or for separation applications a less selective process than expected. In more extreme cases of molecular sieves for fluid separation, the transport properties become predominant to assess the performance. One can leverage the difference of the molecules' diffusion coefficients to selectively filter gas mixtures through a nanoporous membrane.[Miandoab_2021] Here, the main subject becomes the transient state and not the equilibrium. This section is thus dedicated to the kinetics of the adsorption process to better model the time required to reach the equilibrium or to study out-of-equilibrium processes such as molecular sieving by nanoporous membranes.

Kinetic properties

In most computational screenings, the diffusion coefficient considered is the self-diffusion coefficient that describes an infinite-dilution case. Other multi-component diffusion coefficients could be considered, but for simplicity and clarity they won't be mentioned in this review. The calculation of the self-diffusion coefficient gives a first estimation of the kinetics in a storage or a separation process in the limit of low adsorption loading.

There are two approaches to estimate the diffusion inside a porous material: the first one relies on molecular dynamics (MD) and the second one on transition state theories. In the first approach, one can analyze the mean squared displacement of the adsorbed molecule moving in the material. In the second, one identifies minimum energy path along the material to identify transition states (TS) to calculate diffusion energy barriers. The MD-based method requires fewer assumptions and is therefore more reliable than the TS-based method, but the latter is computationally more efficient in the case of low diffusion rate (diffusivity lower than 10^{-11} m^2 s^{-1}).

State-of-the-art MD simulations could calculate rather accurate diffusion coefficients, but the computational cost scales quickly with the number of structures. To use this method on a large dataset without spending too much computation time, Watanabe and Sholl prescreened the pore sizes of 1,163 MOFs to select only the structures within a certain range of PLD (pore limiting diameters).[Watanabe_2012] A restricted list of 359 MOFs was then used to carry out MD simulations to calculate diffusion coefficients. The results of this final screening are then used to extract the most promising structures for further experimental or computational investigation. Similarly, Qiao et al. used a multistage screening to find the best membrane material within about 130,000 hypothetical MOFs for a $CO_2/N_2/CH_4$ separation.[Qiao_2016] They started to select materials based on pore geometry analysis; then they calculated Henry's coefficient and diffusion coefficients at infinite dilution; finally, they compared the binary permselectivities to extract 24 promising MOFs for ternary adsorption and diffusion calculation at the desired pressure and temperature conditions.

Another approach replaces MD simulations with more computationally efficient TS-based methods to determine diffusion coefficients. Haldoupis et al. developed an algorithm to identify diffusion paths by exploiting an energy grid with a clustering algorithm. The diffusion paths are then analyzed to identify the pores and the channels, and to calculate key geometric (the PLD or the largest cavity diameter) and energetic (Henry's constant, diffusion activation energy) features.**Haldoupis_2010** As illustrated in Figure **??**, they found a clear dependence of the diffusion energy barrier to the PLD. As one of the first TS-based screenings, it is still subject to many development perspectives. For instance, the approach is limited to spherical adsorbates and rigid frameworks. Moreover, the diffusion coefficients are approximated using a simplistic hopping model for a qualitative analysis. This method is highly efficient, but the accumulation of approximations makes a quantitative systematic analysis of diffusion coefficients out of reach.

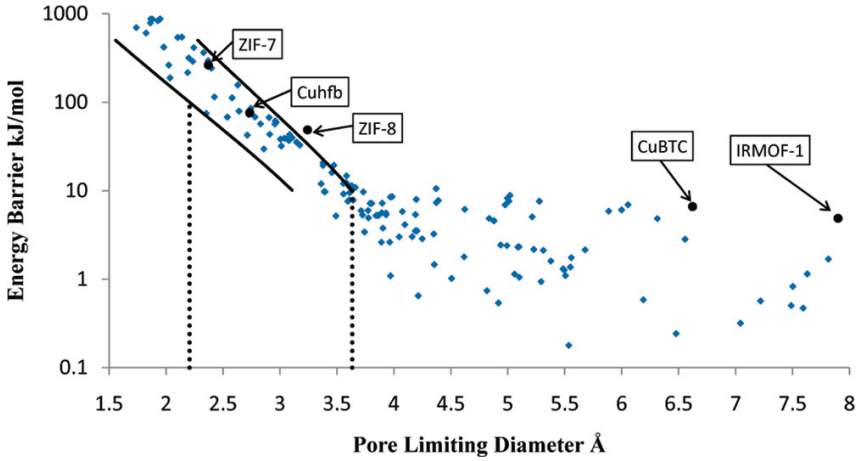


*Figure 1.5: Calculated energy barrier for the diffusion of CH_4 in 216 metal–organic frameworks (MOFs), shown as a function of the pore-limiting diameter. The solid lines represent statistical upper and lower bounds on the energy barrier, in a transition state theory approach. Reprinted with permission from Ref. [***Haldoupis_2010***]. Copyright © 2010 American Chemical Society.*

Later, Kim et al. introduced a flood fill algorithm to obtain all the points within a given energy.**Kim_2013** These points are then identified as channels or blocked regions. Along the channels, local minimums of energy are defined as lattice sites and transition states are defined perpendicular to the diffusion direction. A random walk is then computed along the lattice sites with hopping rates defined according to the activation energy. A diffusion coefficient is then calculated in each three directions of the space and an average diffusion coefficient is finally determined. A comparison with the MD method on the IZA zeolite structures shows good agreement, but there are still some discrepancies explained by correlated hops in the case of rapid diffusion or by the presence of complicated channel profiles. Inspired by this work, Mace et al. developed a similar method that progressively fill the energy grid to detect transition states, hence removing the previous restriction to orthogonal cells only.**Mace_2019** The diffusion coefficient is now computed using a kinetic Monte Carlo simulation allowing the adsorbate to jump freely in all directions instead of restricting it in a single dimension. This new method, called TuTraSt, handles very complex diffusion paths (like in the AEI zeolite). This new approach seems to be promising as it is in good agreement with MD simulations, while being 2-3 orders of magnitude faster. However, the time performance could improve tremendously by translating it from Matlab to C++ and by implementing parallelization procedures.

Very recently a massively parallel GPU-accelerated string method has been implemented and shared publicly to compute very efficiently diffusion coefficients based on the transition state theory.[Zhou_2021] The recent developments in the prediction of diffusion coefficients in nanoporous materials point towards a promising future for the screening of transport properties applied to even larger databases. Going further, Bukowski et al. reviewed thoroughly diffusion in nanoporous solids as an attempt to connect theory to experiments.[Bukowski_2021]

Membrane materials

In separation application, the study of the transport properties can evaluate the feasibility of the thermodynamic equilibrium, crucial for any bed separation process. If this separation is not feasible, kinetic separation or partial molecular sieving are to be considered. Some notable examples are air separation in zeolites using pressure swing adsorption,[ruthven1990air] N_2/O_2 separation in carbon molecular sieves,[Reid_1999] or N_2 removal from natural gas.[Wang_2019] In kinetic separation, the valuable metric is not the selectivity anymore, but the permselectivity, i.e. the product of the selectivity and the permeability (ratio of diffusion coefficients). Therefore, the screening of diffusion coefficients gives complementary information to the thermodynamic selectivity screenings. Here, we give some examples of such screening and the main descriptors that partially explain the computed figures of merit.

To give an overview on the potential of computational screenings to predict transport properties, we are now going to focus on the membrane separation applied to natural gas upgrading. The separation of CH_4 from N_2 and CO_2 is a crucial step of this upgrading process. In 2016, a large-scale high-throughput screening (see Figure **??** for the approach) of hypothetical MOF membranes for upgrading natural gas has been performed using MD simulations.[Qiao_2016] Qiao et al. confirmed the existence of MOF materials beyond the upper bound for N_2/CH_4 and CO_2/CH_4 separations determined by Robeson on a large set of polymeric membranes.[robeson1991correlation] This Robeson's upper bound is systematically crossed by MOF materials in computational screenings, see as an example the Figure **??**. This can be explained by the fact that MOFs perform better than polymeric frameworks and the simulations at this level of theory. They also identified 24 MOFs suitable for the ternary $CO_2/N_2/CH_4$ separation using a multistage screening described in the previous section.

Two years later, Qiao et al. used the same approach to study this ternary separation on a database of synthesized structures.[Qiao_2018] Applying machine learning techniques to their data, they performed a QSPR analysis. Using a principal component analysis, they notably found that the permeability is higher when materials have high PLD and void fraction coupled with low density and percentage of pores within a characteristic range. The opposite was found to be true for high membrane selectivity for the CO_2/CH_4 separation. Using decision tree algorithms, they gave objective procedures of selecting the best separation membranes based on some key descriptors. Finally, they studied in detail some top performing materials found by a support vector machine algorithm.

Altintas and Keskin later performed a screening on the same database for CO_2/CH_4 membrane separation to identify the best performing materials and perform more computationally demanding simulations.[Altintas_2018] The simulations in rigid structures at infinite dilution show numerous structures above the Robeson's upper bound as shown in the Figure **??**, this crossing of the upper-bound can be explained by either a better performance of MOF membranes compared to the polymeric membranes used by Robeson, or an overestimation due to oversimplified

assumptions (infinite dilution, rigidity). But when higher pressures and flexibility are considered, the selectivity values are dropping down closer to the upper boundary, hence confirming the overestimation of the performance in screenings based on rigid approximations at infinite dilution. But the best performing materials are still above the Robeson's upper bound and can therefore be used in mixed matrix membranes with polymeric membranes. Budhathoki et al. developed a screening methodology for MOFs in mixed matrix membranes for carbon capture applications by estimating permeation values in these composite materials using a Maxwell model.[Budhathoki_2019] The authors even proposed a pricing for each material compared to their relative performance. Similar studies have been carried out on different materials, Yan et al. showed the influence of decorating COFs with different chemical compounds on the membrane selectivity.[Yan_2018]

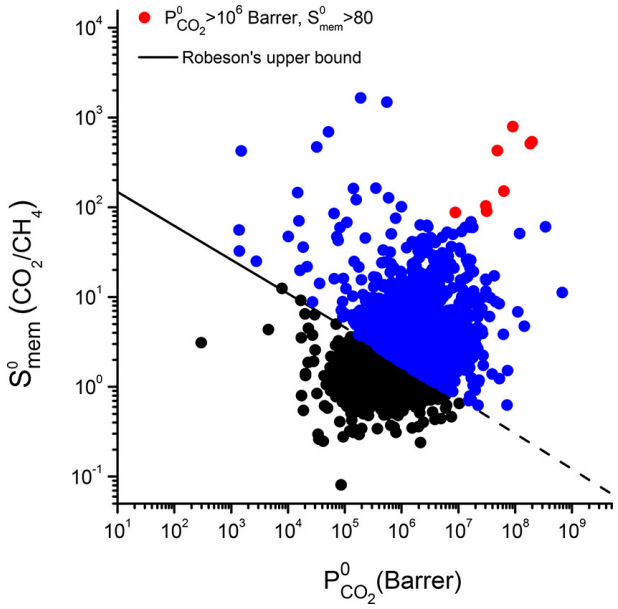


Figure 1.6: Selectivity and permeability of metal–organic framework (MOF) membranes for CO_2/CH_4 separation computed at infinite dilution by combining Grand Canonical Monte Carlo and molecular dynamics simulations.[Altintas_2018] *The black solid line represents the Robeson's upper bound.*[robeson1991correlation, Robeson_2008] *MOFs that can exceed the bound are shown in blue, and the 8 top-performing MOF membranes are shown with red symbols. Reprinted with permission from Ref. [**Altintas_2018**]. Copyright © 2018 American Chemical Society.*

The transport properties screening is based on the calculation of diffusion coefficients at infinite dilution and in rigid molecules. There are different methods to calculate them (mainly MD and TS-based methods). Flexibility and pressure dependence are very hard to incorporate directly in the screening procedures. Researchers usually consider these factors at the end of the screening on the most promising structures because of the computational complexity of the corresponding simulations. To take account of pressure dependence, we need an MD simulation of several adsorbates that takes much more time than running single component simulations,[Keskin_2007, Keskin_2009] which makes it harder to include in a high-throughput screening. Flexibility could be taken account by calculating snapshots and running multiple MD simulations, or by using flexible force fields, which means in both cases an increase in computational run-time. Some faster methods of quantitatively predicting the impact of

flexibility on diffusion are being investigated in ZIFs and could give an interesting alternative to these expensive methodologies.**Han_2020**

1.2.3 Thermodynamic adsorption properties

In its early development, computational screening was mainly used to predict thermodynamic properties in adsorption processes. Three main applications have been identified in the associated literature: gas storage (for energy or medical applications), gas separation (noble gas, hydrocarbons, carbon dioxide, etc.) and post-combustion CO_2 capture. These applications are closely linked to urgent environmental and energy issues that are yet to be solved. Screening can guide the development of better performing materials by shedding light upon unknown structure-property relationship, probes possible theoretical limitations (unreachable targets) and identifies potential candidates that need to be experimentally tested.

Gas storage

One can leverage the high surface density of the nanoporous materials, especially the MOFs, to stock in very low-density gas. In the field of energy storage or transportation, natural gas (mainly methane) or hydrogen are considered plausible alternative fuels to replace conventional ones for transport. The US Department of Energy (US DOE) recently financed research programs and set targets for methane and hydrogen storage. Nanoporous materials could reduce energy, infrastructure and security cost due to the required compression and cooling. In this section, we are focusing on high-throughput screening for methane storage in nanoporous materials, before broadening the scope hydrogen and other perspectives.

One of the pioneering works in computational screening was published in 2012 by Wilmer et al..**Wilmer_2012** They performed a large-scale screening of 137,953 hypothetical MOF structures to estimate the methane storage capacity of each MOF at 35 bar and 298 K based on the US DOE standards. Back then, the US DOE set a target methane capacity value of 180 $vol^{STP}vol^{-1}$ (which has since been achieved by several materials reported in the literature). In their large-scale analysis, Wilmer et al. found over 300 hypothetical MOFs that meet the targeted requirements and the best one can store up to 267 $vol^{STP}vol^{-1}$, surpassing the state-of-the-art of the time. From their large dataset, a preliminary structure-property relationship analysis revealed that void fraction values of approximately 0.8 and gravimetric surface areas in a range 2500-3000 $m^2\,cm^3$ resulted in the highest methane capacities. Optimal pore size is also shown to be around the size of one or two methane molecule(s). Maximization of gravimetric surface area was a common strategy in the MOF design for storage applications, but this study showed the existence of an optimal range of surface area values. Computational screenings can draw clear relationships between structural descriptors and performance. Later, a more quantitative relationship was drawn by Fernandez et al. using ML models as illustrated on Figure **??**. Beware not to over-interpret the relation given by the response surface, since the identified maxima do not always have a physical reality, especially where there is no training data in the area pointed by the red arrows. However, it highlights promising unexplored feature space and shows potential research directions.

Since then new materials above the target have been found and the US DOE decided to set a higher target of 315 $vol^{STP}vol^{-1}$. Until now, this new target is not yet reached. This is why the recent developments have focused on assessing the feasibility of such a target by accelerating the screening methods so that more data can be screened, and by interpreting the QSPR models

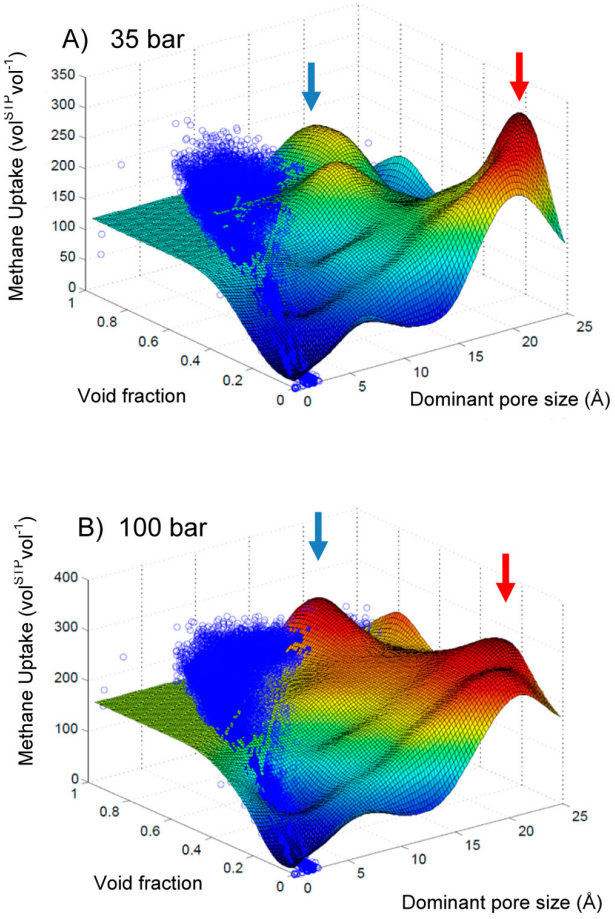


*Figure 1.7: Two-dimensional response surfaces of the support vector machine (SVM) models trained by Fernandez et al. for methane storage at (A) 35 bar and (B) 100 bar using void fraction and dominant pore size. The blue dots represent the GCMC simulated uptake values. The color of the surface represents the methane storage value, from blue (the lowest values) to red (the highest values). Blue and red arrows indicate maxima on the response surface. Reprinted with permission from Ref. [***Fernandez_2013***]. Copyright © 2013 American Chemical Society.*

to extract important knowledge for the design of novel materials. For instance, Gómez-Gualdrón et al. showed that even by artificially quadrupling the Lennard-Jones interaction factor ϵ and by increasing the delivery temperature by 100 K, the newly set target is only reached by a handful of MOFs.[**Gomez_Gualdron_2014**] This study suggests the impossibility to reach the DOE target using a preconceived (experimentally or theoretically) material to store methane. However, this theoretical limitation can be overcome by increasing the surface density of sites with high affinity with methane and by increasing the delivery temperature.

Later, a larger-scale screening on methane storage was carried out by Simon et al. on 650,000 experimental and hypothetical structures of zeolites, MOFs, and PPNs. This study confirmed that the classes of materials currently being investigated were unlikely to meet the new target. The authors suggested that it wasn't surprising since the target was based on economic arguments, while the screening is based on thermodynamic arguments.[**Simon_2015_EES**] This example illustrates the power of large-scale screening to settle questions of physical feasibility (if simulations are accurate) and hence avoiding experimental efforts spent on impossible tasks.

More recently, a dataset containing trillions of hypothetical MOFs have been screened for methane storage.[Lee_2021] Lee et al. developed a methodology using machine learning combined with genetic algorithm to perform the largest screening until now. In addition to confirming most of the results (theoretical limits and QSPR) found by previous screenings, 96 MOFs were found to outperform the current world record. This study shows the scaling potential of ML-assisted screenings in handling "Big data".

Similarly, computational high-throughput screenings have been applied to other storage applications such as hydrogen storage. Computational screenings showed that cryogenic storage of hydrogen can meet the DOE target of 50 g L^{-1}.[Gomez_Gualdron_2016, Bobbitt_2016, Thornton_2017] Anderson et al. performed a large-scale screening based on neural networks to test out multiple pressure/temperature swing conditions to find that the maximal deliverable capacity cannot exceed 62 g L^{-1}.[Anderson_2018] Compared to the density of liquid hydrogen (72 g L^{-1}), this upper limit seems reasonable since the adsorbent material takes at least 10-20% of the tank. Here, we only showed some flagship results of the field. For a more detailed meta-analysis, Bobbitt and Snurr wrote a very complete review on computational high-throughput screening of MOFs for hydrogen storage.[Bobbitt_2019]

1.2.4 Gas separation

As a representative example of what could be done in the field of gas separation, we are going to focus on Xe/Kr separation. Nanoporous materials can be used as a safer, cheaper and less energy-intensive option for this gas separation. However, experimental design of top-performing materials can be cumbersome. Computational screenings is an ideal tool to kick-start the development of this new technology by identifying rapidly the best candidates.

Small-scale screenings

Metal–organic frameworks, and later other supramolecular porous materials like covalent organic frameworks (COFs), have been proposed for applications in separation of noble gases for a decade. With no aim of being exhaustive, we highlight some milestones in that area, both from experimental and computational point of view.

In 2012, Liu et al.[Liu_2012] published an experimental study of two MOFs, HKUST-1 and Ni/-DOBDC, for adsorption of Xe and Kr at ppm (part-per-million) levels in air. The target application was the removal of Xe and Kr from nuclear fuel reprocessing plants. The same group later proposed a two-column method for the separation of Kr and Xe from processed off-gases,[Liu_2014] based on MOF materials. At about the same time, Bae et al.[Bae_2013] combined a computational Grand Canonical Monte Carlo (GCMC) study with experimental breakthrough measurements of the separation of a Xe/Kr mixture on MOF-505 and HKUST-1.

Parkes et al.[Parkes_2013] studied sixteen different MOF materials for Kr, Ar, and N_2 adsorption and separation, through GCMC simulations. They concluded on the potential of MOFs for separation, and a general correlation between the Henry's constant and the isosteric heat of adsorption for the three gases studied. A year later, in 2014, Chen et al.[Chen_2014] demonstrated, again through a combined computational and experimental study, the potential of porous organic cages for selective binding of xenon over krypton.

Later experimental work expanded these early separation studies to different types of MOF materials. Xiong et al.[Xiong_2015] studied a flexible zinc tetrazolate framework for xenon selec-

tive adsorption over krypton, argon and nitrogen. Thermodynamic analysis of the adsorption isotherms at various temperatures confirmed the occurrence of a "breathing" structural transition upon Xe uptake, contributing to a high working capacity for a pressure swing adsorption (PSA) cycle. Lee et al.[Lee_2016] compared the selective adsorption properties for Xe/Kr mixtures on three highly studied MOFs, namely UiO-66(Zr), MIL-100(Fe) and MIL-101(Cr), and confirmed a high potential of UiO-66(Zr) for separations under dynamic flow conditions. These authors also assessed the hydrothermal and radioactive stability of the material, a test seldom performed in the existing literature, and found it to be good. In a further study,[Lee_2018] they demonstrated that Xe/Kr selectivity could be further improved by ligand substitution.

In parallel, computational studies were published to provide insight at the microscopic level into the mechanisms behind good (and bad) separation properties. Wang et al.[Wang_2014] studied 6 MOFs and COFs for adsorption of Xe and Xe/N_2 separation, through GCMC simulations looking at the impact of pressure (and therefore pore filling) on selectivity. Anderson et al.[Anderson_2017] combined GCMC and biased MD simulations to elucidate the nature of adsorption- and diffusion-based Kr/Xe separation mechanisms in four archetypal nanoporous materials: SAPO-34, ZIF-8, UiO-66, and IRMOF-1. These authors draw a couple of general conclusions, including the fact that diffusion selectivity for krypton dominates membrane separation selectivity, and large pore cages and stiff pore windows are desirable — however the scope of these conclusions is inherently limited by the small number of materials actually studied.

In a different family of materials, Tong et al.[Tong_2017] have surveyed the structure–property relationships of covalent organic frameworks (COFs) for noble gas separation, by GCMC simulations of 187 different materials for Kr/Ar, Xe/Kr and Rn/Xe separations. These authors included in their calculations some adsorption figures of merit (AFM), representative of the conditions of industrial vacuum (VSA) and pressure swing adsorption (PSA) processes.

One area that has been particularly explored is the tuning and improvement of separation properties through the presence and nature of coordinatively unsaturated sites (or open metal sites) in MOFs. In 2016, Vazhappilly et al.[Vazhappilly_2016] used density functional theory (DFT) calculations of host–guest binding energies to probe the impact of the metal atoms in a specific framework (MOF-74) on Xe and Kr adsorption. Later, Zarabadi-Poor et al.[ZarabadiPoor_2018] investigated — again through computational methods — a series of metal–BTC MOFs for recovering xenon from exhaled anesthetic gas, i.e., mixtures of CO_2, O_2, and N_2.

Large-scale computational screening

In its early stage, computational screening has been used on a small series of nanoporous materials to generate specific knowledge on some subclasses of materials. These small-scale screenings combined with experiments helped faster identification of good performing candidates, but they failed to establish general rules of design or to explore the unknown. Larger-scale screenings overcame these limitations by trying to exhaustively cover the whole spectrum of nanoporous materials.

The first large-scale computational screening on Xe/Kr adsorption-based was performed by Sikora et al. based on the same approach previously developed for methane storage by their group at the Northwestern University.[Sikora_2012] This study was based on the same 137,000 structures of hypothetical MOFs.[Wilmer_2012] They calculated the Xe/Kr selectivity using Monte Carlo molecular simulations on the whole database by iteratively increasing the number of

steps and selecting the best materials similar to the approach on Figure **??**. By analyzing the relationships between pore sizes and selectivity, they confirmed a hypothesis from a smaller scale study that the pores should be between the size of 1 to 2 xenon molecules.**Ryan_2010** Tube-like channel was also found to favor better selectivity. Moreover, they found that top performing materials could have a selectivity around 500; but we can only conclude on the order of magnitude of the theoretical limitation of the Xe/Kr selectivity, considering the statistical uncertainty of the simulation.

Seizing the opportunity of a formidable expansion of the nanoporous materials database triggered by the Materials Genome Initiative, Simon et al. screened 670,000 experimental and hypothetical nanoporous material structures for Xe/Kr separation (see Figure **??**).**Simon_2015** It is one of the largest-scale screenings performed in this area. Inspired by the work of Fernandez and co-workers,**Fernandez_2013** they used ML algorithms to train a model on a diverse subset of 15,000 structures. This method allowed them to run time-consuming molecular simulations only on this training set, before applying the ML model to predict the selectivity values on the larger set of structures. On top of analyzing the links between pore descriptors and selectivity, they rationalized it using theoretical pore models of spherical and cylindrical geometries to confirm the findings of Snurr and co-workers.**Ryan_2010**, **Sikora_2012** By comparing the structural descriptors of good-performing and bad-performing structures, they concluded that geometrical descriptors wasn't enough to explain the performance. The analysis of a few top candidates suggests that different chemical insights could explain their good performance. For SBMOF-1 or KAXQIL,**KAXQIL** an experimental MOF, its higher performance was explained by the tubelike 1D channel with a very favorable binding site formed by carbon aromatic rings. This nanoporous material was later tested using breakthrough experiments and proved to be one of the most promising candidates.**Banerjee_2016** This close collaboration between computation and experimentation is a testimony of the potential of computational screenings to find nanoporous materials for any targeted application.

The experimental work on Xe/Kr separation on SBMOF-1 revealed discrepancies between the selectivity values obtained experimentally and computationally.**Banerjee_2016** The assumption of rigid crystal structures in the molecular simulations could partially explain the difference observed. Witman et al. proposed that the flexibility of the materials that weren't considered in the screening of Simon et al. could explain the lower selectivity observed experimentally.**Witman_2017** In this study, they screened the Henry regime separation of about 4,000 MOF structures of the CoRE MOF 2014 database,**Chung_2014** and found that intrinsic flexibility, i.e. the thermal vibration of the material, can make the pore size derive from the ideal value for the separation and hence lower the selectivity. This study further confirms the importance of the pore size by highlighting the effect of its evolution over time.

In 2019, Chung et al. screened the most extensive simulation-ready and experimentally synthesized MOF structures for Xe/Kr separation.**Chung_2019** This study pointed out the potential of coordinated solvent molecules to fine-tune the selectivity for any separation application, since their presence can enhance selectivity in some cases. The results of their screening confirm the potential of structures such as SBMOF-1 found by Simon et al., but they also described a few structures with similar selectivity but with better xenon uptake. The authors emphasize the importance of considering other figures of merit such as the adsorption capacity. Other factors should be taken into account to find the best trade-off between all the relevant figures of merit;

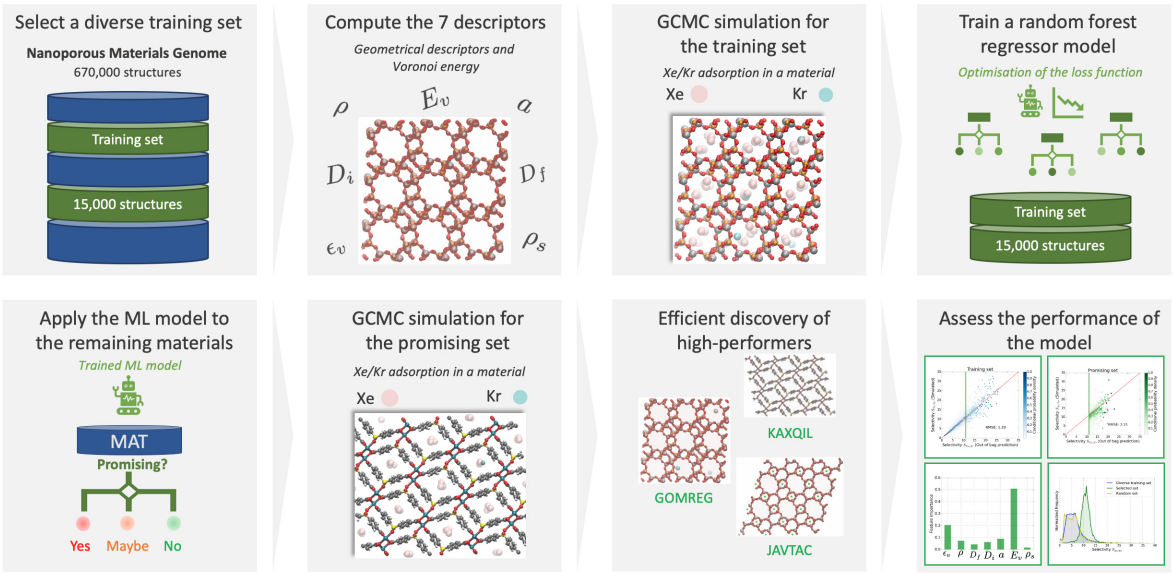


Figure 1.8: Schematic representation of large-scale screening of nanoporous materials for Xe/Kr adsorption-based separation by Simon et al.,[Simon_2015] *based on a combination of Grand Canonical Monte Carlo simulations and machine learning algorithm (Random Forest Regressor). The main goal of this screening is to find high-performing materials in a large dataset of both experimental and hypothetical materials. Adapted with permission from Ref. [**Simon_2015**]. Copyright © 2015 American Chemical Society.*

we could think of the kinetics of such a separation, the effect of flexibility on the performance, the stability of the materials (especially in radioactive environment), the financial aspects, and more.

After this quick overview of the different screening studies in the field of xenon/krypton separation, we are now going to detail its industrial context, the foreseen top materials that could fulfill the industrial separation and what further studies are needed to better understand the process while discovering new materials.

1.3 Separation of Xenon from Krypton

In this section, we will try to see how we can apply the above-mentioned screening methodologies to help us understand the origins of the Xe/Kr separation and identify promising materials for industrial applications.

1.3.1 Industrial applications

The industrial interest for noble gases lies first in the many applications attached to them. For instance, xenon has multiple applications in the medical (e.g. anesthesia, painkiller, imaging),[cullen1951anesthetic] aeronautical,[Patterson_2002, Coxhill_2005] lithographic,[Abramov_2018] microelectronic[Chang_1995] or lighting sectors,[Jarman_1974, Tanaka_2019] just to cite a few. To meet the demand for these noble gases, one should consider all available sources, the most obvious one being the air we breathe. Xenon and krypton have both very low atmospheric concentrations; out of a thousand liters of air, we would extract at most one tenth of a milliliter of xenon and one of krypton.[kerry2007industrial] Nevertheless, direct extraction from the air remains the main

production mean for xenon and krypton along with chemical plant off-gases that contains a higher concentration of inert gas (e.g., amonia purge gas). In these cases, the industry more commonly uses cryogenic distillation to extract xenon and krypton, which requires a compression and cooling of the gas mixture at very low temperatures. The separation process can be broken down into three steps: first the condensation of all gas with a boiling point higher than the oxygen, then the purification of oxygen resulting in a 20-80 xenon/krypton mixture, and finally the separation of xenon from krypton. In 1997, several cases of explosion of separation units were caused by the reaction of non-filtered dangerous hydrocarbons with purified liquid oxygen produced in the second step of this long process.[distill_accident, distill_accident2] The extreme chemical and physical conditions required for cryogenic distillation support the need for less energy-intensive and safer alternatives.

Industrial application	Xe/Kr composition
Extraction from ambient air[kerry2007industrial]	20/80
Spent nuclear fuel[auerbach2003handbook]	90/10
Molten Salt Reactor[Riley_2019]	?

Table 1.1: Composition of the Xe/Kr mixture in different applications.

The role of a dispatchable source of low-carbon energy can only be fulfilled by batteries charged by renewable energies (wind or solar) or by nuclear plants. However, one of the major criticisms on this source of energy concerns the management of the radioactive waste. As promising technologies in gas separation emerge, there is an increasing need for a solution for the release of very small amount of radioactive off-gases like Kr_{85} from nuclear spent fuels.[Blomeke_1969] Furthermore, stable xenon isotopes are also produced in these spent nuclear fuels, which can be used in all the above-mentioned applications. In the context of a regained interest in nuclear energy, the fourth generation nuclear plants are projected to be built on other technologies such as the light water or the molten salt technologies.[LeBlanc_2010] Molten salt reactors would continuously produce xenon and radioactive krypton in the electricity generation process.[Riley_2019] The development of gas separation units in these facilities would represent a promising source for xenon production. Yet, we can laboriously imagine deploying standard cryogenic distillation units in a nuclear facility for obvious security reasons. Consequently, nanoporous materials are considered as the alternative technology for xenon/krypton separation. Zeolites are already used as a pre-purification system,[kerry2007industrial] and they are now projected to be used as a standalone separation system.

Banerjee et al. proposed a two-bed system with a first bed filled with MOFs designed for xenon separation and then a second one for radioactive krypton capture.[Banerjee_2014] The authors proposed some examples of material that could be used for this separation unit; more research is needed to find out what the best materials for these separations are. In the following section, we will review the most promising materials for this separation and the structural explaining their high performance.

1.3.2 Promising materials for the separation

Several experimental reports used the strategy outlined by computational screenings to improve separation properties, as well as tuning the chemical nature of the organic linkers. The main

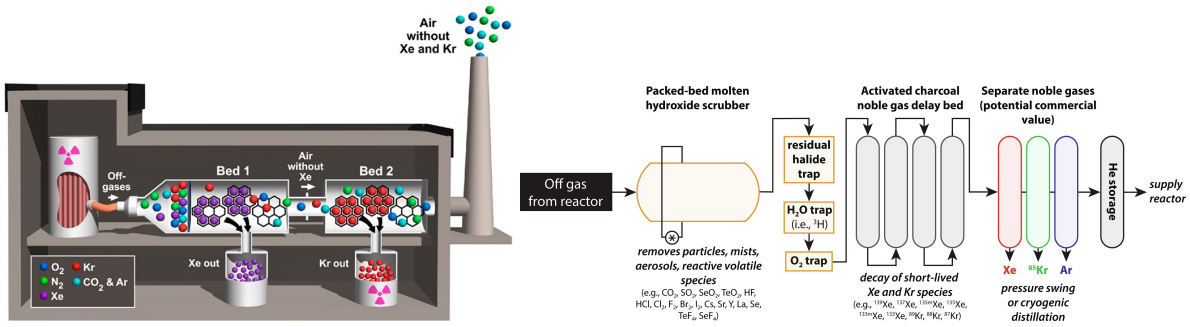


*Figure 1.9: Representation of xenon/krypton separation process using porous materials in a nuclear fuel reprocessing plant and in a molten salt reactor. Reprinted with permission from Ref. [***Banerjee_2014***] copyright © 2014 American Chemical Society and Ref. [***Riley_2019***] copyright © 2019 Elsevier.*

criteria outlined by the different studies on xenon/krypton separation call for pore size tailor-made for xenon and also for maximized interactions with the framework atoms obtained either through the chemical nature or the shape of the cavities.

In the early phase of the experimental design of materials for the xenon/krypton separation, Wang et al. synthesized a cobalt MOF $Co_3(HCOO)_6$ with a selectivity of 12 that present rather narrow pores (around 5 Å) connected by zig-zag segments.[**Wang_2014**] Later, Chen et al. synthesized a selective porous cage material by not only focusing on the pore size but more importantly on the shape of the cavity, the selectivity of around 20 was considered record high at that time. For instance, the cage windows are open for small noble gases such as krypton, whereas they close around the xenon hence maximizing the interaction.[**Chen_2014**] Mohamed et al. also designed a material with a similar selectivity,CROFOUR-1-Ni. However the performance was now explained by the chemical nature of the chromium oxide ligands that interact more strongly with the more polarizable xenon than the krypton molecules.[**Mohamed_2016**] Finally, Banerjee et al. tested a previously synthesized[**KAXQIL**] MOF after it was identified through high-throughput screening[**Simon_2015**] for its outstanding theoretical selectivity around 70. However, experimental measurements showed that its selectivity was not exceeding the one of the previous top materials. Similar emphases were made on the ideal pore size coupled with highly attractive framework atoms.[**Banerjee_2016**]

More recently, Li et al. proposed a rigid squarate-based MOF with "perfect pore size" (comparable with the kinetic diameter of Xe), and an internal pore surface decorated with very polar hydroxyl groups. This material experimentally demonstrated record-high Xe/Kr selectivity of 60.6 at low pressure (0.2 bar) and ambient temperature.[**Li_2019**] Later, Pei et al. discovered even better performing materials with Xe/Kr selectivity of 74.1 and 103.4 in the same conditions 0.2 bar and 298 K. In addition to the perfectly tailored pore size, the structure features two oppositely adjacent open metal sites that strongly clamp the adsorbed xenon molecule.[**Pei_2022**] These studies clearly show the potential of polar sites that preferentially interact with the more polarizable xenon over the krypton, hence explaining these record-breaking separation performances.

1.3.3 From the computer to the test tube

To connect back our study to computational screenings, we are now going to present one of the rare cases of direct contribution of high-throughput screenings to the lab testing of a material. In 2015, Simon et al.[Simon_2015] analyzed the Nanoporous Materials Genome,[Simon_2015_EES, Boyd_2017] a database of about 670,000 experimental and hypothetical porous material structures, including MOFs, zeolites, PPNs, ZIFs, and COFs, for candidate adsorbents for xenon/krypton separations. This study led to the rediscovery of SBMOF-1, a promising nanoporous material that was presented one year later.[Banerjee_2016]

It is possibly the largest-scale study performed in this area, both by the sheer number of frameworks involved and by the diversity of their nature. Because such a set is too big for brute-force screening with GCMC simulations, they proposed a multiscale modeling strategy combining machine learning algorithms (trained on a diverse subset of 15,000 materials) with molecular simulations (used both to generate the ML training data, and to refine the separation properties for the top performers obtained by the ML predictor). Without going into details (see Fig. **??** for more details), the ML model they trained was mainly based on geometric structural descriptors, with the addition of a single energy-based descriptor: the Voronoi energy (i.e., the average energy of a xenon atom at the accessible nodes in the Voronoi partition of space). In addition to identifying and describing some top performing materials, the authors also analyzed the correlations between high Xe/Kr selectivity and the geometric properties of the frameworks, in order to "rationalize the strong link between pore size and selectivity". In particular, by developing theoretical pore models of spherical and cylindrical geometries, they could highlight the general geometrical trends observed, but also the fact that there is a wide diversity of performance beyond the geometrical features of the frameworks, which suggests the key role of the chemical nature of the cavities.

By looking at the distribution of the most selective materials ($s \geq 14$) compared to the less selective ones ($s < 14$) in the Figure **??**, Simon et al. established a first profile of the selective materials. These materials have pore sizes of a specific diameter very close to the kinetic diameter of xenon around 4.4 Å depending on how it is defined. They have rather low surface areas and porosities (void fractions) unlike what we would normally expect since the adsorbable surface is a key reason for using nanoporous materials in adsorption applications. This behavior can be rationalized by the fact that small pores of the order of a few Å drives mechanically to smaller pore volumes and surface areas (the framework atoms have more space). The crystal density is therefore also a bit higher for these reasons. Moreover, the pore's shape is also a crucial factor since a shape closer to a sphere would interact with the xenon with more atoms, hence increasing its affinity and the selectivity. Finally, last but not least, the Voronoi energy described the physical nature of the binding between the xenon and the pore atoms, the more negative it is and the more selective the material will be. To wrap up, the ideal materials have a pseudo-spherical shape (a complete sphere would stop the diffusion of the adsorbates) with a size close to the diameter of a xenon which is rather dense and not very porous.

The chemical nature of the cavities was best described using the Voronoi energy descriptor they developed. This descriptor gives an idea of the xenon adsorption isosteric heat of the material. Given these results, more studies should focus on describing the adsorption thermodynamic quantities such as the adsorption enthalpy but also the Henry adsorption constants. This study finally leads to the synthesis and testing of one of the top performing materials in the field.

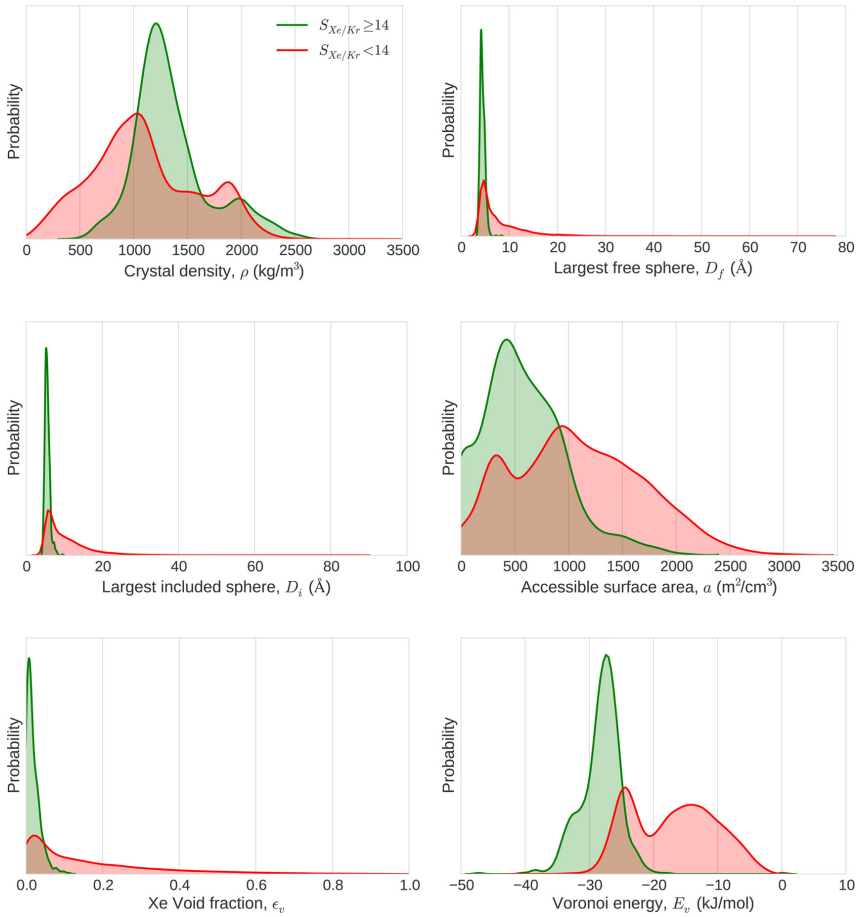


*Figure 1.10: Statistical analysis of the adsorption separation of xenon/krypton mixtures by nanoporous materials. The graphs represent the distributions of structural descriptors explored by highly selective (green) and poorly selective (red) materials separately. Reprinted with permission from Ref. [**Simon_2015**]. Copyright © 2015 American Chemical Society.*

However, we cannot stop but wondering why there is a discrepancy between the theoretical selectivity of around 70 of SBMOF-1 and its actual experimental selectivity of 16. In the final chapter of this thesis, we will try to give an explanation for this. In the future, such close collaboration between experimental and computational teams are crucial even if they are still too rare. A recent paper suggests that these collaborations are rare across all nanoporous fields and a lot of improvements are needed to foster cooperation between the labs.[**Li_2022**]

1.3.4 The future of screening

Despite the progress made, important drawbacks of the current methodologies remain. High-throughput screenings rely too much on oversimplified assumptions such as the rigidity of the framework, the absence of defects, the use of Lennard-Jones potentials and inaccurate charges. For instance, the rigidity of the framework only takes into account one conformation of the framework. Yet, thermal agitation induces a “breathing” movement of the framework with an amplitude dependent on its intrinsic flexibility. The pores of the framework can change depending on the number of adsorbates to interact more optimally with them, which can be induced by a change in pressure. The issue of flexibility is rarely tackled, and when considered, it is only on the few most selective structures given by an inaccurate screening based on the

rigid crystal approximation. One can wonder about the results obtained if it is applied to larger sets of structures. Witman et al. found that flexibility applied to top performing materials can decrease the selectivity, because the pore does not have an optimal size anymore.[Witman_2017] In some cases, the selectivity of a well-performing material can even increase to become a top performing one. Computational screenings can be closer to predict experimental values of selectivity, diffusivity, and other key performance metrics.

Many open problems remain for the design of efficient high-throughput computational screenings. The connection between different properties for a given application is not systematically integrated in the screening procedures. For example, in methane storage, the working capacity of the material is the main property to optimize, but the kinetics of the adsorption/desorption or the mechanical resistance to compaction among others also need to be considered. Designing a nanoporous material is in fact a multivariate optimization problem with tacit constraints, for example the synthesizability. For instance, the diffusion coefficients of adsorbates in a xenon/krypton separation problem can help us better understand the breakthrough simulations and eventually the whole separation process in pressure or temperature swing adsorption beds. For this reason, studying transport properties along with uptake capacities and thermodynamic selectivity of the xenon/krypton separation can give a more complete picture of the industrial process we ultimately want to model.

Moreover, the transferability of the methodology to a broad range of materials is often achieved at the expense of accuracy in specific cases. And one can rightly question the universality of depending on faster but less elaborated models, which boils down to a trade-off problem between prediction accuracy and computational cost (or complexity). For instance, classical forcefields are broadly used in rigid materials for adsorption properties, but the switch to more costly *ab initio* methods or the addition of flexibility can result in a more accurate description at the expense of computational resources. The addition of polarization could be very promising since several top performing materials harbor open metal sites and highly polar sites that explain the acute affinity to xenon adsorbates.

The development of ML-assisted screenings is paired with the advances in data science techniques and algorithms. Recent advances in deep learning have enabled the development of transformer-based (the technology at the foundation of ChatGPT) machine learning models to predict adsorption properties.[Kang_2023, Cao_2023] More importantly, the construction of descriptors tailored to the many possible applications is also an ongoing work. This construction work cannot be dissociated to the physical and chemical intuition of the scientists. Topological, chemical, electronic and other descriptors have been developed on top of the more common geometrical and thermodynamic descriptors, which displays the importance of strong physical chemistry knowledge. Recently Shi et al. highlighted the key role of energy histograms in predicting adsorption properties.[Shi_2023] The discovery of novel relevant descriptors remains the main lever for increased performance of the ML models and is closely related to a rigorous theoretical work. For these reasons, we worked during this thesis on more accurate and faster ways of calculating these interaction energies to extract valuable energy/thermodynamic descriptors.

The development of databases is another key aspect in the promotion of data science in the field of materials science in general, and nanoporous materials chemistry in particular. The

diversity of materials, the inclusion of experimental data (successful or failed), the addition of understudied classes of materials (e.g., amorphous) are all key aspects to upgrade the existing database. Even if existing attempts to create a centralized database have been initiated by the materials project,[MaterialsProject] this database does not contain all the existing information on each material. Furthermore, this high amount of data will need to be efficiently explored, and non-supervise deep learning algorithms have been developed to do so.[Park_2023] Coupled with synthesis robot, these methods can navigate through the unexplored databases to find the few most interesting candidates for a given targeted application.

In the future, computational high-throughput screening could be integrated more tightly into the design process of nanoporous materials, hence further improving its efficiency. The computational prescreening can be coupled with automated screenings of the most promising materials to finally identify candidates for further studies. This automated design process is described by Lyu et al. in their paper on "Digital Reticular Chemistry" and set out promising perspectives for computational screenings in the field.[Lyu_2020] Some studies are already pioneering this new research area by combining high-throughput characterizations, active learning algorithms and robotic synthesis.[Greenaway_2018, Moosavi_2019] Another step towards faster industrialization would integrate process modeling to enrich the purely atomistic approach.

2

THERMODYNAMIC EXPLORATION OF XENON/KRYPTON SEPARATION

2.1 Characterization of Adsorption Equilibrium Properties 33
2.1.1 Geometrical descriptors. 33
2.1.2 Intermolecular interaction energies 35
2.1.3 Mixture adsorption: Grand Canonical Monte Carlo. 37
2.1.4 Infinite dilution adsorption: Widom insertion 39
2.1.5 The thermodynamics behind adsorption-based separation . . 42
2.2 Preliminary Analyses of the Separation Performance 44
2.2.1 Structure–selectivity relationships 44
2.2.2 Thermodynamic quantities correlations at infinite dilution . 51
2.3 Selectivity Drop between Two Pressure Regimes 56
2.3.1 Thermodynamic origins 56
2.3.2 Detailed investigation 61
2.3.3 Toward the development of new screening tools 67

2.1 Characterization of Adsorption Equilibrium Properties

2.1.1 Geometrical descriptors

Before going into the details of the adsorption properties themselves, we will first introduce the different simulation techniques used to characterize the internal pore structure of a material key in interpreting the adsorption properties obtained using more complex molecular simulations. All the geometrical descriptors used in this thesis have been calculated using the Zeo++ software;**Zeo++** other tools exist,**First_2013**, **PoreBlazer** but the use of Voronoi decomposition of the volume speeds up the calculation making it the preferred tool for this task (efficiency gain mainly on volume calculation).**Rycroft_2009**

Pore size

There are a multitude of pore sizes depending on the point where we measure it, all these pore sizes compose what we call a pore size distribution. Some specific values are, however, uniquely defined and used to put a single value to characterize pore sized. The diameter of the largest sphere that can diffuse freely in the structure is called D_f. The diameter D_{if} corresponds to the diameter of the largest included sphere along a free diffusion path; the diameter of the largest included sphere (not necessarily in a free diffusion path) is denoted D_i. The Figure ?? illustrates the difference between these pore sizes. In thermodynamic studies we will often use the term "largest cavity diameter" (LCD) instead of the largest included sphere D_i. And, the term "pore limiting diameter" will be used instead of D_i especially when studying the transport effects with the nanopores.

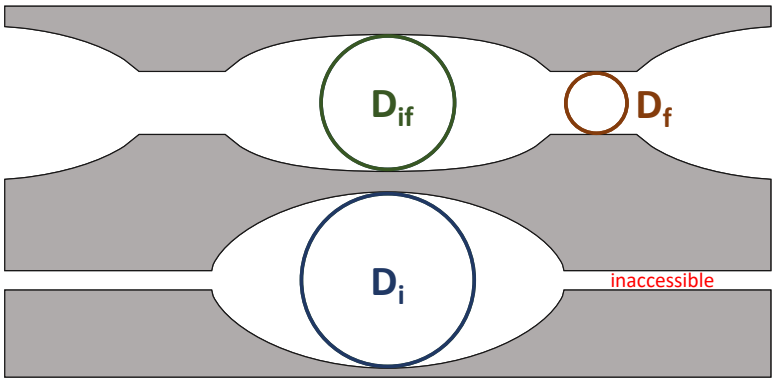


Figure 2.1: Illustration of the different pore sizes D_f, D_i and D_{if}. Note that in some materials D_{if} is equal to D_i, when the largest included sphere is also accessible through a free diffusion path.

To define these pore sizes, we need to first set the radii of the framework atoms that will shape the surrounding pores. These radii can be defined using different methods, the default mode uses the Cambridge Crystallographic Data Centre's (CCDC) radii. This method is most commonly used in the literature. We also introduced another set of radii based on the universal forcefield[rappe1992] we use for molecular simulations. The determination of these radii are inspired by an approach developed by Hung et al.[Hung_2021] The atomic radii are defined as the distance where the LJ potential reaches $3k_BT/2$, for T = 298 K. This type of definition can be more easily compared to the molecular simulations. For instance, we will use indexes to tell apart both methods. For example, LCD_{CCDC} corresponds to the standard definition of the LCD that uses the CCDC radii to run the Zeo++ software, while the LCD_{UFF} is associated with the definition of the atomic radii dependent on the UFF forcefield. In this chapter, I will mainly use the forcefield-based definition — the largest cavity diameter will be noted LCD_{UFF} and the studies on void fraction and surface areas are also defined using this set of radii.

Surface area

The surface areas are calculated using a random sampling over the surface of the different atom surfaces. The algorithm counts only the points that do not overlap with another atom. For each atom we can therefore calculate an adsorbable surface. By summing up all the surfaces, we finally have the surface area. This "rolling ball" algorithm has been developed since 1973 by Shrake and Rupley.[Shrake1973] the By defining a probe, the Voronoi tessellation defines the accessible and the non-accessible areas of the structure. Depending on where the surfaces are,

they are either counted in the accessible or the non-accessible surface areas. In this chapter, we will use the accessible surface area defined by a probe of 1.2 Å; this is a computational equivalent of the experimental N_2 BET surface area.

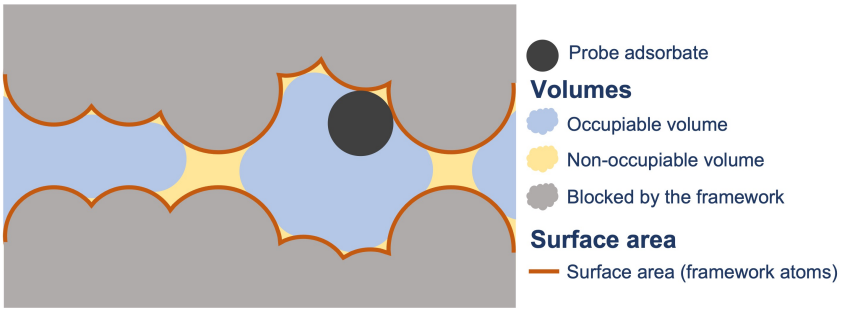


Figure 2.2: Illustration of the pore surface area and volume in a nanoporous material. As illustrated, there are different definitions of the pore volume: we can either consider the whole volume of the pores or the volume occupiable by a given probe. The surface area also changes depending on the definition. Studies have shown that occupiable volume has a better accordance with experimental data.[vol_Ongari2017]

Pore volume and porosity

The pore volume is calculated by random sampling of the accessible and inaccessible Voronoi cells. Similar algorithm, random sampling over a regular mesh. If the probe does not overlap with a framework atom, then it is counted in N. The ratio of this final N and the total number of points sampled gives the fraction of the pore volume. This ratio is called the void fraction or the porosity. Using the Voronoi decomposition, we can also define the accessible and non-accessible Voronoi cells to reduce the space the Monte Carlo simulation need to sample for the surface area and the void fraction calculations.

2.1.2 Intermolecular interaction energies

In most of the studies in this thesis, we will consider rigid structures interacting with guest adsorbates; the intramolecular interactions will not play any role in the simulations since the ionic, chemical or metallic bonds between the atoms of a same molecule are predefined at a given set of distances and remain the same. As discussed in the final chapter, this approximation can generate discrepancies between the theoretical model and the experiments. However, since the goal is to achieve screening approaches like the ones introduced in the first chapter, adding flexibility in the intramolecular interactions would reduce considerably the size of the database that could be screened. For these reasons, the term "interaction energy" will designate the guest–host and guest–guest intermolecular interactions mainly — host–host interactions would compromise the assumption on the rigidity of the framework.

In classical theory of molecular physics, the intermolecular interactions can be categorized in three different types according to their strength: (i) the ion–dipole and ion–induced dipole forces (40–600 kJ mol^{-1}), (ii) the hydrogen bonding (10–50 kJ mol^{-1}), and (iii) the van der Waals forces (1–10 kJ mol^{-1}). Note that these energy values are only indicative and the interaction depends on the nature of the molecules, but it allows us to rank the different foreces according to their strength; and to complete the molecular interactions picture, we can add that the ionic and covalent bonding will always be stronger than any intermolecular interactions (over 100 kJ mol^{-1}). The generic term "van der Waals interactions" actually regroups three

different concepts usually called Keesom, Debye and London interactions. The Keesom interaction focuses on the electrostatic interaction between permanent multipoles (representing the electronic density around the molecules),[keesom1915second] while the Debye induction force corresponds to the interaction between a multipole of a molecule and an induced multipole of another one,[Roberts_1938] and the London dispersion interaction occurs between instantaneous multipoles created by natural fluctuations in the electron density around polarizable atoms.[london1930theorie, polanyi1932section] To quantify these interactions, we can consider dipole interactions since they are the most influential in the multipole expansion of the electron density. The Keesom interaction potential U_K can therefore be reduced to the dipole–dipole interaction, which depends on the inverse third power of the distance for fixed dipoles; but in fluid phases we are interested in, the average over all the angles is rather given by the inverse sixth power as described in the equation **??** below:

$$U_K = -\frac{\mu_1^2 \mu_2^2}{(4\pi\epsilon_0\epsilon_r)^2 r^6} \times \frac{1}{1.5 k_B T} \tag{2.1}$$

where μ_1 and μ_2 are the dipole moments of the molecules 1 and 2, ϵ_0 the vacuum dielectric permittivity and ϵ_r relative permittivity of the surrounding material, k_B the Boltzmann constant, T the temperature and r the intermolecular distance. The Debye interaction potential U_D being reduced to the permanent dipole–induced dipole interactions can now be expressed using the electric polarizability α_1 and α_2 of the molecule 2 as shown in equation **??**.

$$U_D = -\frac{\mu_1^2 \alpha_2 + \mu_2^2 \alpha_1}{(4\pi\epsilon_0\epsilon_r)^2 r^6} \times \frac{1}{k_B T} \tag{2.2}$$

Finally, the London dispersion interaction potential U_L is now the fluctuating dipole–induced dipole interaction and can be expressed as follows:

$$U_L = -\frac{\alpha_1 \alpha_2}{(4\pi\epsilon_0\epsilon_r)^2 r^6} \times 1.5 \frac{I_1 I_2}{I_1 + I_2} \tag{2.3}$$

where I_1 and I_2 are the first ionization energies. We can note that the van der Waals potentials are all negative (attractive interaction) and depend on the inverse sixth power of the distance — considering only the dipole moments. Before moving to the computational modelization of these long-distance intermolecular forces, we need to specify the repulsive force that occurs at very short distances; this force can be explained by the impossibility for electrons of both atoms to occupy the same quantum space as stated by Pauli in his exclusion principle.

For the system we are studying in this thesis, the adsorption of noble gases in nanoporous materials, the guest–guest and guest–host interactions can be described by the induction and dispersion interactions only. We will use a simplistic model, the Lennard-Jones (LJ) potential U^{LJ},[LJ_1924] that relies on a repulsive term for the Pauli exclusion principle and an attractive term to model the attractive van der Waals component of the interaction, as shown below:

$$U^{LJ} = 4\epsilon \left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right) \tag{2.4}$$

where ϵ is the depth of the well (minimal energy) and σ is the distance at which the potential is zero. The forcefield defines the LJ parameters ϵ and σ for either all atom pairs or only for

the same type of atoms. For the commonly used universal forcefield (UFF),[rappe1992] only the parameters for atoms of the same nature are defined and the parameters of the pair atoms can be induced using combining rules. In this thesis, we will use the UFF forcefield (because it performs well compared with *ab initio* forcefield[McDaniel_2015]) and the Lorentz-Berthelot mixing rules to combine the LJ parameters — it makes an arithmetic average of the σ values (Lorentz rule) and a geometric one of the ϵ values (Berthelot rule). Finally, to reduce the computation time, we usually set a cutoff distance at which the LJ potential can be considered negligible. At this cutoff distance, we can for example apply a shift so that the energy equals zero at the cutoffs (discontinuity of energy), just truncate (discontinuity of the force), or use a tail switching function to make the tail converge smoothly to zero near the cutoff. To make it simple, in most of the simulations in the screenings, a shifting strategy combined with a cutoff of 12 Å have been used.

Usually in adsorption simulations of other gas molecules containing partial charges, we need to calculate the Coulomb interaction between the partial charges of the host framework and those of the adsorbate — in periodic systems, the Ewald summation can be used. For noble gases, these ion–dipole and dipole–dipole interactions do not exist due to the perfect neutrality of the molecules. However, we can argue that the ion–induced dipole is not taken into account in a simplified LJ potential. To complete the whole picture of the intermolecular interactions, we would need to study the energy inducted by the surrounding framework atoms' charges on the adsorbate. Several approaches have been developed in the literature to improve the description of the intermolecular interactions by coupling the LJ potentials with an induction potential.[Lachet_1998, Becker_2017]

To wrap up this small section on the modelization of the intermolecular interactions in the adsorption simulations, I want to emphasize once more on the main modelization assumptions that could alter the accuracy of our method. First, the framework remains rigid during the whole simulation, which avoids the need for a molecular dynamics simulation of the framework to save time but also hides the effects of a known phenomenon.[Witman_2017] Second, the polarizability of the adsorbate is not fully taken into account since the interaction with the charges of the framework are not considered; the difference in polarizability between xenon and krypton can be further exploited to enhence the selectivity even more, experimental studies suggest the key role of polar groups and open metal sites.[Li_2019, Pei_2022, Perry_2014] And finally, the complex induction and dispersion interactions are described with a two-parameter model, which cannot capture all the subtlety of the differences between a same atom in different environments for instance; it could be possible to fine-tune these parameters in very specific cases, but the overall good performance[McDaniel_2015] of the UFF forcefield needed in a screening process can induce small errors when looking at specific cases. These assumptions have been made to find the right trade-off of the speed of computation versus a detailed description of the physical phenomena at stake.

2.1.3 Mixture adsorption: Grand Canonical Monte Carlo

As explained before, we can think of adsorption as a gas–solid or liquid–solid interfacial phenomenon, the adsorbate phase fills the accessible pore volumes depending on the physical conditions the material is put under. No simple model can predict how the adsorbates would interact with the pore surface, how many of them can fit in, what configuration is the most

stable, etc. To answer these questions, we can evaluate all possible adsorption configurations that would undeniably have different numbers of adsorbate, and only keep the most thermodynamically plausible ones. To do so, these configurations will have to follow a given probability distribution, the grand canonical ensemble probability for instance, because it allows the variation of the number of molecules (adsorbate molecules) and the total energy. With the help of a Monte Carlo simulation, we can vary the energy and the loading inside the pores so that the distribution of configurations c follows this probability law:

$$\mathrm{P}_c = \frac{1}{\Xi} e^{-\beta(\mathrm{E}_c - \mu \mathrm{N}_c)} \tag{2.5}$$

where E_c and N_c are respectively the energy and the number of adsorbate particles in the configuration c. Normally the energy and the number of molecules of all particles should be considered, but for now, since the whole system is considered rigid we will only focus on the adsorbate molecules. The chemical potential μ and the temperature T inside β correspond to the ones of the gas phase in equilibrium with the adsorbent material. And the pressure and volume V are considered fixed under the rigidity assumption. The grand canonical partition function $\Xi(\mu, \mathrm{V}, \mathrm{T})$ will then be the following sum over all possible configurations:

$$\Xi(\mu, \mathrm{V}, \mathrm{T}) = \sum_c e^{-\beta(\mathrm{E}_c - \mu \mathrm{N}_c)} \tag{2.6}$$

This multiplicative constant does not need to be known in the Monte Carlo simulation we will describe now.

Beyond these theoretical considerations, the grand canonical Monte Carlo simulation, referring to a Metroplis-Hastings Monte Carlo algorithm in the context of the grand canonical thermodynamic ensemble, will need several key characteristics in order to fulfill the previous claims on the probability distribution of the configurations. Monte Carlo (MC) refers to the randomness inherent to the gambling games of the eponymous casino on the azure coast of Monaco. The MC simulations are therefore relying on randomly generating atomic configurations; however in order to do it efficiently, we need to stay as much as possible in the physically possible atomic space, while exhaustively exploring all the chemical configurations. In molecular simulations, to do so, only the initial configuration c_0 is really randomly generated, but then the algorithm has different rational moves to change the configuration with a controlled amount of randomness. The second key feature (acceptance or rejection condition) was introduced by Metropolis and co-workers that allows to reproduce any distribution with an unknown multiplicative prefactor.**Metropolis1949** The configuration c_1 resulting of the random move is evaluated by calculating the transition probability (like in a Markov chain) or acceptance rate $acc(c_0 \to c_1)$:

$$acc(c_0 \to c_1) = \min\left(1, e^{-\beta\left(\mathrm{E}_{c_1} - \mathrm{E}_{c_0} - \mu\left(\mathrm{N}_{c_1} - \mathrm{N}_{c_0}\right)\right)}\right) \tag{2.7}$$

Any move that has a greater probability of occurring is always accepted, if the probability is lower than the acceptance rate depends on the probability ratio. The multiplicative prefactor has no influence on the algorithm, we do not need to know the chemical space to explore beforehand, which is a valuable simplification. This sequence of a Markov-type chain can then be used to approximate the probability distribution of the grand canonical ensemble we seek to describe in the equation **??**.

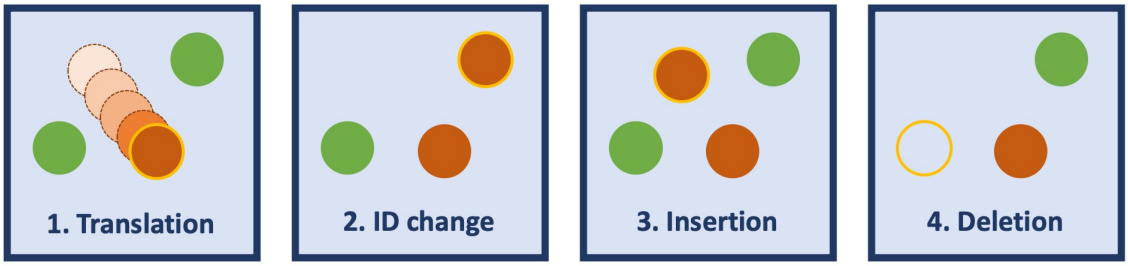


Figure 2.3: MC moves in a system of two types of monoatomic atoms (green and orange). The modification on the first box is highlighted by the yellow circle and the dragging pattern is represented by a set of dashed circles. The boxes 2 to 4 represent the moves going from the initial state represented in box 1, the corresponding move is highlighted by a yellow outer circle.

To complete the description of the grand canonical Monte Carlo (GCMC) simulation we are interested in, let us now consider the different MC moves used to generate a configuration from another. Depending on the parameterization, these moves have different probabilities of occurring. For monoatomic atoms only four moves are relevant (see Figure **??**): (i) the translation of a randomly chosen molecule with a displacement randomly chosen within a given radius, (ii) the change of identity of a randomly chosen molecule into another one, (iii) the insertion of an adsorbate molecule and (iv) the deletion of an adsorbate molecule. We deliberately omitted the rotations of the adsorbate because of the spherical symmetry of noble gases and the change of volume since the flexibility of the material framework is neglected.

By using a GCMC algorithm, we can now generate a set of configurations according to their probability of occurrence. Because the probability law is directly taken from equation **??**, the series of configurations describe the thermodynamic equilibrium state of a nanoporous material in contact with a reservoir of a xenon-krypton mixture at a given composition, pressure and temperature. Different thermodynamic quantities can be derived from ensemble averaging: the averaging loading or uptake at a given pressure (several pressures give the isotherm) and the isosteric heat of adsorption of each adsorbate (Xe and Kr). The ratio of the uptakes q informs on the selectivity s of the thermodynamic separation process:

$$\boxed{s = \frac{q^{\mathrm{Xe}}}{q^{\mathrm{Kr}}} \times \frac{y^{\mathrm{Kr}}}{y^{\mathrm{Xe}}}} \tag{2.8}$$

where y^{Xe} and y^{Kr} designate respectively the mole fractions of Xe and Kr in the gas phase reservoir.

To characterize a separation process, we theoretically only need to perform a GCMC calculation at every pressure conditions that we are interested in. However, this type of simulation requires a lot of time to converge since we need to test out a lot of insertion/deletion moves to accurately estimate the number of adsorbed molecules and the composition of the mixture. Hence, faster methods (machine learning) are developed to estimate the selectivity at any pressure conditions.[**Simon_2015**, **Kang_2023**] If we are interested in the infinite dilution case, faster methods are already available, we are now going to introduce the Widom insertion

that can estimate the adsorption performances at infinite dilution by estimating the Henry adsorption constant.

2.1.4 Infinite dilution adsorption: Widom insertion

In 1963, the professor B. Widom introduced a simple method of calculation of thermodynamic properties in a material or fluid mixture.[Widom1963] Generally, this method allows accessing to the difference of internal energy before and after the insertion of a Widom test particle while fixing all other particles, therefore comparing the N-particle and (N+1)-particle states. This difference of energy $\Delta\Phi$ can then be used to deduce the excess free energy associated to it $\Delta F_{exc} = -k_B T \ln\left(\langle \exp(-\beta\Delta\Phi)\rangle\right)$, which corresponds to the excess chemical potential induced by the addition of a particle. In the domain of fluid phase equilibrium, Widom insertion is the most straightforward method to calculate a chemical potential value; however it has drawbacks in liquid-like phases because the insertable space is very narrow, and no relaxation is implemented to account for the reorganization of the surrounding particles.[Nezbeda_1991] In our case, I will only focus on the insertion from 0 to 1 particle, where no problems of overlapping between adsorbate particles can happen. Widom insertion is in our case only a random insertion of an adsorbate into an empty nanoporous framework. By randomly sampling the void space, we obtain a distribution of interaction energies $\mathcal{E}_{int}$, the average of the Boltzmann weights associated is directly proportional to the adsorption free energy ΔG_{ads} and the Henry adsorption constant K_H. By taking the Boltzmann average of the interaction energies, we can also compute the adsorption enthalpy ΔH_{ads}. All these quantities stay only valid at infinite dilution, for higher quantities of adsorbates the previously described GCMC technique should be used.

In the infinite dilution case, this test particle insertion technique is similar to a random sampling of the adsorbable space inside a material. If the sampling is thorough enough, we can derive the following definitions of ΔG_{ads} (equation **??**), K_H (equation **??**) and ΔH_{ads} (equation **??**) based on a complete sampling of the interaction energies $\mathcal{E}_{int}$.

The adsorption Gibbs free energy ΔG_{ads} is equal to the excess free energy previously calculated in a Widom insertion since the structure is rigid and PV does not fluctuate (G = F + PV).

$$\boxed{\Delta G_{ads} = -RT \ln\left(\langle \exp(-\mathcal{E}_{int}/RT)\rangle\right)} \tag{2.9}$$

To derive the Henry constant K_H, we need to consider an ideal gas. The number of adsorbed molecules n_{ads} can be expressed using the bulk density of the gas $\rho_{ads,bulk}$ and the volume of the pores V_{pore}:

$$n_{ads} = \rho_{ads,bulk} \times V_{pore} \tag{2.10}$$

The pore volume can be seen as the continuous sum of each voxel times the Boltzmann probability of presence, which is represented by the following integral of the Boltzmann factors. This integral can then be changed to the average of the Boltzmann factors:

$$V_{pore} = \int_V \exp\left(-\mathcal{E}_{int}(\mathbf{r})/RT\right) d\mathbf{r} = V\langle \exp\left(-\mathcal{E}_{int}/RT\right)\rangle \tag{2.11}$$

Let us apply the equation **??** and the perfect gas equation of state $P = \rho_{ads,bulk} RT$ on the bulk gas in equilibrium, we can change the equation **??** to:

$$\frac{n_{ads}}{V} = \frac{P}{RT}\langle \exp\left(-\mathcal{E}_{int}/RT\right)\rangle \tag{2.12}$$

If we now consider the gravimetric loading L_{ads} (in mmol g^{-1}), we need to divide the equation by mass density of the framework ρ_f:

$$L_{ads} = \frac{n_{ads}}{V\rho_f} = \frac{\langle \exp\left(-\mathcal{E}_{int}/RT\right)\rangle}{\rho_f RT} P \tag{2.13}$$

Since the Henry's law is described by $L_{ads} = K_H \times P$, we have the final relation between the Henry adsorption constant and interaction energy distribution.

$$\boxed{K_H = \frac{\langle \exp\left(-\mathcal{E}_{int}/RT\right)\rangle}{\rho_f RT} = \frac{1}{\rho_f RT} \exp\left(-\frac{\Delta G_{ads}}{RT}\right)} \tag{2.14}$$

Note that the ρ_f factor comes from the use of a gravimetric loading expressed in mmol g^{-1} and is not always present in the different derivations of the literature.**PoreBlazer** The RT factor comes from the perfect gas assumption we made, which is a good approximation in the noble gas case.

Finally, if we consider an adsorption equilibrium (e.g., $Xe_{(g)} \rightleftharpoons Xe_{(ads)}$), we can define an equilibrium constant $K_{ads} = x_{ads}/y_{gas}$ where x_{ads} is the mole fraction in the adsorbed phase and y_{gas} the mole fraction in the gas phase for a given compound (e.g., Xe). For a pure gas ($y_{gas} = 1$) at infinite dilution, we can apply the Henry's law to derive the following relation:

$$K_{ads} = \frac{n_{ads}}{n_{site} y_{gas}} = \frac{K_H P \rho_f V}{n_{site}} = PV \frac{\langle \exp\left(-\mathcal{E}_{int}/RT\right)\rangle}{n_{site} RT} \tag{2.15}$$

where n_{site} is the number of sites considered constant since it is much higher than n_{ads} at infinite dilution.

Now by applying the Van't Hoff equation to this infinite-dilution adsorption equilibrium constant K_{ads}, we can derive an expression of the adsorption enthalpy at infinite dilution:

$$\Delta H_{ads} = -R \frac{d \ln\left(K_{ads}(T)\right)}{d(1/T)} \tag{2.16}$$

Then by decomposing the logarithm on the fraction of equation **??**,

$$\Delta H_{ads} = -\frac{d \ln\left(PV/n_{site}\right)}{d(1/T)} - R \frac{d \ln\left(\langle \exp\left(-\mathcal{E}_{int}/RT\right)\rangle\right)}{d(1/T)} - R \frac{d \ln\left(1/T\right)}{d(1/T)} \tag{2.17}$$

Then, PV/n_{site} being a constant, we can reduce the expression to two derivatives, the first one being the logarithmic derivative of itself (1/T) and the second being the logarithmic derivative of the sum of the exponential terms.

$$\Delta H_{ads} = 0 - R\frac{d\ln\left(\left\langle \exp\left(-\mathcal{E}_{int}/RT\right)\right\rangle\right)}{d(1/T)} - RT \tag{2.18}$$

Knowing that for any function f the logarithmic derivative equals the quotient of its derivative f', $\frac{d\ln(f)}{dx} = f'/f$, we can calculate the derivative of the average of the Boltzmann factors $\langle \exp(-\mathcal{E}_{int}/RT)\rangle$:

$$\Delta H_{ads} = -R\frac{1}{\frac{1}{N}\sum e^{-\frac{\mathcal{E}_{int}}{RT}}}\frac{1}{N}\sum\frac{d\exp\left(-\mathcal{E}_{int}/RT\right)}{d(1/T)} - RT \tag{2.19}$$

where N corresponds to the number of points where the Widom particle has been inserted. The exponential derivative makes the energy factors come out, and we get the following expression:

$$\Delta H_{ads} = -R\frac{1}{\sum e^{-\frac{\mathcal{E}_{int}}{RT}}}\sum -\frac{\mathcal{E}_{int}}{R}e^{-\frac{\mathcal{E}_{int}}{RT}} - RT \tag{2.20}$$

With some simplification, we can express the adsorption enthalpy ΔH_{ads} as a Boltzmann average of the interaction energies minus a term RT that comes from the ideal gas assumption (perfect gas equation of state).

$$\boxed{\Delta H_{ads} = \frac{\sum \mathcal{E}_{int}e^{-\frac{\mathcal{E}_{int}}{RT}}}{\sum e^{-\frac{\mathcal{E}_{int}}{RT}}} - RT} \tag{2.21}$$

From the values of the adsorption free energy and enthalpy, we can now deduce the adsorption entropy ΔS_{ads} using the definition of the Gibbs free energy (G = H – TS):

$$\boxed{\Delta S_{ads} = \frac{1}{T}\left(\Delta H_{ads} - \Delta G_{ads}\right)} \tag{2.22}$$

We already defined the selectivity as the ratio of the proportion of Xe/Kr in the adsorption phase to the proportion in the gas phase in the equation **??**. At infinite dilution, we can rewrite the selectivity using the Henry's law ($q^i = V\rho_f K_H^i y^i P/n_{tot}$) and simplifying the constant term $PV\rho_f/n_{tot}$:

$$\boxed{s = \frac{K_H^{Xe}y^{Xe}}{K_H^{Kr}y^{Kr}} \times \frac{y^{Kr}}{y^{Xe}} = \frac{K_H^{Xe}}{K_H^{Kr}}} \tag{2.23}$$

By extrapolating at the zero loading regime, the Xe/Kr selectivity can be simply expressed as the ratio of the Henry adsorption constant of xenon and krypton.

In this section, we saw that simple thermodynamic quantities such as the adsorption Gibbs free energy, enthalpy and entropy can be derived from the study of a simple adsorption equilibrium equation. In the next one, we will explore a thermodynamic characterization of the adsorption-based separation process using another equilibrium.

2.1.5 The thermodynamics behind adsorption-based separation

Now that the main simulation tools used to describe the competing adsorption of Xe/Kr binary mixtures are introduced, let us rationalize the separation process by modeling the process within a theoretical "exchange" equilibrium that corresponds to the exchange of gas phase Xe and Kr on a model adsorption site representing all the most attractive sites for a given pressure condition:

$$Xe_{(g)} + Kr_{(ads)} \rightleftharpoons Xe_{(ads)} + Kr_{(g)} \tag{2.24}$$

The equilibrium constant associated to the Equation (**??**) at any pressure for a given composition is simply the selectivity s, defined in the equation**??**, because the gas phase activities of $Xe_{(g)}$ and $Kr_{(g)}$ correspond the partial pressures y^{Xe} and y^{Kr}, and the adsorption phase activities of $Xe_{(ads)}$ and $Kr_{(ads)}$ correspond the mole fractions q^{Xe} and q^{Kr}. The Gibbs free energy at equilibrium can be directly defined using the equilibrium constant, by applying this relation to the exchange equilibrium we can define an exchange Gibbs free energy $\Delta_{exc}G$:

$$\boxed{\Delta_{exc}G = -RT\ln(s)} \tag{2.25}$$

This exchange equilibrium can be seen as the substraction between the adsorption equilibria of xenon and krypton. So by applying the Hess's law of constant heat summation, we can derive an expression of the exchange enthalpy as the difference of the adsorption enthalpies between xenon and krypton within the mixture.

$$\boxed{\Delta_{exc}H^{Xe/Kr} = \Delta_{ads}H^{Xe} - \Delta_{ads}H^{Kr}} \tag{2.26}$$

Moreover, the adsorption enthalpies $\Delta_{ads}H$ can be obtained in a GCMC calculation using a formula derived from the fluctuation theorem in statistical mechanics (see a derivation in this online article**github_simon_gcmc**):

$$\Delta_{ads}H^{Xe} = \frac{\langle EN\rangle - \langle E\rangle\langle N\rangle}{\langle N^2\rangle - \langle N\rangle^2} - RT \tag{2.27}$$

where E corresponds to the energy of the adsorbates and N the total number of adsorbates at every step of the simulation. Note that this equation remains only valid for $N \gg 1$, because the first step of the derivation is based on a first order Taylor expansion $\langle E\rangle(\langle N\rangle + 1) - \langle E\rangle(\langle N\rangle) = \frac{\partial\langle E\rangle}{\partial\langle N\rangle}$. This expression of the adsorption enthalpy echoes with the one at infinite dilution (equation **??**), where for $N \to 0$ we now have $\Delta H_{ads} = \langle E\rangle(1) - \langle E\rangle(0) - RT = \langle E\rangle(1) - RT$.

Now that we defined the exchange Gibbs free energy and an exchange enthalpy at any pressure, we can now use the same approach as for the equation **??** to derive the exchange entropy:

$$\boxed{\Delta_{exc}S = \frac{1}{T}(\Delta_{exc}H - \Delta_{exc}G) = \frac{1}{T}\Delta_{exc}H + R\ln(s)} \tag{2.28}$$

Before concluding this methodological section, we need to note that the thermodynamic quantities associated to this newly defined adsorption exchange equilibrium can be defined at different pressures and different methodologies can be used to calculate them. At infinite dilution, we would preferably use Widom insertions and the adsorption free energies and enthalpies to deduce these exchange quantities; at higher pressure, we would use the GCMC

calculation to define a free energy (via the loading values) and the isosteric adsorption heat to define them. In the following study, we will focus on only two pressures: the ambient pressure (at 1 atm) and the limit of zero loading (infinite dilution). At 1 atm, the previously defined quantities will have an index 1 to differentiate them from the infinite dilution case where an index 0 will be used; for example, $\Delta_{ads}H_1^{Xe/Kr}$, $\Delta_{exc}G_1^{Xe/Kr}$ or $s_1^{Xe/Kr}$ at 1 atm, and $\Delta_{ads}H_0^{Xe/Kr}$, $\Delta_{exc}G_0^{Xe/Kr}$ or $s_0^{Xe/Kr}$ at the low-pressure limit. One final note on the simulation details, to run the GCMC calculations and the Widom insertion, we used the RASPA software developed by Dubbeldam et al.**dubbeldam2016** And, the intermolecular van der Waals interactions were 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),**rappe1992** and the guest atoms (xenon and krypton) have their LJ parameters taken from a previous screening study.**Ryan_2010** All the MOFs described here are taken from the CoRE MOF 2019 database.**Chung_2019**

2.2 Preliminary Analyses of the Separation Performance

As we have seen above in the existing literature, the computational screening of the nanoporous materials — both existing frameworks and hypothetical structures — for targeted adsorption properties has been the object of many studies, and several of those high-throughput screening studies have focused on noble gas separation, and Xe/Kr separation, in particular. For large-scale studies we have found that, in addition to the testing and validation of methodological developments, the screening aimed in most cases at one of three objectives: (i) to identify top performing materials for synthesis and/or characterization; (ii) to better understand the limits of possible performance, and the relationships and trade-offs between various metrics of performance (selectivity, uptake, etc); (iii) identify structure–property relationships, correlating separation performance with structural properties of the materials that can be more easily determined (i.e., at low computational cost). In this initial screening study of the thermodynamic quantities, we performed a screening of around 9,700 tridimensional structures of a preprocessed version of the CoRE MOF 2019-ASR (all solvent removed) database that are publicly available — only the non-disordered structures and the structures with a cell volume smaller than 20 nm^3 (to limit the overall calculation time) were considered. We will focus on the different relationships of the Xe/Kr selectivity has with structural descriptors based on geometrical analyses, and then with different thermodynamic descriptors (free energy, enthalpy, entropy).

2.2.1 Structure–selectivity relationships

An adsorption separation process is primarily characterized by a pivotal performance metric called the selectivity we defined in equations **??** and **??**. By comparing this selectivity to geometrical descriptors calculated by the Zeo++ software,**Zeo++** we want to characterize the materials that will most likely be selective for a separation of a 20:80 Xe/Kr mixture (to compare with most literature screenings on a mixture extracted from the air). Three structural descriptors have been calculated: the accessible surface area of a N_2-sized probe of 1.2 Å, the void fraction occupiable by a 2.0 Å radius probe (roughly the size of a xenon),**vol_Ongari2017** and the diameter of the largest included sphere (D_i) using specially designed atom radii. Inspired by a recent work on the comparison of pore limiting diameters and self-diffusion coefficients,**Hung_2021** we defined a list of van der Waals radii to be read by the Zeo++ software

(more details in https://github.com/eren125/zeopp_radtable). All Zeo++ calculations use an atomic radius that corresponds to the distance where the LJ potential reaches $3k_BT/2$, for T = 298 K.

Xenon uptake and selectivity

Before digging deeper in the structure–selectivity relationship, we will look at the relation between the xenon uptake (the number of adsorbed xenon in the GCMC simulation) and the selectivity at 1 atm. For instance, the xenon uptake could also be very important in a separation process, because it defines the working capacity of xenon produced by adsorption/desorption cycles. According to the Figures **??** and **??**, first high xenon uptakes are associated with a high selectivity, but at some point very high selectivity is associated to lower uptakes. There seem to be an area where materials have selectivity over 100 and Xe uptake around 3 mmol g^{-1}, whereas an uptake over 6 mmol g^{-1} can only be obtained for a selectivity between 10 and 20. One cannot maximize both uptake and selectivity metrics at the same time, a trade-off needs to be made when designing nanoporous materials for xenon/krypton separation.**Zhang_2022** Different strategies have been implemented to optimize both metrics using mixed metrics such as the adsorbent performance score (APS).**Solanki_2020** This trade-off can be rationalized by using the different structural descriptors (pore size, surface area and volume) we presented earlier.

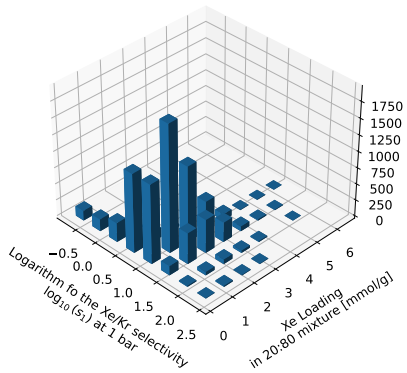


Figure 2.4: 3D histograms of in a bidimensional space formed by the Xe/Kr selectivity and the xenon uptake. The z-axis represents the number of structures with characteristics close to the one specified in x and y-axis. A base-10 logarithm has been applied to the selectivity values.

Furthermore, even if we know that it is possible to optimize either the xenon uptake or the Xe/Kr selectivity, these very successful materials are very rare inside a given diverse dataset. In the histogram presented in Figure **??**, the number of very selective materials is very low, same for the high-capacity materials. The most frequent materials have a selectivity between 1 and 10 and an uptake below 3 mmol g^{-1}. These values can be considered the standard values of nanoporous material for Xe/Kr separation, which sets reference values to compare the various performance metrics and build a chemical intuition. A selectivity above 20 is therefore considered rather high (even if the top materials have a much higher selectivity**Pei_2022**) and a xenon uptake above 4 mmol g^{-1} is also a pretty good adsorption capacity. The rarity of these top-performing materials gives the impression of searching a needle in a haystack, which has prompted some computational studies to design their algorithm to focus on finding the best materials rather than to describe equally all materials.**Deshwal_2021**, **Glasby_2021**

Surface area and selectivity

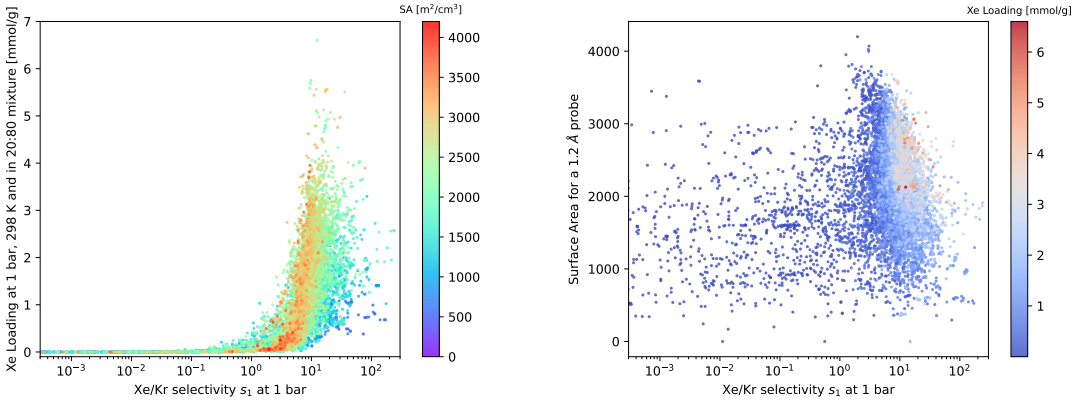


Figure 2.5: On the left: scatterplot of the xenon uptake as a function of the selectivity and labeled by the values of the surface area. On the right: scatterplot of the selectivity and the surface area labeled by the quantity of xenon adsorbed. The selectivity and uptake are calculated by a GCMC simulation of a 20:80 Xe/Kr mixture.

As demonstrated by other studies on methane storage applications by Wilmer et al.[**Wilmer_2012**] and then later by Fernandez et al.,[**Fernandez_2013**] the methane uptake is maximal for a specific optimal range of surface area values (2500-3000 m^2 cm^3). Higher values of surface area will not yield to higher values of methane uptake. This limitation also occurs for the selectivity as we can see in the right plot of the Figure **??**. Materials with a selectivity around 5 will have any surface areas from 0 to 4000 m^2 cm^3, whereas the ones with a selectivity above 40 will have a surface area below 2500 m^2 cm^3. The optimal surface area for xenon uptake would on the other hand be between 2000 and 3000 m^2 cm^3. The relationship between selectivity and surface area is quite complex, we cannot clearly state a precise range of surface areas that guarantees a high selectivity. This structural descriptor cannot characterize the selectivity, it needs to be coupled with other descriptors.

Looking at the 3D histogram on the Figure **??**, we can see the breakdown of the surface area ranges for different categories of selectivity. For selectivity values higher than 92, the surface areas are most likely to be under 2000 m^2 cm^3; between 92 and 35, there is a slightly wider range that goes to 2500 m^2 cm^3; between 35 and 13, the interval goes even further to 3500 m^2 cm^3 but is mostly centered between 1000 and 2500 m^2 cm^3. This split view of the distributions gives a better grasp on what the best materials look like; however the surface area will never be a deterministic variable — we will never be available to deduce the selectivity by simply looking at the surface area, because a surface area between, for instance, 500 and 1000 m^2 cm^3 have a relatively good chance of being selective but it concerns a lot of materials and it even has a higher chance of having a selectivity between 5 and 35 than a selectivity higher than that.

Void fraction and selectivity

A similar analysis of the relationship with the void fraction was also carried out by Wilmer et al. (Figure 5 of Ref. [**Wilmer_2012**]) and an optimal value of the void fraction was found at around 0.8. As we can see on the plots of the Figure **??**, the materials with the highest value of Xe uptake have void fraction values around 0.5, whereas the ones with the highest value of

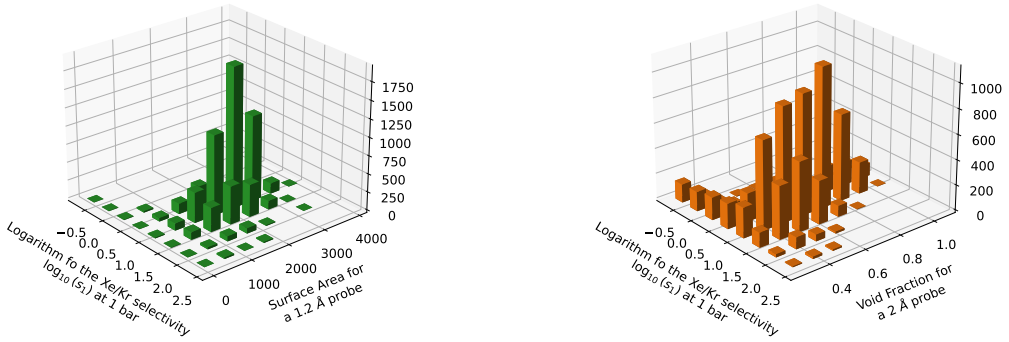


Figure 2.6: 3D histograms of in a bidimensional space formed by the Xe/Kr selectivity and the surface areas (on the left) and formed by the Xe/Kr selectivity and the pore void fractions (on the right). A base 10 logarithm has been applied to the selectivity values. Bin size increased by 2.4 (on log scale) for the selectivity, by about 500 m^2 cm^3 *for the surface areas and by* 0.125 *for the void fraction.*

selectivity have much lower void fractions around 0.1. The optimal range of void fraction for uptake would be between 0.2 and 0.6, whereas the one for selectivity is completely dissociated and is below 0.2. We can characterize a bit more finely the selectivity using the void fraction than using the surface area, even though they both give very similar results. Both descriptors describe a rather dense material with a "microporosity", in the sense of the IUPAC,**Sing_1985** which is characterized by medium-low pore volume and surface area.

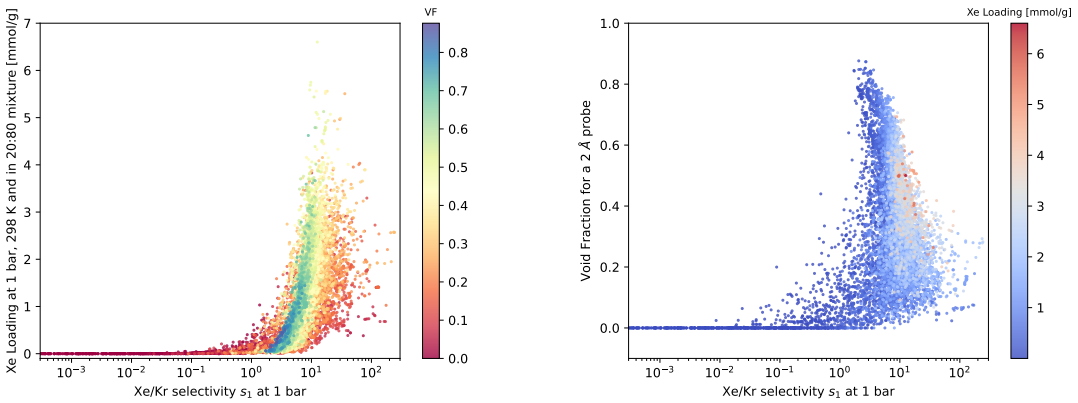


Figure 2.7: On the left: scatterplot of the xenon uptake as a function of the selectivity and labeled by the values of the void fraction. On the right: scatterplot of the selectivity and the void fraction labeled by the quantity of xenon adsorbed. The selectivity and uptake are calculated by a GCMC simulation of a 20:80 Xe/Kr mixture.

By carrying out a similar analysis than for the surface areas bu for the void fraction using the Figure **??** (right), we can also identify different intervals of void fractions that correspond to highly selective materials. For instance, selectivity values above 92 correspond to materials with a porosity between 0% and 37.5% (with a higher peak between 12.5% and 25%); selectivity values between 92 and 35 can be found in materials with a void fraction between 0% and 50.0% and much more frequently found for void fraction between 12.5% and 37.5%; selectivity values

between 35 and 13 can be found in materials with a void fraction between 0% and 75.0% in a bell distribution centered around 31%. This center of distribution shifts toward higher values of the void fraction as lower selectivity values are considered, which suggests that a rather low porosity (below 25%) is preferable for selectivity performance. However, as we mentioned for surface areas, the void fraction is not a deterministic variable neither, we cannot predict the performance of the material solely based on this descriptor. Let us investigate whether adding another variable like the pore size as a joint variable wan better characterize the material's performance.

Pore Size and selectivity

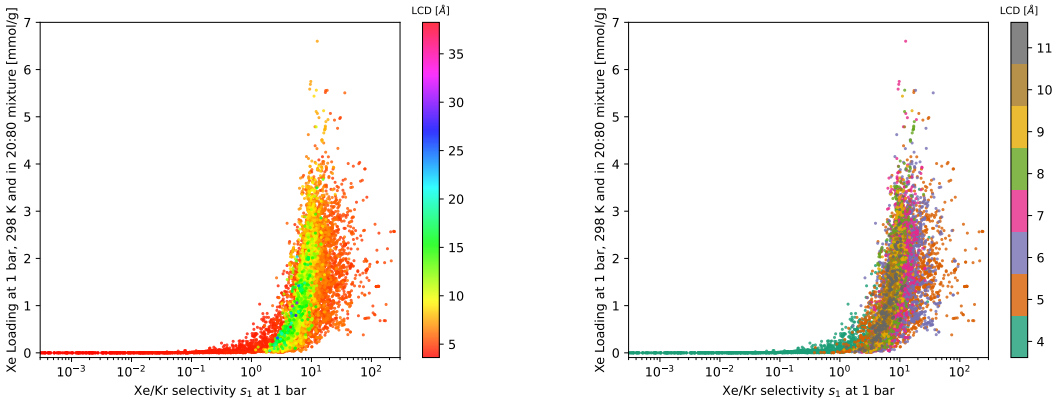


Figure 2.8: Scatterplot of the xenon uptake as a function of the selectivity (20:80) and labeled by the values of LCD_{UFF} (left). The same scatterplot restricted to values of D_i between (3.6 and 11.6 Å) and labeled using a different color code to distinguish the most selective materials from the least selective ones. The most selective materials are colored in orange corresponding to a pore size adapted for xenon adsorption (around 5 Å). The least selective ones are in green, with a pore lower than the size of a xenon hence preventing its adsorption.

If we now look at the joint effects of the void fraction and the largest cavity diameter (D_i) on the selectivity, we can note that the most selective materials are located in a very particular domain of this bidimensional descriptor space. On Figure **??**, the structures with a selectivity over 10 are very likely to have a void fraction under 0.4 with a rather wide range of D_i. However, as we can see on the filtered version of the plot (on the right), the most selective materials (over 40) exist, on the other hand, in a very narrow range of D_i values between 4.8 and 6 Å approximately. This can be explained by the size of a xenon atom being very close to these D_i values, which allows a maximal stabilization of the xenon that we want to separate from krypton. The krypton being slightly smaller, the interaction with the pores are less favorable, hence explaining the higher selectivity we observe.

As presented in the previous chapter, Simon et al. found that the most selective materials have a pseudo-spherical shape with a size close to the diameter of a xenon which is rather dense and not very porous. By taking a slightly different approach and including the xenon uptake as a co-metric alongside the selectivity, we found very similar results by identifying specific intervals of the cavity diameter and the pore volume. However, this structure–property relationship can only be used as a description of selective materials, and no prediction could be achieved by only using structural descriptors.

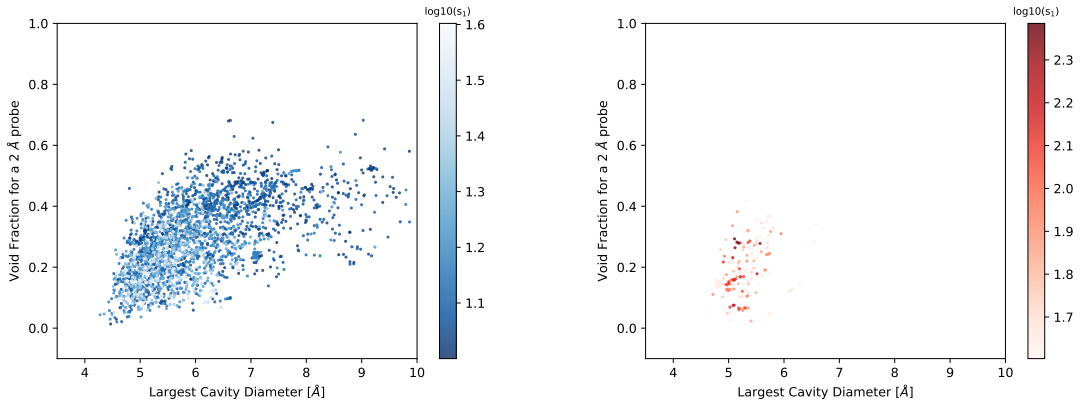


Figure 2.9: Scatterplots of the void fraction as a function of the LCD_{UFF} and labeled by the $\log_{10}$ of the selectivity values. On the left, only the materials with a selectivity between 10 *and* 40 *are considered; and on the right, selectivity values over* 40*.*

Effect of the composition

Finally, we previously only tackled one type of composition (20:80) associated to the extraction of xenon and krypton from a cryogenic distillation from the air (see section **??**). We are now going to investigate the effects of the composition by looking at the case of xenon/krypton separation in spent nuclear fuel off-gases. In nuclear applications, the mixture has a 90:10 Xe/Kr ratio, which is much richer in xenon in comparison to the previous one. For this reason, the quantity of xenon adsorbed in the materials will mechanically be higher than previously. However, the second quotient in the formula of selectivity in equation **??** compensates the first one that will be logically higher. Here, we want to evaluate these two effects to see if they cancel each other out or there are different trends depending on the composition value.

As shown on the Figure **??**, selectivity values of both compositions are quite close. However, we can note a slight decrease in performance when increasing the proportion of xenon in the mixture for some materials that are moderately selective (s between 2 and 50). This loss in performance could be explained by the fact that the materials display pores with different xenon affinities. With a lower proportion of xenon, the Xe adsorbates would access preferentially the most favorable pores and the small quantity of xenon is concentrated there. Whereas when there is a higher content of xenon, these most favorable sites begin to saturate and the xenon need to compete with krypton in much less favorable sites hence decreasing slightly the selectivity. We will see that later, we could use very similar reasoning to explain another variable that could change the selectivity.

We will now see the effect of the composition on the different analysis we carried out on the different structural descriptors. One of the major changes when considering a mixture with much more xenon is the values of the xenon uptake. The nanopores of selective materials ($1 < s_1 \leq 50$) are much more saturated in Xe and the maximum amount of xenon is therefore much higher. By comparing the Figures **??** and **??**, the maximum uptake is now 11.7 mmol g^{-1} instead of 6.6 mmol g^{-1} (for the 20:80 composition). For moderately selective materials, at 20:80, the xenon competes with krypton mainly in the most selective nanopores, but when we increase the xenon content it has to compete with krypton in much less favorable sites since the other sites are already saturated. We previously stated that the most selective ($s_1 > 50$)

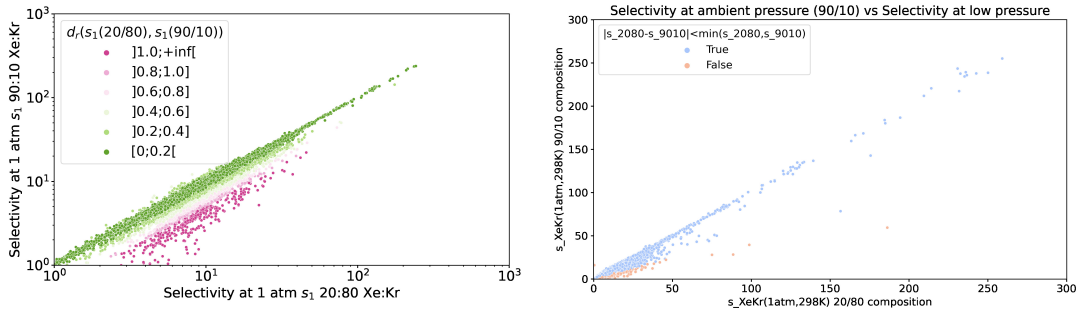


Figure 2.10: Illustration (scatterplot) of the difference of selectivity ($s_1(20:80)$ and $s_1(90:10)$) for two different Xe/Kr mixture compositions 20:80 (x-axis) and 90:10 (y-axis) at 1 atm *and* 298 K*. On the left, the axis is in log scale and the relative difference of selectivity between the two compositions is particularly high for the points labeled in purple. On the right, the axis is in linear scale and we only label the point to differentiate the materials with relative difference under and over* 1*.*

materials could reach up to 4.0 mmol g^{-1} of xenon uptake, for a composition with a higher xenon content, and it reaches up to 4.2 mmol g^{-1} again, which is not a big change. For the most selective materials, the previous conclusion on the maximum uptake can still stand. Because of the extremely high selectivity, the change in composition does not change the nature of the adsorbed state and about the same quantities of xenon are present in the pores. Finally, we can say that a higher content in xenon does not influence a lot the most selective materials' performance but it could alter the selectivity and increase a lot the xenon uptake for some moderately selective materials.

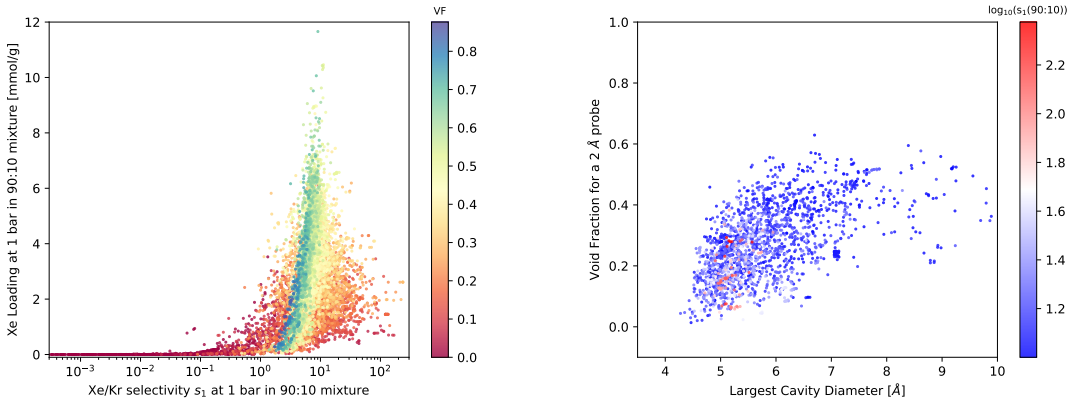


Figure 2.11: Illustration of the effect of the composition by representing the same figures as in **??** *and* **??** *but for a 90:10 composition. On the left: scatterplot of the xenon uptake as a function of the selectivity ($s_1(90:10)$) and labeled by the values of the void fraction. On the right: scatterplots of the void fraction as a function of the LCD_{UFF} and labeled by the selectivity ($s_1(90:10)$) values superior to* 10 *in log-scale.*

Finally, the composition does not affect the previously determined structural characteristics that a material needs to be very selective. As we can see on the right plot of the Figure **??** (right), the most selective materials still have pore size around 5 Å and porosity under 40%. This structural domain constitutes a necessary condition a selective material should have but being in this domain is not enough since less selective materials can also display these characteristics.

Now that we described the geometrical conditions needed to have a good selectivity, we will focus on the thermodynamic origins of the selectivity by focusing on energy-based quantities and the different correlations between them.

2.2.2 Thermodynamic quantities correlations at infinite dilution

In this section, our goal is not directly to address the structure–property relationships, but rather to map out the details of the thermodynamic features of Xe/Kr adsorption and separation in nanoporous materials. We used the high-throughput screening methodology as a way to map out the space of thermodynamic properties, going beyond the usual quantities of selectivity and uptake, to focus more specifically on the role of adsorption enthalpy and entropy, the differences between Xe and Kr adsorption thermodynamics, and between selectivity at low and high pressure.

To evaluate the performance of a given nanoporous material for separation in the low loading (or low pressure) limit, Henry's constants are often calculated from linear fits of low-pressure adsorption isotherm data — both experimentally and computationally. In this section, we investigate the thermodynamics of Xe and Kr adsorption at low pressure. Here, we have calculated the low-pressure adsorption properties by using the Widom insertion method**Widom1963**, **frenkel2001widom** on 9,668 structures from the dataset selected. It has higher accuracy than the fitting of isotherms, where it can be difficult to know what the extent of the linear adsorption regime is. With these simulations, we could obtain for each material the Henry's constant K and the adsorption enthalpy $\Delta_{ads}H_0$ (at the zero loading limit) for both xenon and krypton. The Xe/Kr thermodynamic selectivity s_0 in the low-pressure limit is then determined by the ratio $s_0 = K^{Xe}/K^{Kr}$ of the Henry's constants for the two gases. In the following, we look at the statistical relationships between the thermodynamic quantities at low pressure: s_0, K^{Xe}, K^{Kr}, $\Delta_{ads}H_0^{Xe}$, $\Delta_{ads}H_0^{Kr}$ and $\Delta_{exc}H_0$.

We display the distribution of thermodynamic properties of materials with favorable thermodynamic Xe/Kr selectivity ($s_0 > 1$) in Figure **??** — we restrict these plots to selectivity above 1, because those are the materials of interest for separation, and doing so removes several outliers with specific geometries or binding sites (but does not change the overall conclusions). We can first see that although the logarithm of the Xe Henry's constant K^{Xe} is weakly correlated to the logarithm of the selectivity s_0, this correlation is stronger for highly selective materials. Therefore, in a multistep screening study to identify the most selective materials, it could be possible to use as a "first filter" criterion based purely on Xe adsorption, discarding materials below a certain threshold (e.g., the materials with $s_0 \geq 30$ are contained in the subset with $K^{Xe} \geq 2.7\,10^{-1}$ mmol g^{-1} Pa^{-1}). The correlation between K^{Kr} and s_0, on the other hand, is weaker.

With regard to Henry's constants, we observe a broad selection of behavior, with K^{Xe} ranging from $2.6\,10^{-7}$ to $7.9\,10^{-1}$ mmol g^{-1} Pa^{-1}, and K^{Kr} ranging from $1.3\,10^{-7}$ to $5.1\,10^{-3}$ mmol g^{-1} Pa^{-1}. We also see that statistically, a high affinity for xenon usually translates into a high (relative) affinity for krypton, which is a general trend for noble gases where the adsorption sites are not strongly specific. In order to look more in detail into the thermodynamics behind this wide diversity in behavior, we plot in Figure **??** the enthalpies involved.

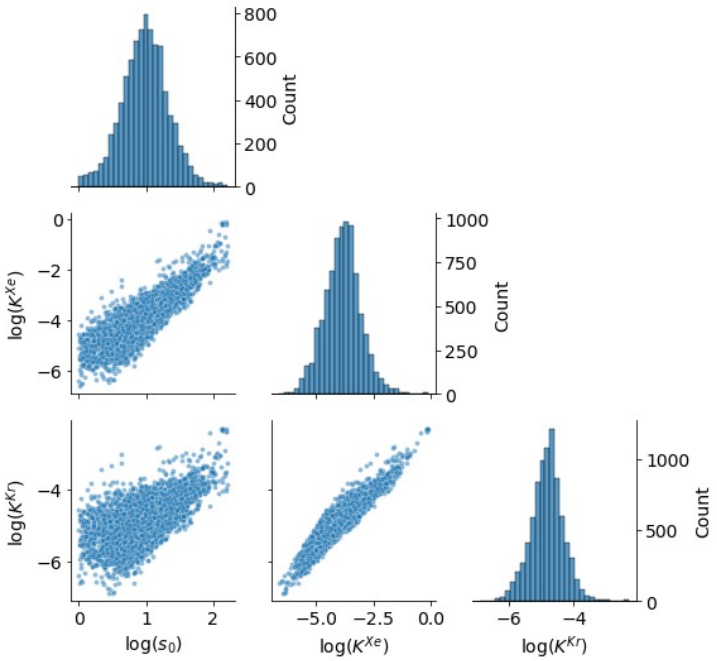


Figure 2.12: For 8,401 MOFs with favorable thermodynamic Xe/Kr selectivity ($s_0 > 1$), pair plots of $\log_{10}(s_0)$, $\log_{10}(\mathrm{K}^{Xe})$ *and* $\log_{10}(\mathrm{K}^{Kr})$ *(the Henry's constants are in* $\mathrm{mmol\,g^{-1}\,Pa^{-1}}$*) in the off-diagonal subplots (note that the y-axis is displayed on the right side) and the distribution of each quantity are on the diagonal (note that the y-axis displayed on the right side corresponds to the count and the x-axis is correctly labeled below each subplot).*

We first observe that the low-loading adsorption enthalpy of xenon ($\Delta_{\mathrm{ads}}\mathrm{H}_0^{\mathrm{Xe}}$) is strongly correlated to that of krypton ($\Delta_{\mathrm{ads}}\mathrm{H}_0^{\mathrm{Kr}}$). Echoing the similar correlation seen between respective Henry's constants, it suggests a rather generic physisorption mechanism is at play in the majority of materials, and that host–adsorbate affinities are mainly determined by the enthalpy. The main driver of Xe/Kr selectivity is neither the xenon nor krypton adsorption enthalpy alone (both are weakly correlated to the selectivity), but as expected their difference, $\Delta_{\mathrm{exc}}\mathrm{H}_0$, which is strongly correlated to $\log(s_0)$. This is further confirmed by the lack of correlation between selectivity and adsorption entropies (*cf.* Figure **??**): the separation is mostly enthalpic in nature, and the entropy causes the dispersion in the correlation between selectivity $\log(s_0)$ and $\Delta_{\mathrm{exc}}\mathrm{H}_0$.

Analyzing the Figure **??** in more detail, the adsorption entropy of xenon and krypton being noticeably correlated, their difference (the exchange entropy) does not have a lot of variations (see Figure **??**) compared to the enthalpy. This thermodynamic quantity plays a minor role in the selectivity performance of the materials. However it seems that the most selective materials do not have any values of exchange entropy but is centered around a value of about $-10\,\mathrm{kJ\,mol^{-1}\,K^{-1}}$. This is not a clean correlation but rather a necessary attribute of a selective material.

To emphasize one more time on the enthalpic nature of the separation by comparing the base-10 logarithm of the Henry constant (proportional to the adsorption free energy) and the adsorption enthalpy for both xenon and krypton. We can see, Figure **??** that the free energy can be almost totally explained by the enthalpy, which confirms the secondary role of the

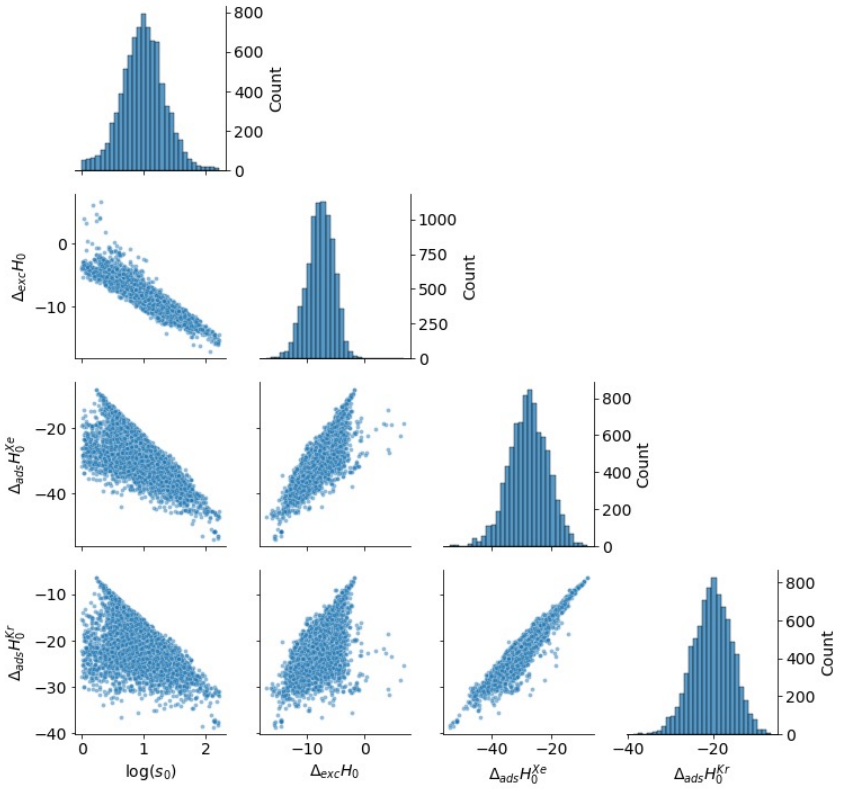


Figure 2.13: For 8,401 MOFs with favorable thermodynamic Xe/Kr selectivity ($s_0 > 1$), pair plots of $\log(s_0)$, $\Delta_{exc}H_0$, $\Delta_{ads}H_0^{Xe}$ *and* $\Delta_{ads}H_0^{Kr}$ *(the enthalpies are in* $\mathrm{kJ\,mol^{-1}}$*) in the off-diagonal subplots and the distribution of each quantity is on the diagonal.*

entropy that explains the variance in this linear relation. The effect of the entropy makes the correlation quite weak for the less favorable adsorption materials, but as we move to more negative values of the adsorption enthalpies, the correlation is stronger and stronger. The most selective materials have an almost negligible entropic contribution in the final free energy value (G = H – TS).

To go a little bit further in the correlation interpretation, the Figure **??** suggests that the entropic effect depends on the pore size. The bigger the size of the pores, the more positive the entropic term, which explains the weaker correlation for less attractive materials.

To check this, we looked at the influence of the pore size and the void fraction on the entropic term $T\Delta_{ads}S_0^{Xe}$ (see Figure **??**). The entropy is clearly related to the pore size here represented by the LCD_{UFF} — the larger the pore the higher the entropy is likely to be. This can simply be explained by the confinement effect of the pore — a small pore gives very little possible adsorbable positions to the xenon, whereas a larger pore opens up the possible sites for the adsorption. The same trend can be observed for the pore volume represented by the void fraction here. A weak linear correlation exists between the void fraction (in log-scale) and the adsorption entropic term of xenon. We can however argue that the whole picture of the entropic behavior is not captured by these simple geometric descriptors, especially for the larger pore sizes; other effects also play a role such as the shape of the channel and cavities (e.g. tortuosity) or the whole picture of the pore size distribution that cannot be simply captured in the LCD_{UFF}.

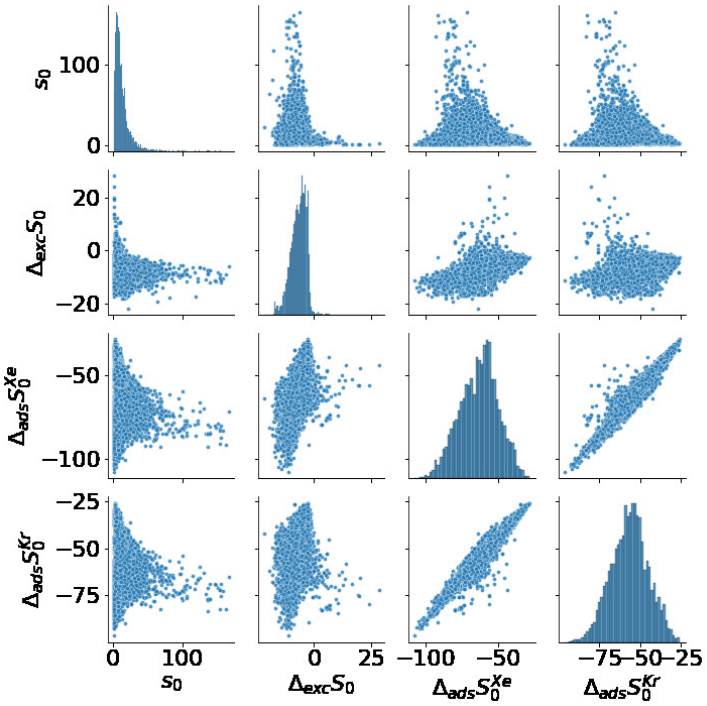


Figure 2.14: For 8,401 MOFs with favorable thermodynamic Xe/Kr selectivity ($s_0 > 1$), pair plots of s_0, $\Delta_{exc}S_0$, $\Delta_{ads}S_0^{Xe}$ and $\Delta_{ads}S_0^{Kr}$ in the off-diagonal subplots and the distribution of each quantity are on the diagonal.

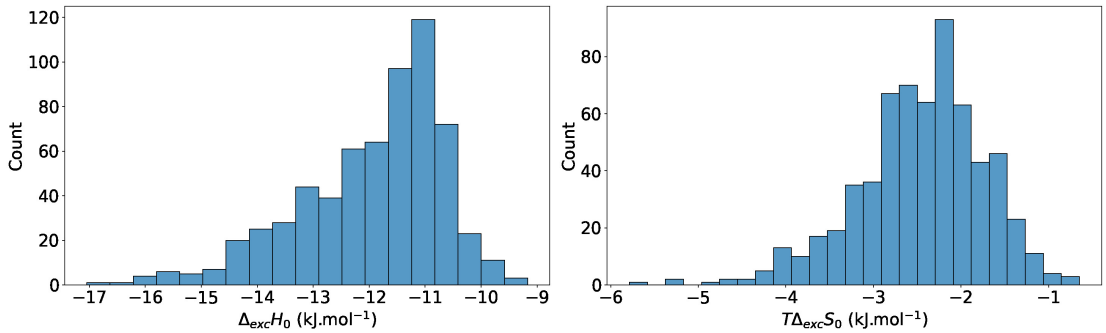


Figure 2.15: Distribution of the enthalpy $\Delta_{exc}H_0$ and entropy $T\Delta_{exc}S_0$ of exchange at low pressure on the 630 most selective structures

If we now croos these results with the previous results obtained on the influence of geometric descrioptors in the section **??**, we can note that the entropic effect goes in the same direction than the enthalpic term for explaining the selecitivity, when the pore size is around the size of a xenon. The confinement of the xenon makes the entropy lower in the adsorbed phase than in the gas phase, which is even more true for pores whose size are tailored for a xenon. The second benefit of this type of pore is the optimal interaction with the surrounding framework atoms, which lowers down the enthalpic term. Both effects are going in the same way and explain the optimal selectivity for this particular pore size value (around 5 Å).

The main takeaway messages of this section are based on two relations. The first one is between the Henry constant of xenon and the selectivity: knowing the performance of xenon we can infer the Xe/Kr selectivity — the most selective materials have very high xenon affinity. The

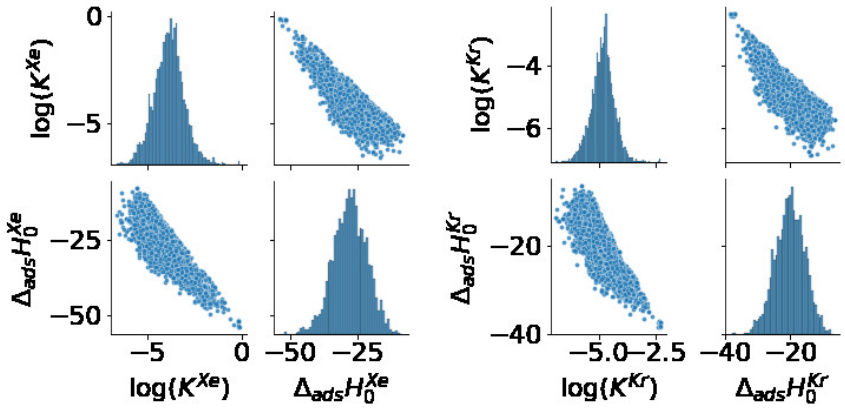


Figure 2.16: For 8,401 MOFs with favorable thermodynamic Xe/Kr selectivity ($s_0 > 1$), pair plots of $\log(K_H^i)$ and $\Delta_{ads}H_0^i$ in the off-diagonal subplots for both i=Xe and i=Kr and the distribution of each quantity are on the diagonal.

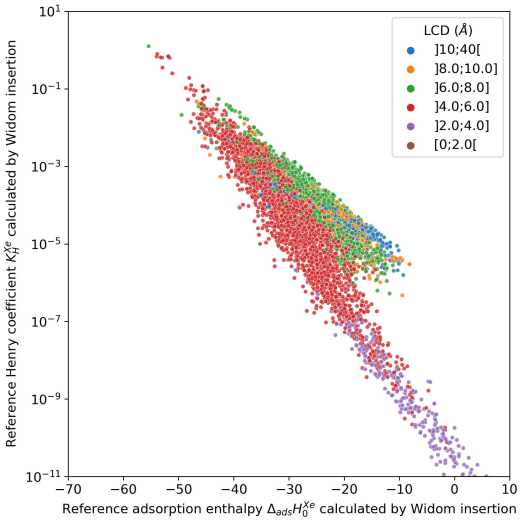


Figure 2.17: Comparison between the Xe Henry constant and Xe adsorption enthalpy labeled by categories of LCD_{UFF} values for the CoRE MOF structures.

second one concerns the relation between enthalpy and selectivity — the separation process has an enthalpic nature as a first-order approximation, which is even more true for the most selective materials. By studying the energy interactions within the material, we can understand most of the performance of it. We only looked at the thermodynamic properties at infinite dilution, in the next section we will focus on the effect of the pressure in the selectivity by focusing on a 20:80 Xe/Kr mixture at 1 atm and 298 K.

2.3 Selectivity Drop between Two Pressure Regimes

2.3.1 Thermodynamic origins

s After looking in the depth of the thermodynamics of the infinite dilution case, we will now focus on the impact of a change of working pressure on the adsorption selectivity, and analyze its thermodynamic origins. This is key to accurately assess the thermodynamics of adsorption

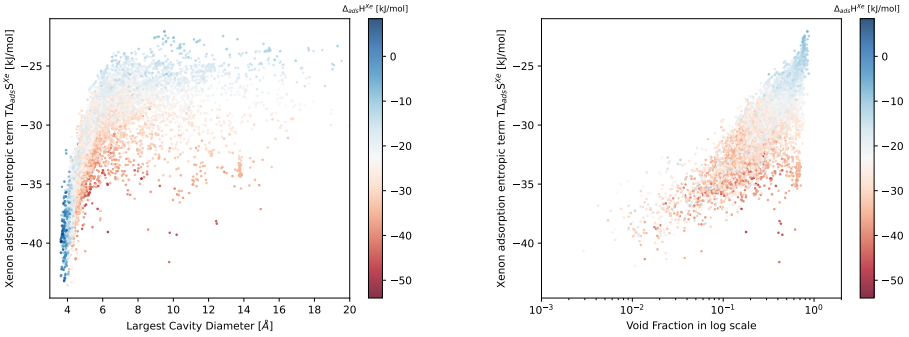


Figure 2.18: Comparison plots of the entropic term $\mathrm{T}\Delta_{ads}\mathrm{S}_0^{Xe}$ *at infinite dilution and two geometric descriptors: the* LCD_{UFF} *(left) and the void fraction (right).*

in different working conditions for specific industrial processes, and any insight into the impact of pressure on selectivity may allow for faster screening limited at selected thermodynamic conditions.

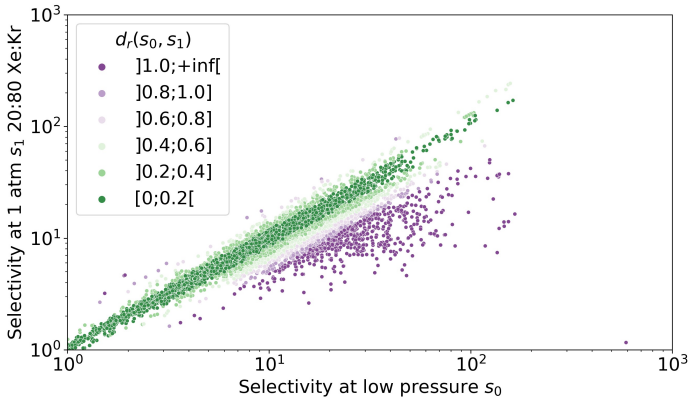


Figure 2.19: Difference of selectivity between low pressure and at a 1013 hPa *pressure for a 20:80 xenon krypton composition. The relative difference between the low-pressure selectivity and the ambient pressure is particularly high for the points labeled in purple.*

We calculated the selectivity s_1 at pressure 1 atm and ambient temperature using GCMC calculations on the entire dataset, with Xe/Kr mixture composition of 20:80 (found in a byproduct stream from air separation[**kerry2007industrial**]) and 90:10 (found in the off-gas streams from nuclear waste[**auerbach2003handbook**]). For high-selectivity materials, we find that the impact of composition appears rather marginal (*cf.* Figure **??**). In the following, we discuss the selectivity for the 20:80 mixture, which is the most commonly studied one in the literature. To measure the difference in selectivity between low and ambient pressures, we consider a relative difference $d_r(s_0, s_1)$ defined in equation **??**.

$$d_r(s_0, s_1) = \frac{|s_0 - s_1|}{\min(s_0, s_1)} \tag{2.29}$$

In Figure **??**, the selectivity at ambient pressure s_1 is plotted against its low-pressure counterpart s_0 (for materials where $s_0 > 1$, as before). The points are color-coded according to the value of $d_r(s_0, s_1)$, in 6 discrete categories for the sake of clarity. There is some broad level of correlation,

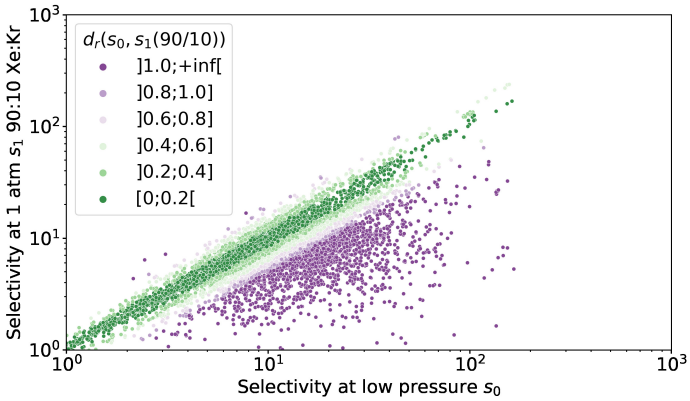


Figure 2.20: Difference of selectivity between low pressure and at a 1013 hPa *pressure for a 90:10 xenon krypton composition. The relative difference between the low-pressure selectivity and the ambient pressure is particularly high for the points labeled in purple.*

see near the diagonal with 61.5% of materials where the difference is below 20% (near the $s_0 = s_1$ line). We also see clearly that there are many more points (74.3% among the materials with $d_r(s_0, s_1) \geq 0.2$) below the first bisector ($s_1 < s_0$) than above: for these materials the selectivity s_1 at 1 atm is significantly lower than the one at low pressure s_0.

This drop in selectivity mainly concerns the materials with a relatively high selectivity $s_0 > 10$ (see Figure **??**), and forewarns that considering solely pure-component Henry's constant (i.e., zero-pressure selectivity) for materials screening could be misleading in some cases. Although it is simpler and faster to calculate, those low-pressure results that can overestimate selectivity by more than 100% in a significant number of materials (646 out of 9,668 in our dataset). By using a thermodynamic approach, we now try to explain the reasons behind these shifts in selectivity.

If we now look at the 90:10 composition, we can note that the drop in selectivity is even more important. The selectivity with a higher proportion of xenon was already found to be higher than the selectivity for 20:80 composition (see Figure **??**); we explained this drop by the presence of more or less favorable adsorption sites. In some materials (labeled in purple), at a low xenon content composition, the xenon and krypton mainly compete in the most favorable sites until these sites are saturated and no xenon is left to compete in the less selective sites. When we increase the Xe/Kr ratio, these less selective nanopores drive the overall selectivity down. Combined with the effect of increasing the pressure, some materials undergo both phenomena and have an exacerbated drop in selectivity compared to the selectivity at low pressure.

To evaluate quantitatively the thermodynamic effects at play in the competitive adsorption in different regimes, we can consider thermodynamic properties of the "exchange equilibrium" predefined in equation **??**. We plot in Figure **??** the exchange entropy at low pressure (plotted as $T\Delta_{exc}S_0$) against the exchange enthalpy $\Delta_{exc}H_0$. In this scatterplot, the points are color-coded according to the selectivity s_0 (with discrete categories for the sake of clarity), which is related to the enthalpy and entropy through Equation **??** — meaning iso-selectivity lines are parallel straight lines in this scatterplot.

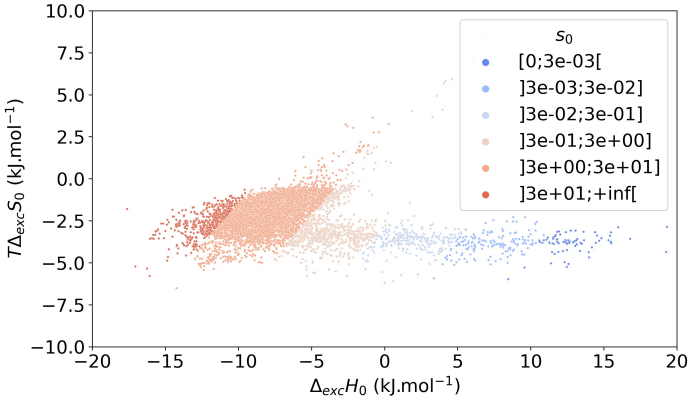


Figure 2.21: The energetic equivalent of exchange entropy $\mathrm{T}\Delta_{exc}\mathrm{S}_0$ *and enthalpy* $\Delta_{exc}\mathrm{H}_0$ *at low pressure labeled using the selectivity* s_0 *at low pressure. The limits between labels follows an affine function of slope* $1/\mathrm{T}$ *and of intercept* $-\mathrm{R}\ln(s_0^{lim})$ *where* s_0^{lim} *is the limit selectivity value (*cf. *Equation (??)). In other words, the iso-selectivity lines are all parallel lines of equation* $y = f(x)$ *where* f *is the affine function described previously.*

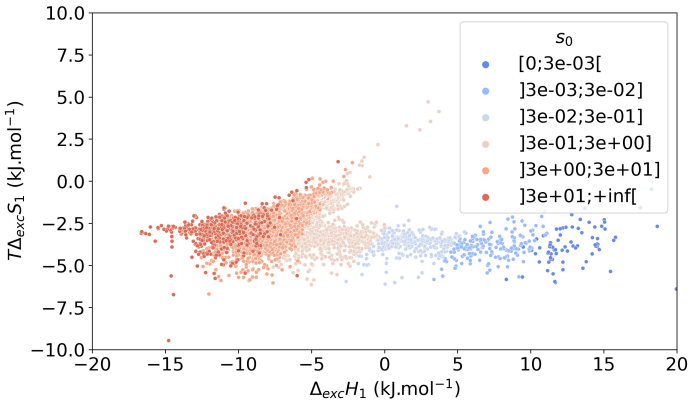


Figure 2.22: The energetic equivalent of exchange entropy $\mathrm{T}\Delta_{exc}\mathrm{S}_1$ *and enthalpy* $\Delta_{exc}\mathrm{H}_1$ *at ambient pressure labeled using the selectivity* s_0 *at low pressure. The points are layered so that the points with higher* s_0 *are always above. To see a split version of this plot, please refer to the Figure ??.*

In the Figure **??**, we display the distributions of the exchange enthalpy and entropy at low pressure. For the 630 most selective materials ($s_0 > 30$), the distribution of the exchange enthalpy $\Delta_{\mathrm{exc}}\mathrm{H}_0$ is centered on $-12.0\,\mathrm{kJ\,mol^{-1}}$ with a standard deviation of $1.3\,\mathrm{kJ\,mol^{-1}}$, whereas the distribution of the exchange entropy (plotted as $\mathrm{T}\Delta_{\mathrm{exc}}\mathrm{S}_0$) is centered on $-2.5\,\mathrm{kJ\,mol^{-1}}$ with a standard deviation of $0.7\,\mathrm{kJ\,mol^{-1}}$. These figures, along with the overall distribution plotted in Figure **??**, further confirms the moderate role of entropy in the low-pressure selectivity: it is equivalent in average to about 20% of the exchange enthalpy at low pressure.

Looking at the Figure **??**, at ambient pressure, we can state the same conclusions on the limited influence of the entropy on the selectivity values. The distribution of the entropic term $\mathrm{T}\Delta_{\mathrm{exc}}\mathrm{S}_1$ is now centered around $-3\,\mathrm{kJ\,mol^{-1}}$, which is also quite small in comparison of the values of $\Delta_{\mathrm{exc}}\mathrm{H}_1$. For the most selective materials, the entropic term represents about 19% of the exchange enthalpy at ambient pressure.

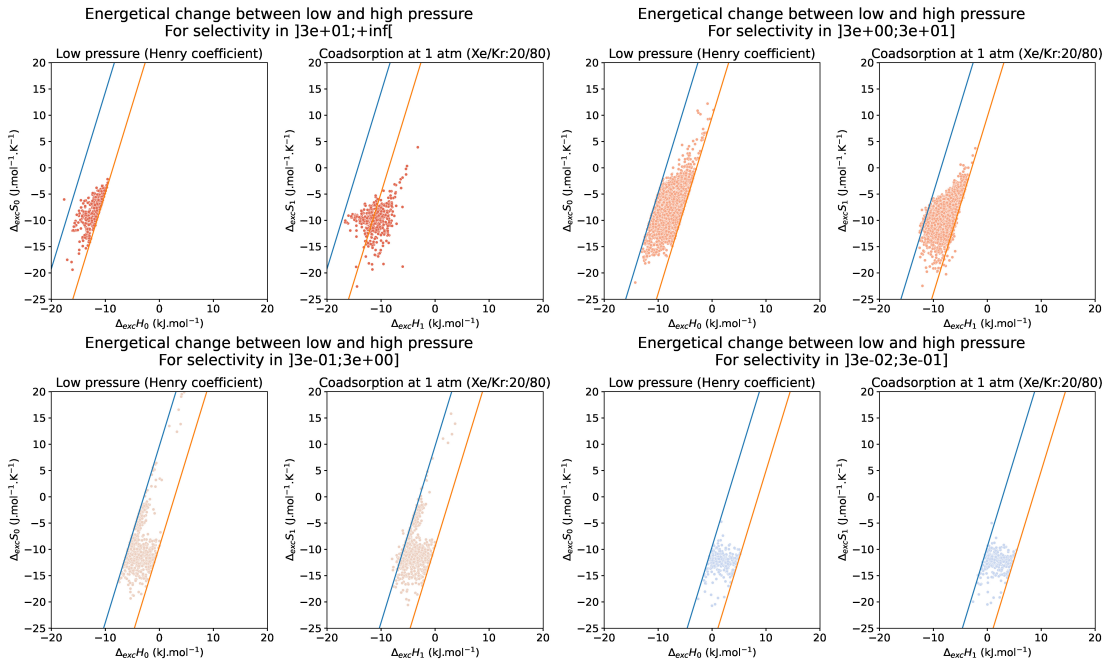


Figure 2.23: Split view of the Figure ?? and ??. The iso-selectivity lines for the limit considered are represented with blue and orange lines. We can clearly see the shift in exchange enthalpy for the structures with a selectivity higher than 30.

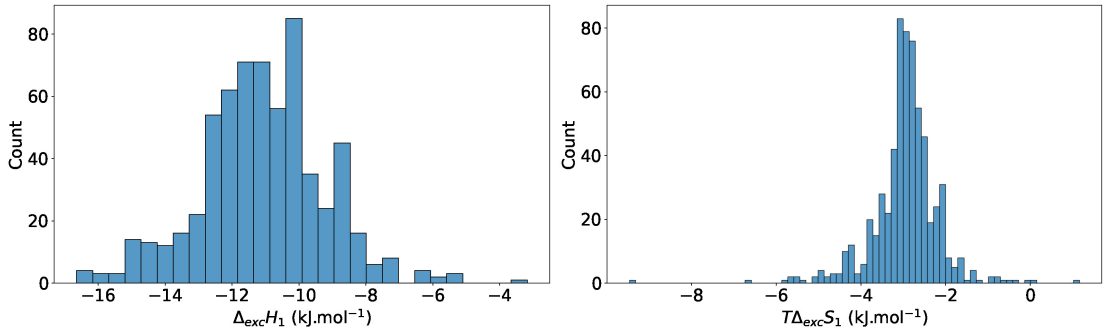


Figure 2.24: Distribution of the enthalpy $\Delta_{exc}H_1$ and entropic term $T\Delta_{exc}S_1$ of exchange at ambient pressure on the 630 most selective structures.

Figure ?? represents a scatterplot of the exchange entropy at P = 1 atm $\Delta_{exc}S_1$ against the exchange enthalpy at ambient pressure $\Delta_{exc}H_1$. To compare it to the Fig. ??, the points are color-coded according to the low-pressure selectivity s_0. Compared to the iso-selectivity s_1 straight parallel lines (*cf.* Figure ??), we can see that many materials with high s_0 have lower s_1 — seen as a migration of points to the right of the plot, compared to Fig. ??. This shift is therefore mainly due to a higher (less favorable) exchange enthalpy, hinting at an important role of enthalpy to determine higher pressure selectivity.

To quantify this change, we consider the distributions of the exchange enthalpy $\Delta_{exc}H_1$ and the energetic equivalent of the exchange entropy $T\Delta_{exc}S_1$ at ambient pressure (Figure ??). The enthalpy $\Delta_{exc}H_1$ is now centered on $-11.1\,kJ\,mol^{-1}$ with a standard deviation of $1.9\,kJ\,mol^{-1}$. Compared to the zero-pressure values, the enthalpy distribution is more dispersed, showing that there are important changes in individual values, and is higher in average — majority of materials have lower ambient pressure selectivity due to enthalpic effects. This can be

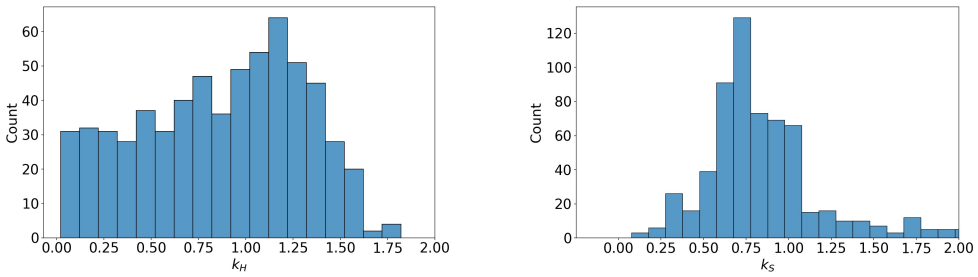


Figure 2.25: Distribution of the enthalpic k_H and entropic k_S contributions to the change of selectivity from low to ambient pressure for the 630 materials with $s_0 > 30$. k_H has a rather uniform distribution, whereas k_S has a bell-like distribution.

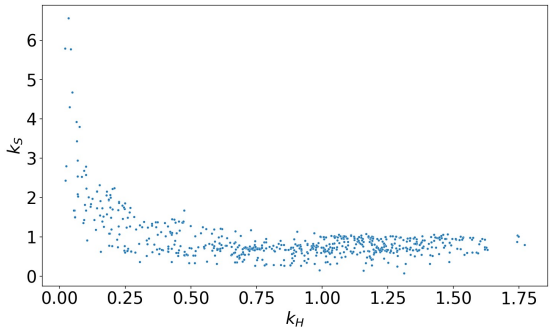


Figure 2.26: Scatterplot of the enthalpic contribution k_H and entropic contribution k_S for the 630 materials with $s_0 > 30$. The entropic compensation occurs when the enthalpic contribution is around 0.1, else its value is around 1 and has little effect on the selectivity change.

explained by the very general increase of adsorption enthalpy upon loading in the gas phase, which is linked to the presence of more adsorbed molecules. In fact, the correlations (Figure **??**) suggest that highly selective materials have high affinity in xenon; therefore they feature significant uptake at 1 atm and the large Xe loading means the most favorable adsorption sites can be saturated, and further adsorption involves weaker host–guest interactions and therefore increases the average adsorption enthalpy at non-zero loading.

The entropic term $T\Delta_{exc}S_1$ is now centered on $-2.9\,kJ\,mol^{-1}$, with a standard deviation of $0.8\,kJ\,mol^{-1}$ (almost unchanged from low-pressure). The entropy is on average lower, which means an overall less favorable separation due to entropic effects: this evolution of the entropic term hints at the potential of a reorganization of the adsorbed molecules inside each material. The difference in distribution of enthalpy has, overall, more impact on the high-pressure selectivity than that of entropy. This suggests that the overall contribution of enthalpy remains more decisive than the role of entropy in the selectivity change, even at ambient pressure. This is an interesting conclusion for screening studies, because evaluation of adsorption enthalpy can be computationally faster than that of the adsorption free energy (or entropy).

To further investigate the thermodynamics of the selectivity change, we quantify in this section the contributions of enthalpy and entropy. The ratio s_1/s_0 is equal to the product $k_H \times k_S$

where k_H and k_S are the enthalpic and entropic contributions to the selectivity change defined as:

$$k_H = \exp\left(-\frac{\Delta_{exc}H_1 - \Delta_{exc}H_0}{RT}\right)$$
$$k_S = \exp\left(\frac{\Delta_{exc}S_1 - \Delta_{exc}S_0}{R}\right) \tag{2.30}$$

As we can see in Figure **??**, the entropic contribution k_S has a bell-like distribution, with a mean of 0.9 and a standard deviation of 0.6. This confirms that k_S is close to 1, and has therefore only a marginal effect on the selectivity change. On the other hand, the enthalpic contribution k_H has a more uniform distribution ranging from 0.1 to 1.5, which means that enthalpy has a crucial role in the selectivity change we observe. There are a significant number of materials with a k_H close to zero, they correspond to the same materials highlighted in the section **??**.

Furthermore, the scatterplot of k_H and k_S (shown in Figure **??**) confirms a rather moderate effect of entropy. For most of the materials with $0.25 \leq k_H \leq 1.75$, we see that k_S is close to 1. The most significant entropic contributions are found for materials where k_H is close to zero (typically below 0.25). If we look in more detail at the 29 materials with $k_S > 2$, the entropic contribution k_S moderately compensate the enthalpic contribution as the average ratio s_1/s_0 is around 0.25. In such cases, the entropy is non-negligible and it can partially compensate the enthalpic contribution to the selectivity change, but the general trend is still given by enthalpy, since the overall selectivity is decreasing as a result.

2.3.2 Detailed investigation

Table 2.1: Enthalpic (k_H) and entropic (k_S) contributions to the selectivity change (s_1/s_0) between low and ambient pressures for some archetypal structures selected for their high s_0 selectivity at infinite dilution. Every structure is identified using a CSD Refcode and a reference the first article that mentions it. The pore size is also characterized using the diameters D_i and D_f in Å.

CSD Refcode	Ref.	s_0	s_1	s_1/s_0	k_H	k_S	D_i	D_f
VOKJIQ	[**VOKJIQ**]	157.17	242.73	1.54	1.46	1.06	5.2	3.2
KAXQIL	[**KAXQIL**]	103.78	132.57	1.28	1.32	0.96	5.2	4.1
JUFBIX	[**JUFBIX**]	106.11	114.83	1.08	1.08	1.00	5.3	3.0
FALQOA	[**FALQOA**]	162.20	171.10	1.05	1.09	0.96	5.1	3.5
GOMREG	[**GOMREG_GOMRAC**]	114.14	73.83	0.65	1.01	0.64	5.8	4.0
JAVTAC	[**JAVTAC**]	117.38	66.93	0.57	0.77	0.74	5.5	4.3
GOMRAC	[**GOMREG_GOMRAC**]	124.11	47.34	0.38	0.58	0.66	5.7	3.7
MISQIQ	[**MISQIQ**]	138.94	37.32	0.27	0.51	0.53	4.6	4.4
BAEDTA01	[**BAEDTA01**]	154.10	37.74	0.24	0.12	1.97	5.7	4.6
VIWMOF	[**VIWMOF**]	81.13	13.24	0.16	0.04	4.30	10.2	5.3
LUDLAZ	[**LUDLAZ**]	165.68	16.42	0.10	0.16	0.63	6.7	4.2
WOJJOV	[**WOJJOV**]	146.32	13.94	0.10	0.06	1.68	8.2	6.8
VAPBIZ	[**VAPBIZ**]	146.73	12.76	0.09	0.06	1.50	6.3	3.7

In this section, we go over some of the most selective materials, as identified at low pressure and listed in Table **??**, and we provide a detailed investigation of the thermodynamic effects

behind their behavior. We can split them into three main categories: materials with a slight increase in selectivity or little change in selectivity ($s_0/s_1 > 0.8$), materials with a slight decrease in selectivity ($0.5 \leq s_0/s_1 \leq 0.8$) and materials with a significant decrease in selectivity ($s_0/s_1 < 0.5$). In this section, we investigate the origins of these different behaviors: all materials are referenced by their CSD refcode.

Table 2.2: Thermodynamic quantities associated for a few archetypal structures. Henry's constant K^{Xe}, K^{Kr} *are in* $\mathrm{mmol\,g^{-1}\,Pa^{-1}}$, *loadings* q_1^{Xe} *and* q_1^{Kr} *are in* $\mathrm{mmol\,g^{-1}}$, *enthalpies* $\Delta_{ads}H_0^{Xe}$, $\Delta_{ads}H_0^{Xe}$, $\Delta_{ads}H_1^{Xe}$ *and* $\Delta_{ads}H_1^{Xe}$ *are in* $\mathrm{kJ\,mol^{-1}}$

CSD Refcode	Ref.	s_0	K^{Xe}	K^{Kr}	$\Delta_{ads}H_0^{Xe}$	$\Delta_{ads}H_0^{Kr}$	s_1	q_1^{Xe}	q_1^{Kr}	$\Delta_{ads}H_1^{Xe}$	$\Delta_{ads}H_1^{Xe}$
VOKJIQ	[**VOKJIQ**]	157	$7.92\,10^{-1}$	$5.04\,10^{-3}$	−53.9	−38.2	243	2.57	0.04	−61.1	−44.5
KAXQIL	[**KAXQIL**]	104	$3.01\,10^{-2}$	$2.90\,10^{-4}$	−44.6	−30.5	133	1.41	0.04	−41.5	−26.8
JUFBIX	[**JUFBIX**]	106	$1.59\,10^{-2}$	$1.50\,10^{-4}$	−45.6	−31.4	115	0.80	0.03	−45.7	−31.3
FALQOA	[**FALQOA**]	162	$2.23\,10^{-2}$	$1.38\,10^{-4}$	−47.3	−32.0	171	0.68	0.02	−48.6	−33.1
GOMREG	[**GOMREG_GOMRAC**]	114	$9.16\,10^{-2}$	$8.03\,10^{-4}$	−44.7	−31.1	74	2.59	0.14	−47.5	−33.8
JAVTAC	[**JAVTAC**]	117	$1.24\,10^{-1}$	$1.06\,10^{-3}$	−47.7	−33.5	67	1.50	0.09	−48.5	−34.9
GOMRAC	[**GOMREG_GOMRAC**]	124	$1.17\,10^{-1}$	$9.45\,10^{-4}$	−45.6	−31.8	47	2.51	0.21	−47.3	−34.8
MISQIQ	[**MISQIQ**]	139	$6.87\,10^{-1}$	$4.94\,10^{-3}$	−51.9	−37.4	37	2.30	0.25	−45.6	−32.8
BAEDTA01	[**BAEDTA01**]	154	$1.39\,10^{-2}$	$9.04\,10^{-5}$	−47.7	−31.7	38	1.05	11	−34.0	−23.1
VIWMOF	[**VIWMOF**]	81	$7.87\,10^{-3}$	$9.70\,10^{-5}$	−46.3	−30.1	13	2.99	0.90	−26.0	−17.8
LUDLAZ	[**LUDLAZ**]	166	$9.04\,10^{-2}$	$5.46\,10^{-4}$	−45.4	−30.9	16	1.59	0.39	−38.3	−28.3
WOJJOV	[**WOJJOV**]	146	$4.19\,10^{-2}$	$2.86\,10^{-4}$	−46.4	−30.7	14	2.82	0.81	−33.0	−24.4
VAPBIZ	[**VAPBIZ**]	147	$3.54\,10^{-2}$	$2.41\,10^{-4}$	−46.4	−30.5	13	2.50	0.78	−34.1	−25.3

Before introducing the different archetypal structures that undergo different changes in selectivity, let us bring in some notions on adsorption isotherms. The isotherms are representation of the adsorbed quantity as a function of the pressure for different components at a given temperature. Here, we will only tackle the case of pure-component isotherms at 298 K. Different models have been developed to interpret these plots,**Al_Ghouti_2020** but here we will only use the Langmuir model as it is the most prominent equation to explain adsorption equilibria. The Langmuir model is a local model of adsorption based on the filling of a monolayer by non-interacting adsorbates. Depending on the distribution and shape of the pores, these isotherms can be either modeled by a 1-site Langmuir or a 2-site Langmuir model. At given temperature, some mono-site materials' isotherm can be described by the following equation:

$$q(\mathrm{P}) = \mathrm{N_{max}}\frac{\mathrm{KP}}{1+\mathrm{KP}} \tag{2.31}$$

where q is the adsorbed quantity of a mono-component gas, K is the adsorption equilibrium constant and P is the pressure. When the material has 2 sites, the isotherm can be described by the following equation:

$$q(\mathrm{P}) = \mathrm{N_{max}}\left((1-\alpha_2)\frac{\mathrm{K_1P}}{1+\mathrm{K_1P}} + \alpha_2\frac{\mathrm{K_2P}}{1+\mathrm{K_2P}}\right) \tag{2.32}$$

where q is the loading of a given mono-component gas, K_1 and K_2 are the adsorption equilibrium constants in the respective sites, α_2 is the proportion of secondary sites, and P is the pressure.

We first study a few examples of the category of materials where ambient-pressure selectivity is close to (or even higher than) the low-pressure value. For VOKJIQ, the selectivity is multiplied by 1.5 between low and ambient pressure. We see that the adsorption enthalpy of xenon $\Delta_{ads}H^{Xe}$ decreases from $-53.9\,kJ\,mol^{-1}$ to $-61.1\,kJ\,mol^{-1}$, whereas for krypton $\Delta_{ads}H^{Kr}$ decreases from $-38.2\,kJ\,mol^{-1}$ to $-44.5\,kJ\,mol^{-1}$ (*cf.* Table **??**). This increased stability of the adsorption sites upon loading is not common in nanoporous materials for rare gas adsorption, and can be linked to a cooperative effect between the adsorbed molecules. The stabilization favors the xenon molecules over the krypton molecules, due to an interatomic distance inside the pores that is a closer match to the energy well for favorable Lennard-Jones potential for xenon-xenon interactions than for krypton-krypton interactions (which is the case for a distance higher than 4.2 Å; see Figure **??**).

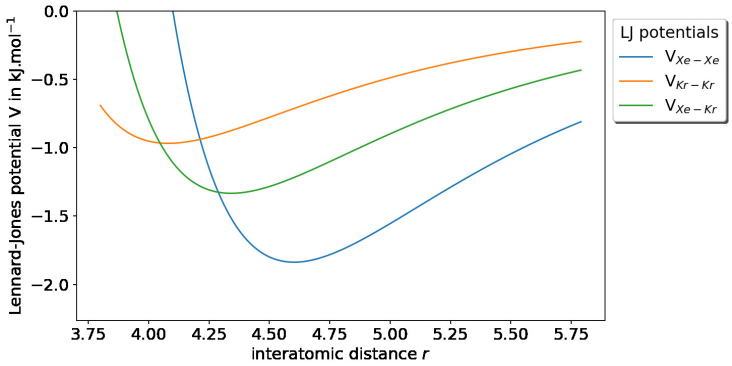


Figure 2.27: The LJ potentials for xenon and krypton interactions. The xenon-xenon interaction is more stabilizing than the krypton-krypton interaction for interatomic distance higher than 4.2 Å.

In the case of KAXQIL, the channels are one-dimensional tubes (see Figure **??**) and the distance between two adsorption sites is approximately the unit cell parameter along the direction of the tube (5.6 Å). There the selectivity increases with pore filling, for enthalpic reasons, which we can explain by relatively simple reasoning. The Lennard-Jones potentials U^{LJ} can be estimated for all species at 5.6 Å: $U^{LJ}_{Xe\text{-}Xe} = -1.0\,kJ\,mol^{-1}$, $U^{LJ}_{Kr\text{-}Kr} = -0.3\,kJ\,mol^{-1}$ and $U^{LJ}_{Xe\text{-}Kr} = -0.5\,kJ\,mol^{-1}$. In a simplistic model where all adsorbed molecules are 5.6 Å apart, the cooperative effect is higher between two xenon molecules, which explains the increased selectivity at high uptake. If we look further at the adsorption enthalpy of both xenon and krypton (*cf.* Table **??**), they both increase: the guest molecules move from the "ideal" adsorption sites, and the guest–guest interactions do not fully compensate. The selectivity change in this material is therefore a consequence of the guest–guest interactions that rearranges the position of the adsorbates inside the nanopores.

To further corroborate the role of the guest–guest interactions, we look at another material with one-dimensional tubelike channels: JUFBIX, a cobalt(II) coordination polymer based on carboxylic acid linkers (see Figure **??**).**JUFBIX** The periodicity along the direction of the tubes is much higher at 7.2 Å. The pair interaction energies corresponding to the LJ potentials at this distance are $U^{LJ}_{Xe\text{-}Xe} = -0.24\,kJ\,mol^{-1}$, $U^{LJ}_{Kr\text{-}Kr} = -0.06\,kJ\,mol^{-1}$ and $U^{LJ}_{Xe\text{-}Kr} = -0.13\,kJ\,mol^{-1}$. By looking at the adsorption enthalpies (Table **??**), these values are too small to affect the position of the adsorbed molecules. At high loading, the distance between adsorbed molecules is high, and every adsorption site is independent of the others. The ambient-pressure selectivity s_1 is therefore the same as the low-pressure selectivity s_0, since every guest–guest interactions are

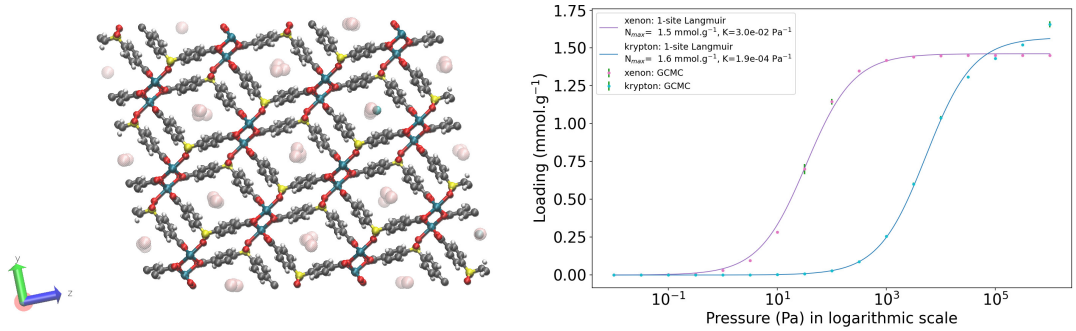


Figure 2.28: KAXQIL: On the left side, an illustration of a clean version (all solvent removed) of the calcium coordination framework [Ca(SDB)]·H_2O, where SDB = 4,4′-sulfonyldibenzoate loaded with xenon and krypton obtained by GCMC calculations. Color code: Ca in dark cyan, C in gray, O in red, H in white, S in yellow; Xe in transparent pink and Kr in cyan for the adsorbates. The mono-component isotherms fitted with a 1-site Langmuir model (Equation ??) for both xenon and krypton at 298 K is represented on the right side.

negligible. It confirms the crucial role of cooperative effects between guest molecules when considering a saturated material.

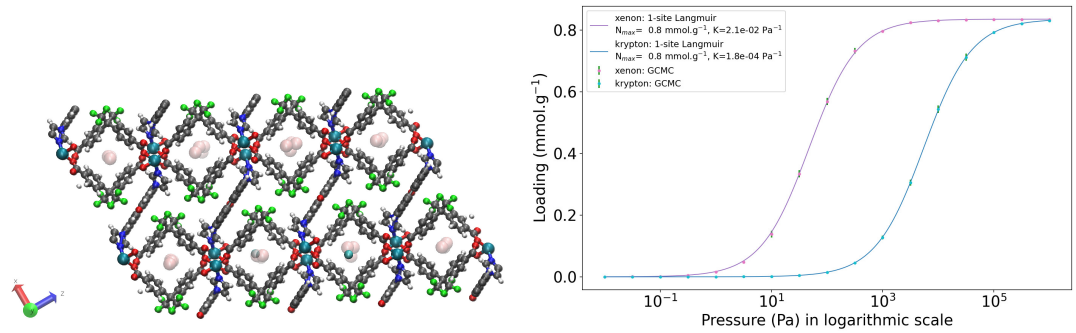


Figure 2.29: JUFBIX: Representation of a clean version (all solvent removed) of the cobalt(II) coordination framework [Co_2(L)(ppda)$_2$]$_2$·H_2O, where the ligand L is 2,8-di(1H-imidazol-1-yl)dibenzofuran and the carboxylic acid ligand H_2ppda is 4,4′-(perfluoropropane-2,2-diyl)dibenzoic acid loaded with xenon and krypton obtained by GCMC calculations. Color code: Co in dark cyan, C in gray, O in red, H in white, N in blue, F in green ; Xe in transparent pink and Kr in cyan for the adsorbates. The mono-component isotherms fitted with a 1-site Langmuir model (Equation ??) for both xenon and krypton at 298 K is represented on the right side.

GOMREG and JAVTAC are frameworks that belong to the second category of materials, with a moderate decrease in selectivity from low to ambient pressure. In GOMREG, the channels are composed of one-dimensional tubes larger than the ones found in KAXQIL or JUFBIX (see Figure ?? and Table ??). The adsorption sites are alternating from left to right inside the channel, and the adsorbed molecules organize in a "zigzag" pattern. Looking at the adsorption enthalpies, we see that both xenon and krypton have lower enthalpies by a similar margin, suggesting an equivalent stabilization for both atoms, hence the enthalpic contribution to the selectivity change is close to 1. Since krypton is smaller and less strongly tied on its adsorption site than xenon, it has more available space inside the pore space. This gives an entropic advantage to the Kr, seen in the entropic contribution k_S of 0.64 in Table ??. This indicates that even if enthalpic considerations mainly explain the observed changes at a statistical level, as

discussed in the previous sections, for individual cases entropic considerations can be a strong factor in pressure-dependent selectivity.

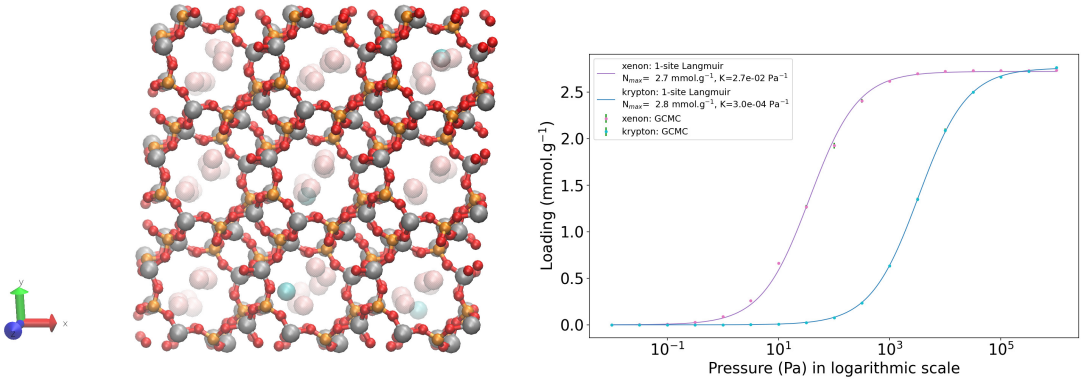


Figure 2.30: GOMREG: Representation of a clean version (all solvent removed) of this aluminophosphate $AlPO_4$-n that has a zeotype LAU topology with one-dimensional 10-ring channels loaded with xenon and krypton obtained by GCMC calculations. Color code: Al in silver, P in orange, O in red ; Xe in transparent pink and Kr in cyan for the adsorbates. The mono-component isotherms fitted with a 1-site Langmuir model (Equation ??) for both xenon and krypton at 298 K is represented on the right side.

The remaining materials discussed here form a third category, with a strong decrease in selectivity from low to ambient pressure. We look at several phenomena that can be at the root of this decrease, which is important for screening studies as it can limit the working performance of a material that appears to be a "top performer" based on zero-pressure screening.

For example, GOMRAC has a similar structure compared to GOMREG (see Figure ??), except for the fact that the pores and channels are smaller (see the values of the D_i, and the D_f, in Table ??). The distances between the adsorbed molecules — in their ideal sites — are then consequently smaller. At such distances, we can assume that the interactions between adsorbates become more stabilizing for krypton than for xenon molecules in GOMRAC (see LJ potentials at distance lower than 4.2 Å in the Figure ??), which translates into an enthalpic contribution k_H of 0.58. Moreover, this is compatible with the equivalent guest–guest interactions in GOMREG, as previously discussed. It explains why the difference between the adsorption enthalpies becomes smaller for GOMRAC, whereas it stays the same for GOMREG (between low and ambient pressure). This further validates the crucial role of the interactions between adsorbed molecules, and their relationship with the guest-guest distances when considering a high loading condition.

If we look at the case of MISQIQ, we see that the pure-component Xe isotherm in Figure ?? cannot be fitted by a single-site Langmuir isotherm, but is well fitted by a two-site Langmuir model (see Figure ??). Visual inspection of the adsorbed density at various loadings shows that this is not a second, separate adsorption site that is populated at high loading: instead, the second step in the isotherm (representing about 20% of the uptake at full loading) is associated with a reorganization of the adsorbate molecules occurs at high loading, accompanying a contraction of the interatomic distances. In this case, the potential for a reorganization of the adsorbate in the material's nanopores leads to the change in selectivity. This reorganization can be detected on the basis of the xenon isotherm alone, and has a major role in the selectivity

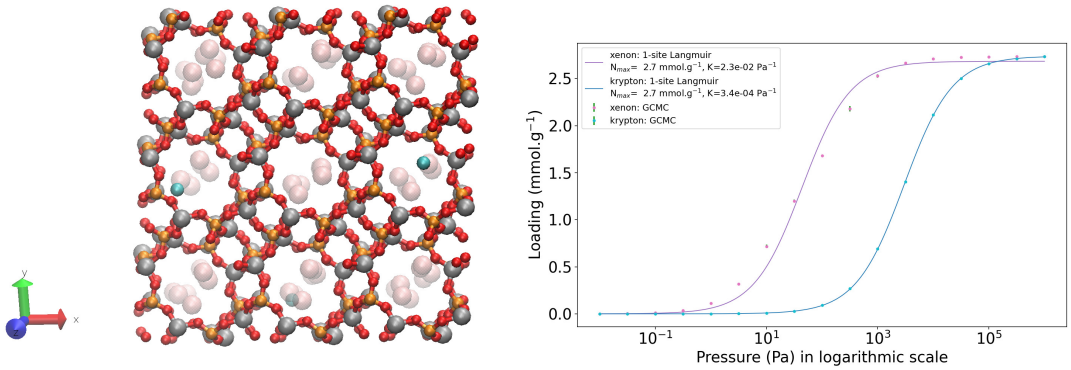


Figure 2.31: GOMRAC: Representation of a clean version (all solvent removed) of this aluminophosphate $AlPO_4$-n that has a zeotype LAU topology with one-dimensional 10-ring channels loaded with xenon and krypton obtained by GCMC calculations. Color code: Al in silver, P in orange, O in red; Xe in transparent pink and Kr in cyan for the adsorbates. The mono-component isotherms fitted with a 1-site Langmuir model (Equation ??) for both xenon and krypton at 298 K is represented on the right side. It seems that this aluminophosphate is just a smaller version of GOMREG.

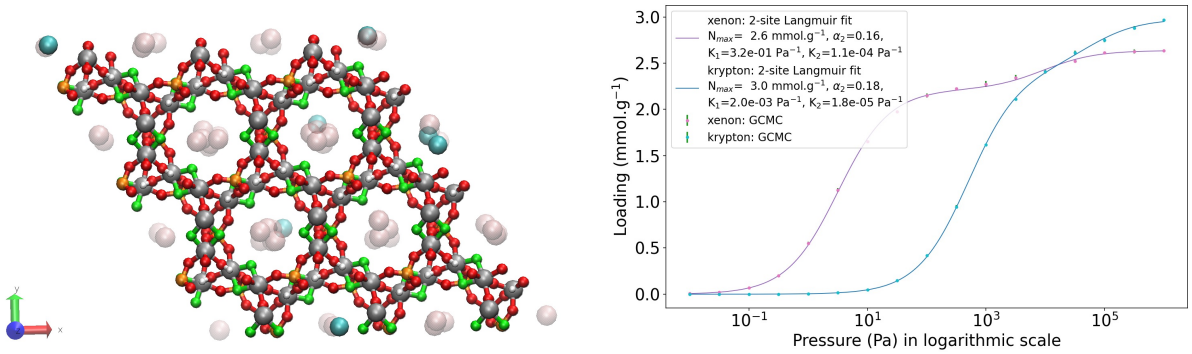


Figure 2.32: Representation of a chiral open-framework fluoroaluminophosphate $[C_4N_3H_{16}]\cdot[Al_6P_3O_{12}F_6(OH)_6]$ denoted AlPO-ƷU89 (referenced MISQIQ in the Cambridge structural database), which has been loaded with xenon and krypton in a GCMC simulation, on the left side.[MISQIQ] *Color code: Al in silver, P in orange, O in red, H in white and F in green for the framework; and Xe in transparent pink and Kr in cyan for the adsorbates. The pure-component isotherms fitted with a 2-site Langmuir model (Equation ??) for both xenon and krypton at 298 K on the right side.*

at ambient pressure. This repacking of the adsorbed phase is linked to a strong entropic effect, and also impacts the enthalpic contribution to selectivity.

More extreme cases of selectivity drop can occur when more than one site is available, as is the case for materials BAEDTA01, VIWMOF, LUDLAZ, WOJJOV, and VAPBIZ. The pure-component isotherms and the representation of the materials loaded in xenon and krypton molecules (presented in the supporting information of the Ref. [**Ren_2021**] Figures S19-23) confirm the existence of at least two distinct adsorption sites in each material. The most selective sites (i.e., the most favorable for Xe) are filled in priority at low loading, and the less selective sites will then be populated when the pressure increases, leading to a net selectivity drop at ambient pressure for these materials. The different types of adsorption sites, and therefore the potential for a drop in Xe/Kr selectivity (at non-zero pressure) is a factor that could be explicitly included

in screening of pure-component isotherms, without the need for explicit multi-component GCMC simulations.

2.3.3 Toward the development of new screening tools

In the current state of the art on Xe/Kr separation by adsorption in nanoporous materials, many studies have focused on the determination of structure/property relationships, the description of theoretical limits of performance, and the identification of top-performing materials, whether for existing experimental structures or for novel hypothetical structures yet to be synthesized. Here, we provide a study based on a high-throughput screening of the adsorption of Xe, Kr, and Xe/Kr mixtures in 12,020 experimental open-framework materials, in order to provide a better comprehension of the thermodynamics behind Xe/Kr separation in nanoporous materials and the microscopic origins of Xe/Kr selectivity at both low and ambient pressure.

The statistical correlation found between Henry's constant for Xe and Xe/Kr selectivity showed that the most selective materials are those with the highest affinity for xenon. To some degree of accuracy, we conclude that directly screening for Kr adsorption or for xenon adsorption free energy may not be necessary for a coarse-grained evaluation of a nanoporous framework selectivity. This could help build more efficient screening methodologies, for example with multistage studies with a first rough selection on Henry's constant at a low computational cost, followed by more expensive GCMC simulations on the selected materials (a gain that can be between 5 and 10-fold in our setup). Furthermore, inspection of the correlations between enthalpy and entropy contributions at low pressure showed that the adsorption-based separation process in the open frameworks studied is mainly enthalpic in nature. We intend to extend the study in the future to other classes of nanoporous materials beyond MOFs, including covalent organic frameworks, porous aromatic frameworks, purely inorganic porous frameworks such as zeolites, but also amorphous porous materials such as porous polymer membranes.

In order to use nanoporous materials to separate xenon from krypton, pressure swing adsorption (PSA) processes have been widely offered: pressure is therefore a crucial thermodynamic variable in the separation cycle. Here, we studied the difference of selectivity between a system under very low pressure (at the zero loading limit, which is calculated at relatively low computational cost) and a system at ambient pressure (closer to working conditions, but obtained at higher simulation cost). We demonstrated that the selectivity could be highly dependent on the pressure, with high low-pressure selectivity that could be maintained in some materials at ambient-pressure selectivity, while in others there would be a large drop in selectivity: a high ambient-pressure selectivity requires high low-pressure selectivity, but the reverse does not hold.

Using a thermodynamic approach to describe the separation selectivity, we showed that the differences in selectivity between the different pressures (and therefore different loading regimes of the frameworks) are mainly explained by the evolution of the adsorption enthalpies for Xe and Kr. By focusing on specific examples, we uncovered the microscopic origins of these selectivity changes, and related them to the relative roles of host–guest and guest–guest interactions. Population of different adsorption sites, or repacking of the adsorbed phase at higher loading, can lead to drastic changes in the overall selectivity. The mechanisms behind selectivity at high pressure are complex and unique to each framework, requiring a good

understanding of the interactions between guest molecules constrained in the nanopores. Nevertheless, our classification of the interactions at play can help in the future to design more efficient high-throughput screening procedures.

For instance, the essentially enthalpic nature of the xenon/krypton separation process supports the need for more efficient ways of sampling the interaction energies and using them as cheap descriptors to tackle more and more numerous structures. In the next chapter, we will go over different ways of evaluating the adsorption enthalpy by comparing the computation time required and the accuracies of each of them. Finally, the influence of the partial pressure through the change in composition or in pressure questions the possible use of infinite dilution thermodynamic quantities to predict the selectivity at any pressure (GCMC). Many studies have focused on predicting the results of GCMC simulations;**Simon_2015**, **Shi_2023**, **Kang_2023**, **Li_2023** by using this new angle, can we achieve good results in predicting GCMC values?

Data Availability: https://github.com/fxcoudert/citable-data/tree/master/132-Ren_FaradayDiscuss_2021

3

ADSORPTION ENERGIES SAMPLING

3.1 Voronoi Sampling 71
3.1.1 Theoretical considerations 71
3.1.2 Implementation in a screening 73
3.1.3 Comparative study of the Voronoi sampling 74
3.1.4 Performance of a Voronoi energy sampling 77
3.2 Rapid Adsorption Enthalpy Surface Sampling (RAESS) 78
3.2.1 Initial implementation 79
3.2.2 Performance improvement of the algorithm 81
3.2.3 Final surface sampling implementation 84
3.2.4 Surface sampling application use cases 88
3.2.5 Perspectives of surface sampling 93
3.3 Grid Adsorption Energies Descriptors (GrAED) 94
3.3.1 Implementation of an efficient grid algorithm 94
3.3.2 Performance on the adsorption equilibrium 96
3.3.3 Performance on the exchange equilibrium 99
3.3.4 Description of the ambient-pressure selectivity 101

3.1 Voronoi Sampling

3.1.1 Theoretical considerations

In mathematics, a tessellation of a given space corresponds to a partition into non-overlapping subspaces. In the Voronoi tessellation, named after Georgy Feodosevich Voronoy, a set of points (seeds) are associated to a tessellation of regions (Voronoi cells) so that each seed has a cell whose points are closer to this seed than any other seeds.[Rycroft_2009] Applied in materials science, the Voronoi cells associated to each atom of the framework can be used to determine key geometrical descriptors (void volume, accessible surface area, pore sizes). This decomposition can also be used to sample adsorption energies as introduced by Simon et al. — an average of the interaction energies was calculated on the accessible vertices of each Voronoi cell.[Simon_2015]

Equal radii

In a tridimensional space, let us consider the positions $(\boldsymbol{x}_k)_{k\in\{1,\dots,n\}}$ of the n points in a box B that could be periodically propagated in the whole space. For every $k \in \{1,\dots,n\}$, we can then define a subspace S_k (also called Voronoi cell) around the atom k so that any point $\boldsymbol{x}$ inside this subspace is closer to the position $\boldsymbol{x}_k$ than to any other points $\boldsymbol{x}_l$ ($l \neq k$).

$$\mathrm{S}_k = \left\{\boldsymbol{x} \in \mathrm{B} \mid \forall\, l \neq k,\ \|\boldsymbol{x}-\boldsymbol{x}_k\|_2 \leq \|\boldsymbol{x}-\boldsymbol{x}_l\|_2\right\} \tag{3.1}$$

The set of all these 3D polyhedral subspaces S_k is then called the Voronoi partition of the space B. The edges and vertices of these polyhedra can then give valuable information of the void space between the adjacent Voronoi cells associated to them. We can quickly use them to determine the accessible and inaccessible points of the void space. For instance, a vertex $\boldsymbol{v}$ of p subspaces $\left\{\mathrm{V}_{i_1},\dots,\mathrm{V}_{i_p}\right\}$ is the closest point to the atomic positions $\boldsymbol{x}_{i_1},\dots,\boldsymbol{x}_{i_p}$ — this can simply be proven by combining the different conditions of equation **??**. The same assessment can be done for any point on an edge adjacent to some subspaces, it will be closer to the atoms associated with these subspaces than to any other atoms.

Figure 3.1: Bidimensional illustrations of a Voronoi decomposition using three types of algorithm: (i) for equally sized circles using the equation **??** *(www.shadertoy.com/view/MslGD8) (ii) for unequally sized circles using the Apollonian Voronoi decomposition condition* **??** *(www.shadertoy.com/view/4sd3D7) and (iii) another algorithm for unequally sized circles using the radical Voronoi condition* **??** *(www.shadertoy.com/view/4tV3z3). Note that the second picture shows the curved boundaries between the Voronoi cells, while the switch to the radical Voronoi decomposition gives straight line boundaries.*

This regular Voronoi tessellation can only be used to separate the space for equally sized atoms because it sets the boundaries at equidistance of all the surrounding atoms, as we can see on the Figure **??**. For unequally sized atoms, this type of definition could be undesired since the boundary can be closer to the surface of an atom than of another. The initial thought behind using a Voronoi decomposition is based on delimiting a region for each atom that is closer to this atom than any other one. The ambiguity of this definition relies on the definition of "closeness". In this regular Voronoi decomposition, we define the closeness using the distance between the center of mass of the different atoms, which is problematic for unequally distributed radii.

Unequal radii

The last definition of the Voronoi decomposition works only for equal-sized atoms because the closest region to an atom is also the closest to its center of mass, which does not apply to the complex atomic structures of nanoporous frameworks. To model the atomic radii $r_1, \ldots, r_n$ of the points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ in the same box B, we can implement a so-called Apollonian Voronoi diagram.[voronoi_apollonian] For every $k \in \{1, \ldots, n\}$, the new subspaces A_k are defined as follows:

$$\mathrm{A}_k = \left\{ \boldsymbol{x} \in \mathrm{B} \mid \forall\ l \neq k,\ \left\| \boldsymbol{x} - \boldsymbol{x}_k \right\|_2 - r_k \leq \left\| \boldsymbol{x} - \boldsymbol{x}_l \right\|_2 - r_l \right\} \tag{3.2}$$

This new definition of the Voronoi diagram has been built on the intuitive property that stipulates that the subspace is the set of points closest to the surface of the sphere around the associated atom. It deals therefore with an unequal distribution of atomic radii, because it is now dependent on the radii. However, as we can see on the Figure **??**, this first implementation has a very convenient definition, but the edges of the subspaces are curved, which makes it harder to use computationally.

For this reason a less intuitive implementation is more commonly used instead, which is called the radical Voronoi tessellation or power diagram or Laguerre-Voronoi diagram.[aurenhammer_1987] As we can see on the Figure **??**, the subspaces obtained using this method are now convex polygons with straight edges instead of curved ones. The condition is way less intuitive because the condition does not rely on a simple definition. The subspaces V_k are now defined by the following condition:

$$\mathrm{V}_k = \left\{ \boldsymbol{x} \in \mathrm{B} \mid \forall\ l \neq k,\ \left\| \boldsymbol{x} - \boldsymbol{x}_k \right\|_2^2 - r_k^2 \leq \left\| \boldsymbol{x} - \boldsymbol{x}_l \right\|_2^2 - r_l^2 \right\} \tag{3.3}$$

In addition to the polyhedral form of the Voronoi cells, this new implementation presents interesting properties for porosity calculations in a framework of unequal spheres like in MOFs of zeolites.[voronoi_radical] First, the boundary between two overlapping spheres corresponds simply to the intersection place between the spheres. Then, the boundary between non-overlapping spheres is always in the void space between the spheres. This can be simply proven by taking a point $\boldsymbol{x}$ in V_k and outside the sphere, we have $\left\| \boldsymbol{x} - \boldsymbol{x}_k \right\|_2 \geq r_k$, which implies $\forall\ l \neq k,\ \left\| \boldsymbol{x} - \boldsymbol{x}_k \right\|_2 \geq r_k$. The point $\boldsymbol{x}$ is therefore also not overlapping with any other atom, which means it is in the void space of the framework.

If we now consider a point $\boldsymbol{v}$ on a boundary between p Voronoi cells $\left\{ \mathrm{V}_{i_1}, \ldots, \mathrm{V}_{i_p} \right\}$, this point would verify all the conditions $\left\| \boldsymbol{x} - \boldsymbol{x}_{i_1} \right\|_2^2 - r_{i_1}^2 = \cdots = \left\| \boldsymbol{x} - \boldsymbol{x}_{i_p} \right\|_2^2 - r_{i_p}^2 = \mathrm{C}$. We can know their minimum distance to the center of mass of all nearby atoms to test for possible overlapping. More specifically, in the Zeo++ software,[Zeo++] the Voronoi diagram is characterized by storing the minimum distance to the closest atoms and the index of the atoms for every vertices and edges (for edges that connect two different periodic images a periodic displacement vector is also stored). This information can be used to speed up the void fraction calculation, by skipping the volume calculations in the non-adsorbable Voronoi cells. It is also a fast way of determining the accessible and non-accessible surface areas and volumes.[Zeo++] Note that if the probe has a radius r_{probe}, then the sphere radii considered are $r_k = r_{\mathrm{atom}} + r_{\mathrm{probe}}$.

3.1.2 Implementation in a screening

The use of the Voronoi decomposition of the pore space of materials for their geometric characterization has been widely employed in computational studies in the last decade,[Willems_2012] especially since it was made easily available as part of the Zeo++ software package.[Pinheiro2013] Its use was extended recently to implement a novel sampling scheme, in a study that introduces the ML-assisted screening of nanoporous materials for xenon/krypton separation. In this article, Simon et al.[Simon_2015] relied on a Voronoi tessellation of the nanoporous materials and assigned the potential adsorption sites (i.e., the sampling points) at the nodes of this decomposition. The Voronoi tessellation identifies the vertices of polyhedra that correspond to the closest regions of each atom of the structure. These vertices (or *Voronoi nodes*) are the points equidistant to at least four atoms of the structure, and they can be associated with adsorption sites since they are positioned near the center of the pores.

The definition of accessibility defined by the Zeo++ software was used in a screening to find the best materials for Xe/Kr separation.[Simon_2015] The interaction energies of xenon were calculated only on the accessible nodes as schemed on the Figure **??**. The average of the energies at the accessible Voronoi nodes gives an approximation of the adsorption enthalpy. However, this sampling assumes that the nodes are close to the real, most favorable, adsorption sites. Or to put it differently, the adsorption sites need to be at the center of the pores, which is only true for structures with pore sizes close to adsorbate size. This newly defined adsorption energy descriptor was found to be one of the most influential descriptors for the ML model developed to predict the ambient-pressure selectivity.

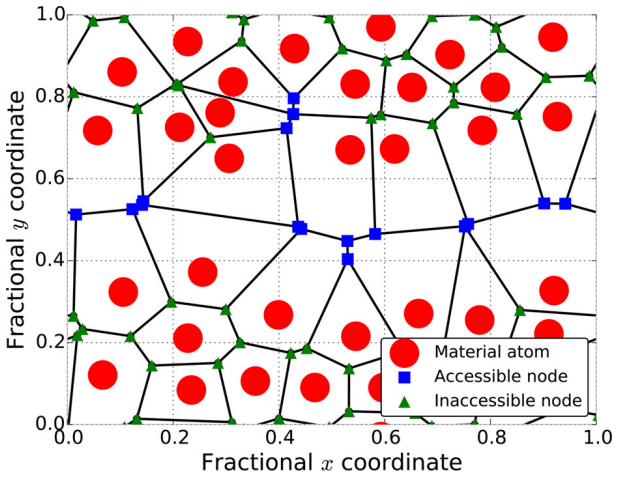


*Figure 3.2: Voronoi network model of void space (2D caricature). The unit cell of a toy material is shown. Red circles represent atoms of the material; accessible and inaccessible Voronoi nodes are blue squares and green triangles, respectively. The black lines are the edges in the periodic Voronoi graph that model the void space. Reprinted with permission from Ref. [**Simon_2015**]. Copyright 2015 American Chemical Society.*

Starting from the initial idea of Voronoi sampling, we want to question the relevance of using a direct average of the interaction energies instead of a Boltzmann averaging in describing the adsorption enthalpy. To better understand the strengths and weaknesses of this methodology, we compared different ways of approximating adsorption enthalpies using the Voronoi sampling

with more accurate infinite dilution and ambient-pressure xenon adsorption enthalpies using Widom insertions and GCMC for a 20:80 Xe/Kr mixture at 1 atm and 298 K.

3.1.3 Comparative study of the Voronoi sampling

We introduced in the previous chapter, the definition of the xenon adsorption enthalpy at infinite dilution (Widom insertion section **??**) and at ambient pressure (GCMC sections **??** and **??**). These methods can be considered accurate methods to calculate the adsorption enthalpy which has been proven to be strongly correlated to the logarithm of the selectivity in our previous study on the thermodynamic exploration of the xenon/krypton separation using a high-throughput screening.

Introduction of the main concepts

The Voronoi energy as initially conceptualized by Simon et al. is based on the average of the xenon interaction energies at the accessible Voronoi nodes. But because we will compare to thermodynamic simulations without blocking pockets, we will not consider the accessible Voronoi nodes but the adsorbable ones since it is closer to the simulation we want to approximate. For simplicity, we define the set of the adsorbable Voronoi nodes A as the Voronoi nodes with a negative energy value, we also apply a condition on the minimum distance (about the radius of a xenon 2 Å) to reduce the number of points to be calculated. This average on the adsorbable Voronoi nodes $E^{Xe}_{\text{voro-A}}$ can be written:

$$E^{Xe}_{\text{voro-A}} = \sum_{i \in A} E_i - RT \tag{3.4}$$

Another interesting energy descriptor could simply be the minimum of the interaction energies among the Voronoi nodes V with a minimum distance to the nearest atom is higher than 2 Å. This minimum Voronoi energy $E^{Xe}_{\text{voro-M}}$ can be written:

$$E^{Xe}_{\text{voro-M}} = \min_{i \in V} E_i \tag{3.5}$$

Finally, to get closer to the definition of the heat of adsorption defined in the previous chapter, we can build an energy descriptor using a Boltzmann averaging. This Boltzmann average of the xenon interaction energies at the Voronoi nodes V written $E^{Xe}_{\text{voro-B}}$ can be expressed as follows:

$$E^{Xe}_{\text{voro-B}} = \frac{\sum_{i \in V} E_i e^{-\beta E_i}}{\sum_{i \in V} e^{-\beta E_i}} - RT \tag{3.6}$$

Note that the −RT term is to make the expression comparable to the one of adsorption enthalpy.

We can intuitively say that the Boltzmann averaging being closer to the definition of the adsorption enthalpy it would be a better candidate as an energy descriptor, hence improving the screening methodology. To test these different methodologies, we will now compare the different energy descriptors to more accurate evaluation of the adsorption heat.

Low-pressure comparison

The Widom insertion is typically used to calculate the infinite dilution adsorption properties such as the adsorption enthalpy, the Henry constant and the selectivity. The evaluation of the interaction energies of xenon at the different Voronoi nodes correspond to a low-pressure averaging and is more comparable to a Widom insertion method. It is, however, biased by the inhomogeneous sampling of the space, which can explain some discrepancies we could observe.

Note that in this chapter we will mainly use the standard pore size definition we find in the literature that is based on the atom radii defined by the Cambridge Crystallographic Data Centre (CCDC). This pore size will only serve a labeling purpose and will help us classify the structures according to their relative size. It has qualitative role mainly, which justifies the fact that we are not using a much more accurate definition based on the forcefield like in the previous chapter.

As we can see on the Figure **??**, the average of the energies is not performing very well and is clearly less correlated to the adsorption enthalpy than the minimum interaction energy or the Boltzmann average of the interaction energies. This is because in a normal average, the high-energy values have a much higher weight than in a Boltzmann average, which makes the average much more important than expected. The Voronoi average descriptor $E^{Xe}_{\text{voro-A}}$ is always higher than the infinite dilution adsorption enthalpy $\Delta_{\text{ads}}H^{\text{Xe}}_0$.

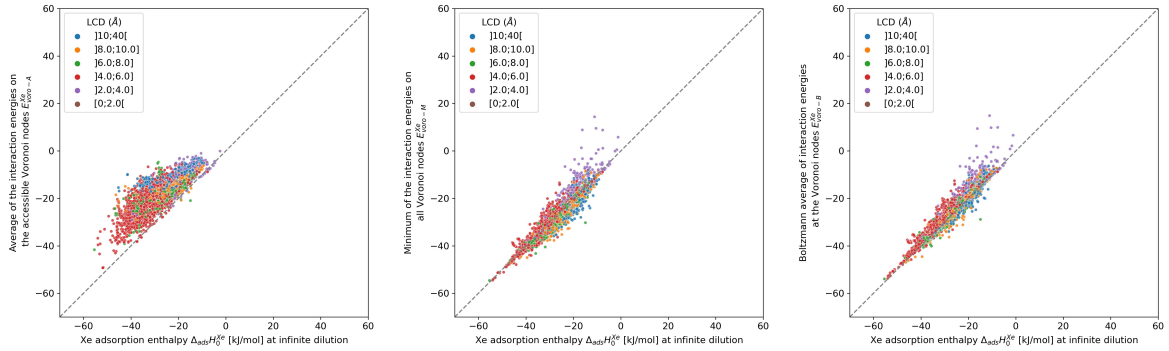


Figure 3.3: Scatterplots of the energy descriptors $E^{Xe}_{voro\text{-}A}$, $E^{Xe}_{voro\text{-}M}$ and $E^{Xe}_{voro\text{-}B}$ calculated by a Voronoi sampling compared to the enthalpies calculated by a 100k-step Widom insertion simulation of xenon in structures of CoRE MOF 2019. The points are labeled according to the largest cavity diameter (LCD_{CCDC}) belonging to one of the intervals.

The Pearson correlation coefficients corroborate our initial observation. The correlation coefficient between $E^{Xe}_{\text{voro-A}}$ and $\Delta_{\text{ads}}H^{\text{Xe}}_0$ is equal to 0.81, whereas for the minimum $E^{Xe}_{\text{voro-M}}$ it is 0.95 and for the Boltzmann average $E^{Xe}_{\text{voro-B}}$ it is 0.97. For this reason, to evaluate the relevance of a Voronoi energy sampling we should consider a Boltzmann average. As we have shown in the previous chapter, the selectivity is correlated to the difference of adsorption enthalpies of xenon and krypton. Better describing the enthalpy is a first step toward a better description of the selectivity. However, we only looked at the selectivity values at low pressure. What would happen for selectivity at higher pressure?

On the Figure **??**, we can see that the selectivity drop between the low-pressure case and the ambient-pressure one also impacts the enthalpy values of xenon. There is a reduction

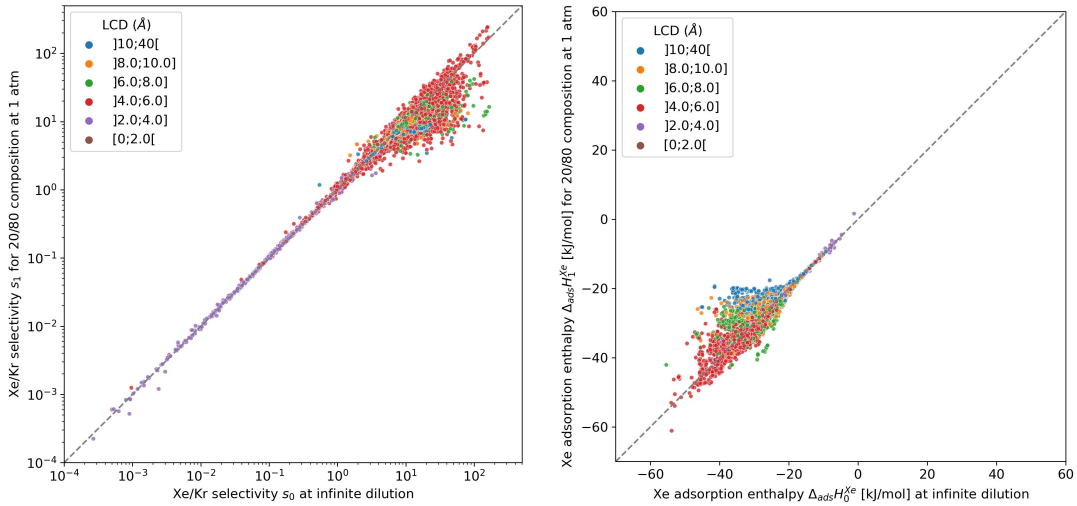


Figure 3.4: Comparison of the ambient-pressure and low-pressure case through two thermodynamic quantities: the Xe/Kr selectivity (left) and the xenon adsorption enthalpy (right).

of the xenon affinity as the pressure goes up. Since the study of Simon et al. focused on the ambient-pressure selectivity prediction, we want to see if the energy descriptor they developed could, however, be used to describe the adsorption enthalpy at high pressure.

Ambient-pressure xenon/krypton separation

If we look at the Figure **??**, it is not so clear which descriptor is better to describe the enthalpy at ambient pressure. The correlation shown by the different scatterplots seems to be equally poor, and this could justify the use of a regular average rather than a Boltzmann average. The correlation coefficient for the average $\mathrm{E}^{Xe}_{\text{voro-A}}$ is now equal to 0.86, which is equivalent to the one of the minimum $\mathrm{E}^{Xe}_{\text{voro-M}}$ and slightly lower than the 0.87 for the Boltzmann average $\mathrm{E}^{Xe}_{\text{voro-B}}$.

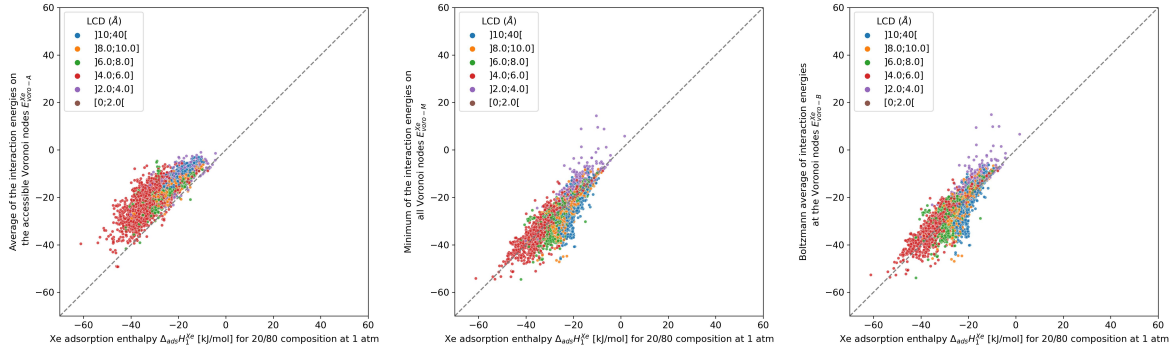


Figure 3.5: Scatterplots of the energy descriptors $\mathrm{E}^{Xe}_{\text{voro-A}}$, $\mathrm{E}^{Xe}_{\text{voro-M}}$ and $\mathrm{E}^{Xe}_{\text{voro-B}}$ calculated by a Voronoi sampling compared to the enthalpies calculated by a 100k-step GCMC simulation of xenon in structures of CoRE MOF 2019. The points are labeled according to the largest cavity diameter (LCD_{CCDC} or D_i) belonging to one of the intervals.

At higher pressure, the adsorption enthalpy has higher values, which degrades the correlation with the Boltzmann average and the minimum of the interaction energies evaluated at the Voronoi nodes. For the averaging, the over-evaluation of the energy values make it closer to the

values at higher pressure. Following this idea, we later come up with another idea of averaging with bigger weights on the higher values, a Boltzmann average with a higher temperature value that will be tested in the next chapter.

3.1.4 Performance of a Voronoi energy sampling

We will now focus on some performance metrics associated to the Boltzmann average at the Voronoi nodes and compare it to our reference sampling, the Widom insertion with 100,000 cycles. The right plot of the Figure **??** compares the enthalpy computed in the Voronoi sampling with the reference adsorption enthalpy (ground truth) — showing at the same time the largest cavity diameter for each porous framework. The correlation between the values of enthalpy is very good only for a restricted number of structures with enthalpy around $-50\,\mathrm{kJ\,mol^{-1}}$. For structures with higher enthalpy, the correlation starts to degrade, and becomes very poor for small-pore structures. For the points in purple, the largest cavity diameter is lower than the kinetic diameter of a xenon, the sampling of the Voronoi nodes is clearly insufficient. In addition, the accuracy loss on the other points (larger pores) can be explained by the fact that the pores are slightly bigger and the center of the pore is not a good approximation of adsorption site position anymore: the adsorption sites are actually closer to the pore surface than to the center of the pore. This conclusion is what prompted us to propose a new sampling scheme based on the molecular surface of the pore space, which we will detail in the next sections.

The root mean squared error (RMSE) and the mean absolute error (MAE) for Voronoi sampling are respectively $6.78\,\mathrm{kJ\,mol^{-1}}$ and $2.01\,\mathrm{kJ\,mol^{-1}}$, if we consider all structures in our set, which seems too high to be useful for screening purposes. However, the non-porous materials would be screened out *a priori* in any high-throughput workflow, as they would not be of interest. We can only consider the structures with large enough cavities, higher than 3.7 Å (a bit lower than 3.96 Å Xe kinetic diameter). Thereby, the RMSE and MAE drop respectively to $2.11\,\mathrm{kJ\,mol^{-1}}$ and $1.55\,\mathrm{kJ\,mol^{-1}}$, which can be considered acceptable for a quick estimation of the guest–host affinity, but not for accurate adsorption enthalpy calculation.

This is reinforced by the very low computational cost of the method. The Voronoi tessellation done by the Zeo++ software is extremely quick and can output the positions of the Voronoi nodes in 0.28 s (measured as an average over all the structures of the CoRE MOF 2019 database), on a typical workstation (a single Intel Xeon Platinum 8168 core at 2.7 GHz). While a simple Python for the energy calculation took around 27 s per structure, we benchmarked that a C++ optimized implementation can perform the Voronoi sampling in around 0.4 s. We only need to remember that this method takes a few hundred milliseconds per structure, while a Widom insertion needs approximately hundreds of seconds per structure. Voronoi sampling is therefore 2 to 3 orders of magnitude quicker than a full sampling of the pore space.

This preliminary study identified a fast method for adsorption enthalpy calculation that can be widely used in screening procedures, but has limited accuracy for quantitative prediction — this sampling technique assumes that the nodes are close to the real, most favorable, adsorption sites, which is not always true. Or to put it differently, the adsorption sites need to be at the center of the pores, which is only true for structures with pore sizes close to adsorbate size. It raised important questions on the importance of selecting sampling points within the pore space of materials, and we wanted to develop an intermediate technique that is both fast

and accurate for the prediction of adsorption enthalpy. For this purpose, we developed and optimized a new sampling technique that focuses the sampling on the surface of the material, which is expected to make up for the main flaws of the Voronoi sampling.

3.2 Rapid Adsorption Enthalpy Surface Sampling (RAESS)

In this section, we describe the development of our surface sampling algorithm, with the goal of being more accurate than Voronoi sampling and faster than Widom insertion. Our initial idea is based on a series of theoretical considerations: (i) the strong adsorption sites are near the surface of the material; (ii) by changing the problem from 3D to 2D sampling we can reduce the complexity; and (iii) the algorithm can scale with the number of unique atoms in the structure (and not with the size of the unit cell), which is efficient because many porous frameworks have high symmetry. The first consideration ensures that this method will be more accurate than a Voronoi sampling, and the last two made us think that a well-optimized code would be fast. To confirm these hypotheses, we will analyze both the accuracy and the speed of this new algorithm and compare them to existing methods.

3.2.1 Initial implementation

We present here our initial implementation of the surface sampling algorithm and its principles. This first implementation is a relatively basic one and already performs well compared to the other methods. In the next sections, we refine it with two additional features that will improve its accuracy and its speed.

This initial implementation speeds up the calculation of adsorption enthalpy in nanoporous materials by sampling interaction energies only near the surface. It is illustrated in Figure **??**. For this purpose, a loop over all unique atoms (as defined by crystalline symmetry) is performed. And for each atom, a sphere around its position is sampled using a uniform distribution around it, these points will be called sampling points, and we can change the number of sampling points. The default radius chosen for the sampling spheres is the distance $r_{\min} = 2^{1/6}\sigma_{ij}$ to the minimum of the LJ potential between atoms of type i (belonging to the framework) and j (the guest), corresponding to the strongest possible pair interaction (although the neighboring atoms will, of course, have an influence). After calculating the interaction energy at each of the sampled points, a Boltzmann average of these energies corresponds to a biased adsorption enthalpy, as described by the equation **??**.

In order to validate the accuracy of the approximation made using this sampling, we applied this algorithm with 300,000 sampling points per unique atom. The results are illustrated by the Figure **??**. There is a good numerical agreement with the reference calculations, the RMSE and MAE are only around $0.90\,\mathrm{kJ\,mol^{-1}}$ and $0.66\,\mathrm{kJ\,mol^{-1}}$ considering all the structures from the database. Moreover, there is no noticeable difference of RMSE when considering the structures with a pore size above 3.7 Å (as determined by the $\mathrm{LCD_{CCDC}}$). Unlike Voronoi sampling, this method gives a consistent accuracy across all the structures of the database with a lower error. The fact that the RMSE error is below $1\,\mathrm{kJ\,mol^{-1}}$ is quite promising, and validates our intuition that this new sampling technique can be an intermediate between to the two previous methods (Voronoi and Widom).

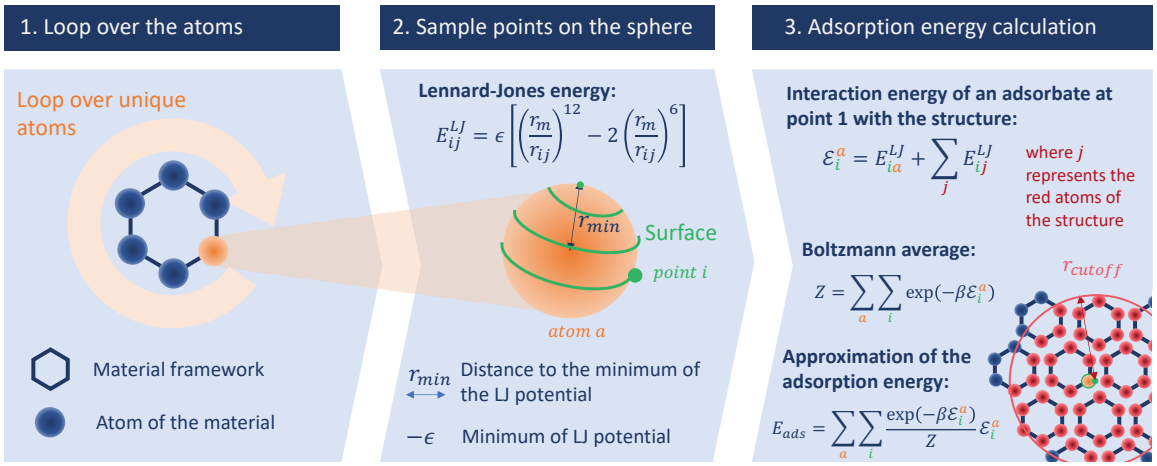


Figure 3.6: Schematic description of our surface sampling based on the three main steps of the algorithm: the loop over the unique atoms, the spiral sampling around each atom, and the energy averaging. The adsorbate is represented by the point i and is moved across all the points around the unique atoms of the structure.

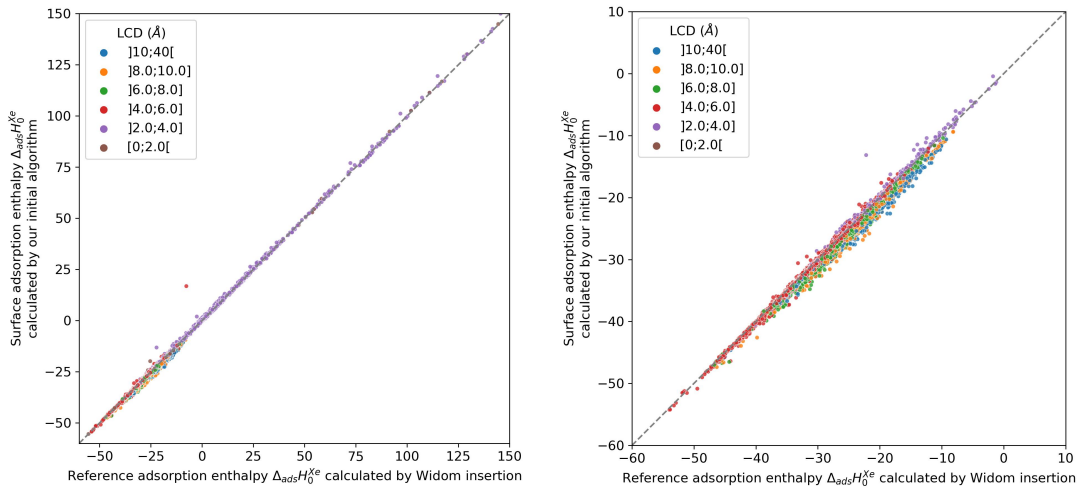


Figure 3.7: Scatterplots of the xenon surface adsorption enthalpy calculated by an initial implementation of the RAESS algorithm as a function of the xenon adsorption enthalpy calculated by a 100k-step Widom insertion simulation using two value windows. The second plot zooms on the negative values corresponding to the most selective materials.

After proving the good accuracy of the method, we are now exploring the computation time required. We see on Figure ?? that the method reaches an RMSE below 1.0 kJ mol^{-1} very quickly for an average CPU time of 1.2 s, corresponding to 2,000 sampling points per atom. This is far less than the 150 s required for a Widom insertion to be near its plateau value (converges to zero), for an RMSE of 0.10 kJ mol^{-1} with 12,000 cycles. Moreover, the Widom insertion needs around 14 s to reach a similar RMSE of 1.0 kJ mol^{-1}, which is still slower than the surface sampling. We can conclude that this initial implementation of the surface sampling is faster than a standard Widom insertion, with a good accuracy.

These results on the convergence speed and the limit values of the error can be simply rationalized by the nature of each sampling. In a surface sampling, the sampled points are biased toward the most attractive points of adsorption for the xenon, which explains the fact that the value converges very quickly because the most influential terms of the Boltzmann average are quickly gathered. However, in a Widom insertion every point of the space has equal chances of being sampled, which is extremely close to the definition of the enthalpy but requires much more time to randomly sample a very attractive adsorption site. The surface sampling by its biased nature, however, will inherently be less accurate since not all points are considered equally and sometimes the most optimal adsorption site could be missed, because it could in some cases be further from the sampled surface.

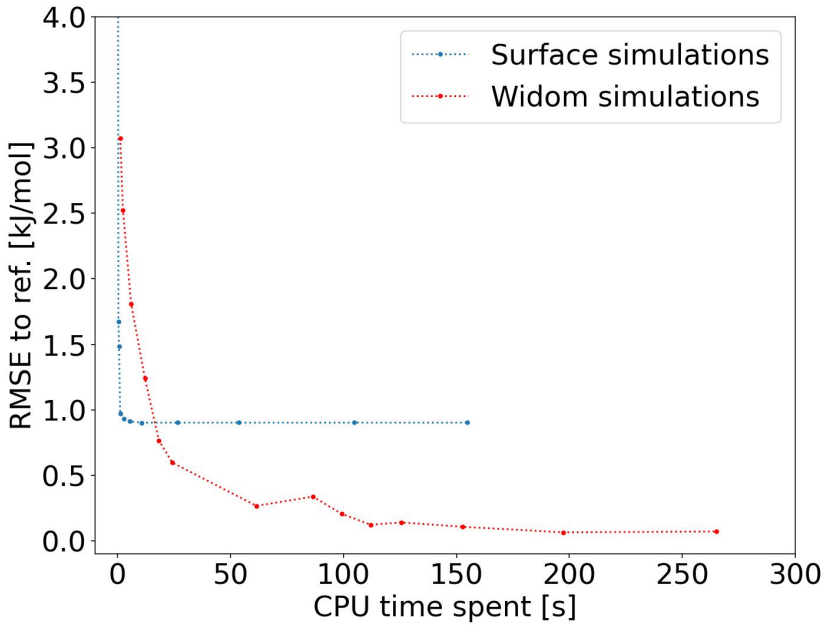


Figure 3.8: Convergence plot of the RMSE on the adsorption enthalpy for our algorithm (blue) compared to a 100k-step Widom insertion simulation (red) for xenon adsorption in all structures of the CoRE MOF 2019 database.

However, this initial implementation of the method is slower than a Voronoi sampling that only needs to sample around 1,600 points on average, instead of 13,000 sampled points on average (if we multiply by the average number of unique atoms). The sampling part would take approximately 0.15 s and the Voronoi nodes generation 0.28 s, so our surface sampling algorithm remains 2 to 3 times slower (implemented in an identical compiled language, in this case C++). In order to improve the accuracy and performance, we have further tweaked the surface sampling method, adjusting the size of the sampling sphere and adopting a fast rejection criterion. The rejection of high-energy points with little contribution to the final enthalpy value can reduce the simulation time, whereas the size of the sampling sphere can improve the accuracy. The initially chosen sphere size is only taking account of the interaction with the closest atom, we therefore chose to set it at the minimum of Lennard-Jones potential. However, the interaction with the neighboring atoms can further stabilize the adsorbate, sampling further from this minimum could in consequence increase the accuracy of our surface sampling method.

3.2.2 Performance improvement of the algorithm

Size of the sampling sphere

The validity of the initial algorithm is based on the assumption that the adsorption site is at the minimum of the Lennard-Jones potential. It will only perform well if the closest atom contributes to almost all the interaction, but in real frameworks other neighboring atoms contribute to the host/guest interaction as well. We have found that in vast majority of materials, the adsorption sites are located farther apart compared to the LJ potential minimum, in order to maximize the contribution of all atoms — and because of the dissymmetry of the interaction potential well. In order to see if this could be introduced in our algorithm, we implemented a parameter λ, and the sampling sphere radius is now defined by $R_\lambda = \lambda\sigma$, where σ is the distance at which the LJ potential is zero. If $\lambda = 2^{1/6}$, we fall back to our initial definition of the sampling sphere, and the adsorbent is at the minimum of the LJ potential of the atom. If $\lambda = 1$, the sampling sphere is at the zero of the LJ potential, and by increasing this parameter, we can check if our intuition was right.

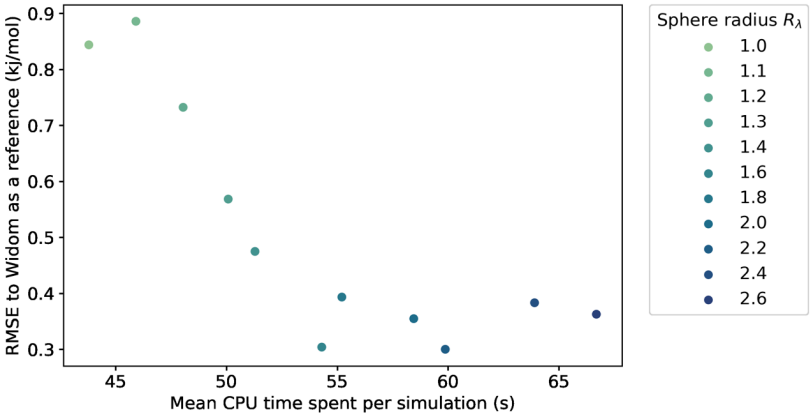


Figure 3.9: Influence of the sampling sphere radius R_λ on the average CPU time required for a simulation of 100k sampling points and the RMSE, compared to the reference adsorption enthalpy. The averaging is done only on the structures with a largest cavity diameter (LCD_{CCDC}) higher than 3.7 Å.

Because we have no physical model that would predict the optimal value of the sampling sphere, we followed a statistical approach. We studied the influence of the λ parameter on both the accuracy and the computation time, and the results are represented on Figure **??**. The RMSE turns out to be relatively high around $0.90\,\mathrm{kJ\,mol^{-1}}$ for radius sphere lower than the $r_{\min}$, it then decreases for larger values of radius to reach a plateau around $0.35\,\mathrm{kJ\,mol^{-1}}$. We confirm that by increasing the sampling sphere radius we can improve the accuracy of our algorithm, and find that for values of λ higher than 1.6, the accuracy is stabilized. We also find that increasing the sphere radius negatively impacts the computational efficiency, since it increases the number of neighbors considered in the energy calculation.

By choosing an optimal sampling sphere, we can more than halve the error, while increasing the computation time by around 20 percent, when comparing the case $\lambda = 1.6$ with $\lambda = 1.1$ (close to $r_{\min}$). In most cases, it will be an acceptable trade-off. However, in a case where the computation time is crucial, like in a rapid screening, the optimal choice might not be to increase the sampling sphere at $\lambda = 1.6$ but to have it lower at $\lambda = 1.4$ or $\lambda = 1.2$, and have

an RMSE around 0.5 kJ mol^{-1} — still quite acceptable. The new scale parameter introduced in this section can therefore be tweaked to serve the users' purpose, whether it is to focus on the accuracy or to optimize the computation speed. If one wants to use it on a completely different database in very different conditions, then one can either choose a default value that works fine (*e.g.* $\lambda = 1.4$) or one can optimize the parameter on a small diverse sample of the unseen data.

Rejection condition

As shown above, our algorithm has better accuracy than Voronoi sampling, but its initial implementation was several times slower, which could be unsuitable for screening applications in high-throughput workflows, where the number of structures to be screened can reach one million or more. To reduce the computational expense, we thought of rejecting the points with little contribution to the final enthalpy, i.e., the largely positive interaction energies that would vanish in the exponential of the Boltzmann average.

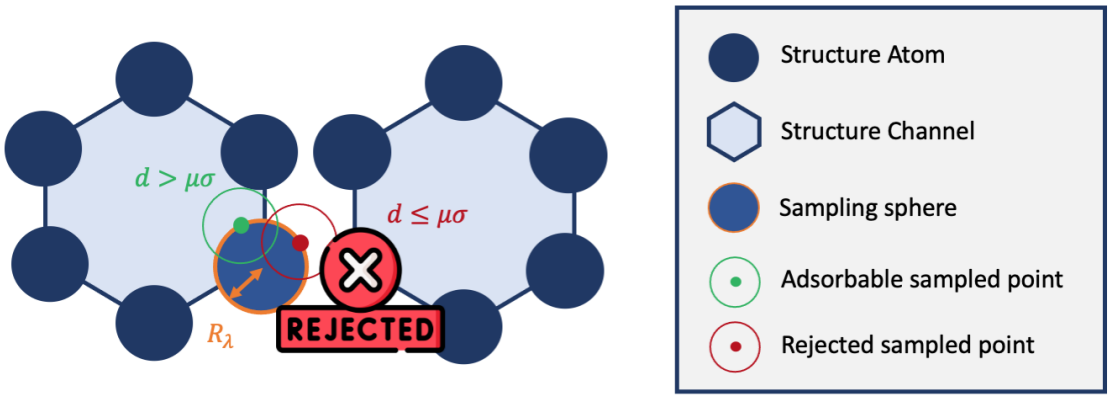


Figure 3.10: Simplified representation of the principle of rejection condition and the concept of sampling sphere inside 2D channels of a nanoporous material.

Inspired by typical methods for accessible surface calculation, we implemented a hard sphere rejection condition based on the distance to neighbors. If the adsorbate is too close to another atom of the structure, the sampling point is rejected, i.e., its energy is not calculated (or considered to be infinite). We based this distance threshold on the σ_{ij} parameter of the Lennard-Jones potential. To determine the optimal threshold, we introduced a factor μ with real values between 0 and 1, that changes the size of the hard sphere rejection condition. If the guest–host distance is lower than $d_\mu = \mu \times \sigma$, then the point is rejected. If $\mu = 0$, then there is no rejection condition. And if $\mu = 1$, we reject all points with a positive energy interaction to at least one atom of the structure. This condition could be a bit strong and points with non-negligible contribution would end up rejected. This rejection condition is schematically represented on Figure **??**.

This rejection condition is expected to speed up the calculation, since the energy calculation is avoided for the rejected sampling points. The energy calculation accounts for the largest portion of the CPU time spent in the surface sampling. For the structure KAXQIL,[Banerjee_2012] the Lennard-Jones potential calculation represents up to 90% of the calculation time for 100,000 sampling points per sphere (with the initial algorithm). The higher the factor μ, the more rejections there would be. But, if too many points are rejected, the accuracy would drop. Here again, we used a statistical analysis to determine the optimal value of μ, making our sampling

faster without compromising the accuracy of the enthalpy calculation. The results are displayed on Figure **??**.

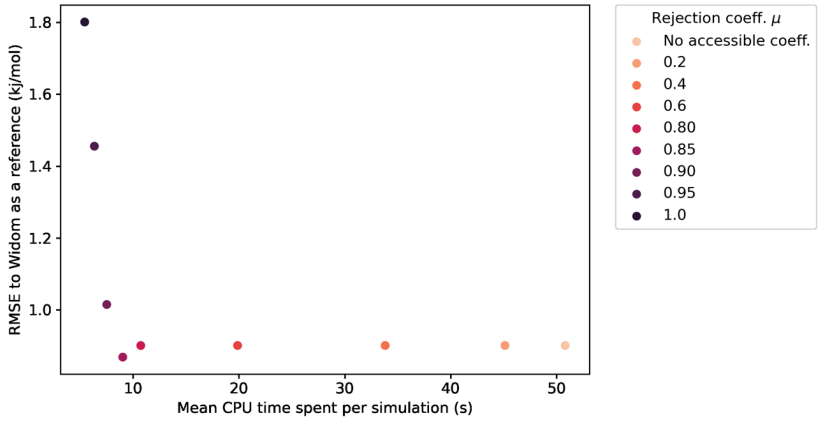


Figure 3.11: Influence of the rejection coefficient μ on the average CPU time required for a simulation of 100k sampling points and the RMSE compared to the reference adsorption enthalpy. The averaging is done only on the structures with a largest cavity diameter (LCD_{CCDC}) superior to 3.7 Å.

The values of RMSE and time on Figure **??** are averaged only on the most interesting structures for xenon adsorption ($LCD_{CCDC} \geq 3.7$ Å). For $\mu \leq 0.85$, increasing the value of μ improves the speed of the calculation without changing the RMSE.[1] For high values of μ, the rejection condition is too strong and we reject points with non-negligible contribution to the overall enthalpy. The RMSE increases as a consequence. If we want to keep the accuracy unchanged, the optimal value is therefore $\mu \simeq 0.85$, because it gives the lowest computation time with a similar RMSE. We note that it would be possible, in specific cases, to explore higher values of μ that trade a bit more accuracy in exchange for further speed gains.

For the simulations considered in Figure **??**, the use of a rejection condition $\mu = 0.85$ makes the simulation four times faster than the standard algorithm. As we will see in the next section, the combination of optimal values for the λ and μ parameters generates an algorithm with very interesting performance compared to Voronoi sampling or Widom insertion.

3.2.3 Final surface sampling implementation

Performance comparison

For the calculation of adsorption enthalpy, our proposed surface sampling method is a good compromise between the accuracy of Widom insertion (full sampling of the porous space) and the speed of a less accurate method such as Voronoi sampling. The performance of our algorithm, including the two new features (sampling sphere scaling and rejection criterion) is illustrated in Figure **??**, where we can see the improvement brought by each feature and how it compares to reference simulations. All CPU times are calculated using the smallest possible number of sampling points so that the respective algorithms reach convergence. With the implementation of a rejection condition, we find that surface sampling is even quicker than

[1] In fact, what we observe is a deterioration of the accuracy for structures with small pores because the probability of rejection in a confined space is really high and all sampled points end up rejected. But these points are not considered if we apply the condition on the cavity size ($LCD_{CCDC} \geq 3.7$ Å).

Voronoi sampling. Moreover, the increase of the size of the sampling sphere makes the surface sampling much more accurate, reaching an RMSE of 0.33 kJ mol^{-1} and an MAE of 0.21 kJ mol^{-1}. The ideal set of parameters, determined for porous materials from the CoRE MOF 2019 database, is (λ = 1.6, μ = 0.85) in order to combine the lowest error and smallest computational cost. By combining both of these new features to the algorithm, we have a final surface sampling method with an RMSE of 0.33 kJ mol^{-1} and an average computation time of 0.34 s per structure. According to the data represented in Figure **??**, it is about 6 times more accurate and 26% faster than Voronoi sampling, and it is also about 430 times faster than a Widom insertion with 12k cycles.

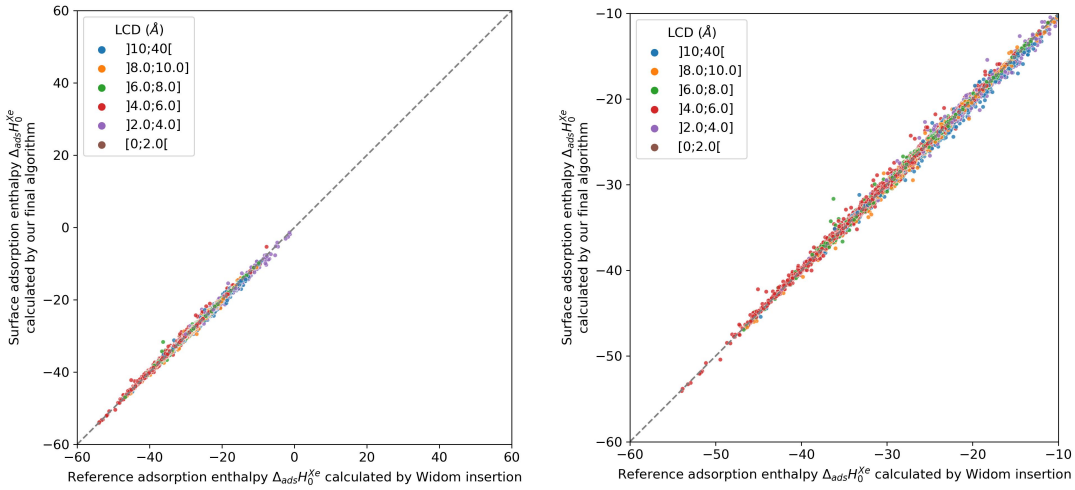


Figure 3.12: Scatterplots of the xenon surface adsorption enthalpy calculated by the final RAESS algorithm (λ = 1.6 and μ = 0.85) as a function of the xenon adsorption enthalpy calculated by a 100k-step Widom insertion simulation using two value windows, in structures of CoRE MOF 2019 with $LCD_{CCDC} \geq$ 3.7 Å at 298 K. The second plot zooms on the negative values corresponding to the most selective materials.

Finally, we suggest that the values of the parameters optimized in this work might need adjustment when applied to other adsorption systems. The optimal μ parameter depends on the size of the adsorbent, and it should be tweaked differently when considering another adsorbent. For instance, the set of structures used for the optimization of μ depends on the size of their cavities, and the 3.7 Å threshold chosen here would need to be changed according to the kinetic diameter of the adsorbate. Furthermore, as aforementioned in the section on the rejection condition, it is possible to trade off a bit of accuracy for faster simulations especially in high-throughput screenings where speed is extremely important. Similarly, in the case of xenon, the cost of increasing the sphere size is around 10 to 20%. On very large databases, one could consider that this increase on the required computational time is not worth the accuracy improvement, and one could decide to keep a smaller sampling sphere. If this method is transposed to different molecular systems, its parameters should be tested on the specific database and adsorbate of interest.

Calculation of Henry constant and surface area

The main goal of our sampling algorithm is to calculate adsorption enthalpy in the zero-loading limit. But the method can also calculate at the same time the Henry constant and surface area

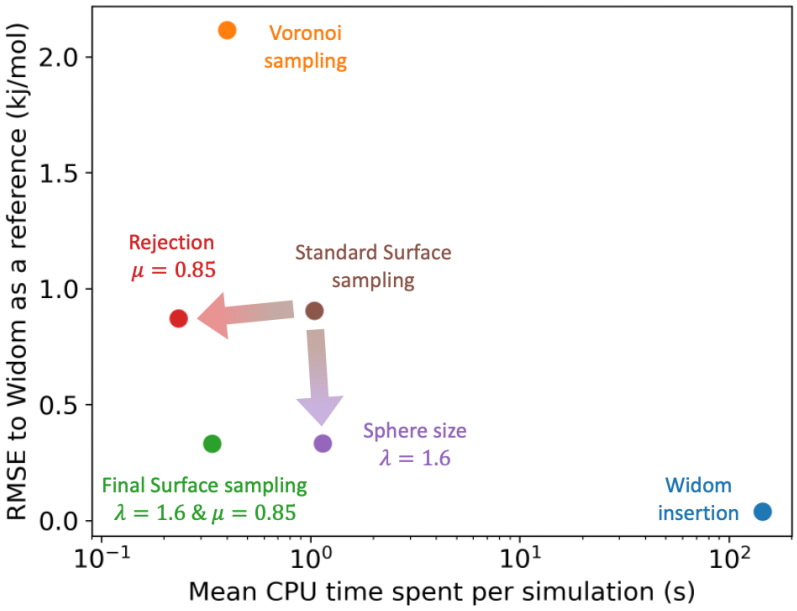


Figure 3.13: Comparison of the RMSE to the reference Widom insertion and the average computation time for different types of enthalpy calculation methods. The surface sampling calculation were all done with 2k sampling points on each sphere and the Widom simulations were done using 12k cycles. These values correspond to the value at the convergence identified using Figure **??**.

of the materials, without significant additional computational cost. The Henry constant is a key metric for assessing the affinity of an adsorbate to a nanoporous structure. The Xe/Kr gas selectivity at low pressure is defined as a ratio of Henry constants of Xe and Kr. This important property can be calculated using Equation **??** in a Widom insertion calculation. Instead of using the interaction energies at the Widom inserted points, we can now use the surface sampled points to get an approximate value for the Henry constant.

Using the optimized set of parameters for surface sampling, we assessed the performance of our algorithm on the values of Henry constant by comparing them to ground truth obtained by 100,000 cycles of Widom insertion. Since the Henry constant corresponds to the exponential of an adsorption free energy, and we are more interested in the precision on the free energy, we are using a log-scale evaluation metric. For surface sampling, the log-RMSE of K_H is equal to 0.2, which means that the order of magnitude of the values are well predicted as we can see on the Figure **??**. If we consider the derived free energy $\Delta F_{ads} = -RT \log(\rho_f RTK_H)$, the RMSE is of the order of 1.1 kJ mol^{-1} reached in about 1 s (Figure **??**). Whereas for Widom insertion, this level of error is also reached in a similar amount of time and 0.1 kJ mol^{-1} of RMSE is reached in about 86 s (Figure **??**). For free energy calculation, surface sampling is still 86 times faster to converge. If consider that the main target is the adsorption enthalpy, the Henry constant is calculated with little additional computational cost and with reasonable accuracy: we get two thermodynamic properties of interest for the price of one.

The same goes for the determination of the surface area. We can adapt our algorithm to count the number of points of the sampling spheres that have a negative energy. These represent the points were a guest molecule can favorably interact, therefore when dividing it by the

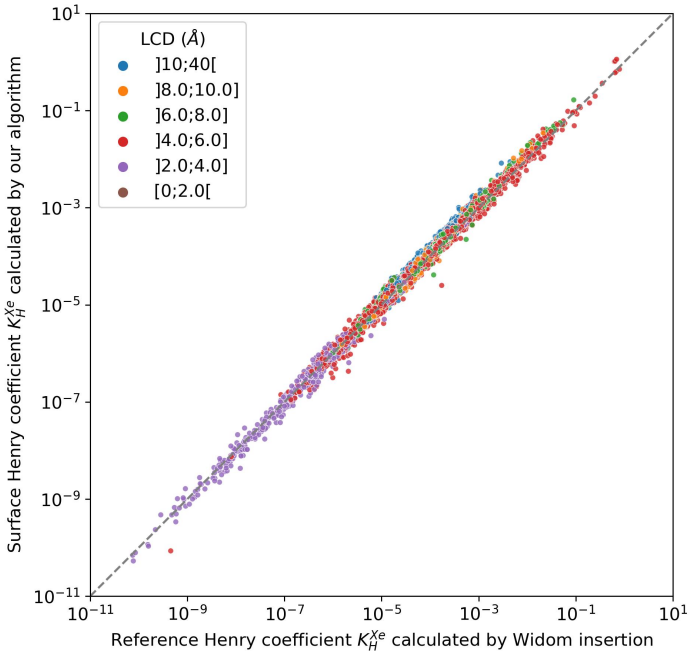


Figure 3.14: Scatterplots of the xenon Henry constants calculated by the RAESS algorithm compared to the ones calculated by a 100k-step Widom insertion simulation using two value windows.

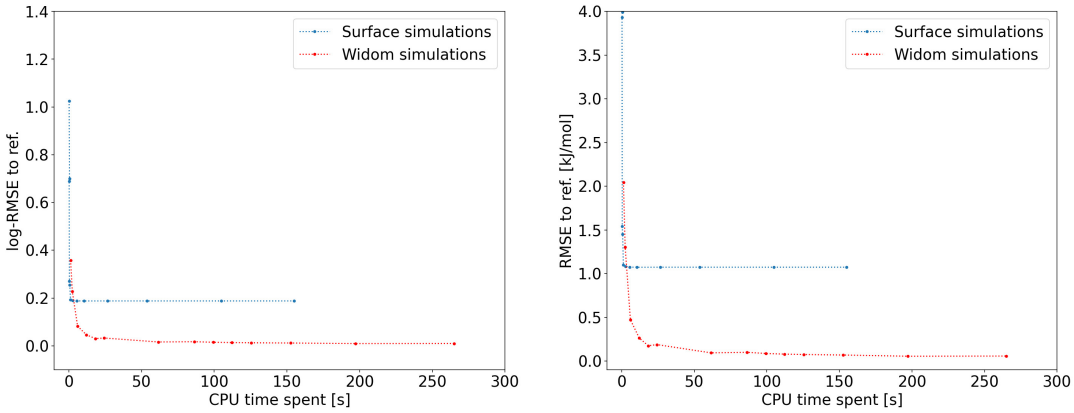


Figure 3.15: Left: convergence plot of the log-RMSE on the xenon Henry constants for both the surface sampling and the Widom insertion. Right: convergence plot of the RMSE on the xenon adsorption Gibbs free energy for the final implementation of the surface sampling and the Widom insertion.

number of sampled points, we obtain a proportion of adsorbable area of the sphere. Summing this over all atoms, we obtain a total surface area. This implementation is summed up in equation **??**:

$$\mathrm{SA} = \frac{1}{\mathrm{V}} \sum_{a \in \mathrm{cell}} \frac{\mathrm{N}_{\mathrm{accessible}}(a)}{\mathrm{N}_{\mathrm{total}}} 4\pi r(a)^2 \tag{3.7}$$

where V volume of the cell a atoms of the cell; $\mathrm{N}_{accessible}(a)$ accessible points around the atom a; N_{total} sampling points; $r(a)$ radius of the sampling sphere around the atom a. When we set $\lambda = 1$, we are sampling spheres that have a radius σ, and it is equivalent as considering hard

spheres all defined by σ (convention used by RASPA to calculate surface areas). If we compare simulation with $\lambda = 1$, we obtain surface areas that are very close to the one obtained by RASPA (see Figure **??** in SI). However, when we consider $\lambda = 1.6$, we lose the perfect accordance previously obtained and the points weakly correlated in log-scale (see Figure **??** in SI). The difference can be explained by the fact that the sphere size is larger, but the proportion of adsorbable points also changes. The relationship between these two adsorption surface areas is not trivial at all. Since the calculation of surface areas is quite cheap, this implementation would not be very useful, except for having a rough idea of the surface area.

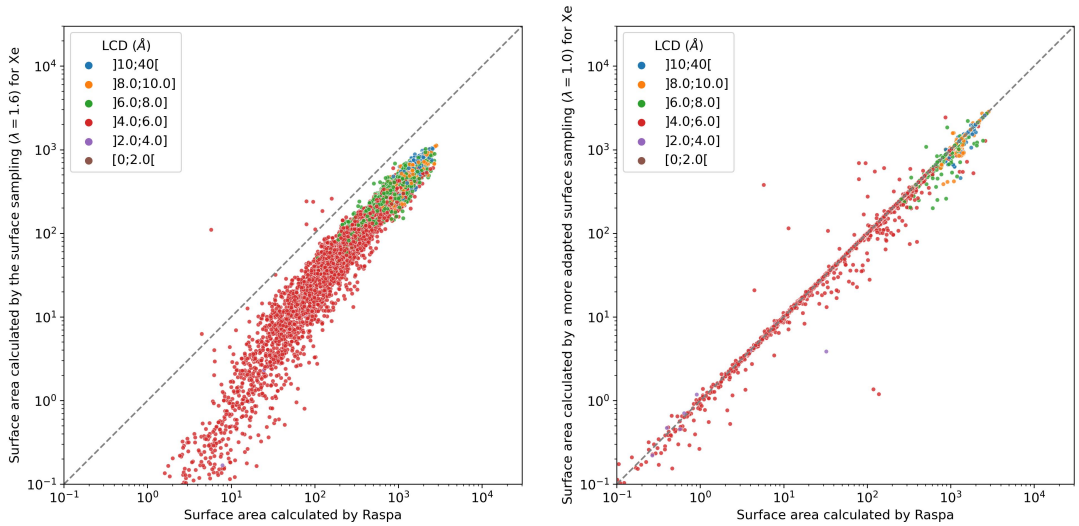


Figure 3.16: Scatterplots of the surface areas calculated by our algorithm with two different parameterizations compared to the surface area given by a Raspa surface area calculation. The left plot corresponds to the surface sampling described in the section **??** *with $\lambda = 1.6$ and $\mu = 0.85$, while the right plot uses a sampling sphere near σ with $\lambda = 1.0$. The second parameterization is much closer to what a Raspa sampling based on the σ parameter of a LJ potential does, hence explaining the much better accordance.*

3.2.4 Surface sampling application use cases

After introducing the performance of our surface energy sampling algorithm for xenon and on specific materials from CoRE MOF 2019 at 298 K, we will explore other conditions to test the transferability of the methodology. First, we will use the algorithm to assess the xenon/krypton selectivity at infinite dilution compared to the standard Widom insertion. Then, we want to compare the influence of the temperature on the performance since it most likely would be much less ideal since the Boltzmann weights are less concentrated on the less attractive points. Finally, we tested our algorithm on databases of different materials.

Selectivity Calculations

The selectivity value is the most important metric in evaluating the Xe/Kr separation performance of a nanoporous material. We want to see if a surface sampling technique can accurately evaluate this metric while being limited by all the approximations inherent to the technique.

A few precautions should be considered before blindly using the algorithm for selectivity prediction. When investigating the calculation of the selectivity, we noted that the rejection

condition on xenon can be high since we are interested in the most favorable materials for xenon adsorption. But for krypton, we want to accurately describe very low Henry constants, because a selective material would also be a material unfavorable to krypton. For these reasons, the parameter μ needs to be chosen wisely and needs to be low enough to have accurate Kr Henry constant and then selectivity values.

As we can see on the Table **??**, the error on the selectivity highly depends on the μ value that excludes the points at $\mu\sigma$ of a framework atom center. Logically, the lower this parameter the higher the sampled energy values can be in the Boltzmann averaging. Another reason is that when dividing by small values any small error on the values can be amplified in the quotient, and this effect can be reduced as we increase the number of points actually sampled.

rejection parameter μ	log10-RMSE to 100k-Widom	log10-MAE to 100k-step Widom
0.85	0.107	0.077
0.50	0.0635	0.0402
0.20	0.0637	0.0403

Table 3.1: Influence of the rejection condition in the krypton surface simulation on the accuracy of the Xe/Kr selectivity calculation. The lower the parameter μ the more accurate the simulations are for the final selectivity calculation.

According to the quick study here, the optimal value is $\mu = 0.5$ since it gives the best accuracy for a minimal amount of time. This value will be used for krypton to make a comprehensive study of the performance on the Xe/Kr selectivity for materials from CoRE MOF 2019. To sum up, in the following study, we will use the RAESS algorithm with $\lambda = 1.6$ and $\mu = 0.85$ for xenon and $\lambda = 1.6$ and $\mu = 0.5$ for krypton.

The selectivity can be compared directly using a log-scale plot and log-scale metrics. If we apply the log10 to the selectivity values, we obtain RMSE of 0.064 and MAE of 0.04. This means that we have an error of about 0.06 when we compare orders of magnitude of the selectivity. For example, if a selectivity is predicted to be $s = 10^{-7}$, then s would be in the interval $[10^{-7.06}, 10^{-6.94}]$.

To be able to give a thermodynamic interpretation, we can use the exchange Gibbs free energy associated $\Delta_{\text{exch}}G_0^{\text{Xe/Kr}}$ to this selectivity defined in the previous chapter (equation **??**). Using this exchange Gibbs free energy, we can assess much more easily the performance of the approach. The RMSE is about $0.36\,\text{kJ mol}^{-1}$. We cannot compare it to the adsorption enthalpy errors, since the ranges and interpretation are very different. Here, the selective materials have a negative $\Delta_{\text{exch}}G_0^{\text{Xe/Kr}}$, and it goes to a maximum value of about $-12.7\,\text{kJ mol}^{-1}$. The relative error is, of course, higher on the Gibbs free energy. This is due to a higher uncertainty on the Henry constant and to the denominator term brought by the krypton.

To know how well the RAESS algorithm would work in real situations, we tried to compare the top 100 most selective materials given by RAESS and a Widom simulation (RASPA). We found that 83 structures of the top 100 given by RAESS are in the top 100 given by Widom insertion. As the correlation is not perfect, it is inevitable that there is a change in the order of the top 100 given by these two methods. This number of 83% proves that the difference is quite narrow. If we enlarge to the top 150 of the Widom simulation, 94 are present in the top 100 of

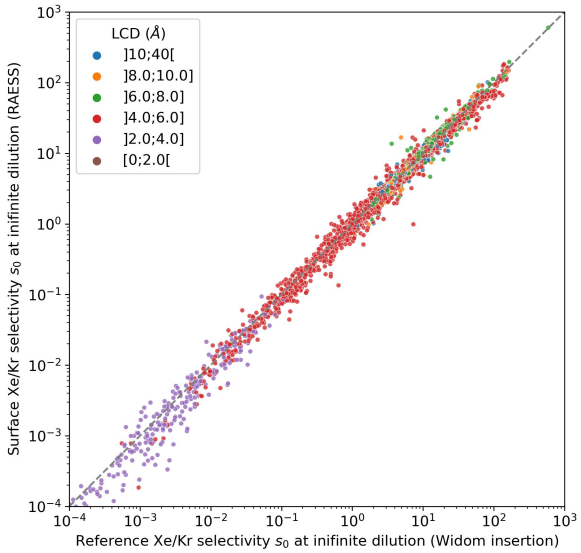


Figure 3.17: Scatterplot comparison of the Xe/Kr selectivity calculated by RAESS algorithm and the one calculated by the Widom insertion (in log scale) and labeled by the cavity size.

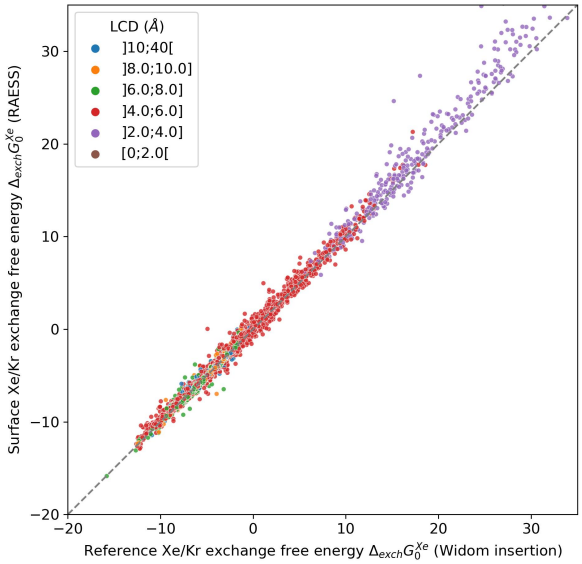


Figure 3.18: Scatterplot comparison of the exchange Gibbs free energy $\Delta_{exch}G_0^{Xe/Kr}$ calculated by the Widom insertion compared to the final implementation of RAESS (RMSE=0.36 kJ mol^{-1} and MAE=0.23 kJ mol^{-1}).

the surface simulation. We can therefore say that a vast majority of the best candidates given by the Widom insertion simulation are found by the RAESS algorithm.

A Higher Temperature

The RAESS method relies on the higher weight of the strong sites close to the surface of the pores. If we increase the temperature, the less attractive sites would play an increasing role and the accuracy of the method would drop. To grasp this limitation of the RAESS algorithm at higher temperature, we compared the results of a screening over the CoREMOF 2019 database.

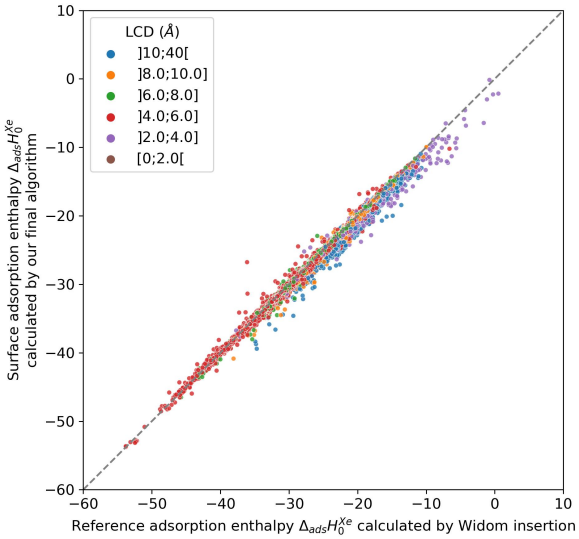


Figure 3.19: Scatterplot of the enthalpies calculated by our final algorithm (λ = 1.6 and μ = 0.85) compared to the enthalpies calculated by a 12k step Widom insertion simulation of xenon in structures of CoRE MOF 2019 with $LCD_{CCDC} \geq 3.7$ Å at 600 K.

The method is as expected less accurate, but it still gives a reasonable correlation on the performance, with an RMSE of 0.70 kJ mol^{-1} and an MAE of 0.41 kJ mol^{-1}. The errors have almost doubled when going from 298 K to 600 K. However, these limitations of the method are not crippling since adsorption processes are usually not performed at very high temperature. High temperatures are commonly used in temperature swing adsorption (TSA) to desorb the adsorbates rather than to adsorb them.

Other databases

ToBaCCo

We randomly selected 1,000 structures from the 13,511 porous frameworks of the ToBaCCo database to test the robustness of the RAESS method on a database other than CoreMOF. Since ToBaCCo contains structures with larger pores as suggested by a Moosavi et al., these materials are more unfavorable for adsorption of small molecules (such as Xe). The correlation is therefore found to be weaker than in the CoRE MOF 2019 database. This lower accuracy should be nuanced by the lack of suitability of these materials for Xe/Kr separation. Moreover, we can note that the points with weaker correlations correspond to the ones with an LCD_{CCDC} greater than 10 Å, which is not ideal to separate Xe from Kr.

The algorithm performs very well on the most adsorptive materials with xenon adsorption enthalpy values lower than −30 kJ mol^{-1}, because for these materials the adsorption sites are located near the surface. For broader pore sizes, it is interesting to note some limitations of the methodology we should be aware of. For determining the most attractive materials, this shortcoming does not influence the final results. Moreover, this limitation is not that damaging since the correlation although weaker still remains.

Amorphous materials

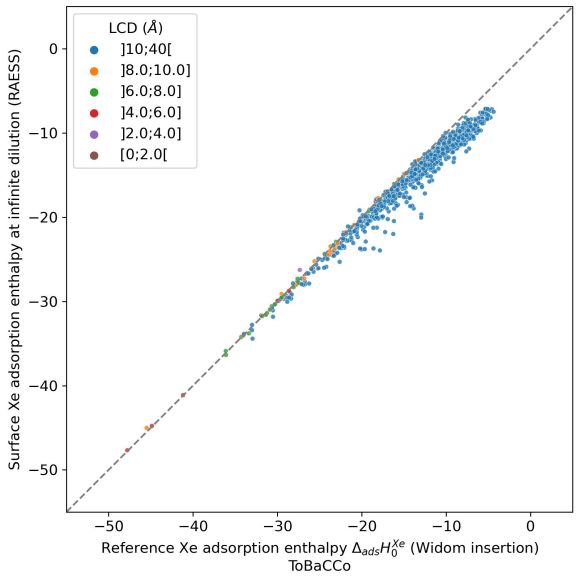


Figure 3.20: Scatterplot comparison of the xenon adsorption enthalpy calculated by the RAESS algorithm and the Widom insertion (RASPA) on the ToBaCCo database. RMSE = 1.79 kJ mol^{-1} *and MAE* = 1.48 kJ mol^{-1}. *915 structures have an* LCD_{CCDC} *greater than* 10 Å.

To further extend the possible uses cases of the RAESS algorithm, we tested our algorithm on the amorphous database.[Thyagarajan_2020] The algorithm found results for 176 structures out of 196. The RASPA software was not able to run on these amorphous structures with our computers, running out of memory due to the large system size: therefore, there is no comparison point with a Widom simulation. However, we used another simulation that is closer to the ground truth since it samples a homogeneously distributed grid using an optimized algorithm we will present in the next section. This grid sampling managed to compute the adsorption energies of 175 structures.

The Table **??** gives the values of the adsorption enthalpies and the Henry constants of a few amorphous materials as well as the time they took to be computed. The sheer number of atoms in each structure makes the CPU time required much higher than for the crystalline structures of CoRE MOF 2019. The time required are, however, quite manageable in a hypothetical screening procedure. If we consider all the 175 structures that could be calculated by our methods, the average time required is about 75 s per structure.

Structure Name	$\Delta_{ads}H_0^{Xe}$ (kJ mol^{-1})	K_H^{Xe} (mol kg^{-1} Pa^{-1})	CPU time (s)
aCarbon-Marks-id035	-63.55	6.98e-01	285.45
HCP-Colina-id016	-30.61	8.85e-05	3.88
Kerogen-Coasne-id010	-44.38	8.02e-03	61.2
PIM-Colina-id012	-26.39	7.00e-05	8.86

Table 3.2: Some amorphous materials' performance according to the RAESS algorithm. The results on the whole amorphous database is given in CSV format on the Github: `github.com/fxcoudert/citable-data/tree/master/154-Ren_ChemSci_2023`.

As we can see on the Figure **??**, the accuracy of the surface sampling is rather high since by comparing it to an unbiased grid-based sampling, we obtain very similar results. The

RMSE is about 0.83 kJ mol^{-1}, which is higher than the one for CoRE MOF structures. This method could very likely be used to evaluate amorphous materials as a fast screening tool especially since the computation time required by the optimized grid sampling is about 623 s. The dimension reduction inherent to the surface sampling makes it one order of magnitude faster than conventional techniques.

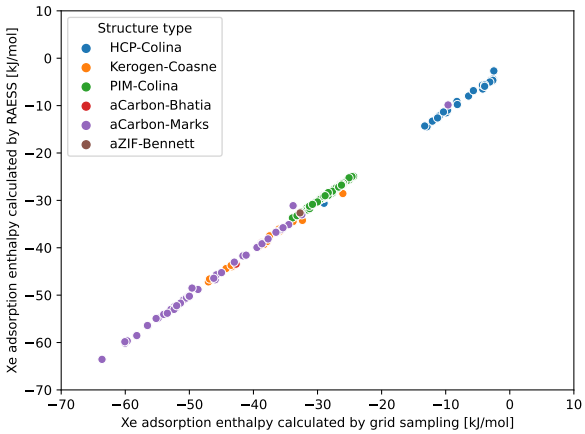


Figure 3.21: Scatterplot comparison of the xenon adsorption enthalpy calculated by the RAESS algorithm and the one calculated by a grid sampling (presented in the next section) on a database of porous rigid amorphous materials.[Thyagarajan_2020] *Raspa simulation could not be run on this database. Only the 175 structures computed by both methods are presented here.*

3.2.5 Perspectives of surface sampling

We described a novel algorithm for the high-speed calculation of adsorption enthalpy in nanoporous materials, that takes a unique approach to reduce the sampling necessary. This new algorithm is based on the core principle of dimensional reduction, from a volume problem to a surface one. The algorithm is proven to be significantly faster than the reference Widom insertion (random sampling of porous space). Moreover, the error associated is found to be of the order of 0.4 kJ mol^{-1}, tested throughout the entire CoRE MOF 2019 database, for xenon adsorption. Even when compared to existing very fast sampling techniques such as the Voronoi sampling, this surface sampling technique requires similar CPU time, combined with a better accuracy.

Based on these results, this algorithm has important potential for applications in the current computational analysis workflows of material databases, such as high-throughput screening studies. For instance, this algorithm can be used to get a fast approximation of the low-loading adsorption enthalpy of a molecule inside nanoporous materials. This cheap evaluation of the enthalpy can be used to screen out the structures with little affinity with the targeted adsorbate molecule. It can also be used as a thermodynamic descriptor for selectivity prediction in a machine learning model, as done by Simon et al.[Simon_2015] The computational speedup brought by this novel methodology can also enable the screening of materials databases at larger scale in the future.

We note, moreover, that the speed of our method resides in the sampling technique itself, rather than in the actual energy calculation. While we have benchmarked it in this work for a simple Lennard-Jones interaction potential, this sampling technique could equally be used to speed up

samplings of space based on more expensive modeling strategies, including polarizable force fields or density functional theory (DFT) calculations. In the literature, the need for cheap *ab initio* grade thermodynamic properties is usually fulfilled by using an importance sampling method based on a classical force field.[Vandenbrande2018] In our method, the description of surface sampling is independent of any force field, and the sampling spheres can be defined according to kinetic radius, van der Waals radius or any other physically relevant distance. Consequently, given a definition of atomic radii, it is possible to define a surface on which to carry out other types of simulations such as neural network potential, DFT or any other force fields. Although the accuracy or relevance of such a sampling remains an open question, the approach will undeniably speed up the simulations. This could even be applied to calculate adsorption enthalpies while considering intrinsic structure flexibility,[Witman_2017] a task whose main drawback is the high computation time required. Since surface sampling is hundreds of times faster than standard methodologies, we could use hundreds of snapshots in a flexibility-aware calculation.

Finally, although the algorithm in its present form can already be applied in a wide range of applications, additional development work could allow us to generalize it to polyatomic adsorbates. For instance, we would need to work on a definition of the molecular radius for non-spherical adsorbates as well as working all the orientation conformation of the adsorbent. We could imagine making the distance to the surface depend on the orientation of the adsorbate or sample a band volume on the surface. Although the best implementation of the surface sampling for polyatomic adsorbates remains an open question, in theory it should be possible to apply it to more complex adsorbates than the spherical noble gas. This would add more complexity to the algorithm but would not change the fundamental speedup due to surface sampling, since these orientation moves are also performed in other standard methodologies. To improve even more the accuracy, we could test hybrid samplings with multiple sampling spheres, or a combination of Voronoi nodes and sampling spheres. Another idea could be to have fractions of spheres that are oriented toward the center of the pore given by the Voronoi node. In theory, having a wider variety of sampling points can only improve the sampling. There are therefore multiple possible sampling techniques that could be built around the method introduced herein. The code is made freely available on the group's GitHub (`github.com/coudertlab/RAESS`), where further development will be released.

Data Availability: `https://github.com/fxcoudert/citable-data/tree/master/154-Ren_ChemSci_2023`

3.3 Grid Adsorption Energies Descriptors (GrAED)

3.3.1 Implementation of an efficient grid algorithm

To build more relevant energy descriptors, we will now go back to the definitions of the adsorption enthalpy and the Henry constant (equation **??** and **??**) that call for a homogeneous sampling of the adsorption space. The easiest way of sampling consists in laying a grid in the 3D space. This method is, however, known to be the most time-consuming on in theory. Inspired by our work on surface sampling, we designed an approach based on a symmetry-respecting grid,generated using the algorithms of the Gemmi Project,[Wojdyr_2022] where the points overlapping with framework are discarded. In our grid adsorption energy descriptor

(GrAED) calculation algorithm, these new features coupled with a grid sampling makes the calculation of adsorption energies much less time-consuming while being very accurate.

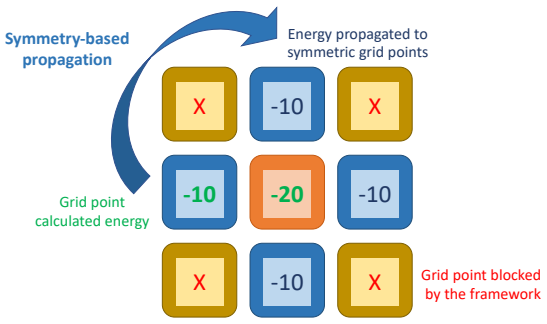


Figure 3.22: Principle of the energy sampling on a symmetry-based grid. On the 9 grid points, 4 points are blocked because they are too close to the framework atoms, 2 points are really calculated using the LJ potential and 3 points are propagated using the inner symmetry of the framework.

The corresponding algorithm has a core structure of a grid algorithm, where we need to evaluate the interaction energy at each point of the preset grid over the structure's unit cell. A naive approach would call for an expensive energy calculation at each grid point. To improve this approach, our algorithm is based on two main simplifications — a quick evaluation of the framework occupied grid points and the exploitation of symmetry. The grid points that overlap with the framework's atoms have highly positive energy mainly due to the interaction with the overlapping atom. High values of energy does not contribute much to any thermodynamic quantities displayed in the section **??**. For this reason by using a rejection parameter similar to the one developed for surface sampling introduced in the section **??**, we can precalculate the interaction energy of the grid points within the sphere of radius $\mu \times \sigma_{g-h}$. If the value of this interaction energy is higher than a preset energy threshold E_{th}, then the associated grid point takes these values as the interaction energy and no further calculation will be performed for it. The symmetry of the grid is based on the symmetry of the structure determined using the Grid definition of the Gemmi Project. By using the symmetry operations on a grid point value, we can propagate it to the other symmetry-equivalent grid points as illustrated in the Figure **??**, which reduces the amount of time we need to calculate the interaction energy of a guest molecule at a given grid node with all the surrounding framework atoms within a given cutoff. Now that we presented the main building blocks of our optimized grid calculation, we will show how it integrates in the implementation of the algorithm.

1. We loop over the framework atoms and the grid points around a sphere of radius $\mu \times \sigma_{g-h}$, 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 closest host atom is calculated and only the grid points with an energy lower than a predefined threshold E_{th} 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 **??**.

2. 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 **??**.

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 control the trade-off between accuracy and computation time, we can vary the spacing between the grid points, whose computation time is theoretically inversely proportional to the cube of the spacing. And, for some values of this spacing, this algorithm can even be faster than the surface sampling on the CoRE MOF database where the symmetry plays an important role (see Table **??**). The full implementation of the GrAED algorithm can be found at the following Github url: `github.com/coudertlab/GrAED.git`.

3.3.2 Performance on the adsorption equilibrium

If we now look at the performance of this new grid sampling algorithm compared to other sampling algorithms introduced in the previous sections. We can see that it is very advantageous to use this new sampling on the CoRE MOF 2019 database because of its accuracy and its speed. The good time performance of the grid sampling on the structures of CoRE MOF 2019 database can be explained by their rather small porosity of the materials and their high order of symmetry. For instance, 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 (most MOFs present symmetry operations). On average, the "blocking" procedure means that only ∼16% of the grid points really need to be calculated. And the "symmetry" procedure implies that only ∼17% of points need to be considered, and the combination of both theoretical 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 while keeping the accuracy level (low error on the Xe adsorption enthalpy 0.014 kJ mol^{-1}) compared to the naive grid approach, as shown in Table **??**. In the grid simulation, with the blocking procedure, the time required is reduced by ∼70.6% compared to the naive approach, and a similar reduction of ∼76.6% is observed for the symmetry-aware grid sampling. By combining both simplifications, the time required of the fast grid sampling is reduced by almost ∼91.6% for a grid spacing of 0.12 Å, which echoes with the fewer points needed to be sampled we mentioned above.

As we can see of the Figure **??**, the approach does not damage the accuracy of the adsorption enthalpy and of the Henry constant. There is an almost perfect accordance between the Widom insertion method and the grid-based approach for a very finely meshed grid (0.12 Å spacing). This was expected since both methods are unbiased sampling of the adsorption energies. Almost no error can be detected if looking at the figure for both adsorption enthalpy and Henry constant. The RMSE on the adsorption enthalpy is only about 0.01 kJ mol^{-1}, while the RMSE on the log10 of the Henry constants (in mmol g^{-1} Pa^{-1}) is also very low at 0.01. This method goes back to the initial definition of these quantities at infinite dilution, the perfect correspondence is therefore not very surprising.

Energy sampling method	RMSE on xenon adsorption enthalpy (kJ mol^{-1})	Average CPU time (s)
Grid – naive – 0.12 Å	0.014	35.4
Grid – blocking – 0.12 Å	0.014	10.4
Grid – symmetry – 0.12 Å	0.014	8.3
Grid – fast – 0.12 Å	0.014	2.96
Grid – fast – 0.2 Å	0.048	0.41
Grid – fast – 0.3 Å	0.21	0.13
Voronoi sampling	2.1	0.40
RAESS[Ren_2023]	0.33	0.34
Widom[Widom1963] (12k cycles)	0.038	150

Table 3.3: 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 an LCD_{CCDC} over 3.7 Å *of CoRE MOF 2019 database with a single Intel Xeon Platinum 8168 core at 2.7 GHz. The GrAED algorithm (with* μ = 0.8 E_{th} =100 kJ mol^{-1}*) is evaluated at different grid spacings (*0.12, 0.20, 0.30*).*

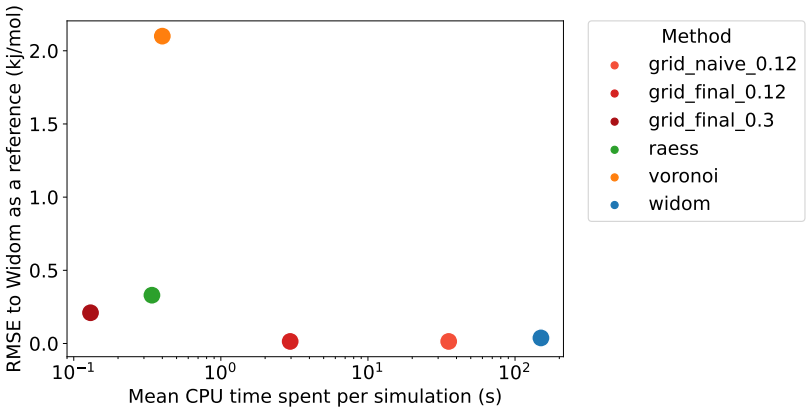


Figure 3.23: Comparison of the RMSE on Xe adsorption enthalpy and the average CPU time required to run a simulation on a structure of CoRE MOF 2019 ($LCD_{CCDC} \geq$ 3.7 Å). The metrics are taken from the Table **??**.

The little computation time required to achieve such an accuracy is, however, much more interesting. If we look at the Table **??**, the very accurate grid sampling reaches a similar accuracy than a 12k-cycle Widom insertion calculated using the Raspa software, while being 50 times faster. On the CoRE MOF 2019 database, by using a looser grid spacing of 0.3 Å, the GrAED algorithm can even be more interesting than the RAESS algorithm since it halves the computation time while being slightly more accurate. This very comparable performance compared to a dimensionally reduced sampling technique can be explained by two factors of the CoRE MOF database: first, the structures have smaller pores which means a greater surface to volume ratio, which increase the RAESS computation time; second, the highly symmetric structures of CoRE MOF reduces considerably the computation time required for GrAED, and

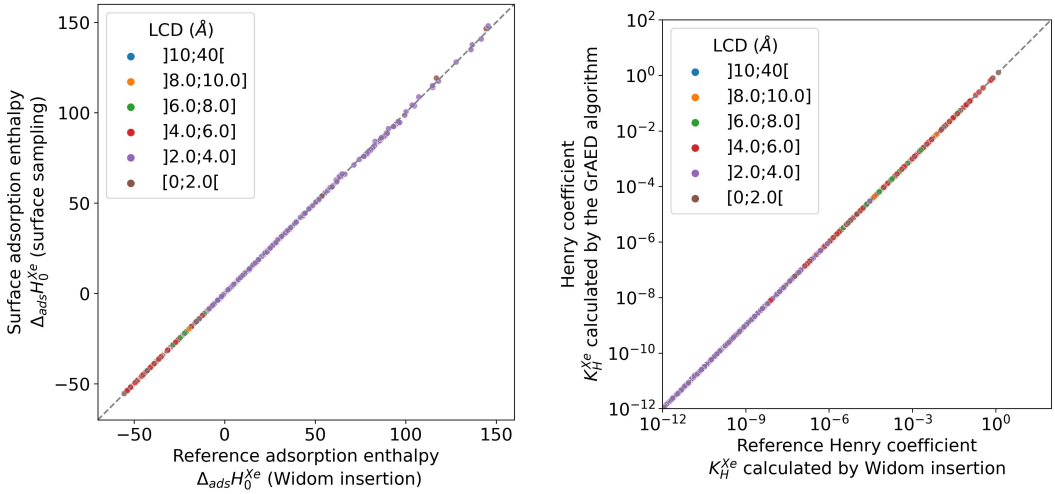


Figure 3.24: Comparison of the xenon adsorption enthalpies (left) and the Henry constants (right) calculated by the optimized grid energy sampling (for a 0.12 Å *spacing, a rejection parameter* $\mu = 0.8$ *and an energy threshold* E_{th} *of* 100 kJ mol^{-1}*) and by the Widom insertion of RASPA with 100,000 cycles on the CoRE MOF 2019 structures (*$LCD_{CCDC} \geq 3.7$ Å*).*

the same cannot be said for other databases such as ToBaCCo or the amorphous database we studied section **??**.

For instance, the computation time required is found to be 750 times higher for a grid sampling than a surface sampling on the amorphous database (see section **??**), and the RMSE is only of 0.83 kJ mol^{-1}. For amorphous databases, the surface sampling would be much better than an exhaustive grid sampling because both symmetry and overlap consideration reduce much less the number of sampled points, and we see the effect of surface sampling dimension reduction much more than on the CoRE MOF database. If we now take a look at the ToBaCCo database,**Colon_2017** the symmetry does not play any role anymore, and the pores are much larger, which means less points blocked by the framework. This impacts directly the performances of the grid sampling compared to the RAESS algorithm. The average time required on the thousand structures of ToBaCCo, considered in section **??**, is now 735 s instead of less than 2 s for a surface sampling. By increasing the grid spacing to 0.3, we can expect to reduce time required to 47 s with a simple application of the rule of three. The accuracy however is much better than for a surface sampling (see Figure **??**), reaching an extremely low RMSE of 0.02 kJ mol^{-1}. Depending on the number of structures and their nature (symmetry, porosity), we need to choose between the more efficient but less accurate RAESS and the GrAED software.

From the energy values of this grid, we can now calculate many useful descriptors of the adsorption process. We have seen the performance on the Xe adsorption enthalpy and the Xe Henry constant. But as mentioned in the section **??**, we can also derive the Xe adsorption Gibbs free energy and the Xe adsorption entropy. If we now consider the krypton in addition to the xenon, we can naturally evaluate the Kr adsorption thermodynamic quantities but also the exchange thermodynamic quantities and especially the Xe/Kr selectivity (the key metric in evaluating the separation process we are interested in).

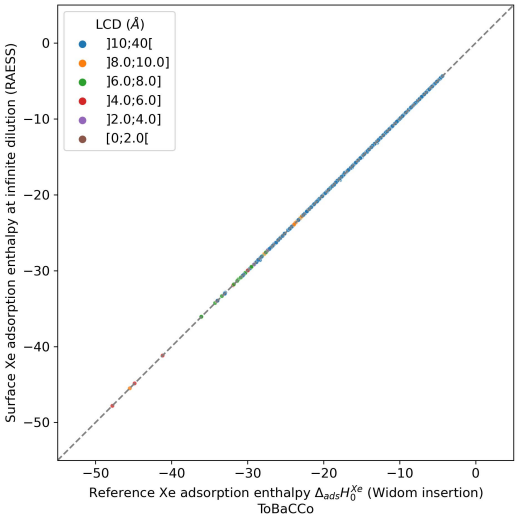


Figure 3.25: Comparison of the xenon adsorption enthalpies (left) and the Henry constants (right) calculated by the optimized grid energy sampling (for a 0.12 Å *spacing, a rejection parameter* $\mu = 0.8$ *and an energy threshold* E_{th} *of* $100\,\mathrm{kJ\,mol^{-1}}$*) and by the Widom insertion of RASPA with 100,000 cycles on 1000 randomly selected structure of the ToBaCCo.*[Colon_2017]

3.3.3 Performance on the exchange equilibrium

To characterize the competitive adsorption of the binary mixture of xenon and krypton we commonly use the Xe/Kr selectivity. Compared to a single component metric like the Henry constant, the relative uncertainty will mechanically increase since the selectivity is a quotient of the Henry constants of the competitive adsorbates. In this section, we want to measure this error and see if it is relevant to characterize the separation with this optimized grid sampling method.

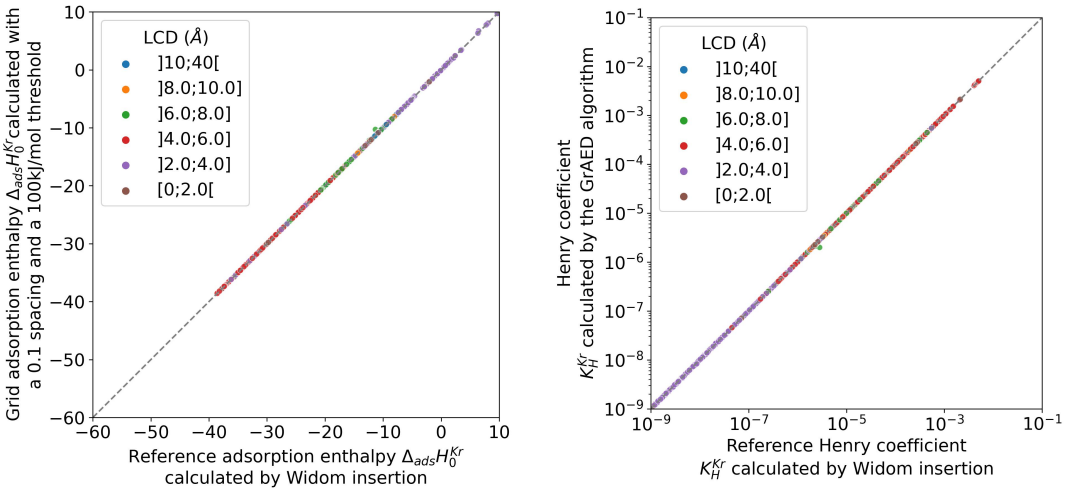


Figure 3.26: Comparison of the krypton adsorption enthalpies (left) and the Henry constants (right) calculated by the optimized grid energy sampling (for a 0.12 Å *spacing, a rejection parameter* $\mu = 0.8$ *and an energy threshold* E_{th} *of* $100\,\mathrm{kJ\,mol^{-1}}$*) and by the Widom insertion of RASPA with 100,000 cycles.*

First, we can start by looking at the adsorption properties of krypton calculated using the same grid spacing of 0.12 Å. The accuracy is more or less equivalent with an RMSE and an MAE on the krypton adsorption enthalpy of about 0.02 kJ mol^{-1} and 0.01 kJ mol^{-1}. As we can see on the Figure **??**, the correlation is very strong for both the adsorption enthalpy (linear scale) and the Henry constant (log scale). The base 10 logarithm of the Henry constant (in mmol g^{-1} Pa^{-1}) has typically an RMSE of 0.002, which is very similar to xenon. The relative error on the adsorption enthalpies of xenon and krypton does not go beyond 0.1% (values of the enthalpy have order of magnitude of dozens of kJ mol^{-1}), and the error on the xenon/krypton exchange enthalpy would therefore be very close to it. There is no reason for a huge impact on the exchange enthalpy. For the selectivity, we need to consider the relative error on the adsorption free energy which is a logarithmic transform of the Henry constant. We can estimate this relative error to around 0.2% (for a Henry constant of 10–4 mmol g^{-1} Pa^{-1}), which also corresponds more or less to the relative error expected on the exchange Gibbs free energy or the log of the selectivity.

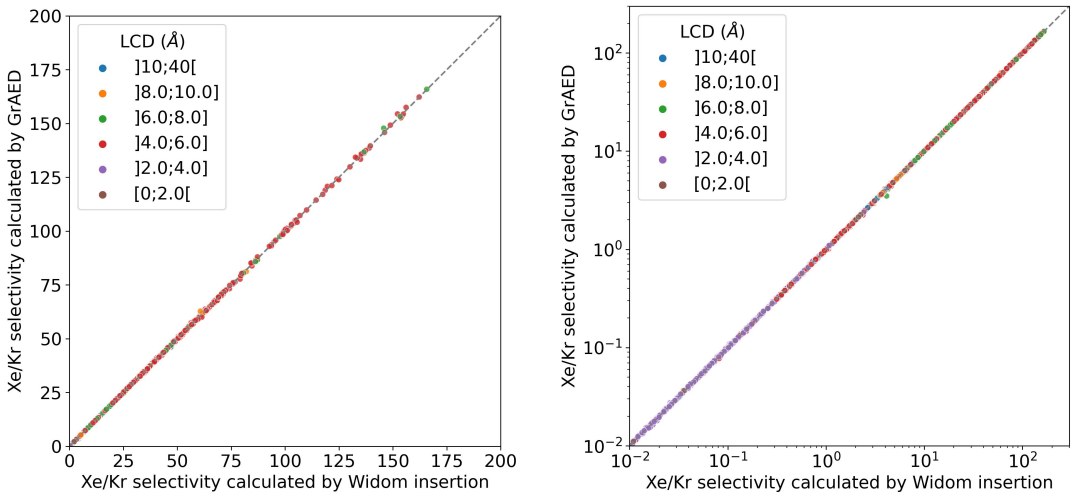


Figure 3.27: Comparison of the Xe/Kr selectivity calculated by the optimized grid energy sampling (for a 0.12 Å *spacing, a rejection parameter* $\mu = 0.8$ *and an energy threshold* E_{th} *of* 100 kJ mol^{-1}*) and by the Widom insertion of RASPA with 100,000 cycles. On the left, the axes are in linear scale, whereas the log scale has been used on the right.*

If we now look at the Figure **??**, the selectivity is as expected very well represented by the new grid sampling, especially for the logarithmic transform. The RMSE and the MAE on the selectivity values are actually of around 0.097 and 0.035, which is very low compared to the values of selectivity we are interested in (over 10). For these selective structures, the relative error is actually below 0.1%. For base 10 logarithm of the selectivity or the exchange Gibbs free energy, the RMSE is around 0.014, which means that we have a very good knowledge on the order of magnitude of the selectivity; if the selectivity were written in powers of ten, the exponent would be known at a precision of $\pm$0.014.

The computation time required to compute the selectivity value of a structure of the CoRE MOF 2019 database is in average 6.5 s, since the krypton takes about 3.5 s to compute. If an algorithm computes both selectivity values at the same time, we can at least save the time required for

initializing the software, which can marginally improve this time. This computation time required is still much lower than what we would need to compute two Widom insertions.

Now that we have proven the high accuracy and the efficiency of the GrAED algorithm for low-pressure selectivity evaluation, we will try to find relations between descriptors obtained using the grid-based algorithm and the selectivity at ambient pressure.

3.3.4 Description of the ambient-pressure selectivity

At first sight, if we look at the left plot of the Figure **??**, the selectivity at ambient pressure seems completely decorrelated to the selectivity at infinite dilution, which would mean that the sampling we have done is completely useless in determining the higher pressure selectivity values. However, the right plot suggests otherwise, the correlation between the logarithm of the selectivity does exist, the complete decorrelation we see at linear scale is actually a phenomenon that only occurs on highly selective materials. This phenomenon is described in detail in the chapter 2 and corresponds to a selectivity drop experienced by some highly (at infinite dilution) selective materials. To say it simply, the saturation of the most selective sites make the remaining sites way less selective for the xenon/krypton separation for these materials.

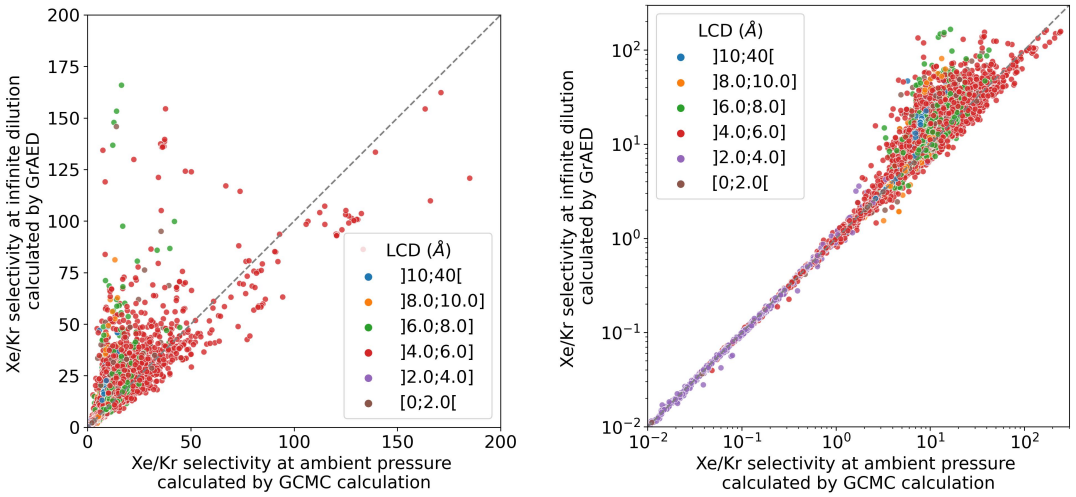


Figure 3.28: Comparison of the low-pressure Xe/Kr selectivity calculated by the GrAED algorithm (same parameters) and the ambient-pressure selectivity calculated by GCMC simulations of RASPA with 100,000 cycles. On the left, the axes are in linear scale, whereas the log scale has been used on the right.

The goal is then to be able to design descriptors that could help distinguish this selectivity dropping materials with the materials keeping their high selectivity at higher pressure. We propose three concepts that could help better understand the origin of this drop of selectivity at higher pressure: (1) other adsorption thermodynamic quantities, (2) higher temperature averaging can also be a good proxy to understand higher pressure adsorption, and (3) statistical quantities derived from the energy distributions. All of these descriptors can be obtained through a grid sampling; however, one should note that the guest–guest interactions that occur at higher pressure cannot be captured by this method.

Thermodynamic quantities

As we saw in the previous chapter, many thermodynamic quantities can be calculated at infinite dilution: adsorption and exchange Gibbs free energies, enthalpies and entropies. If we consider the xenon and krypton separation, we can generate 9 different descriptors. We will look at the relation between these quantities and the high-pressure selectivity values. In the introduction, we already tackled the relation between the exchange Gibbs free energy, since it is the logarithmic transform of the infinite dilution selectivity plotted on the Figure **??**. This quantity is the most important descriptor as it sets an initial reference value to understand the problem. The selectivity at high pressure can be seen as the selectivity at infinite dilution plus or minus a shift due to the specificity of the adsorption at higher pressure in a given material.

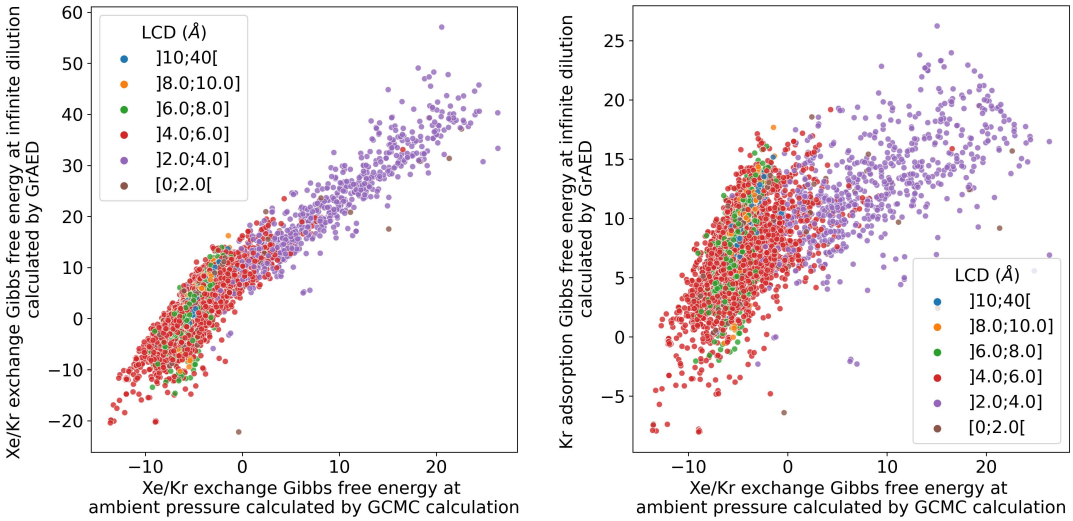


Figure 3.29: Comparison of the ambient-pressure Xe/Kr exchange Gibbs free energy calculated by GCMC simulations of RASPA with 100,000 cycles and the low-pressure adsorption free energies of xenon (left) and krypton (right) in kJ mol^{-1} *calculated by the GrAED algorithm (same parameters).*

Without any surprise, the Figure**??** confirms that a good adsorption of xenon is a good indication for the efficiency of the separation from krypton since we have a very strong correlation between the Xe/Kr exchange Gibbs free energy and the xenon adsorption Gibbs free energy. The correlation with the xenon adsorption Gibbs free energy is very weak and still positive, this means that a good material for the Xe/Kr separation would not be very bad for krypton adsorption but rather pretty average. Put differently, it is not possible to find a material very good for xenon adsorption and very bad for krypton, which explains an apparent theoretical limitation to the selectivity theoretically capped under 200 (in our level of theory for materials of CoRE MOF 2019, of course, see Figure **??** and **??**). Experimentally, no material has exceeded a selectivity value of 100.

The same statement on the importance of the adsorption attractiveness of xenon holds true when looking at the adsorption enthalpies from the Figure **??**. The correlation is very strong for the most selective materials; however, for less selective materials, the xenon adsorption enthalpy is not enough in predicting the exchange Gibbs free energy at ambient pressure. The solution would, of course, be to include the krypton into account first, the difference of both adsorption enthalpies give the exchange enthalpy which is better quantity for comparison. The

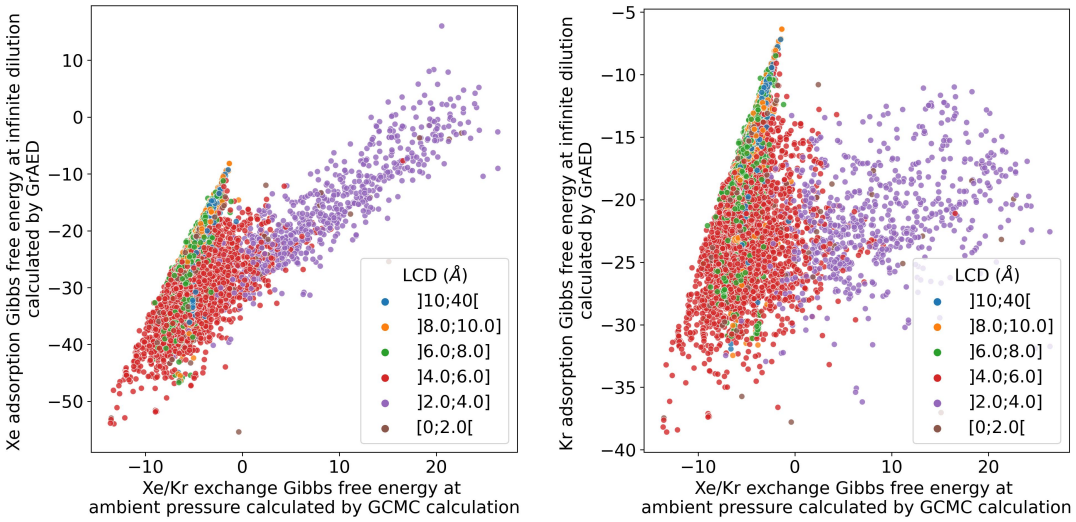


Figure 3.30: Comparison of the ambient-pressure Xe/Kr exchange Gibbs free energy calculated by GCMC simulations of RASPA with 100,000 cycles and the low-pressure adsorption enthalpies of xenon (left) and krypton (right) in $\mathrm{kJ\,mol^{-1}}$ *calculated by the GrAED algorithm (same parameters).*

comparison to the krypton adsorption enthalpy alone is not sufficient either, the very loose correlation suggests that it is not the main explanatory factor in the separation process.

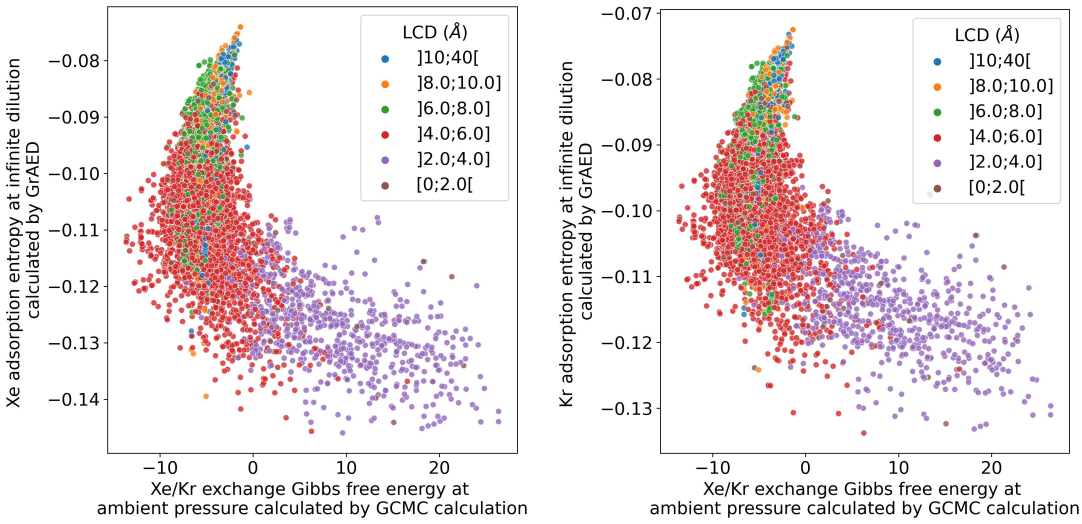


Figure 3.31: Comparison of the ambient-pressure Xe/Kr exchange Gibbs free energy calculated by GCMC simulations of RASPA with 100,000 cycles and the low-pressure adsorption entropies of xenon (left) and krypton (right) in $\mathrm{kJ\,mol^{-1}\,K^{-1}}$ *calculated by the GrAED algorithm (same parameters).*

The entropy values can explain why the adsorption free energy of xenon is well correlated to the Xe/Kr exchange free energy at ambient pressure and the adsorption enthalpy of xenon is weakly correlated. The difference between enthalpy and free energy being the entropic term (G = H – TS), this entropic term has small influence on the correlation as we can see on the Figure **??**. The values of the entropy are rather stable (between −0.15 and −0.11 $\mathrm{kJ\,mol^{-1}\,K^{-1}}$), but for some structures with an ambient-pressure exchange free energy between −10 and

0 kJ mol^{-1}, there is a swing in entropy values going from −0.11 and −0.07 kJ mol^{-1} K^{-1} for structures with very similar values of enthalpy. This means difference between Gibbs free energy and enthalpy that could reach a span of 12 kJ mol^{-1}, which explains the points that go off the diagonal in the left plot of the Figure **??**.

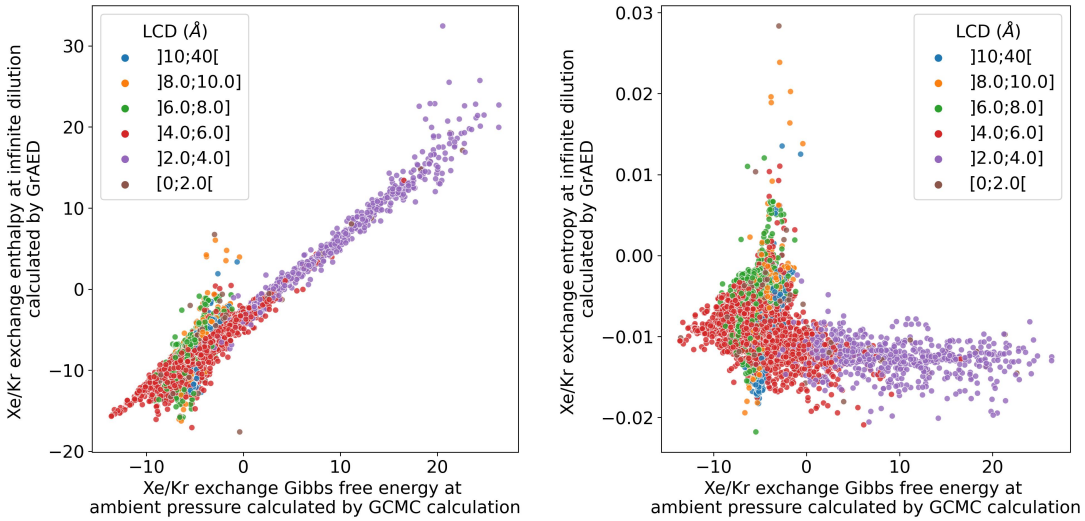


Figure 3.32: Comparison of the ambient-pressure Xe/Kr exchange Gibbs free energy calculated by GCMC simulations of RASPA with 100,000 cycles and the low-pressure exchange enthalpy (left, in kJ mol^{-1}*) and entropy (right, in* kJ mol^{-1} K^{-1}*) calculated by the GrAED algorithm (same parameters).*

Now, if we go back to the exchange thermodynamic quantities that are more relevant to our context, we can see a good correlation between the exchange enthalpy at low pressure calculated with GrAED and the exchange Gibbs free energy at ambient pressure calculated by GCMC on Figure **??**. Some discrepancies can be detected around −10 and 0 kJ mol^{-1} of the ambient-pressure exchange free energy, which can also be explained using the exchange entropy that is stable −0.01 kJ mol^{-1} K^{-1} and experiences a peak for structures near the previously mentioned range of ambient-pressure exchange free energy. The overall strong correlation can be explained using the previously identified enthalpic nature of the separation process of xenon from krypton (see chapter 2). And, the problematic range, where the correlation is weaker, also corresponds to the one associated with a selectivity drop in the Figure **??**. These exchange thermodynamic quantities can help distinguish the materials to better model the drop phenomenon, a more quantitative approach will be developed in the next chapter.

High-temperature quantities

Although the previous quantities are valuable in modeling the adsorption at ambient pressure, they are still not enough, since they describe a state where the atoms adsorbe on the most attractive sites only. The ambient-pressure state is actually characterized by an adsorption on a more diverse set of sites and with an increasingly important role of the guest–host interaction, which are also the main factors that contribute to the difference in selectivity between both pressure conditions and identified in the previous chapter.

In this section, we introduce a descriptor that better describes the energy distribution in the ambient-pressure case by increasing the weight of the more energetic adsorption sites. To

do so the easiest way was to increase the temperature in the Boltzmann averaging for both the Gibbs free energy and the enthalpy defined in equations **??** and **??**. Many temperatures were tested, and the one that yielded the higher correlation coefficient when comparing the adsorption enthalpies (infinite dilution and ambient pressure) was chosen.

A temperature of 900 K was found to be the optimal temperature to describe an ambient-pressure adsorption enthalpy of xenon across the structures of CoRE MOF 2019, and gives a minimal error (RMSE) of 1.76 kJ mol^{-1} instead of 2.87 kJ mol^{-1} for the 298 K case. This improvement impacts the exchange free energy metrics as well as the adsorption enthalpy associated with the separation of xenon from krypton. The exchange Gibbs free energy and the adsorption enthalpy of xenon at ambient pressure are better correlated to the equivalent values at lower pressure and higher temperature (900 K) than to the 298 K case. These observations support the use of higher temperature averaging to describe the ambient-pressure selectivity.

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 **??**). As we can see in the Figures **??** and **??**, the averaging at higher temperature performs better on the most selective materials, while degrading the description of the less selective materials.

On the Figure **??**, we can clearly see what an increase in the averaging temperature does: the low-temperature averaging gives a better description of the xenon adsorption enthalpy, the points are more centered around the $y = x$ axis, but the correlation is not perfect. We now also have a bigger uncertainty on the points that were initially very well predicted as badly performing materials. The high dispersion around the correlation is probably due to the guest–guest interactions that are not described in the high temperature averaging and plays a non-negligible role in the ambient pressure case.

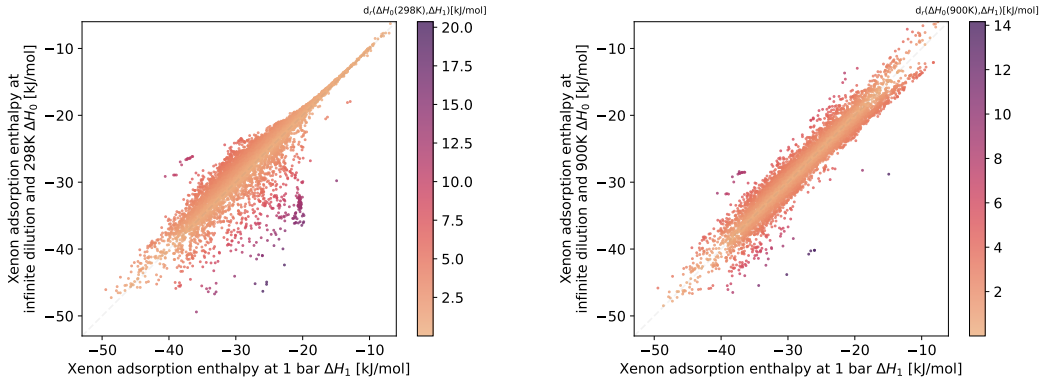


Figure 3.33: Scatterplots of the low-pressure xenon adsorption enthalpy at 298 K *(left) and at* 900 K *(right) calculated by the GrAED algorithm against the ambient-pressure xenon adsorption enthalpy at* 298 K*. Using a higher temperature Boltzmann averaging, the correlation with the ambient-pressure case of interest is much higher, the R2 coefficient improves from* 0.80 *to* 0.92 *for instance. The RMSE also decreased from* 2.87 kJ mol^{-1} *to* 1.76 kJ mol^{-1}*.*

On the Figure **??**, we can see that this improvement in the adsorption enthalpy of xenon is not directly translated into the performance on the exchange Gibbs free energy. The correlation is better overall between the exchange free energy at 298 K and infinite dilution and the one at ambient pressure (298 K). However, we can argue that the exchange free energy at 900 K is,

however, slightly better describing the materials that experience a selectivity drop as shown in Figure ??.

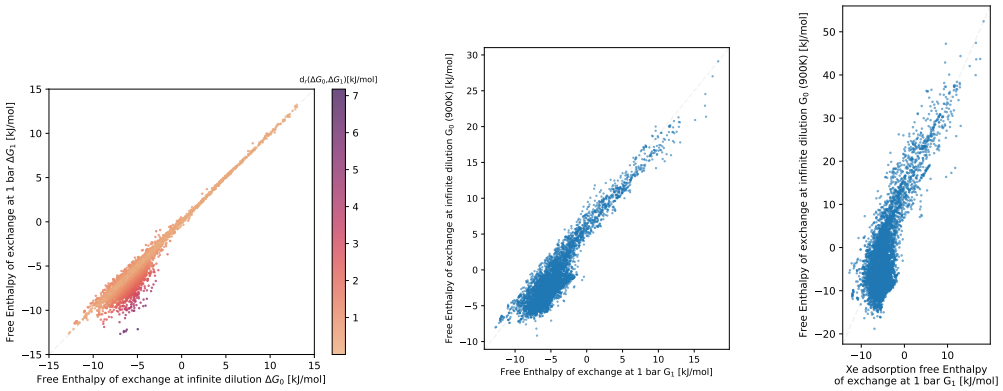


Figure 3.34: Comparison plot between the low-pressure exchange free energy at 298 K *(left) and* 900 K *(right) calculated by the GrAED algorithm and the ambient-pressure exchange free energy at* 298 K *calculated by Widom insertion.*

This high-temperature averaging can be used to better model the selectivity of the selective materials that loses selectivity between low and ambient pressure. Without spoiling the incoming chapter on how we can exploit these descriptors to quantitatively predict the selectivity at ambient pressure, we can qualitatively see how a case disjunction between the problematic structures and the other and an exploitation of the values obtained can help us identify these problematic structures.

Statistical characterization of the energy distributions

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. A statistical analysis can let us glimpse the complexity of the pore adsorption process at higher pressure by showing the diversity of the energy values and how they are distributed (the quantity of the higher energies in comparison to the lower ones, for example). The grid sampling we present here can use all the energy values from the sampled point to draw a histogram that can be used as an energy distribution and studied to retrieve interesting statistical information.

These statistics include the 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 energy distribution can be weighted uniformly or with the Boltzmann weights, both methods are explored here. The average of the Boltzmann weighted distribution is typically the definition of the adsorption enthalpy and it won't be tackled in this subsection.

Boltzmann weighted distribution

The Boltzmann weighted distribution consists in assigning a weight $\exp(-\beta E)$ according to the energy E of each point sampled by the GrAED algorithm. By doing so, we put a much higher weight on the most negative values of the energy (the most favorable adsorption sites) than the others. The unfavorable adsorption sites can be considered negligible since the exponential scaling crushed the importance of these points in the Boltzmann weighted distribution. This distribution has already been used to calculate the adsorption enthalpy (first-order moment or average) and also indirectly the Henry constant (sum of the weights used in the normalization of the distribution). And in this section, we will focus on other statistical quantities that could be derived from the distribution and are not commonly used in describing the thermodynamics of the system.

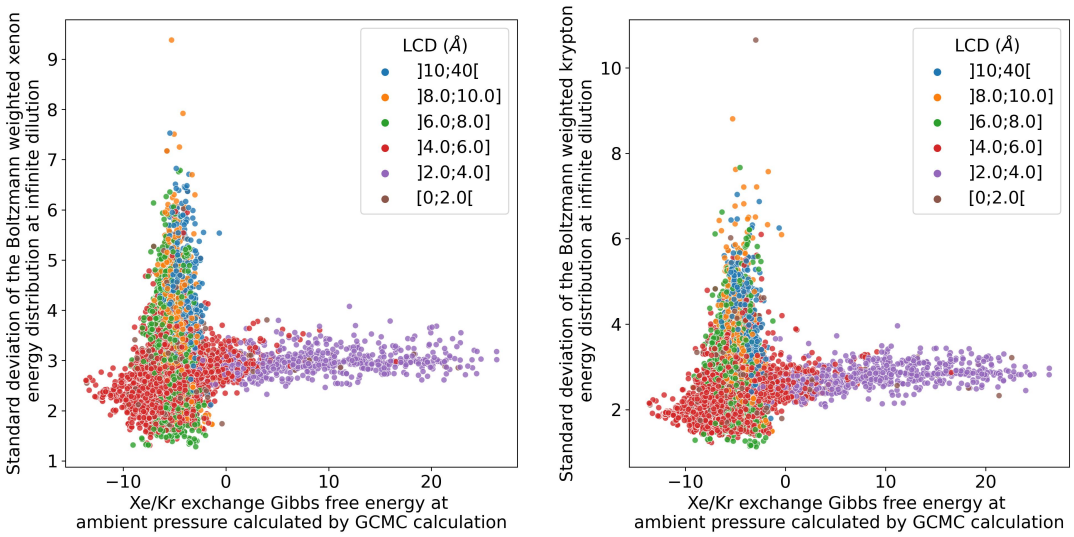


Figure 3.35: Comparison of the ambient-pressure Xe/Kr exchange Gibbs free energy calculated by GCMC simulations of RASPA with 100,000 cycles and the standard deviations of the Boltzmann weighted energy distribution of xenon (left) and krypton (right) calculated by the GrAED algorithm at 298 K.

For instance, the standard deviation of the Boltzmann weighted energy distribution is one of the statistical quantities that could be interesting in evaluating the drop of selectivity. In the previous chapter, we identified the diversity of site attractiveness as one of the main factors that could explain the drop in selectivity. The standard deviation of the energies can therefore be useful characterization of the diversity of nature between the different adsorption sites. On the Figure **??**, this standard deviation is calculated for both xenon and krypton, and the higher variation of its values is concentrated in a range of values corresponding to the one we identified for the entropy change on one hand but more importantly to the range where the selectivity drop is observed on the Figure **??** (between −10 and 0 $\mathrm{kJ\,mol^{-1}}$). Qualitatively, it is understandable that the standard deviation informs us on the pore diversity which could help characterizing the origins of the selectivity drop; and, the higher the diversity the higher probability of experiencing a selectivity drop. The challenge is then to quantify this probability through a model, a simple theoretical model would never be accurate enough, this is the reason we will be looking into machine learning models in the next chapter.

We introduced two other statistical quantities that could be interesting in describing the distribution: the skewness that characterizes the asymmetry of the distribution and the kurtosis

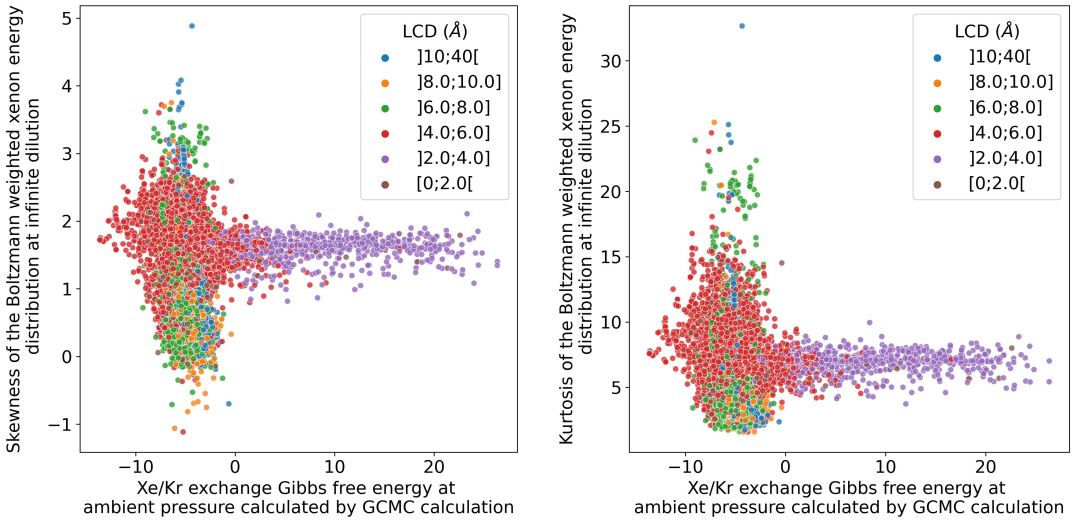


Figure 3.36: Comparison of the ambient-pressure Xe/Kr exchange Gibbs free energy calculated by GCMC simulations of RASPA with 100,000 cycles and the skewness (left) and the kurtosis (right) of the Boltzmann weighted energy distribution of xenon calculated by the GrAED algorithm at 298 K.

that quantifies the "tailedness" (the number of values in the tail of the distribution). The first one is a standardized third moment while the second is a standardized fourth moment of the distribution. For a random variable X, whose mean value is m and standard deviation σ an n-th order moment $\mathrm{M}_n(\mathrm{X})$ is defined as:

$$\mathrm{M}_n(\mathrm{X}) = \mathbb{E}\left[\left(\frac{\mathrm{X} - m}{\sigma}\right)^n\right] \tag{3.8}$$

These quantities bring in another information on the distribution. For instance, if the distribution is skewed toward the most negative pore energies it would favor the adsorption at higher pressure, while the opposite would explain a bigger drop in selectivity. The overall shape of the distribution needs to be defined by more than the standard deviation and the mean value in order to capture the origins of the drop in selectivity. Ideally, it would be better to have the whole information on the distribution, but visually speaking, it would be far too complex to compare the structures according to this multidimensional descriptor and the statistical quantities compress in a way the complex data on the energy distribution.

The Figure **??** further highlights the range of selectivity value we discussed across this section. The materials within this range can be sorted out using the different statistical quantities we introduced throughout this discussion. Depending on these skewness or kurtosis values it could be theoretically possible to identify a link with the selectivity drop previously identified, but without a model or a framework it is impossible to find the right relationship using our bare eyes.

Uniformely weighted distribution

To finish this overview on the thermodynamic/energetic descriptors we can deduce from the grid sampling we developed, we will study a more uniformly weighted energy distribution, and since the much higher energy values corresponding to the overlap with a framework atom does not interest us, these values are removed naturally by our sampling. Since a threshold

value of 100 kJ mol^{-1} has been used for the grid sampling here, the overlapping is defined very largely and for an energy below this value the energy point is taken into account. This distribution corresponds to the adsorbable sites (no overlap) that are weighted according to their sole occupancy of the void volume.

We studied the mean value and the standard deviation of this distribution. For the mean value, we can observe a weak correlation with the exchange Gibbs energy at ambient pressure, but the correlation stops for the materials with larger pores and therefore a more diverse set of energy values, the mean value is lower since the void fraction is larger and the weight on the more negative values increase, but it does not mean that very attractive sites are present since there is no Boltzmann weight — the exchange free energy is not continuing the trend and comes back to more positive values. We can note that the values are very high, this is because an energy threshold value of 100 kJ mol^{-1} is used and a lot of points have values within zero and this threshold value, which shifts the mean toward these values. For this reason, we did not continue further than the standard deviation the statistical study of this distribution, and a better design of the distribution needs to be considered to focus more on the negative values — it could be through lowering the threshold or by using a less skewing Boltzmann average (averaging at higher temperature for instance).

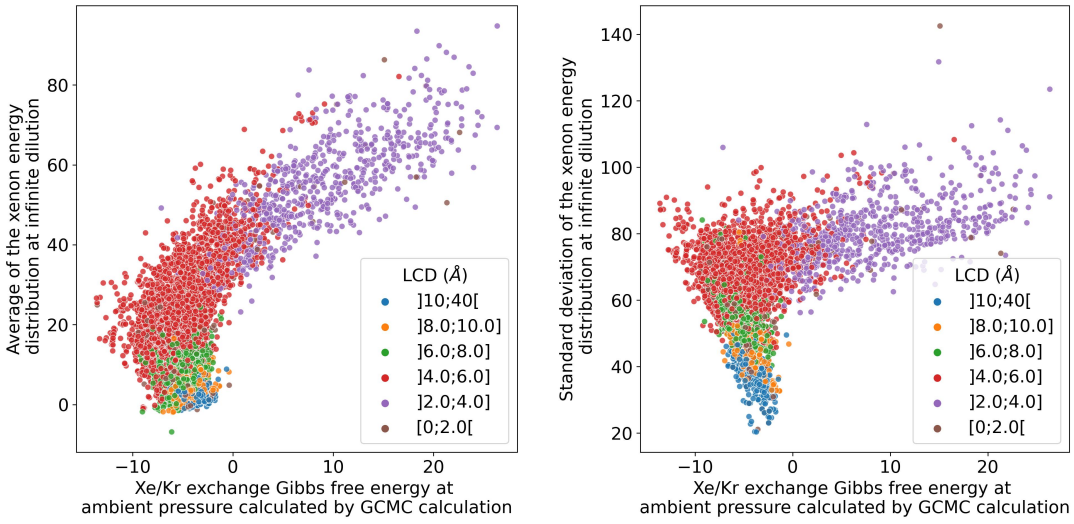


Figure 3.37: Comparison of the ambient-pressure Xe/Kr exchange Gibbs free energy calculated by GCMC simulations of RASPA with 100,000 cycles and the mean values (left) and the standard deviation (right) of the uniformly weighted energy distribution of xenon calculated by the GrAED algorithm at 298 K.

The standard deviation of this distribution has a lower value for materials with larger pores, see Figure **??**. This could be explained by a high number of points near the average value of the energy (around 0 and 10 kJ mol^{-1}) according to the left plot of the figure. For the materials with a standard deviation between 20 and 40 kJ mol^{-1}, we can very confidently say that they have an exchange free energy that could not exceed −8 kJ mol^{-1}, which mean they are not in the top selective materials. The standard deviation presented here can help us identify materials that could be promising at low pressure but are not really promising in reality.

To improve this approach, we can explore the use of a higher temperature Boltzmann average for the weights of the distribution, not necessarily using a temperature of 900 K. We can also test out an average on a distribution with a lower energy threshold (zero or the mean kinetic energy of a gas $1.5k_BT$). The idea being we want to characterize a higher energy state like the one happening at ambient pressure. And of course, higher orders of moments can also be tested to give a more complete picture of the distribution.

In this section, we tried to single out each of the thermodynamic descriptor we can derive from the grid sampling; the approach is limited by the visualization dimension limitation (bidimensional). Although we can see some examples, how it can help sorting the materials according to some properties that could at the end predict the selectivity (for instance, a given range of standard deviation of the uniformly weighted energy distribution can give a range of possible selectivity), it is still difficult to mix the different descriptors and to give a prediction. To achieve just this, many modern approaches use statistical learning to learn this relationship from a large enough set of structures with some computed properties, and this approach will be tested in the next chapter. Finally, every computation detail on the GrAED sampling technique is available online at `github.com/coudertlab/GrAED`.

4

STATISTICAL LEARNING OF ADSORPTION PROPERTIES

4.1 Machine Learning Models 113
4.1.1 From algorithm to machine learning 113
4.1.2 Introduction to supervised learning 115
4.1.3 Machine learning models 123
4.2 Prediction of the Ambient-pressure Selectivity 130
4.2.1 Data Preparation 130
4.2.2 Feature engineering 131
4.2.3 Model training 137
4.2.4 ML model performance 139
4.3 Opening the Black Box 140
4.3.1 Global interpretability 141
4.3.2 Local interpretability 144
4.3.3 Conclusions and perspectives 146

4.1 Machine Learning Models

Machine learning (ML) models have been widely used to characterize adsorption, transport, catalytic or mechanical properties, just to cite a few. It can in some cases replace very time consuming simulations with simpler calculation of key descriptors that can help the model predict the desired properties. In other cases, it is used to describe the structure–property relationships learned by the ML model. However,machine learning is not a silver bullet, we cannot blindly apply it on any applications; it requires a thorough work on understanding the key variables that will improve the prediction. By using our work on the thermodynamic descriptors and our knowledge on the effect of pressure on the selectivity, we will build a machine learning model to characterize the separation of xenon from krypton at ambient pressure.

4.1.1 From algorithm to machine learning

To understand how machine learning, first, we need to understand how a computer accomplishes a task. The human operator plays a key role in the process — after designing the solution through theoretical considerations, he needs to write down a list of instructions, called an algorithm, that specifies all needed actions given the circumstances so that the computer achieves the desired outcome. In physical or chemical sciences, these algorithms usually articulate the different components of a theoretical model, which can be an equation without analytical solutions, an analytical expression or a probabilistic problem, just to cite a few. The previous chapters typically presented such algorithms for the simulation of adsorption processes; for instance, the GCMC simulations are based on the statistical physics of the phase equilibrium between a gas phase and adsorption phase inside a nanoporous material, and a Monte Carlo model is used to reproduce the statistics associated to the grand canonical ensemble. The energy sampling algorithms, along with the Widom insertion, are also good examples of how the computer can help the theoretician model the systems under specific chemical and physical conditions.

A machine learning model is also based on an algorithm, but the goal is very different from the above-mentioned examples — it doesn't aim at giving all the details of the computation according to proven theoretical principles. As implied by the name, the ambition would be to learn underlying relations within the input data so that it performs the task itself. The machine learning (ML) algorithm is then the list of instructions that specifies how the machine is going to learn from the data. For example, clustering algorithms can distinguish different classes of elements within a disordered dataset so that new concepts emerge; this type of machine learning algorithm is called unsupervised learning because we do not predefine or pre-label the data and the machine helps us understand the underlying structure in the data; we will not go deeper in the details of this type of algorithm since it goes beyond the frame of this thesis. The class of algorithm we want to study is rather the supervised learning model, which learns from labeled data the relation between the label and the characteristics (called features or descriptors) from a given set of data points, and can predict the label from unlabeled data using the characteristics. For example, if we want to predict the weather of tomorrow, the model could use the past weather of similar dates to infer if it will rain tomorrow; the history of the weather forms the features of the ML model and the future weather is the target variable or the label of the data.

To articulate the differences between a standard algorithm and an ML algorithm, let me introduce a fascinating board game called Go. This game is traditionally played with 2 players on a 19 by 19 board, where each player places black/white pieces to control the maximum of boxes. Based on these simple rules, different algorithms have been developed to make the computer play the game. The first Go program was written in the late 60s to mimic the pattern recognition of Go players when estimating the "score" through an influence function,**zobrist1970feature** and from the 80s to the beginning of the 21st century the first Go programs capable of playing were releases. These programs were based on simple alpha-beta search algorithms that seeks at testing every possible move (while pruning the less promising ones); while they were working very well in other games like chess (IBM's Deep(er) Blue beat the world champion of chess in 1995), in Go these types of programs were only at the level of a novice player. The difference of performance lies in the combinatorics behind both games, the game of chess has a number of

legal positions lower than 10^{47},[website_labelle] while for the game of Go there are approximately 10^{171} legal positions.[Tromp_2007, github_tromp_go] The state space to explore is incomparably greater and a boost in the computing power that improved the performance for computer chess is not going to make a difference for Go. A drastic reduction of the space to be explored is needed for a computer program to work. The biggest improvement came, when in 2007, Coulom introduced a Monte Carlo tree search.[Coulom_2007] This algorithm uses heuristics to distinguish between bad a good move according to human perception of the game, a probability of selection is assigned to the moves according to their potential (policy), potential moves are randomly picked according to this probability; the average outcomes associated with a parent move gives the value of the move. The computer Go is now more efficient in the evaluation of the moves using a Monte Carlo sampling, and it can now play with average amateur players, but it is nowhere near surpassing them. Up until now, the algorithms are based on human knowledge that the programmer implements directly in the computer using machine instructions. Statistics and randomness are used to orient the machine toward the best moves and reduce their predictability, but the statistics that identifies the moves are based on human heuristics that are usually not generalizable. The big revolution brought about by machine learning in the field aims at better evaluating these statistics using the data from already played games. By using a dataset of 30 million moves, the Alpha Go is based on the same Monte Carlo tree search framework but it replaces the formulas behind the probability of searching a move by a machine learning model called the "policy network" and the one behind evaluating the confidence in winning of the position by a value "network".[Silver_2016] Alpha Go was the first computer program to beat a world champion in 2016. One year later, to further emancipate from human knowledge, an improved version, Alpha Go zero generates its own data by playing games against itself to train a similar machine learning structure than presented before. This new version beats 100 times out of 100 the former version,[Silver_2017] which marks a new era of domination of computer go over the best player in the world, and the defeat of another top player just confirms the advent of this new era.

In this example, we can see how the machine learned the value of each moves by compiling the knowledge of huge datasets in a deep neural network. The main difference between conventional approaches to algorithmic and machine learning is very well illustrated in the previous example; the goal is not to tell the computer how to play using player knowledge implemented in formulas and explicit instructions, but it is to give an explicit framework with flexible parameters that the model needs to learn using a database. In other words, the parameters of a model are fitted to match the values of a database, while being capable of generalizing in situations outside of the database (this notion of generalizability will be further discussed in the following sections). In this section, the goal is not to give a complete overview of all existing models but rather to introduce the main concepts of ML through the example of the model we will use for our problem of selectivity performance prediction.

4.1.2 Introduction to supervised learning

In this thesis, we will focus on the most common way of statistically learning from data, which is the supervised learning. As previously introduced, supervised learning corresponds to the extraction of a relationship between the labels of a set of data points and some of their known characteristics or features. This relationship can be called the model or the predictor and should ideally generalize to similar but unseen data. In this section, we will formalize the goal of

the learning algorithm when given a set of labeled data in order to introduce more complex notions in machine learning like the bias–variance tradeoff and also more specific models that will be used in this chapter like the tree-based models. Different books have been used for the conception of this section, mainly the Elements of Statistial Learning[Hastie_2009] and an Introduction to machine learning (in French) from Azencott.[azencott2022introduction]

Theoretical considerations

In supervised learning, the algorithm learns from a set of data noted $\mathcal{D}_n = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$ with n observed data points, where $\mathbf{x}_i$ represents an input observable which is a vector of $\mathbb{R}^p$ ($p = 1$ for scalars) and y_i is the label of the data point i that belongs to a set $\mathcal{Y}$ (numerical, categorical or vectorial). We can model the characteristics observed by a random variable X and the label by another random variable Y. The dataset only gives a fragmented view of the joint probability (see equation **??**) and the goal will be to extrapolate the relation to unseen data. (X, Y) represents all possible combinations of data points seen an unseen.

$$\forall\ \mathbf{x} \in \mathbb{R}^p,\ y \in \mathcal{Y},\ \mathbb{P}(\mathrm{X} = \mathbf{x}, \mathrm{Y} = y) = \mathbb{P}(\mathrm{X} = \mathbf{x})\mathbb{P}(\mathrm{Y} = y|\mathrm{X} = \mathbf{x}) \tag{4.1}$$

The challenge of supervised learning lies in the fact that the data at our disposal does not give a complete picture of the probability law. And the goal is to give the most probable label y for a data point characterized by $\mathbf{x}$, which is determining the conditional expectation $\mathbb{E}[\mathrm{Y}|\mathrm{X} = \mathbf{x}]$ of Y given the observable $\mathbf{x}$, which depends on the conditional probabilities $\mathbb{P}(\mathrm{Y} = y_i|\mathrm{X} = \mathbf{x}_j)$ seen across all data points $i, j \in \{1, \ldots, n\}$.

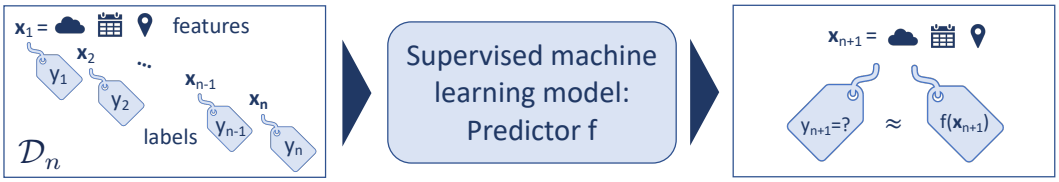


Figure 4.1: Illustration of the core principle of supervised learning.

To do so, the learning algorithm uses a "predictor" f that can be defined as the function that associates values (features) from $\mathcal{X} = \mathbb{R}^p$ to values of $\mathcal{Y}$. By changing the learning model (subsection **??**) or by changing the feature space $\mathcal{X}$ we can define different domains $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ where we search for the prediction function f. The domain $\mathcal{F}$ can either be too restrictive and the optimum function found is too far from the theoretical one, or be too large and the optimization problem is nearly impossible to solve or the solution is way too close to the data, which raises question of fitting that will be discussed later.

This predictor can be interpreted as the function that gives the most probable outcome y for a given input $\mathbf{x}$. To evaluate the quality of the predictor, we can introduce a loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^p$ that compares the predicted value $f(\mathbf{x})$ to the true value y on the dataset at our disposal $\mathcal{D}_n$. This loss function needs to increase when $f(\mathbf{x})$ moves away from y. To extend the definition of the loss to the entire possible space, we can introduce the theoretical risk $\mathcal{R}$ of a predictor h using the random variables X and Y so that $\mathcal{R}(h) = \mathbb{E}[\mathcal{L}(h(\mathrm{X}), \mathrm{Y})]$. However,

since we do not know the exact mapping of the random variables, we will rather evaluate empirically the risk $\mathcal{R}_n$ on the known dataset $\mathcal{D}_n$:

$$\mathcal{R}_n(h) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}\left(h(\mathbf{x}_i), y_i\right) \tag{4.2}$$

The goal is therefore to find a function that minimizes the risk function across the known data, and this optimal predictor f^* can be defined as:

$$f_n^* = \arg\min_{f\in\mathcal{F}}\mathcal{R}_n(f) \tag{4.3}$$

Many loss functions can be used and depending on the definition, we emphasize more or less on the large errors. For instance, a quadratic cost function will penalize a lot the outlier so that a few medium errors are better than one great error, which is not the case for an absolute cost. Since we only use regression models in this thesis, we will not go into the details of classification loss functions and will rather focus more on the regression loss functions. The quadratic loss or squared error loss $\mathcal{L}_{\text{SE}}(f(\mathbf{x}), y) = 0.5(y - f(\mathbf{x}))^2$ of a predictor f on a data point $(\mathbf{x}, y)$ is simply defined as the squared difference between the prediction and the true label. The multiplicative 0.5 coefficient is here to simplify in the derivatives. This loss is similar to the mean squared error (MSE) used to compare two quantities across a dataset, the risk function actually corresponds to half of the MSE on the predictions $\mathcal{D}_n$:

$$\mathcal{R}_{\text{SE}}(f) = 0.5\frac{1}{n}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2 \tag{4.4}$$

A second very common loss function is the absolute loss, which is associated with the mean absolute error (MAE) used in error evaluation. The loss can be expressed as $\mathcal{L}_{\text{AE}}(f(\mathbf{x}), y) = |y - f(\mathbf{x})|$, and the risk function associated is simply the MAE across the dataset predictions:

$$\mathcal{R}_{\text{AE}}(f) = \frac{1}{n}\sum_{i=1}^{n}|y_i - f(\mathbf{x}_i)| \tag{4.5}$$

It is also possible to make this loss function flatter near the minimal error by introducing a parameter ϵ. The ϵ-insensitive loss corresponds to a modified absolute loss $\mathcal{L}_\epsilon(f(\mathbf{x}), y) = \max(0, |y - f(\mathbf{x})|)$.

Finally, it is possible to combine the less outlier-sensitive absolute loss with the smoothness of the quadratic loss near the minimal error domain by using a Huber loss. For a given δ, the Huber loss is defined as:

$$\mathcal{L}_\delta(f(\mathbf{x}), y) = \begin{cases} \frac{1}{2}(y - f(\mathbf{x}))^2 & \text{for } |y - f(\mathbf{x})| \leq \delta \\ \delta\left(|y - f(\mathbf{x})| - \frac{1}{2}\delta\right) & \text{otherwise.} \end{cases} \tag{4.6}$$

And a risk function $\mathcal{R}_\delta$ can also be determined using this loss function. The Huber loss is considered as a robust loss function since it is less sensitive to the outliers (high values of error) and it has a very smooth gradient near the low values of error like the squared error. It can be seen as a combination of the advantages of both the absolute and squared errors as illustrated by the Figure **??**.

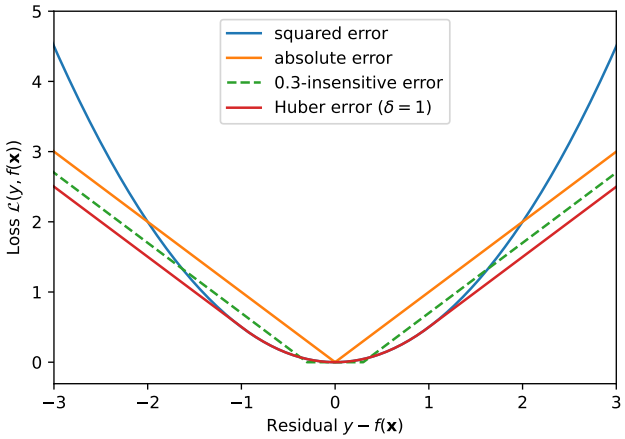


Figure 4.2: Comparison of different loss functions (quadratic loss, absolute loss, ϵ-insensitive and the Huber loss).

Through these theoretical considerations, we demystified the process of a machine learning from data by simply formulating this leaning process as an optimization of a cost function, which is a common tool in any scientific field. However, this optimization problem is challenging in the sense that the variable is a function that lies in a high dimension space and approximations are needed to reduce the space. This is why most of the engineering breakthrough happens in the conception of the architecture of the ML model that defines the form of the prediction function f. Another difficulty of machine learning is to deal with an ill-posed problem, which means that one of the three conditions of Hadamard is not verified. These conditions being the existence and unicity of a solution and its continuity with regard to the initial conditions. This issue is usually tackled using regularization techniques such as the one introduced by Tikhonov in the second half of the 20_{th} century. Furthermore, the minimization of the empiric risk is not always consistent with the minimization of the more global risk (considering all possible observations), since minimizing $\mathcal{R}_n$ does not always give the same solution as the minimization of $\mathcal{R}$. Therefore, the complexity of the risk optimization problem depends on the loss function chosen but also the domain $\mathcal{F}$ defined by the model and different techniques can be used to construct a solution without anything that guarantees the optimality of it. One of the biggest challenges of ML is to overcome the problem of generalizability, which will be the topic of the next discussion.

Generalization and overfitting

As previously discussed, the optimization problem is ill-defined and we have no guarantee for the model to work on other data points as n goes toward infinite. The generalizability of model consists in ensuring the predictability of unseen data so that the solution does not simply correspond to the minimal risk for the data $\mathcal{D}_n$ but also for other m data points $\{(\mathbf{x}_{n+1}, y_{n+1}), \ldots, (\mathbf{x}_{n+m}, y_{n+m})\}$ all different from the previous set. One of the main phenomena that explain this discrepancy between the solution f_n^* and the ideal solution f^* (considering an infinite amount of data) is the noise in the dataset. The data is not perfectly measured, and the uncertainty attached to each $\mathbf{x}_i$ and y_i values can create a residual noise that needs to be ignored in the learning process. Moreover, sometimes, the p explanatory variables considered are not enough to model the target phenomenon. To train a generalizable model, we would

need to learn enough to capture the inner relation between X and Y, but it should not fit the data too closely and capture the noise along the way, otherwise we say that the model overfits the data. If the model is very inaccurate even on the training data, we say that the model underfits, and generally it means that the model is too simplistic (not enough features or too low-level architecture).

This problem of overfitting can be summarized in the fundamental notion of bias–variance tradeoff in machine learning and more generally in statistics. The error can be broken down in two types: the bias error measures the error made on the available data $\mathcal{D}_n$, while the variance error measures the sensibility to small variations in the input values. A high bias error corresponds to an underfitting, we did not learn enough from the data; and a high variance error means an overfitting, we learned too much even superfluous relations. To formalize these errors, we can go back to the empiric risk function $\mathcal{R}_n(f)$ that models the error of our predictor $f \in \mathcal{F}$; to know if we reach the ideal optimum we would need to compare it to the minimal risk that a predictor with infinite knowledge would get $\mathcal{R}^* = \min_{h \in \mathcal{Y}^{\mathcal{X}}} \mathcal{R}(h)$. This excess error $\mathcal{R}_n(f) - \mathcal{R}^*$ can then be broken down in two errors that could be interpreted as the bias and the variance errors:

$$\mathcal{R}_n(f) - \mathcal{R}^* = \left[\mathcal{R}_n(f) - \min_{h \in \mathcal{F}} \mathcal{R}_n(h)\right] + \left[\min_{h \in \mathcal{F}} \mathcal{R}_n(h) - \mathcal{R}^*\right] \tag{4.7}$$

The first term of the above-written sum corresponds to a bias error, because it measures how far the current predictor f is off of a minimum (there can be several in an ill-posed problem) risk predictor f_n^* determined using the n data points. The second term, on the other hand, is the residual error associated with the choice of the predictor domain $\mathcal{F}$ and the fact that only a finite amount of data is accessible to the prediction model. With an infinite amount of data, the model f^* associated with the risk $\mathcal{R}^*$ would not be influenced by the noise since several data points with similar features but with small noises would give a similar prediction. The difference of loss between this ideal function f^* and the current function we are testing f would correspond to an overfitting of the noise that could not be distinguished in the finite case, if we consider the domain $\mathcal{F}$ defined by the model suitable. On the contrary, if there is a problem of model, this error also measures the approximation error due to the choice of a given set of features with a given model architecture.

In general, if the model is very complex in comparison to the amount of data we have, we would fit too closely to the data and have a very high change to overfit. The opposite is also true, a simplistic model would yield to a poor bias error and the model would be underfit. This principle is represented on the Figure **??** and should guide us in the design of a new ML model. The complex art of fitting a model to a dataset consists in finding the right balance between the bias and the variance. Fortunately, some optimization tools can help us reduce the variance error by changing the loss function itself, and we are going to look into them in the next part of our discussion.

Regularization to fight against overfitting

Regularization consists generally in adding implicit or explicit constraints on the optimization problem to find not only the most accurate solution (minimal loss) but also the simplest. This criterion of simplicity is crucial in the generalization of the problem. We typically don't need

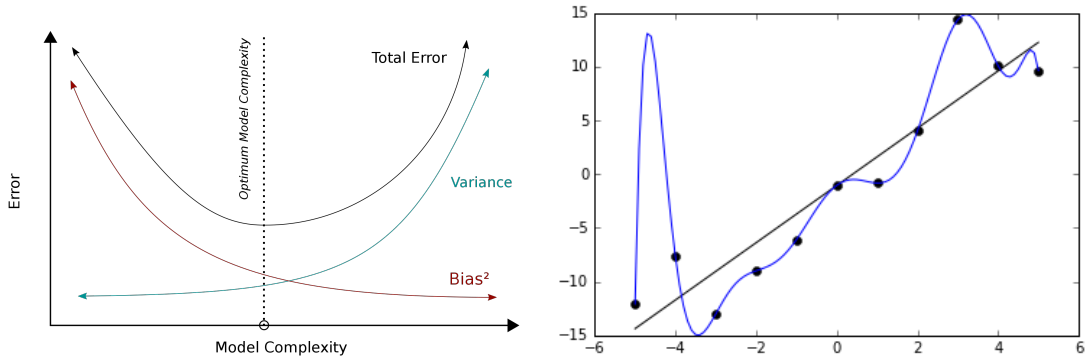


Figure 4.3: On the left, theoretical relation between the bias, variance and total errors and the model complexity taken from Wikimedia Commons under absorbers CC BY-SA 4.0 license. This is an illustration of what happens for instance on the left plot (taken from Wikimedia Commons under a CC BY-SA 4.0 license), when considering different degrees of polynomials. The lower degree linear function is more generalizable that the biased high degree polynomial that fits perfectly the data.

a high degree polynomial when a linear function is a more suitable solution as shown on Figure **??**.

The explicit regularization technique consists in penalizing the complexity of a model by adding to the global loss function an error term that scales with the complexity of the model. The error associated to a predictor f can be expressed with an additional regularization term $\Omega_n(f)$:

$$\mathcal{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}\left(f(\mathbf{x}_i), y_i\right) + \Omega_n(f) \tag{4.8}$$

And, depending on the expression of the regularization term $\Omega_n(f)$, the regularization will have more or less influence on the optimization problem.

Since the regularization is a model-specific function (depends on f), we need to define a model to study more specific expressions of regularization. Let's consider a multilinear model so that $f(\mathbf{x}) = \boldsymbol{\beta}\mathbf{x}^{\mathrm{T}}$, where $\boldsymbol{\beta} = (\beta^{(1)}, \ldots, \beta^{(p)})$ is a vectorial representation of the weights of the p features contained in $\mathbf{x}$ in the linear regression. In a standard multilinear regression, with a quadratic loss, the risk function to minimize can be expressed as:

$$\mathcal{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{\beta}\mathbf{x}_i^{\mathrm{T}} - y_i\right)^2 \tag{4.9}$$

and y_i is now a scalar in a regression problem ($\mathcal{Y} = \mathbb{R}$). One of the earliest regularization tools introduced by Tikhonov to deal with ill-posed optimization problem is the L2 regularization. Used in linear regression, this new type of model is called the ridge regression and consists simply in adding a L2-norm penalty on the model weights in the risk function as expressed in the following equation:

$$\mathcal{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}\left(f(\mathbf{x}_i), y_i\right) + \lambda_2\|\boldsymbol{\beta}\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{\beta}\mathbf{x}_i^{\mathrm{T}} - y_i\right)^2 + \lambda_2\sum_{k=1}^{p}\left|\beta^{(k)}\right|^2 \tag{4.10}$$

where λ_2 is the parameter of the L2-regularization, it controls the importance of the regularization term in the optimization process. This parameter controls the complexity of the model

and needs to be tweaked to find the optimum between accuracy and generalizability as shown on the Figure **??**. If we now consider a polynomial function, the vector $\mathbf{x}_i$ represents the vector of different exponentiations of a scalar x_i so that $\mathbf{x}_i = \left(x_i^0, \ldots, x_i^{n-1}\right)$, and the coefficients $\boldsymbol{\beta}$ are just the polynomial coefficients of the polynomial function f. This is a clear illustration of how the complexity of the model is penalized since regularization terms directly penalize the number of terms used and their influence on the fitting process. Note that this regularization can be adapted to other types of models, given that we manage to define a L2-norm of the prediction function f.

A second very common regularization term is based on the L1-norm of the prediction function. A L1-regularized least square linear regression is called a LASSO (Least Absolute Shrinkage and Selection Operator) regression, it allows for a sparser selection of the model weights by allowing zero weights in the model, which is not the case for a L2-regularization. The risk function associated with this regression model can be expressed as:

$$\mathcal{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}\left(f(\mathbf{x}_i), y_i\right) + \lambda_1 \left\|\boldsymbol{\beta}\right\|_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{\beta}\mathbf{x}_i^{\mathrm{T}} - y_i\right)^2 + \lambda_1 \sum_{k=1}^{p}|\beta^{(k)}| \tag{4.11}$$

where λ_1 is the L1-regularization parameter that controls its importance. The L1-norm can be defined differently depending on the model we consider, but the core idea is that it is a function of the absolute values of the weights of the model.

Finally, if we combine both L1 and L2-regularization, the linear regression becomes an elastic net regression and the risk function becomes:

$$\mathcal{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}\left(f(\mathbf{x}_i), y_i\right) + \lambda_{1,2}\left(\alpha\left\|\boldsymbol{\beta}\right\|_1 + (1-\alpha)\left\|\boldsymbol{\beta}\right\|_2^2\right) \tag{4.12}$$

where $\alpha \in [0, 1]$ defines the relative weight of L1 and L2 regularization term and $\lambda_{1,2}$ governs the importance of the combined regularization term. This regularization technique simply combines both L1 and L2 regularization, and the different regularization parameters can be tweaked to find the best bias–variance tradeoff for the final model. These parameters are also called hyperparameters in machine learning, because it changes the parameters at the higher model level.

Finally, implicit regularization corresponds to other forms of control of the complexity of the model. For instance, it could be the early stopping in a learning process so that we do not converge completely to the minimal error with the data. It could be discarding outliers that prevent the model from learning properly on the relevant data. It could also be in the architecture of the model, for instance the random forest is an ensemble approach that aims at reducing the overfit, and it will be presented in the next section. The learning rate in the gradient boosting is also a regularization parameter that smoothes the learning process and will be tackled in the dedicated section. The implicit regularization is related to the construction of the model itself and will therefore be explained in more details in the section on machine learning models.

Learning strategies

We previously identified the theory behind the bias–variance tradeoff, which boils down to the generalization of model that has a partial glimpse of all the available data. Yet, in practice we

need to evaluate the generalization error $\mathcal{R}_n(f) - \mathcal{R}^*$. To achieve that, the common strategy is to randomly split the available data into two sets a training set $\mathcal{D}^{\text{train}} = \left\{(\mathbf{x}_{i_1}, y_{i_1}), \ldots, (\mathbf{x}_{i_{\text{N}}}, y_{i_{\text{N}}})\right\}$ and a test set $\mathcal{D}^{\text{test}} = \left\{(\mathbf{x}_{j_1}, y_{j_1}), \ldots, (\mathbf{x}_{j_{n-\text{N}}}, y_{j_{n-\text{N}}})\right\}$ so that $\mathcal{D}_n = \mathcal{D}^{\text{train}} \cap \mathcal{D}^{\text{test}}$. The training set is used to perform the optimization problem as defined in equations **??** and **??**, and the test set is used to evaluate the generalization error since it is unseen data for the model. In practice, we choose a ratio of test data $n - \text{N}/n$ (e.g. 20%) that defines the size of the test set from an initial dataset, and the randomness of the split ensures that the data from both sets are similar yet not exactly the same. However, in some cases, one should be aware that outliers can be present in the test set, which makes the performance on the test set worse than expected. Or in some cases, the dataset is too small and every data point is very different from each other, and the test set is very different from the training set, which makes it impossible for the model to predict on the test with the piecemeal information given by the training set. The percentage of the train/test split should therefore be chosen wisely and according to the dataset so that it remains representative of the training set.

The main property of the test set is that it is a completely unsee dataset, which means that the training of the model should be independent of this set except for the very final evaluation. But in some cases, we want to compare different models with each other or change a "hyperparameter" like the regularization parameter within the same model architecture. To evaluate these models, we cannot evaluate for every model the generalization error on the test set, because it would compromise the independence of the test set with the training process. Hence, we introduce validation sets within the initial training set. We could do a simple training/validation split like we did for the test set. However it would weaken the model even more since there is less data available, and, furthermore, it does not use all the potential of the training set. A very common technique to test the performance of a model on a training set is the cross-validation. The idea is to use different training/validation splits to test the model in multiple configurations to have a better evaluation of the model performance by averaging the different performances.

The most used method is the k-fold cross-validation technique which consists in partitioning the training set $\mathcal{D}^{\text{train}}$ in k equal size subsets $\mathcal{S}_1, ..., \mathcal{S}_k$. The model is then trained on the union $\bigcup_{l \neq m} \mathcal{S}_l$ of all subsets but one subset $\mathcal{S}_m$ that will be used as a validation set for all $m \in \{1, \ldots, k\}$. The principle of the k-fold approach is illustrated on the Figure **??**. The approximate generalization error of the model is then the average of every loss calculated on the validation subsets. This tool provides a way of comparing different models without using the test set, which is extremely useful especially in the parameterization of the ML model.

Other cross-validation techniques exist and are used in specific cases, for example the stratification cross-validation consists in ensuring the same repartition of the labels y_i in each subset, which is useful in classification problems. We can also make the validation process even more exhaustive by increasing the k in the k-fold validation. However it requires k training of the models, which increase the computation time required. When increase to the maximum k is equal to the size of the training set and the method is called leave-one-out cross-validation. Finally, for time series, the cross-validation technique usually requires to sort the data according to the time history so that the training set is always prior to the validation set, which creates a whole new approach to cross-validation. The core idea of cross-validation is to find multiple training/validation splits to evaluate the model from more than one point of view, and different strategies exist depending on the training problem before us.

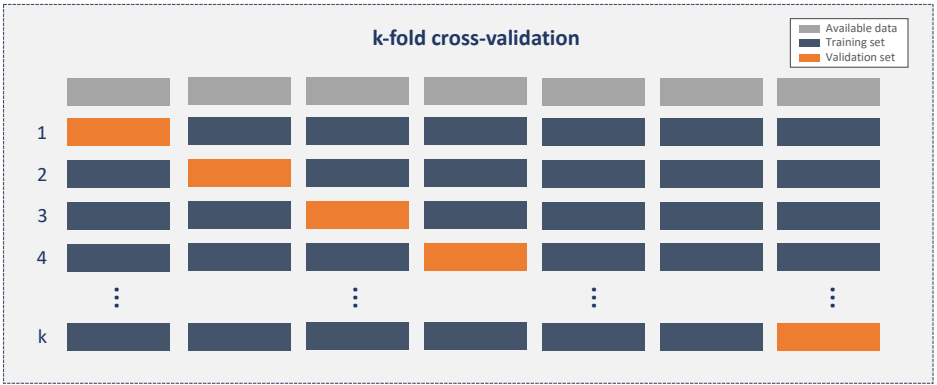


Figure 4.4: Illustration of a k-fold cross-validation. At each step, the machine learning model learns from the training set and is tested on the validation set. The average performance on all validation sets gives an approximation of the generalization error.

4.1.3 Machine learning models

In this chapter, we will go from the basic components of the model (decision tree) to the more complex ensemble model (e.g. random forest), in order to finish with the final stochastic gradient boosting model we used in this work. The discussion will be mainly focused on regression problems and not classification problems since the goal is to predict a continuous variable (the xenon/krypton selectivity).

Regression tree

Tree-based models are usually used in classification problems where depending on a set of "yes" or "no" questions the tree classifies the data points into the different predefined categories. The questions are in fact associated to threshold values of the p features or characteristics $C_1, \ldots, C_p$; for example, a node of the tree could ask the question, "Is C_1 higher than 3?", which splits the space into two categories the "yes" and the "no". This is why we can see a decision tree model as a splitting of the space into rectangles (in 2D) or an equivalent of a rectangle in p-dimensional feature space. To adapt this type of model into a regression problem, we can regroup different label values y into categories that are represented by the average label value. To sum up, a decision tree for regression splits the feature space into a set of pseudo-rectangles (volumes separated by limited hyper-surfaces) defined by the nodes of the tree, and in each of these subspaces are given the average of the different points present in this subspace. To clarify the terminology, a splitting node correspond to a separation between regions, while a terminal node or leaf corresponds to the region itself.

The CART[Breiman_2017] algorithm developed by Breiman et al. is usually presented as the archetype of a decision tree model. The algorithm is pretty straightforward to understand, three steps are required: (i) Examine every split allowed on each feature C_i, (ii) select and use the best split according to a loss function (squared error or absolute error usually), and (iii) stop splitting a node when a stopping rule is satisfied (e.g. minimum samples split).[Dension_1998] We could split indefinitely the decision tree so that each data point has its own region, but this would be a textbook case of overfitting, any new data point would never be correctly predicted with such a model. To prevent this from happening, the decision tree has a regularization parameter called minimum samples split n_{min} that only allows a further split; if the node

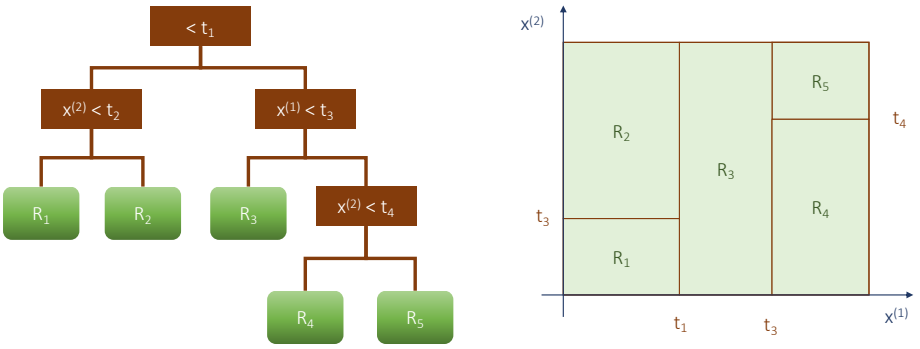


Figure 4.5: Illustration of the decision tree and the region splitting performed by a CART[Breiman_2017] *algorithm. Adapted from an illustration of the book "Elements of Statistical Leanrning" [***Hastie_2009***].*

contains less than a given number $n_{\min}$, then it is necessarily a terminal node. The decision trees are known to be very prone to overfitting, another useful regularization parameter can be used to prevent an ever-growing tree is the maximum depth of the tree, which can also stop the iterative process of tree growing. Finally, to further regularize the tree, a process called tree pruning simplifies the tree and outputs the final model. We won't go into the details of tree pruning as it is not the main subject (see Ref. [**Hastie_2009**] for further details). The final tree f_{tree} can be expressed as a function of the different regions $\mathrm{R}_1, \ldots, \mathrm{R}_\mathrm{M}$ carved by the splitting process:

$$f_{\text{tree}}(\mathbf{x}) = \sum_{m=1}^{\mathrm{M}} c_m \mathbb{1}(\mathbf{x} \in \mathrm{R}_m) \tag{4.13}$$

where c_m is the value of the leaf corresponding to R_m, and $\mathbb{1}$ is an identity function that returns 1 if the argument is true and 0 otherwise. The coefficients c_m of this function are actually equal to the average of the labeling values of the dataset $\mathcal{D}_n$ in the region R_m, i.e. $c_m = \underset{i \in \mathcal{D}_n}{\text{ave}}\,(y_i | \mathbf{x}_i \in \mathrm{R}_m)$. To put it simply, the tree function returns the average value of y (in the dataset) in the region where x (could be new data) is located.

The decision tree has the main advantage of being very easily interpretable as defined in the book of C. Molnar [**molnar2020interpretable**]. This interpretability boils down to the easily understandable binary decision at the root of the decision tree — we can easily see the explaining characteristics (R_m) of a predicted value, we can easily imagine different predictions depending on the value of $\mathbf{x}$, and for small trees we can even run the model in our own head. However, this model has a reputation of being very inefficient in finding simple linear relations resulting in a step-like function. The model is not very smooth, slight changes in the input x can have a big impact on the predicted value (typically near the separation between two regions) and some change (noise) in the training data can totally change the structure of the tree. This instability of a single tree makes it very hard to generalize over unseen data.[molnar2020interpretable] To improve this single decision tree, Breiman introduced bagging predictors in 1996 to improve the accuracy of models that are unstable with regard to small changes in the learning set.[Breiman_1996] This new approach is at the origin of the random forest, and will be presented more in depth in the following subsection.

Random Forest

The core idea behind random forest is that a collection of weak learners, called an ensemble model, is better than a single strong learner, this assumption relies on a proven theorem that states that the minimal error of a forest is lower than the error of a single tree (theorem 11.2. of Ref. [**Breiman_2001**]). The strength of model depends on the amount of information we feed into the model and its complexity. To achieve a diverse forest of weaker decision trees, we need to introduce two concepts; the first one is the bootstrap aggregating (bagging) and second one is the random column subsampling. Both methods ensure a diversity in the generated trees by using random selections and also a relative weakness of the trees by reducing the amount of information it can access.

The bagging method consists in generating a set $\{\phi_b\}_{b\in\{1,\ldots,\mathrm{B}\}}$ of B weaker learners from different bootstrap datasets $\{\mathcal{D}_b^{\text{train}}\}_{b\in\{1,\ldots,\mathrm{B}\}}$. Each bootstrap dataset $\mathcal{D}_b^{\text{train}}$ is generated by randomly selecting t elements of $\mathcal{D}^{\text{train}}$ using a sample with replacement — note that each bootstrap sample has the same number of elements than $\mathcal{D}^{\text{train}}$ but data points can appear several times in it. The number of times a data point $(\mathbf{x}_i, y_i)$ appears represents the weight of this data point in the bootstrap learning set. To simplify, we can say that each tree model ϕ_b learns on the $\mathcal{D}_b^{\text{train}}$ dataset that have randomly defined weights on the different data points, which means that every model will pay attention to different parts of the training data. We can also evaluate the generalization error of the model since some trees have never seen some data points, we can evaluate the generalization error on the unseen data for every tree (similar to cross-validation), this error is called the out-of-bag error.

The second technique consists in randomly choosing a subsample of the features on which to find the best split (second part of the CART tree growing algorithm). This technique is inspired from the one developed by Ho in 1998, where each tree of a forest is trained only on a randomly chosen feature subspace.[**Tin_Kam_Ho_1998**] The only tweak in the procedure lies in the fact that the feature space changes at each iteration of the tree growing process instead of between each tree generation. This method also improves the generalizability of the method by weakening each tree so that they don't overfit, the accuracy is achieved by the aggregation of all the trees.

The random forest as formulated by Breiman combines these two randomness-based techniques to train a forest.[**Breiman_2001**] The algorithm starts by looping over the number of trees B in the forest, for each tree b a bootstrap sample $\mathcal{D}_b^{\text{train}}$ is randomly drawn (with replacement) and this data is used to grow the tree (training procedure). In the training, a modified CART algorithm is applied to grow the tree by splitting recursively on each node: (i) instead of testing all features for the best splitting, only a random selection of m variables is considered among the p features, (ii) the best split point is selected among the m variables, and (iii) the node is split in two until the minimum leaf size n_{min} is reached. The size of the column subsample defines the number of features to randomly consider at each split; this is another implicit regularization parameter associated with the random forest along with the previously identifies regularization parameters of the decision tree such as the minimal leaf size n_{min} or the maximal depth of a

tree. Finally, we have a set of B trees $\{\phi_b\}$ that can be used to make an ensemble model Φ so that:

$$\Phi(\mathbf{x}) = \frac{1}{\mathrm{B}} \sum_{b=1}^{\mathrm{B}} \phi_b(\mathbf{x}) = \frac{1}{\mathrm{B}} \sum_{b=1}^{\mathrm{B}} \sum_{m=1}^{\mathrm{M}_b} c_{m,b} \mathbb{1}(\mathbf{x} \in \mathrm{R}_{m,b}) \tag{4.14}$$

Note that each tree has an equal amount of say in the prediction, and they are trained on different random samples of the initial training data. Random forest is known to be less prone to overfitting because it is the produce of a sort of cross-validation process called bootstrapping. However, the algorithm does very little improve the accuracy (bias error) of the model it relies on the belief that each tree will naturally compensate their mutual weaknesses in the final ensemble model. In the next section, we will see another algorithm that focuses on guiding each tree based on the prior knowledge of previous trees to perform better — this new technique is called boosting.

From boosting to gradient boosting

In the previous approach, the bootstrap dataset is a random selection of the samples in the training set $\mathcal{D}^{\text{train}}$ and each tree has an equal amount of say in the final ensemble decision. In a boosting algorithm,**drucker1997improving** the paradigm changes, data samples are (i) selected according to how they were predicted by the previous trees in order to focus our attention on the poorly predicted sample points, and (ii) the tree ϕ_b that is trained on this weighted dataset $\mathcal{D}_b^{\text{train}} = \left\{ \left(w_i^{(b)}, \mathbf{x}_i, y_i \right) \right\}$ is also evaluated using a measure of confidence α_b that depends on the error made (the higher the error, the lower the confidence), this measure of confidence is used to define the ensemble model:

$$\Phi_{\mathrm{B}} = \frac{1}{\sum_{b=1}^{\mathrm{B}} \alpha_b} \sum_{b=1}^{\mathrm{B}} \alpha_b \phi_b \tag{4.15}$$

To train each individual tree ϕ_b of this forest, we use the CART algorithm (described in the previous sections), but by minimizing a weighted risk function instead of the standard one:

$$\mathcal{R}(\phi_b) = \sum_{i=1}^{\mathrm{N}} w_i^{(b)} \mathcal{L}\left(\phi_b(\mathbf{x}_i), y_i\right) \tag{4.16}$$

where $w_i^{(b)}$ is the normalized weight corresponding to the error $\mathcal{L}\left(\Phi_{b-1}(\mathbf{x}_i), y_i\right)$ made by the previous ensemble Φ_{b-1} on each data point $(\mathbf{x}_i, y_i)$; for $b = 1$, there is no previous model so the weights are just equidistributed across the samples, i.e. $\forall i,\ w_i^{(1)} = 1/\mathrm{N}$. In practice, to simulate the weighting process, a random selection with probability $w_i^{(b)}$ on each sample is performed to draw an equal-sized N training dataset $\mathcal{D}_b^{\text{train}}$ for ϕ_b.

I voluntarily did not go into the details of the confidence rate α_b, because several implementations of it exist. But generally, it is a decreasing function of the total error of the tree on the weighted dataset. The AdaBoost algorithm typically uses the half the opposite of the logit transform function $\alpha_b = 0.5 \log\left((1 - \mathcal{R}(\phi_b))/\mathcal{R}(\phi_b)\right)$ that goes to $+\infty$ for very small errors and $-\infty$ for very large ones.**Freund_1997**, **schapire2013explaining** Gentle AdaBoost would give equal amount of say to each tree independently of their performance, which can in some cases allow better generalization performance than a regular AdaBoost. The very high values of α_b could in some cases make the model overfit, because a very good performance on the weighted dataset

could mean a good fit on noisy data points.[schapire1998improved] To prevent overfitting early stopping procedure with a cross-validation (k-fold usually) training procedure is performed to determine the ideal number of trees required to stay generalizable while having reduced the bias error — like always in machine learning it is a question of bias–variance tradeoff.

AdaBoost in its original implementation uses stamps which are trees composed of a unique splitting node and two leaves, but boosting algorithms can be applied to any tree depth. This tree depth hyperparameter is very important in tree-based models since it defines the complexity/strength of each learner tree. The smaller the tree the less overfitting can occur (see link between complexity and variance on Figure **??**), and the AdaBoost algorithm uses the smallest possible tree in order to compensate the very aggressive learning procedure used. The main takeaway from this study is that boosting focuses on the training trees that compensate the errors of previous trees, and it can play with tree-based hyperparameters (e.g. tree depth, number of trees) to control the variance error.

Actually boosting can be reformulated as a gradient descent problem as formulated by Mason et al.[mason1999boosting] We can prove that AdaBoost is simply a gradient boosting algorithm with an exponential loss (same loss and derivative) and with the steepest gradient descent logic.[mason1999boosting, azencott2022introduction]

Each additional tree ϕ_b in a gradient boosting can be interpreted as a contribution to a predictor Φ_b to minimize an objective function $\mathcal{R}(\Phi_b)$. And the weight $w_i^{(b)}$ which measures how badly each sample i are predicted, can be expressed as a derivative of a differentiable loss function $\mathcal{L}$ since the minimum is reached for a zero derivative.

$$w_i^{(b)} = - \left. \frac{\partial \mathcal{L}\left(y_i, \hat{y}_i\right)}{\partial \hat{y}_i} \right|_{\hat{y}_i = \Phi_{b-1}(\mathbf{x}_i)} \tag{4.17}$$

where $\hat{y}_i$ is a derivation variable describing the ensemble tree prediction, here evaluated at Φ_{b-1}. Instead of predicting the y_i values, we can predict the weight, also called the pseudo-residual, $w_i^{(b)}$ that measures how far the previous model Φ_{b-1} is from the ideal Φ (zero weights everywhere in an ideal world) and compensate it using a tree ϕ_b. In other words, we use the CART framework to grow a tree ϕ_b that predicts the gradients $w_i^{(b)}$ from the features $\mathbf{x}_i$, which iteratively improves the model Φ_b compared to Φ_{b-1}:

$$\Phi_b = \Phi_{b-1} + \eta \phi_b \tag{4.18}$$

where η is the learning rate or shrinkage, introduced by Friedman in his stochastic gradient boosting, to slow down the learning process in order to limit overfitting.[Friedman2002] In a steepest descent step, the values of this learning rate η actually minimizes the risk function $\mathcal{R}(\Phi_{b-1} + \eta\phi_b)$ associated to the output model Φ_b. If $b = 1$, the first estimator Φ_1 is simply a constant function that minimizes the risk over the training set $\Phi_1(\mathbf{x}) = \arg\min_{c \in \mathbb{R}} \sum_{i=1}^{\mathrm{N}} \mathcal{L}(y_i, c)$. For a quadratic loss function, this constant corresponds simply to the average of the y_i values over the training set.

In the particular case of a quadratic loss $\mathcal{L}_{\mathrm{SE}} = \frac{1}{2}(y_i - f(\mathbf{x}_i))^2$ that is used in this chapter, the gradient boosting algorithm can be simply broken down into the three steps below:[Friedman2002]

1. Initialization at $b = 1$ with a constant:
   $\Phi_1(\mathbf{x}) = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} y_i$

2. For $b = 2$ to B:

   (a) Compute the pseudo-residuals that are actually real residuals in the case of a quadratic loss $\quad \forall i \in \{1, \ldots, \mathrm{N}\},\ w_i^{(b)} = y_i - \Phi_{b-1}(\mathbf{x}_i)$

   (b) Train the weak tree ϕ_b on the dataset $\{(\mathbf{x}_i, w_i)\}_{i \in \{1,\ldots,\mathrm{N}\}}$.

   (c) Update the model using a fixed learning rate $\eta \in [0, 1]$ instead of finding $\eta = \underset{c \in \mathbb{R}}{\arg\min} \sum_{i=1}^{\mathrm{N}} \mathcal{L}\left(y_i, \Phi_{b-1}(\mathbf{x}_i) + c\phi_b(\mathbf{x}_i)\right)$ through a minimization problem (steepest gradient descent). $\quad \Phi_b = \Phi_{b-1} + \eta\phi_b$

3. Output the final ensemble model Φ_{B}

Up until now, I showed the different ways of using decision trees to perform a prediction on a training dataset $\mathcal{D}^{\text{train}}$ by focusing on mainly two ensemble models: random forest and gradient boosted trees. The reason why I went through these models is to be able to present the model we used in our prediction model that aggregates techniques from both ensemble models. This model, called eXtreme Gradient Boost or XGBoost, was introduced by Chen et al. and is an improvement compared to similar methodologies due to its scalability. We won't go into the details of the implementation improvement (for more details see Ref. [**chen2016xgboost**]), but we will rather focus on the basic framework it uses to better understand the model we used.

XGBoost model parameterization

The XGBoost model is basically a gradient boosting model as decribed in the previous section but with a few regularization parameters that could be fine-tuned to improve its generalizability. In a learning problem with N learning examples and p features/descriptors, we can express the predictor Φ as the sum of weaker tree learners ϕ_b:

$$\Phi(\mathbf{x}) = \sum_{b=1}^{\mathrm{B}} \phi_b(\mathbf{x}) = \sum_{b=1}^{\mathrm{B}} \sum_{m=1}^{\mathrm{M}} c_m^{(b)} \mathbb{1}(\mathbf{x} \in \mathrm{R}_m^{(b)}) \tag{4.19}$$

where M is the maximal number of leaves a tree can have, in our implementation — this number is fixed using the maximum depth $\max_{\text{depth}}$ in the algorithm since $\mathrm{M} = 2^{\max_{\text{depth}}}$, and B is the maximum number of estimators in the ensemble model. The number of estimators is usually determined using an early stopping in the k-fold cross-validation.

Then, we used a quadratic loss function regularized with L1 and L2-regularization terms on the M leaf weights c_m of a model *phi* so that the loss function $\mathcal{L}$ can simply be expressed as:

$$\mathcal{L}\left(y, \phi(\mathbf{x}_i)\right) = \tfrac{1}{2}\left(y - \phi(\mathbf{x}_i)\right)^2 + \lambda_1 \sum_{m=1}^{\mathrm{M}} \left|c_m^{(b)}\right| + \lambda_2 \sum_{m=1}^{\mathrm{M}} \left|c_m^{(b)}\right|^2 \tag{4.20}$$

where λ_1 and λ_2 are the L1 and L2-regularization coefficients that control the importance of each regularization term.

The risk function $\mathcal{R}$ of a tree ϕ_b with M leaf weights $c_m^{(b)}$ at the iteration b of the gradient boosting process can then be written:

$$\mathcal{R}(\phi_b) = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \tfrac{1}{2}\left(w_i^{(b)} - \phi_b(\mathbf{x}_i)\right)^2 + \lambda_1 \sum_{m=1}^{\mathrm{M}} \left|c_m^{(b)}\right| + \lambda_2 \sum_{m=1}^{\mathrm{M}} \left|c_m^{(b)}\right|^2 \tag{4.21}$$

where $w_i^{(b)}$ is the pseudo-residuals of the previous model on the dataset. This expression of the risk is typically used in the tree-splitting process of the step 2.(b) of the gradient boosting algorithm (see previous subsection **??**) to find the best tree to predict the pseudo-residuals. As previously explained, the pseudo-residual is simply the difference between the observed value y_i and the previously predicted value $\Phi_{b-1}(\mathbf{x}_i)$, also known as the residual in regression problems, in the case of a quadratic loss:

$$w_i^{(b)} = -\left.\frac{\partial \mathcal{L}\left(y_i, \hat{y}_i\right)}{\partial \hat{y}_i}\right|_{\hat{y}_i=\Phi_{b-1}(\mathbf{x}_i)} = y_i - \Phi_{b-1}(\mathbf{x}_i) \tag{4.22}$$

The learning rate η used to update the ensemble model is also a key component of the final model that we will need to tweak in order to maximize the generalizability of the model. The presence of this parameter means that we are converging slower to the solution, which is better for the bias–variance tradeoff. Small values below 0.1 are usually used.

Variable name in XGBoost	Variable in this work	Description of the hyperparameter
"n_estimators"	M	Number of trees in the final ensemble model
"max_depth"	$\simeq \log_2(\mathrm{T})$	Maximum number of levels allowed for each tree that can be expressed as a function of T the number of terminal nodes or leaves
"alpha"	λ_1	L1-regularization parameter
"lambda"	λ_2	L2-regularization parameter
"learning_rate"	η	The shrinkage or learning rate used to update the ensemble model with each basic tree.
"subsample"	$\mathrm{N}_{\mathrm{sample}}/\mathrm{N}$	The ratio of data points randomly sampled (without replacement) for the training of each tree ϕ_b
"colsample_bytree"	p_{tree}/p	The ratio of features randomly sampled per tree iteration (on $b = 1$ to B)
"colsample_bylevel"	p_{level}/p	The ratio of features randomly sampled per level iteration (on $k = 1$ to M, this would be on the leaves really but to simplify)

Table 4.1: Hyperparameters of XGBoost relevant to our work.

We also used three other parameters that are very close to the ones implemented in a random forest to add randomness in the gradient descent procedure. Augmented by these techniques, the model can now be called stochastic gradient boosting as described in the Ref. [**Friedman2002**]. At each iteration, a subsample of the training data is drawn at random (without replacement) according to a parameter $\mathrm{N}_{\mathrm{sample}}/\mathrm{N}$. This parameter has a similar effect as the bagging procedure of the random forest, it narrows the attention of each weak learner on a portion of the learning set, which reduces overfitting like in a cross-validation procedure. The different trees learn from different parts of the training set, which means that the ensemble model can never overfit on the whole dataset. This provides a handy solution to the infamous overfitting problem of standard gradient boosting. Another procedure concerns the random selection of the feature columns. This idea was developed in the Ref. [**Tin_Kam_Ho_1998**] and randomly

extracts a subsample of the features for training of one tree; a parameter needs to be chosen in order to determine the size of the portion of features p_{tree}/p used to train each tree. A similar idea is to make the column sampling at each level instead of each tree, and a proportion p_{level}/p can be defined like this. Similarly we can make a feature selection at the node level but this parameter was not used.

Finally, we compiled all the parameters used in the construction of the final model in the table **??**. This table contains a tree-specific parameter "max_depth", but also an ensemble specific one "n_estimators", in addition to very general regularization parameters inspired by linear models "alpha" and "lambda", as well as more gradient boosting specific parameter "learning_rate" and more randomness-based hyperparameters inspired by random forest such as "subsample", "colsample_bytree" and "colsample_bylevel". This model can be considered as a mixing pot of a variety of ideas coming from all corners of the data science field. Using this machine learning model, we will try to solve the selectivity drop problem that puzzled us in the previous chapter.

4.2 Prediction of the Ambient-pressure Selectivity

Before diving deep into the model of our work, let us review the different literature contributions to xenon/krypton separation screenings. 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.**Simon_2015** 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.**Rycroft_2009** 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 (see section **??**). 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,**Shi_2023** 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 have previously shown that selectivity could drop between the low and ambient pressure cases in the Xe/Kr separation application (see chapters 2 and 3), and it was mainly attributed to the presence of different pore sizes and potential reorganizations due to adsorbate–adsorbate interactions.

We combined grid-based descriptors described in the previous chapter (section **??**) to statistical characterizations of the pore size to propose a set of useful ML descriptors for fast and

accurate ambient-pressure selectivity prediction using an optimized XGBoost model. And we highlight its performance on the case of xenon/krypton separation in the CoRE MOF 2019 database.[Chung_2019]

4.2.1 Data Preparation

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 RASPA software[dubbeldam2016] to run GCMC calculations (introduced in section **??**) 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),[rappe1992] and the guest atoms (xenon and krypton) have their LJ parameters taken from a previous screening study.[Ryan_2010] The study only focuses on a given Xe/Kr composition usually obtained by cryogenic distillation of ambient air[kerry2007industrial] as a first step toward predicting other mixtures at different physical conditions (*e.g.* Xe/Kr mixtures out of nuclear off-gases).

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 selectivity values 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 selectivity values. This approach distributes more evenly the efforts on all the different values of selectivity. Moreover, this logarithmic transform is related to a thermodynamic quantity that we elaborate later in the section **??**; it can therefore be easily compared with the energy descriptors we introduced in this chapter.

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.[Chung_2019] After removing the disordered and the non-MOF structures as well as the ones with a large unit cell volume of 20 nm^3, we obtained a set of 9,748 structures. Then we analyze the string information given by the Zeo++ software[zeopp_Willems2012] 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,[Diwu_2010] Np-MOF KASHUK,[Martin_2017] U-MOF ABETAE[Jouffret_2011] or Th-MOF ASAMUE[Liang_2016]) — 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 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 (see section **??**).

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.

4.2.2 Feature engineering

Geometrical and chemical ML descriptors

Looking at a number of different research papers on supervised ML for the prediction of adsorption properties,[Fernandez_2013, Simon_2015, Fanourgakis_2020, Anderson_2020, Pardakhti_2020] we see that some descriptors are recurrent: (i) geometrical descriptors obtained by software like Zeo++[zeopp_Willems2012] such as the surface area (SA), the void fraction (VF), the largest cavity diameter (LCD) and the pore limiting diameter (PLD); and (ii) 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,[Hung_2021] we defined a list of vdW radii to be read by the Zeo++ software (more details on `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 $3k_BT/2$, for T = 298 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.[vol_Ongari2017]

Because our goal is to predict the difference between the low-pressure selectivity and the ambient-pressure one (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 chapter 2 sketched the role of a diverse distribution of pores with different xenon affinities. 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).

Pore size statistics

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.[poresize_Pinheiro2013] 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, BC = $(\gamma^2 + 1)/k$, represents a simple quantification of how far we are from the unimodality based only on skewness and kurtosis.[Tarba_2022]

Finally, to further measure the diversity of pores, we introduced an effective number $n_{\text{eff}} = \mathrm{N}^2/\sum n_i^2$ 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,,[neffposci_Laakso1979] in ecology the inverse Simpson's index evaluates the species diversity in an ecosystem,[neffbio_Simpson1949] or in quantum physics the inverse participation number measures the degree of localization of a wave-function.[neffphys_Kramer1993] This effective number of pore sizes gives an idea of the diversity of pore sizes (considering a binning of 0.1 Å). A highly 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 carry 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.

Grid-based and geometrical descriptors

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 (section **??**). 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 **??**. 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 calculated by grid sampling 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.

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

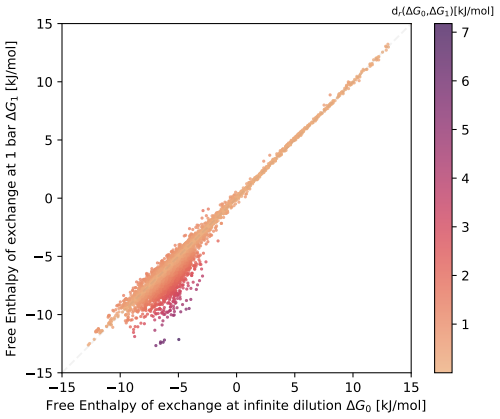


Figure 4.6: Comparison between the Gibbs free energy of exchange at low pressure ΔG_0 and ambient pressure ΔG_1 labeled by the relative distance between them. This plot is equivalent to a logarithmic plot of the selectivity at these two pressure conditions.

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. All these quantities have calculated and compared to the ambient-pressure selectivity in the previous chapter (section **??**).

As explained in the previous chapter, Boltzmann averaging at higher temperature gave better result in describing the ambient-pressure selectivity description. 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 **??**). We used this descriptor to build several descriptors presented in the Table **??**. As we can see in the Figure **??**, 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 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.

Feature name	Description
"ASA_m2/cm3_1.2"	Volumetric surface area accessible to a nitrogen probe (1.2 Å) in $m^2\ cm^{-3}$
"delta_VF_18_20"	Difference of void fraction occupiable by a krypton (1.8 Å radius) and a xenon (2.0 Å radius) probe. Always positive due to the difference of probe radii.
"PO_VF_2.0"	Void fraction occupiable by a xenon probe of 2.0 Å radius
"D_i_vdw_uff298"	Largest cavity or largest included sphere diameter (LCD). Structures atom radii are defined using the UFF forcefield[1]
"D_f_vdw_uff298"	Pore Limiting Diameter (PLD) or largest free sphere diameter defined similarly to the LCD
"pore_dist_mean"	Mean value of the pore size distribution or the average pore size
"delta_pore"	Difference between the LCD and the average pore size: "delta_pore" = "D_i_vdw_uff298" – "pore_dist_mean"
"pore_dist_std"	Standard deviation of the pore size distribution
"pore_dist_skewness"	Skewness (third order standardized moment) of the pore size distribution
"pore_dist_kurtosis"	Kurtosis (fourth order standardized moment) of the pore size distribution
"pore_dist_neff"	Effective number of data associated to the pore size distribution: N_{eff} = sum(weights)^2 / sum(weights^2)
"pore_dist_modality"	Sarle's bimodality coefficient (BC) of the pore size distribution: BC = kurtosis – skewness^2
"C%"	Percentage of carbon (C) in the MOF structure
"O%"	Percentage of oxygen (O) in the MOF structure
"N%"	Percentage of nitrogen (N) in the MOF structure

Table 4.2: Description of geometrical and chemical features used in the ML model.

[01] Using the approach of Ref. [**Hung_2021**]

Feature name	Description
"G_0"	Low-pressure Xe/Kr exchange Gibbs free energy defined using the low-pressure selectivity: $\Delta_{exc}G^{Xe/Kr} = -RT\ln\left(s^{Xe/Kr}\right)$
"G_Xe_900K"	High temperature Xe adsorption Gibbs free energy defined using the Henry's constant: $\Delta_{ads}G^{Xe}(T_h) = -RT_h\ln\left(RT_h\rho_f K_H^{Xe}(T_h)\right)$
"G_Kr_900K"	High temperature Kr adsorption Gibbs free energy: $\Delta_{ads}G^{Kr}(T_h)$
"G_900K"	High temperature Xe/Kr exchange Gibbs free energy: $\Delta_{exc}G^{Xe/Kr}(T_h) = -RT_h\ln\left(K_H^{Xe}(T_h)/K_H^{Kr}(T_h)\right)$
"delta_G0_298_900"	Difference of exchange free energies between the ambient temperature and high temperature: $\Delta_T H^{Xe/Kr} = \Delta_{exc}G^{Xe/Kr}(T_h) - \Delta_{exc}G^{Xe/Kr}(T_0)$
"delta_H0_Xe_298_900"	Difference of Xe adsorption enthalpy between the ambient temperature and high temperature: $\Delta_T H^{Xe} = \Delta_{ads}H^{Xe}(T_h) - \Delta_{ads}H^{Xe}(T_0)$
"delta_TS0_298_900"	Difference of exchange entropic term between the ambient temperature and high temperature: $\Delta_T(-T\Delta_{exc}S^{Xe/Kr}) = \Delta_T(\Delta_{exc}G^{Xe/Kr}) - \Delta_T(\Delta_{ads}H^{Xe} - \Delta_{ads}H^{Kr})$
"enthalpy_std_xenon"	Standard deviation of the Boltzmann weighted Xe energy distribution
"enthalpy_std_krypton"	Standard deviation of the Boltzmann weighted Kr energy distribution
"enthalpy_skew"	Skewness of the Boltzmann weighted Xe energy distribution
"enthalpy_modality"	Bimodality coefficient of the Boltzmann weighted Xe energy distribution
"mean_grid_xenon"	mean value of the xenon interaction energy distribution
"mean_grid_krypton"	mean value of the krypton interaction energy distribution
"std_grid_xenon"	standard deviation of the xenon interaction energy distribution
"std_grid_krypton"	standard deviation of the krypton interaction energy distribution

Table 4.3: Description of the 15 energy-based features used in the ML model. Thermodynamic descriptors are always defined at low pressure since they are derived from an interaction energy grid. Temperatures are defined as follows: T_0=298 K *and T_h=900* K*. All these energy values are defined in* kJ mol^{-1}.

4.2.3 Model training

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.[chen2016xgboost] Compared to more standard tree-based algorithms such as random forest (commonly used in the field[Simon_2015]), the boosting component of the algorithm means that it learns from previous mistakes and puts more efforts 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 is defined later in a section dedicated to the interpretation of the model. The hyperparameters of the model were fine-tuned using random searches to design the best performing final model. Finally, the influence of the preselected descriptors on the final model is interpreted using a unified approach.

Hyperparameter optimization

A hyperparameter search consists in finding the best model to optimize the generalization error as defined in equation **??**. To do so, the most common strategy consists in doing cross-validations to evaluate different model configurations. This process is called the hyperparameter search or optimization. Here, we used the randomized search algorithm to find the best parameters within a predefined reasonable range of parameters. The Python code below allowed us to find a set of optimal hyperparameters (please refer to Table **??** for the meaning of the parameters) for our final ML model.

```
import xgboost as xgb
from sklearn.model_selection import RandomizedSearchCV

xgbr = xgb.XGBRegressor()
params = {
            'n_estimators': [1500],
            'max_depth': [5,6],
            'learning_rate': [0.02,0.04,0.06,0.08],
            'colsample_bytree': np.arange(0.6, 1.0, 0.05),
            'colsample_bylevel': np.arange(0.6, 1.0, 0.05),
            'alpha': np.arange(0, 4, 0.2),
            'subsample': np.arange(0.6, 0.95, 0.05),
          }
clf = RandomizedSearchCV(estimator=xgbr,
                         param_distributions=params,
                         scoring='neg_mean_squared_error',
                         n_iter=30000,
                         verbose=1)
```

```
clf.fit(X_train, y_train.to_numpy())
print("Best parameters:", clf.best_params_)
print("Lowest RMSE: ", np.sqrt(-clf.best_score_))
```

We used the training data to perform a random search of hyperparameters, with a 5-fold cross-validation to evaluate the root mean squared errors (RMSE) of the model. After 30,000 iterations in the hyperparameter space defined by the ranges above, we found a set of optimal hyperparameters that gives an average RMSE of 0.36 kJ mol^{-1} to be used in our final model.

```
optimal_params = {
    'objective':'reg:squarederror',
    'n_estimators': 1500,
    'max_depth': 6,
    'colsample_bytree':0.85,
    'colsample_bylevel':0.65,
    'subsample':0.7,
    'alpha': 0.4,
    'lambda' :1,
    'learning_rate': 0.04,
    }
```

To confirm the relevance of the model we performed another 5-fold cross-validation to obtain a convergence plot of the XGBoost model with this set of parameters shown in Figure **??**. 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.

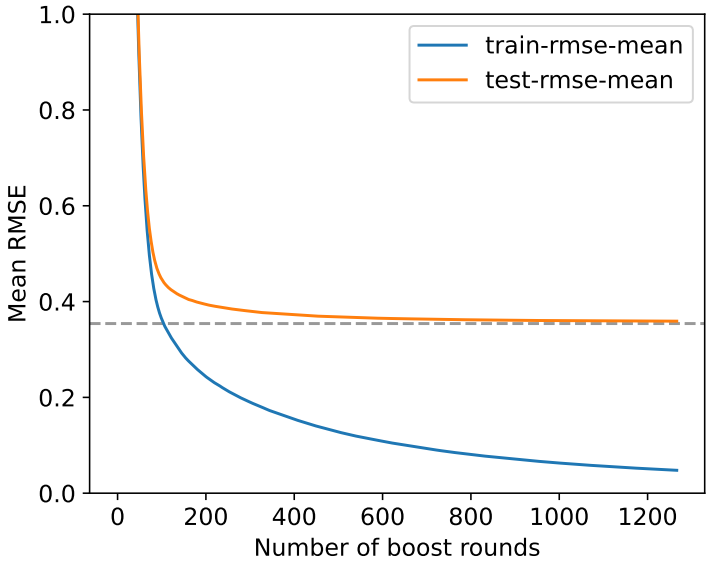


Figure 4.7: Convergence plot of the cross-validation training of our ML model. limit value: 0.36 kJ mol^{-1}

4.2.4 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.

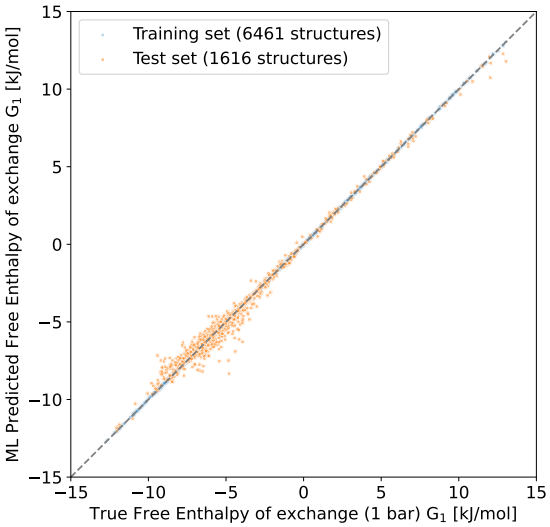


Figure 4.8: Scatter plot of the exchange free energy predicted by the model. There is a good agreement between the predicted and true values. On the test set, there is an RMSE of 0.37 kJ mol^{-1} *and an MAE of* 0.21 kJ mol^{-1}*. This plot is equivalent to the scatter plot between the logarithm of the ambient-pressure selectivity (see Figure* **??***). The corresponding errors for the ambient-pressure 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 the section **??**. By using this parameterization, our XGBoost model has an RMSE of 0.37 kJ mol^{-1} and an MAE of 0.21 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 kJ mol^{-1} with a standard deviation of 0.01 kJ mol^{-1}, which is consistent with the performance given by a standard train/test split.

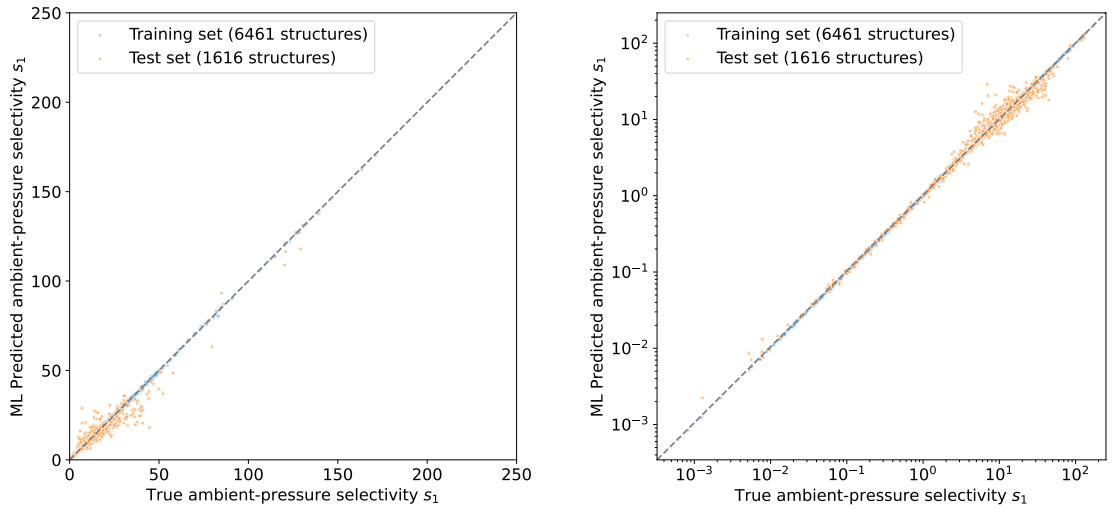


Figure 4.9: Scatter plot of the selectivity derived by the model

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 **??**, 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.

4.3 Opening the Black Box

To better understand the intuition behind this selectivity drop, we used the SHAP[**SHAP**, **molnar2020interpretable**] 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[**shapley1953value**] 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;[**SHAP**] 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

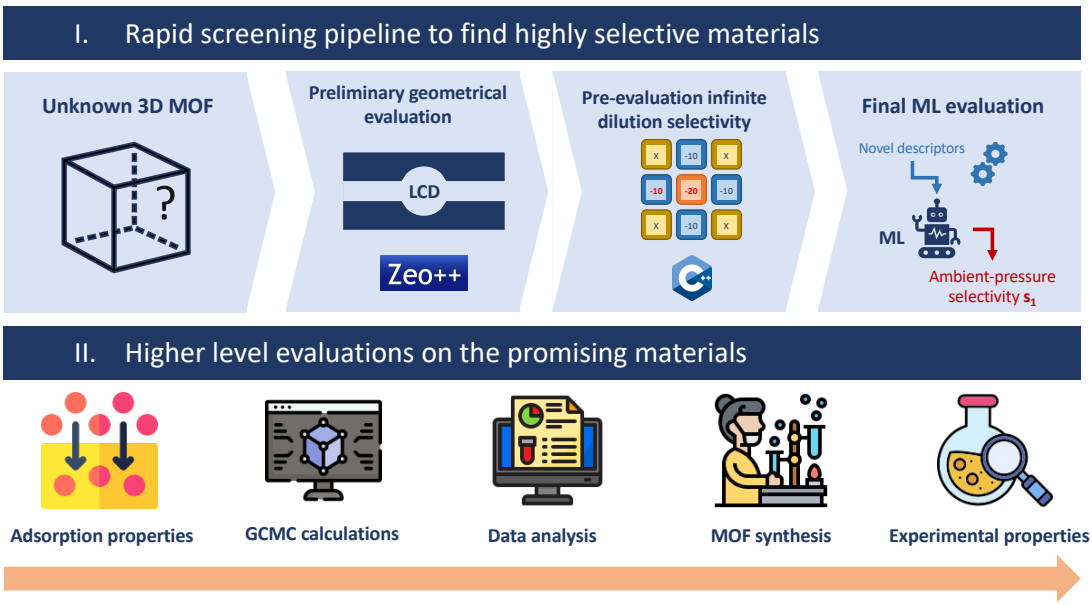


Figure 4.10: An illustration of the screening procedure that could be used to find highly selective materials.

can use the mean absolute Shapley values of each feature on the training set to measure the feature importance (see Figure **??**).

Explainable AI

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,[shapley1953value] 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.[molnar2020interpretable] The SHAP tool developed by Lundberg and Lee[SHAP] 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.

4.3.1 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 **??**. 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 **??** can be reached; the other descriptors play a major role in moving the outliers of the figure closer to the diagonal line.

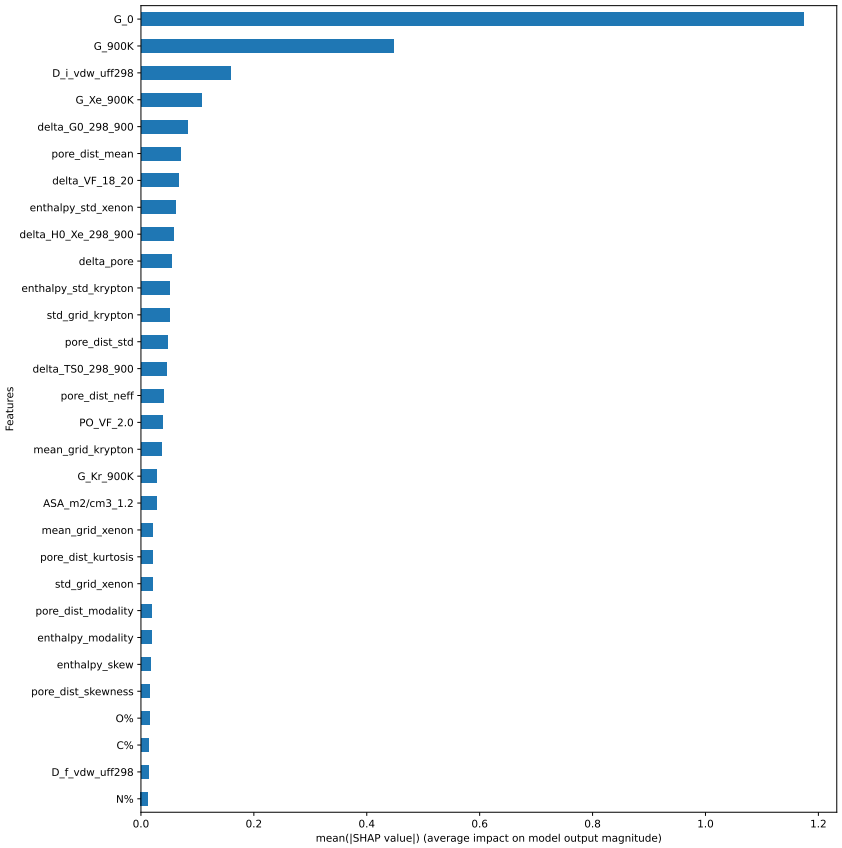


Figure 4.11: Barplot of the features importance for all the descriptors of our final model. The descriptor labels used in this section are explained in more detail in Tables **??** *and* **??**.

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),**molnar2020interpretable** 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 **??**), 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 previous chapter (Figure **??** and **??**), 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 **??**), 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 **??**) 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.**Hung_2021** Moreover, values of "delta_G0_298_900" between 4 and $6\,\mathrm{kJ\,mol^{-1}}$ (see Figure **??**) 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 **??**), 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 **??**); 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.

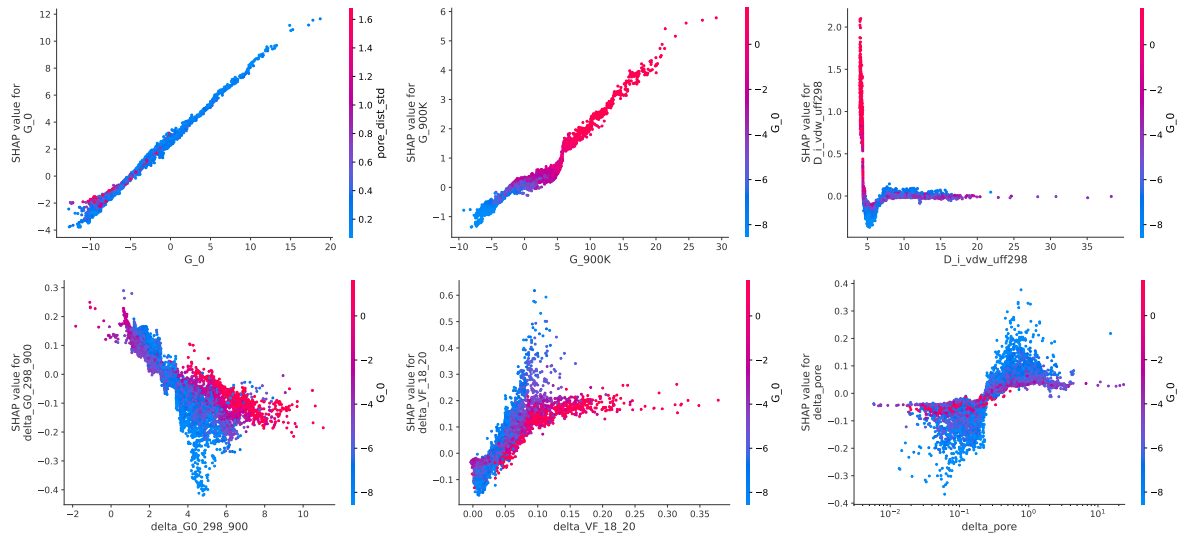


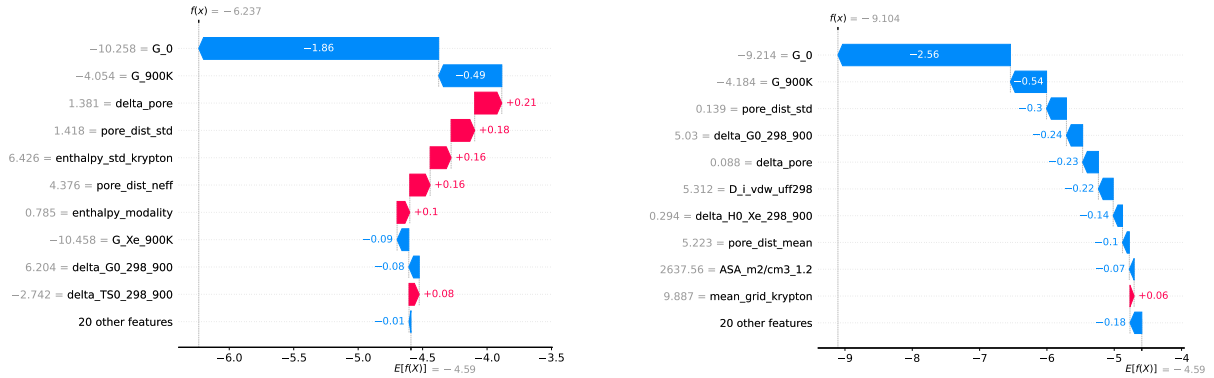
Figure 4.12: Some SHAP dependence plots that are analyzed here.

4.3.2 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.

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 ??), 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 ??: "pore_dist_std" is above the threshold of 0.4, "enthalpy_std_krypton" is above 2.5 kJ mol^{-1}, "pore_dist_neff" is above 2.3, "delta_TS0_298_900" is below 3 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. 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 (chapter 2).



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

Figure 4.13: Main contributions of the descriptors on the selectivity prediction of two archetypal examples. The descriptor labels used are detailed in the Table **??** *and* **??**.

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 kJ mol^{-1} (see Figure **??**). 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. 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 **??**) 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 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.

4.3.3 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 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 (chapter 2) found that a high diversity of pore sizes and channel sizes that favor adsorbate reorganizations could be at the origin of this phenomenon. 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.[Li_2019, Pei_2022] Moreover, the structures can be made flexible using flexible force fields with adapted simulation methodologies[Bousquet2012] or by using multiple rigid simulations of snapshots from NPT simulations.[Witman_2017] 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 chapter. The quest of ever-faster evaluation tools will allow us to investigate more complex properties and uncover structures with ever more interesting characteristics.

Data Availability:

5

XENON AND KRYPTON TRANSPORT PROPERTIES

5.1 Modeling the Diffusion Process 149
5.1.1 Molecular dynamics 150
5.1.2 Lattice kinetic Monte Carlo 153
5.2 Screening of transport properties 154
5.2.1 Diffusion in a selective material 154
5.2.2 Structure–diffusivity relationships 155
5.2.3 Identification of interesting materials 156
5.3 Fast diffusion calculation algorithm 156
5.3.1 Implementation in C++ 156
5.3.2 Preliminary results 156
5.3.3 Visualization tool 156
5.3.4 Development of a first prediction model 156

In separation processes, diffusion can either be the main performance metric or a neglected secondary parameter. There are actually two different use cases for separation using nanoporous materials: the adsorption-based separation that is mainly a thermodynamic process and the nanoporous separation membranes that relies on the kinetic properties. In a membrane-based process, the gas is sieved through a membrane material that blocks some molecules (e.g. Xe) and let other molecules (e.g.) diffuse freely. The performance of the separation is measured with the ratio of the diffusion coefficients instead of the thermodynamic selectivity we defined in chapter 2. The process of interest is, however, the adsorption-based separation performed industrially by pressure and/or temperature swing adsorption, and even if the thermodynamic selectivity is the main performance metric, the kinetic performances can improve our understanding of the adsorption process. For instance, in breakthrough experiments (a lab equivalent of a pressure swing adsorption) used to characterize the comparative adsorption performances of a gas mixture, the shape of the curve can be explained by diffusion processes. The goal of this chapter is to explore this often neglected diffusion parameter in an adsorption-based Xe/Kr separation process.

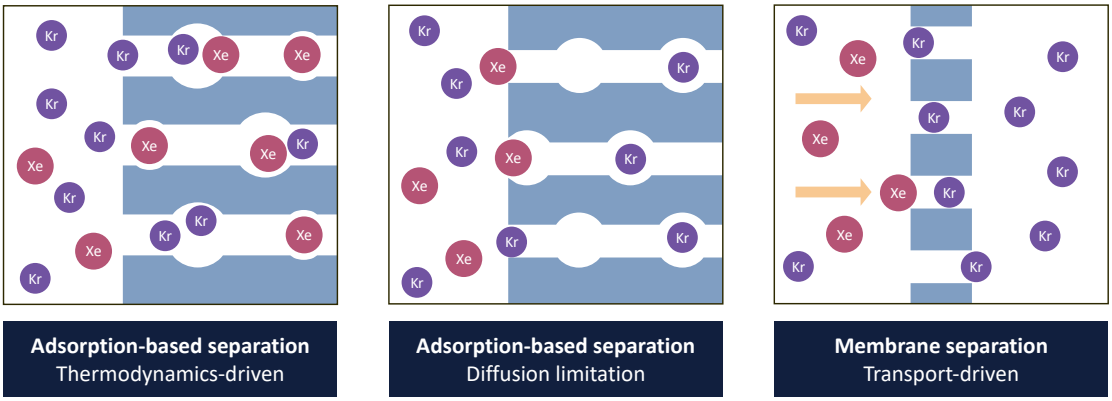


Figure 5.1: Illustration of the comparative role of the thermodynamic and transport properties for Xe/Kr separation in nanoporous materials. From the transport dominated process of membrane separation to the thermodynamically equilibrated separation processes in the nanopores, different more nuanced cases could emerge where the diffusion imposes kinetic limitations.

5.1 Modeling the Diffusion Process

Since the pollen motion observations of the botannist Brown in 1826, scientists have observed and studied the seemingly erratic movement of particles in a static bulk medium. Later, Fick proposed a macroscopic model of this so-called brownian motion by introduced the coefficient D of a diffusion equation **??** (1D) based on experimental measures of the concentration ϕ.**Fick_1855** According to this law, at the macroscopic level, the particles tend to go move from the most concentrated area of the bulk to the less concentrated one.

$$\frac{\partial \phi}{\partial t} = \mathrm{D}_x \frac{\partial^2 \phi}{\partial x^2} \tag{5.1}$$

To better understand the brownian motion of suspended particles on a liquid Einstein derived a microscopic model of the diffusion motion based on the molecular-kinetic theory of heat on the miraculous year of 1905.**einstein1905motion** To determine the so-called self-diffusion coefficient, he followed the motion of a particle assumed independent from other particles and time steps large enough to consider mutually independent two consecutive motions. By using the particle distribution of N independent diffusing particle, he redefined the diffusion coefficient as a function of the mean squared displacement (MSD) of a particle. In a tridimensional space, we have the following Einstein relation:

$$\left\langle r(t)^2 \right\rangle = 6\mathrm{D}t \tag{5.2}$$

where $r(t)$ is the displacement of a particle from the time 0 to t. The brackets represent the average over all independent trajectories (different particles and different time origins). This equation can be generalized to the diffusion of an adsorbate in the adsorbed phase, which describes how easy it is for a particle to move inside a nanoporous material. A low diffusion coefficient means a limited access to the pores of the structure as illustrated on Figure **??**.

Using molecular simulations of the adsorbate displacements, we will try to model the diffusion coefficient of xenon and krypton inside a nanoporous material. Although other approaches

like the Green-Kubo equation exist the relatively less complex Einstein law is prefered for self-diffusion calculations, as shown by the following comparative study [**Maginn_2020**]. In this section, we will focus on the different simulation techniques that can be used to evaluate the diffusion in high-throughput screenings. We will try to present different ways of accessing the MSD of a diffusing particles, by beginning from the most straight-forward molecular dynamics to faster methodologies more suitable in screenings such as machine learned surrogate models and kinetic monte carlo simulations.

5.1.1 Molecular dynamics

Molecular dynamics are used to simulate the motion of molecules in a given system. It is usually used to calculate thermodynamic averagings.[give some examples] Here, we are going to focus on the calculation of diffusion coefficients of monoatomic molecules.

Simulation details

Molecular dynamics (MD) aims at describing the motion of particles subjected to the forces of the surrounding particles. It can therefore be seen as a complex integration of the Newton's law of motion. A particle i of position $\mathbf{r}_i$ and mass m_i subjected to a force $\mathbf{F}_i$ resulting of the cumulated interactions with its surrounding is accelerated according to this equation:

$$m_i \frac{\mathrm{d}^2 \mathbf{r}_i}{\mathrm{d}t^2} = \mathbf{F}_i \tag{5.3}$$

In a classical modeling, the forces are determined using the well-named forcefield that was previously introduced in the chapter 2. Back there, we only considered intermolecular interactions simply modeled by the Lennard-Jones (LJ) interaction potential between atom pairs, which is also what we will use in this section (of course, it is not the only way of defining a forcefield but just a simplification). Using the LJ potentials U^{LJ} (defined in equation **??**), we can derive a vectorial force $\mathbf{f}_{ij}$ between two atoms i and j.

$$\mathbf{f}_{ij} = -\left.\frac{\mathrm{dU}_{ij}^{\mathrm{LJ}}}{\mathrm{d}r}\right|_{r=r_{ij}} \frac{\mathbf{r}_{ij}}{r_{ij}} = 24\epsilon_{ij}\left(2\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right)\frac{\mathbf{r}_{ij}}{r_{ij}^2} \tag{5.4}$$

where ϵ_{ij} and σ_{ij} are the LJ parameters of the pair of atoms ij. And the resulting force is simply the sum of the forces $\mathbf{F}_i = \sum_j \mathbf{f}_{ij}$ exerted by the surrounding atoms j. To reduce the computation time required, molecular simulations only consider the atoms within a given cutoff radius.

Now that we defined the force $\mathbf{F}_i$, we can put a molecule in motion by integrating the equation **??** from a time t to a time $t + \delta t$. There are different methods to integrate equation of motion such as the Euler or velocity-Verlet scheme presented in the book of Frenkel et al.[**frenkel2001md**] Here, we will focus on the *leap frop* integration implemented in the RASPA[**dubbeldam2016**] software that we used for our simulations. The position $\mathbf{r}_i$ and the velocity $\dot{\mathbf{r}}_i$ are between each time step δt using the following equations:

$$\begin{aligned} \dot{\mathbf{r}}_i\left(t + \tfrac{1}{2}\delta t\right) &= \dot{\mathbf{r}}_i\left(t - \tfrac{1}{2}\delta t\right) + \tfrac{1}{m_i}\mathbf{f}_i \\ \mathbf{r}_i\left(t + \delta t\right) &= \mathbf{r}_i\left(t\right) + \dot{\mathbf{r}}_i\left(t + \tfrac{1}{2}\delta t\right)\delta t \end{aligned} \tag{5.5}$$

From the initial conditions $(\mathbf{r}_i(0), \dot{\mathbf{r}}_i(0.5\delta t))$, we can translate the center of mass of the molecule i to any position $\mathbf{r}_i(t_n = n * \delta t)$. We will skip the rotation step required for polyatomic molecules since we are restricting the study on the monoatomic noble gas. The different positions $\{(t_n, \mathbf{r}_i(t_n))\}_{n=0,\ldots,\mathrm{N_{tot}}}$ constitute the total trajectory of the MD simulation (to simplify I do not mention the velocities). It is possible to use this total trajectory to derive an average of MSD that could be analysed to calculate the diffusion coefficient.

Diffusivity calculation using an MD trajectory

I used the MSD sampling technique implemented in RASPA[dubbeldam2016] that was presented in an article [**Dubbeldam_2009**] by a few authors of the adsorption simulation software. The approach is based on a modified approach of the order-**n** algorithm described in the book of Frenkel and Smit [**frenkel2001msd**] I will focus, therefore, on the so-called multiple window algorithm used to calculate the diffusion coefficients of xenon and krypton in this chapter.

To understand it, I will start by explaining what a window algorithm would do and how it generalizes to the multiple window algorithm we are interested in. First, let us consider a single MD trajectory of duration $t_{\mathrm{tot}} = \mathrm{N_{tot}}\delta t$. This trajectory can be used to generate displacement of any size τ. Naively, we can start by taking $\|\mathbf{r}_i(\tau) - \mathbf{r}_i(0)\|^2$ as a square displacement of a sub-trajectory $\mathcal{T}(0 \rightarrow \tau)$ of duration τ. However, it is not enough to make a statistically meaningful average of the MSD as described in the Einstein equation**??**. Using the hypothesis of independence between two movements of the same particle separated by a time δt used in Einstein's paper [**einstein1905motion**], a shift of the origin of time by δt would generate another trajectory. We can repeat this process i times while $\tau + i\delta t \leq t_{\mathrm{tot}}$. This would be very accurate, but also very inefficient in the case where $\tau \gg \delta t$ since two consecutive sub-trajectories $\mathcal{T}(i\delta t \rightarrow \tau + i\delta t)$ and $\mathcal{T}((i+1)\delta t \rightarrow \tau + (i+1)\delta t)$ would be very similar.

To efficiently sample the trajectory into sub-trajectories that are independent we can use a sampling time step of $\delta\tau \lesssim \tau$ chosen to be in the same order of magnitude as τ. To do so, the window approach would first define a value $\delta\tau$ and generate $\mathrm{N}_\tau = \lfloor (t_{\mathrm{t\,ot}} - \tau)/\delta\tau \rfloor$ different sub-trajectories $\{\mathcal{T}(0 \rightarrow \tau), \mathcal{T}(\delta\tau \rightarrow \tau + \delta\tau), \ldots, \mathcal{T}(\mathrm{N}_\tau\delta\tau \rightarrow \tau + \mathrm{N}_\tau\delta\tau)\}$ of duration $\tau = k\delta\tau$, where k is an integer between 1 and K that defines the time window we want to sample. By doing so, we get the MSD $\langle r(\tau)^2 \rangle$ for duration values τ equal to $\delta\tau, \ldots, \mathrm{K}\delta\tau$. The relation $\langle r(\tau)^2 \rangle$ can then be fitted to the equation **??** to obtain the diffusion coefficient if the relation is linear. The trajectory generation of the window approach is illustrated on the Figure **??** for a decomposition into sub-trajectories of a duration $\tau = 3\delta\tau$ shifted by $\delta\tau$.

The major drawback of this method is that we need to define a timescale $\{\delta\tau, \ldots, \mathrm{K}\delta\tau\}$ beforehand. In order to be able to access the different timescales in a single simulation, we can perform a multiple window algorithm developed by Dubbeldam et al. and used in the RASPA software to compute mean squared displacements (MSD) in a molecular dynamics simulation.

The different time windows are defined in a recursive way using the default parameters of RASPA. The first time window is defined by K = 25 displacements of duration $\delta t, 2\delta t, \ldots, \mathrm{K}\delta t$ with a shift of δt (the default shift value δt of the first window can be changed with the parameter SampleMSDEvery). The second window is now based on a sampling duration $\delta\tau_1 = \mathrm{K}\delta t$ and the sub-trajectories will have durations of $\left(\tau_1^{(1)} = \delta\tau_1\right), \ldots, \left(\tau_1^{(k)} = \mathrm{K}\delta\tau_1\right)$. If we repeat the process recursively until no window can be generated anymore, and the i_{th} window would have a sampling duraction of $\delta\tau_i = \mathrm{K}^i\delta t$ and sub-trajectories durations of

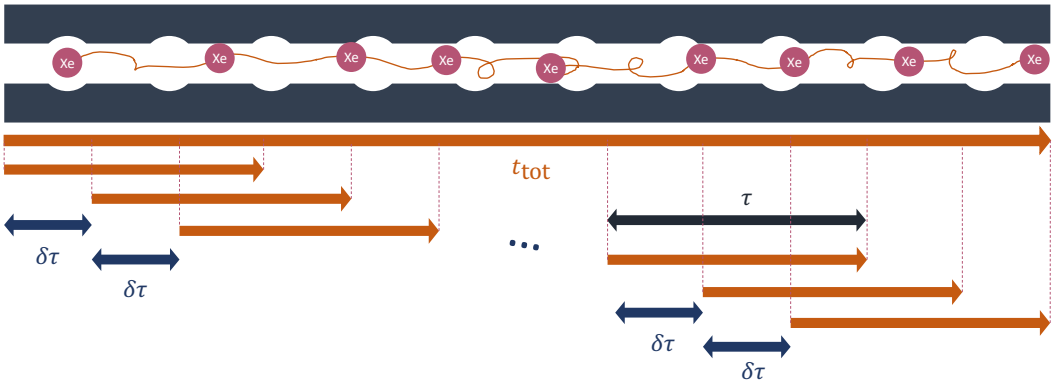


Figure 5.2: Illustration of the generation of trajectories of size τ by shifting the origins of multiple durations $\delta\tau$.

$\left(\tau_i^{(1)} = \delta\tau_i\right), \ldots, \left(\tau_i^{(k)} = \mathrm{K}\delta\tau_i\right)$. The time scale $\delta\tau_i = \mathrm{K}^i\delta t$ we sample follows a geometrical progression and very different time scales can be accessed using this method in order to find the time scale corresponding to the diffusion regime (linear relation between the MSD and the duration of the sub-trajectories used in the averaging). For example on Figure **??**, we can see the different timescales and the exponent value b of a fit to a function of type $x \mapsto ax^b$ for the different time windows — values of b near 1 can be associated to a diffusion regime. The determination of the diffusion coefficient is now reduced to a simple fitting problem that will be explained in more details in the presentation of the diffusion coefficient screening in section **??**.

This methodology can then be used replicated to thousands of structures to characterize the diffusion properties of these materials. Several screenings have already been carried out in the literature as presented in the chapter 1 in the section dedicated to transport property screenings. We will now dive a little deeper in the prediction of these quantities using machine learning.

ML MODELING

In a very recent study, Daglar et al. used an ML model to predict the diffusion coefficient of a 100 thousand hypothetical MOFs using the data for about 5000 CoRE MOF structures.**Daglar_2022** Along with very standard geometrical descriptors, they used chemical composition descriptors and the heat of adsorption as the input features of their machine learning model to predict the diffusion coefficients of H_2, CH_4, N_2 and He in the different MOF materials of CoRE MOF 2019 (training) and of hMOF (testing). The combination of kinetic data with thermodynamic data for the characterization of MOf materials is a very interesting approach. However, the major drawback of most of the approaches in the literature is the lack of structure–property relationship to understand the microscopic origins of the diffusion coefficient values.

Simarly to what we have done on the thermodynamic screening (chapter 2–4), in our approach to tranport property screening, we will also start by drawing structure–property relationships between the diffusion coefficient and the geometrical descriptors of the MOF structures. And in an attempt to have a deeper understanding of the diffusion process, we will try to evaluate the diffusion activation energy using energy grid-based methods described in the literature. All these techniques aim at better predicting the diffusion coefficients either in a direct calculation or

in an ML surrogate model. To achieve that, we will start by introducing the kinetic Monte Carlo approach that is less accurate than the MD approach but is indeed much more efficient.

5.1.2 Lattice kinetic Monte Carlo

The lattice kinetic Monte Carlo method relies on a pre-defined lattice of stable points corresponding to adsorption sites. Each site connected to another if there is a diffusion path (narrow channel) that connects them. To calculate the probability of transition from one site to another, we need to define a transition state in the narraow channel which correspond to the highest energy point on the minimal energy diffusion path (the saddle point). The probability of transition would therefore be defined with regard to the energy barrier to overcome in order to cross the channel. Once the lattice defined, we only need to propagate an adsorbate from one site to another using the different transition probabilities, which gives a coarse-grained trajectory compared to the one obtained in a MD simulation, but is perfectly usable to compute the MSD and calculate a diffusion coefficient.

[Illustration]

Transition state theory

The transition state theory is usually used in chemistry to explain the kinetics of a reaction. To do so it compares the energy of the reactants before reacting and the one of the transition state to calculate the rate of the reaction. And it defines a reaction kinetic rate that explains the kinetics of a reaction.

In our case, we want to study the transition from a pore to another one by transitioning through a channel. It is also possible to define a diffusion rate that connects the adsorption sites of the pores using Bennet-Chandler [Bennet-Chandler approach][BENNETT1977]

[what is a transition state for diffusion?]

Construction of the lattice

[Detect TS, and bassins.]

Fast kinetic Monte Carlo tutrast[Mace_2019] autre étude avant aussi

[Generation of trajectories + similar techniques than previously presented to find the MSD]

5.2 Screening of transport properties

To complete the thermodynamic screenings that we performed in the chapters 2–4, we also carried out a transport property screening. In this section, we will provide a description of the screening approach as well as the analysis of the diffusion coefficients compared with typical geometrical descriptors.

5.2.1 Diffusion in a selective material

Before going into the details of the screening study, we will present the approach adopted for the diffusion coefficient calculation using MSD values, on one example, SBMOF-1.[Banerjee_2016] This preliminary study will help us calibrate the time parameters (time step, maximum time) that will be used in the final screening study.

First, I ran a molecular dynamics simulation of 500 million steps (about 1–2 days of simulation) with a thousand initialization steps and 100 thousand equilibration steps to model a xenon diffusing in the KAXQIL[Banerjee2012] MOF at infinite dilution. To be at infinite dilution, we set the box size so that no interactions occur between the different adsorbates. We tested three different time steps δt as defined in equation ??.

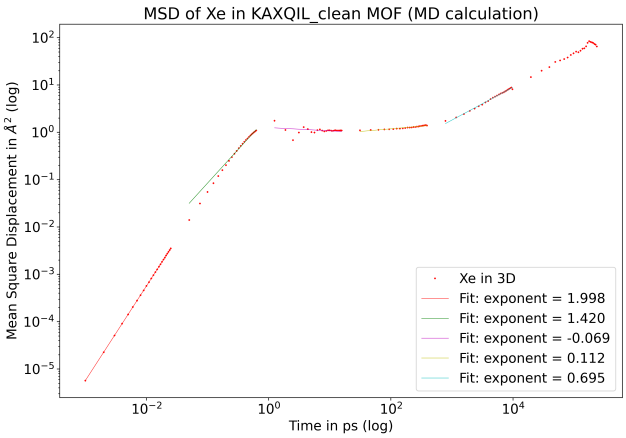


Figure 5.3: [MSD, show the fit at higher time scale/ Change (10-300) yellow / Add (300-1000) il cyan]

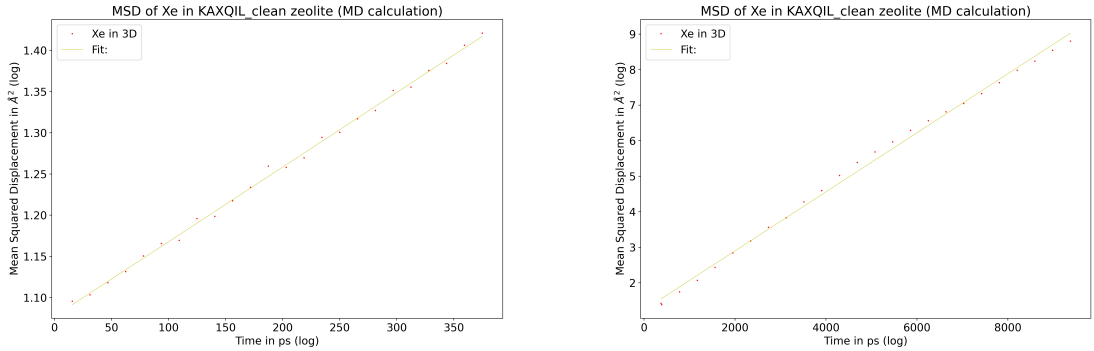


Figure 5.4: [MSD, show the fit at higher time scale/ Change (10-300) yellow / Add (300-1000) il cyan]

The diffusion regime seems to happen at a time scale of tens of nanoseconds. But the number of data points is not high enough at this higher time scale. [Confirm with real data]

1.5e-8 cm2/s (0.01 – 0.4ns) 1.4e-8 cm2/s (0.4–9 ns)

[uncertainty discussion]

[1.77 +/- 0.35 10-8cm2/s] on 10 different trajectories (20% error) [If we don't have enough time steps > can't reach the diffusion regime] [Remodeler] To access timescales where the diffusion is much more linear according to the plot without increasing the calculation time that is already of the order of 1–2 days per structures, we can increase the MD time step. This is generally not recommended as we try to be half the period of the fastest vibration (Nyquist-Shannon sampling theorem). 1–2fs is very common in most MD. Even in "rigid' nanoporous materials, typical time steps are ∼1fs is used, but no one really questioned the relevance of this time step

value.[Bukowski_2021] In this rigid system, the only moving particle is xenon, hence in theory there is no vibration limitations. However, to access higher time steps, we need to check the stability of the diffusion values.

[Put data for 5fs]

0.8 ns – 1.9 ns

2 ns – 46 ns

[kinetic KAXQIL (water loaeded structure) problem]

[Worse if we consider xenon and krypton]

5.2.2 Structure–diffusivity relationships

500 000 000 cycles with restart on the server. [We have generated diffusion coefficients for the ... best materials]

[correlation with VF/SA/Selectivity/permselectivity]

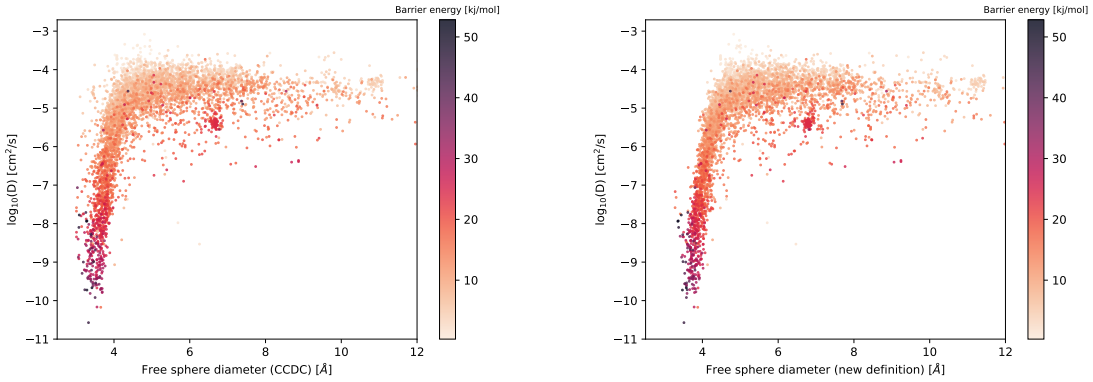


Figure 5.5: [correlation with PLD without labeling]

5.2.3 Identification of interesting materials

CCSD ref. code	LCD (Å)	s_1	s_0	PLD (Å)	Ratio Diff.	Coeff. de Diff. Xe
QOZDOY[**Zhang_2001**]	5,3	37	52	4,7	0.4	7×10^{-5} cm^2 s^{-1}
GUMDEZ[**Yin_2014**]	5,3	42	56	4,8	0.5	7×10^{-5} cm^2 s^{-1}
ADOGEH[**Peikert_2012**]	12,7	10	49	5,1	14	5×10^{-5} cm^2 s^{-1}

Table 5.1: Performances de structures identifiées par un screening prenant en compte les coefficients de diffusion.

5.3 Fast diffusion calculation algorithm

5.3.1 Implementation in C++

Breadth-first search

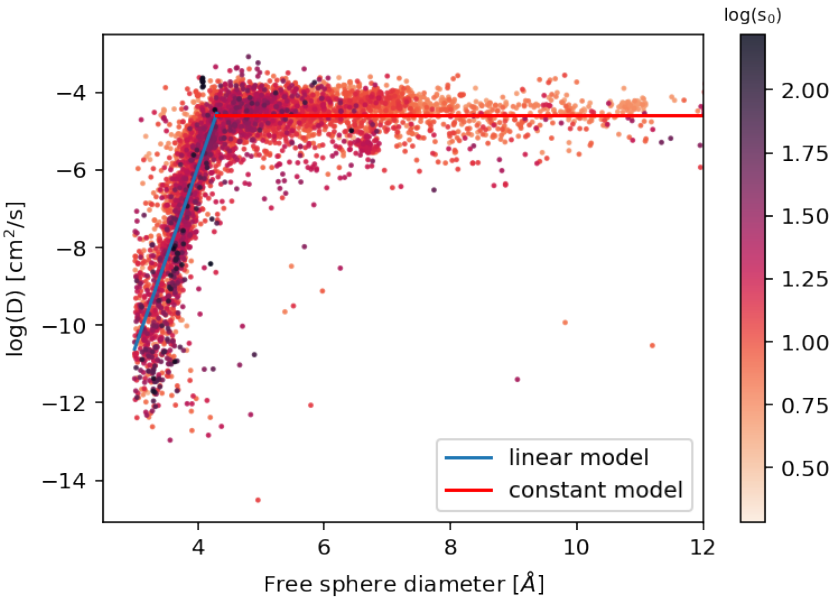


Figure 5.6: Logarithme du coefficient de diffusion du xénon à la limite des basses pressions en fonction du diamètre de la plus petite sphère libre dans le matériau (LCD). Les points sont colorés en fonction du logarithme de la sélectivité à basse pression $\log(s_0)$.

5.3.2 Preliminary results

5.3.3 Visualization tool

[Take some examples for the vizualisation with comparison to the pore size and diffusion coefficient]

5.3.4 Development of a first prediction model

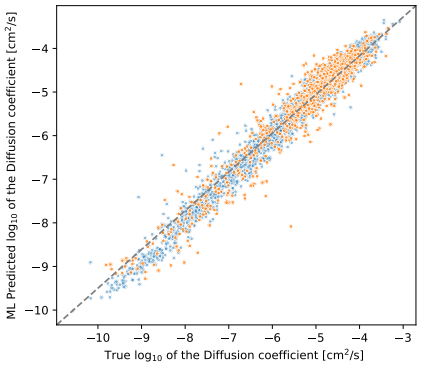


Figure 5.7

ML descriptors

Next steps

[Broaden to the study of collective diffusion, Maxwell-stefan, Onsager, etc.]

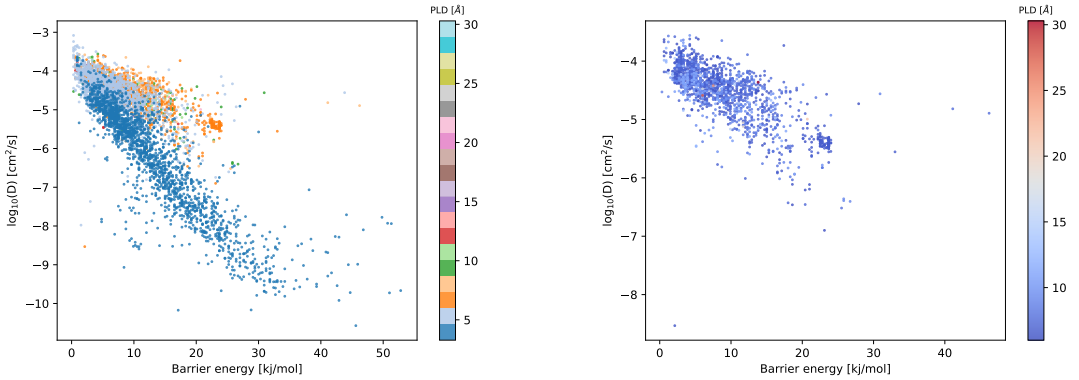


Figure 5.8

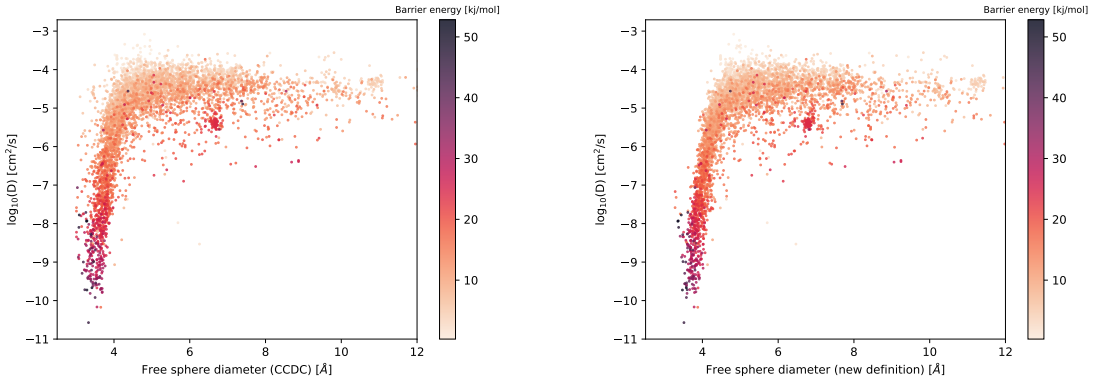


Figure 5.9

[Can be used in breakthrough simulations using RUPTURA[Sharma_2023] and equations that link the diffusion coefficient to the axial dispersion coefficient a key parameter in the breakthrough modeling.]

6

TOWARD THE NEXT GENERATION OF SCREENINGS

6.1 Limits of Current Screening Methodologies 161
6.2 Flexibility in a Screening 163
6.2.1 Problem, literature 163
6.2.2 163
6.2.3 Database approach: Snapshot method 163
6.3 Noble Gas Polarizability 163
6.3.1 Problem definition 163
6.3.2 Studying the polarization 164
6.3.3 Perpectives 164

6.1 Limits of Current Screening Methodologies

Comme présenté dans notre article de revue sur les différentes techniques de screening de performances de matériaux,[Ren_2022] les méthodologies standard de screening sont en silo : certains se concentrent sur les propriétés thermodynamiques, d'autres sur les propriétés de transports et la flexibilité n'est étudiée que de manière embryonnaire du point de vue du screening. L'ambition de la troisième année est de mettre en lien ces différents screenings afin de prédire au plus près les performances expérimentales tout en parcourant de larges bases de données.

Pour voir les limites de l'approche standard, nous allons partir d'un exemple problématique. Si on considère le matériau SBMOF-1,[Banerjee_2016] la sélectivité prédite est certes très élevée (de l'ordre de 70,6 pour Simon *et al.*) mais la sélectivité mesurée expérimentalement est seulement de 16. On pourrait expliquer ce phénomène par la qualité du champ de force, mais également par des mécanismes non décrits par les modèles sous-jacents comme la flexibilité ou la cinétique d'adsorption.

Dans un cristal poreux macroscopique, il y a une distribution plus ou moins diverse de tailles de pore qui dépend d'une part de la flexibilité intrinsèque du matériau et d'autre part de l'interaction des molécules de gaz avec les pores. Or, dans notre modélisation on prend

CCSD ref. code	Adsorbat	LCD (Å)	Sélectivité	PLD (Å)	Coeff. de Diff. Xe
KAXQOR[Banerjee2012]	aucun	4,51	22	4,04	7×10^{-06} cm^2 s^{-1}
QUXRIM[Banerjee2016hydro]	hexane	4,75	52	4,31	3×10^{-05} cm^2 s^{-1}
QUXRUY[Banerjee2016hydro]	hexane	4,91	96	3,57	9×10^{-10} cm^2 s^{-1}
KAXQOR01[Yeh2012]	aucun	4,99	101	3,66	3×10^{-09} cm^2 s^{-1}
QUWYEO[Banerjee2016hydro]	butane	4,99	100	3,65	5×10^{-09} cm^2 s^{-1}
QUXROS[Banerjee2016hydro]	hexane	5,00	99	3,66	5×10^{-09} cm^2 s^{-1}
QUXREI[Banerjee2016hydro]	hexane	5,02	101	3,67	7×10^{-09} cm^2 s^{-1}
QUXRAE[Banerjee2016hydro]	hexane	5,03	100	3,68	7×10^{-09} cm^2 s^{-1}
KAXQIL[Banerjee2012]	H_2O	5,12	104	3,77	3×10^{-08} cm^2 s^{-1}
QUXQUX[Banerjee2016hydro]	butane	5,17	103	3,83	1×10^{-07} cm^2 s^{-1}
UQEFAZ[Banerjee_2016]	krypton	4,53	23	4,08	5×10^{-06} cm^2 s^{-1}
UQEFED[Banerjee_2016]	xenon	4,89	63	3,54	1×10^{-11} cm^2 s^{-1}

Table 6.1: Performances de structures similaires à SBMOF-1 décrite par Simon et al.[Simon_2015, Banerjee_2016] *pour la séparation Xe/Kr. Les deux dernières correspondent aux structures de la publication de Simon, Banerjee* et al. *Différentes tailles de pore induisent différentes sélectivités et différentes diffusivités.*

uniquement en compte une configuration des pores, donc une unique taille de pore à travers tout le matériau nanoporeux. Cette hypothèse simplificatrice peut être remise en cause notamment sur l'exemple de SBMOF-1. On trouve en effet une dizaine de structures différentes décrivant pourtant le même matériau dans la base de données CoRE MOF 2019. Ces structures sont différentes du point de vue de la structure, notamment de la taille des cavités adsorbantes ("Largest Cavity Diameter" – LCD), ainsi que la taille des canaux les reliant ("Pore Limiting Diameter" – PLD).

Pour simplifier, on peut identifier deux types de conformations sur la table **??** : 1) la cavité est petite (autour de 4.5 Å) et les canaux restent larges (autour de 4.0 Å) ; 2) la cavité est grande (autour de 5 Å) et les canaux sont étroits (autour de 3.5 Å). Dans le premier cas, la diffusion à travers le matériau est légèrement entraver sans blocage cinétique mais la sélectivité théorique est moyenne autour de 20. Tandis que dans le deuxième cas, la sélectivité prédite est très élevée mais la diffusion est tellement entravée qu'il y aurait un blocage cinétique de l'adsorption. À l'aune de ces résultats, la différence entre la théorie et l'expérience pourrait en effet s'expliquer par la flexibilité qu'induit l'adsorbat et les conditions de l'expérience (température). Une telle diversité de taille caractéristiques nous amène à nous interroger sur l'influence de la flexibilité sur les résultats du screening. Sachant que la flexibilité peut changer grandement les valeurs de sélectivité, le fait que la structure existe sous différentes conformations nous invitent à interroger l'hypothèse de rigidité.[Witman_2017]

D'autre part, le cas de UQEFED, qui représente la structure de SBMOF-1 adsorbé par du xénon déterminée par Banerjee *et al.*,[Banerjee_2016] allie une forte sélectivité (63) mais avec une très faible diffusivité (1×10^{-11} cm^2 s^{-1}). Bien que la sélectivité thermodynamique est très bonne, cet équilibre ne peut être atteint que très lentement car les molécules de xénon n'ont pas le temps de se déplacer dans la structure. Ainsi, ce blocage cinétique pourrait également expliquer que l'expérience note des valeurs de sélectivités en décalage avec la théorie. En effet, rien ne nous

dit que l'équilibre est atteint lors du tracé des isothermes expérimentales. Une pénétration lente du xénon à l'intérieur du matériau donnerait des quantités de xénon adsorbées plus faibles que prévu par la thermodynamique, ce qui mettrait en défaut les simulations théoriques.

Fort de ces deux constats, il est plus probable que le matériau SBMOF-1 réel ait une certaine flexibilité permettant le déplacement du xénon quand les sites ne sont pas déjà occupés par du xénon. Mais lorsqu'il y a un xénon qui occupe le site, les canaux deviennent plus étroits et induisent un blocage cinétique. Ces phénomènes ne pourraient pas être révélés par une étude purement thermodynamique et montrent l'intérêt de tenir en compte de propriétés clés comme la cinétique ou la flexibilité. C'est pourquoi, il est maintenant crucial d'aller au-delà des méthodes de screening standard décrites dans la littérature et d'essayer de prendre en compte d'autres phénomènes physiques dans le screening. En considérant la flexibilité et les effets de transport on peut ainsi identifier des matériaux qui garderont leur performance lors de l'expérimentation.

[put graph comparing experiment and UFF]

The beginning of some explication

6.2 Flexibility in a Screening

Final screening step, easy integration into the workflow of current sreenings

6.2.1 Problem, literature

6.2.2

6.2.3 Database approach: Snapshot method

On peut distinguer deux types de flexibilité pour les matériaux poreux : la première flexibilité intrinsèque correspond simplement à la vibration thermique du matériau et la deuxième flexibilité induite dépend de l'interaction avec l'adsorbat. Pour décrire complètement la flexibilité, il faut utiliser des champs de force flexibles ce qui démultiplie considérablement le temps de calcul, ce qui est inimaginable pour un screening. Une approche plus raisonnable consiste donc à ne tenir en compte de la flexibilité intrinsèque en générant un ensemble de conformations "snapshots" d'une même structure vibrant via des simulations NPT *ab initio* ou classique. La publication de Witman *et al.* montre un changement des performances de sélectivité Xe/Kr lorsqu'on tient en compte de ce phénomène de vibration thermique.**Witman_2017**

Cette approche peut donc être généralisée pour calculer des sélectivités en système flexible à partir du code d'échantillonnage de surface présenté dans la deuxième partie de ce rapport. Couplé au code de calcul de diffusion (en cours de développement), il serait possible de calculer des coefficients de diffusion en système flexible sur des bases de données de milliers de structures. En piste de recherche potentielle, il pourrait être intéressant d'insérer un adsorbat comme le xénon à des points stratégiques de la structure pour voir la déformation induite et ainsi étudier la flexibilité induite du matériau qui aurait été négligée dans la première approche.

6.3 Noble Gas Polarizability

6.3.1 Problem definition

Best materials use polarization effects[Li_2019, Pei_2022]

Talk about the order of magnitude of the different interactions > charge–(induced dipole high magnitude)

Standard methods failing to describe oms[Perry_2014]

[faire référence à 2-thermo partie sur les interactions??]

6.3.2 Studying the polarization

Inspired by works on the subject[Lachet_1998, Becker_2017]

[essayer d'ajouter la polarisabilité pour PEI et al. et Li et al.]

Xe/Kr difference of polarisability Open Metal Sites/polar groups [20220421_pres]

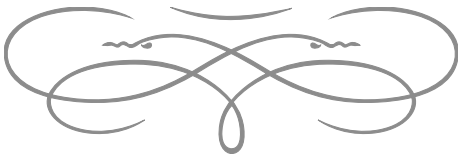
[Not the best material, but interesting discussion on open metal site effect] Tao et al.[Tao_2020] looked at tuning (and improving) the selective adsorption of Xe over Kr by MOF open metal sites in the UTSA-74 framework structure.

6.3.3 Perpectives

GENERAL CONCLUSIONS

The work presented in this thesis is

This work opens perspectives for



LIST OF PUBLICATIONS

Peer-reviewed papers

Preprint

RÉSUMÉ EN FRANÇAIS

Introduction 171

Introduction

[5 à 10 pages]

Les matériaux poreux sont des matériaux